*Ansgar Beckermann*

# Aufsätze

## Band 1

## Philosophie des Geistes

**Universitätsbibliothek Bielefeld 2012**

**Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

**1. Auflage 2012**

**Universitätsbibliothek Bielefeld**
Universitätsstraße 25
33615 Bielefeld
E-Mail: publikationsdienste.ub@uni-bielefeld.de



ISBN 978-3-943363-01-2

Zugleich online veröffentlicht auf dem Publikationsserver der Universität Bielefeld
URL http://pub.uni-bielefeld.de/publication/2508111
URN urn:nbn:de:0070-pub-25081115
[http://nbn-resolving.org/urn:nbn:de:0070-pub-25081115]

## Vorwort

Vor 40 Jahren wurde mein erster Aufsatz „Die realistischen Voraussetzungen der Konsenstheorie von J. Habermas“ veröffentlicht. Seitdem sind etwa 100 weitere Artikel erschienen – an sehr unterschiedlichen, zum Teil auch etwas entlegenen Orten. Aus diesem Grund scheint es mir sinnvoll, einige dieser Arbeiten in zwei Sammelbänden zusammenzufassen, um damit denen, die an diesen Arbeiten interessiert sind, den Zugang zu erleichtern. Dieser erste Band enthält Aufsätze zur Philosophie des Geistes.

In diese Sammelbände habe ich die Aufsätze aufgenommen, von denen ich heute noch überzeugt bin, dass sie einen interessanten Beitrag zur Diskussion leisten, und die meine eigenen Positionen, wie mir scheint, besonders prägnant zum Ausdruck bringen. Die einzelnen Beiträge wurden im Wesentlichen wörtlich übernommen. Nur offensichtliche sprachliche und sachliche Fehler habe ich korrigiert. Bei einigen Beiträgen handelt es sich um deutsche Fassungen von Aufsätzen, die bisher nur auf Englisch veröffentlicht wurden. Bei einem Beitrag handelt es sich um eine überarbeitete Fassung einer früheren deutschen Veröffentlichung. All dies ist jeweils in Fußnoten zu Beginn jedes Beitrags deutlich gemacht.

Aus zwei Gründen habe ich mich für eine elektronische Publikation entschieden. Erstens ist es gar nicht leicht, einen renommierten Verlag für so ein Projekt zu gewinnen. Sammelbände dieser Art verkaufen sich nicht besonders gut und sind daher für diese Verlage nicht attraktiv. Zweitens bin ich aber auch ein entschiedener Anhänger der Idee des *open access*. Mir ist natürlich klar, dass durch diese Idee die etablierten Wissenschaftsverlage in eine schwierige Lage kommen können. Aber dass Beiträge zu wissenschaftlichen Diskussionen möglichst leicht und zu geringen Kosten zugänglich sein sollten, scheint mir das höhere Gut zu sein. Wenn etwa JStore für einen einzigen Aufsatz ein Entgelt von etwa $ 30 in Rechnung stellt, scheint mir das vollkommen absurd – zumal die Autoren wissenschaftlicher Texte (anders als Autoren, die Romane oder Sachbücher schreiben) in aller Regel ja selbst gar kein Geld für ihre Veröffentlichungen haben möchten.

Ich hoffe also, dass trotz dieser im Moment noch ungewöhnlichen Veröffentlichungsweise sich doch einige für meine Aufsätze interessieren. Denn für einen Wissenschaftler ist nichts so wichtig wie, dass seine Arbeiten rezipiert und diskutiert werden.

Bielefeld, im Juli 2012

## Inhaltsverzeichnis

# Physikalismus

# 1

# Ein Argument für den Physikalismus*

## 1

Es gibt nicht *das* Problem des Naturalismus, sondern – wie die Beiträge in diesem Band zeigen – eine ganze Familie von mehr oder weniger stark miteinander verbundenen Teilproblemen. In diesem Aufsatz soll es nur um eines dieser Teilprobleme gehen – das Problem des *ontologischen* Naturalismus. Oder, um es genauer zu sagen, um eine spezifische Variante dieses Teilproblems – das Problem des *ontologischen Physikalismus*. Die Grundthese des ontologischen Physikalismus lautet einfach:

(PH) Alles, was es gibt, ist physischer Natur.

Aber diese Formulierung ist in mehrfacher Hinsicht erläuterungsbedürftig. Was zum Beispiel soll hier ‚Alles' heißen? Wenn wir uns auf einige grundlegende ontologische Unterscheidungen beschränken, heißt es sicher: alle Dinge, alle Eigenschaften und alle Ereignisse. Somit zerfällt die Grundthese des ontologischen Physikalismus in (mindestens) drei Teilthesen:

($PH_1$) Alle Dinge sind physische Dinge.

($PH_2$) Alle Eigenschaften sind physische Eigenschaften.

($PH_3$) Alle Ereignisse sind physische Ereignisse.

Im folgenden werde ich nur auf die ersten beiden Thesen eingehen – (a) weil ich denke, daß die dritte These aus den ersten beiden folgt, und (b) weil ich Zweifel daran habe, daß eine Position, die nur durch die Thesen ($PH_1$) und ($PH_3$) gekennzeichnet ist, eine hinreichend starke physikalistische Position darstellt.[1] Damit stellt sich als nächstes die Frage, was unter

---

* Erstveröffentlichung in: G. Keil & H. Schnädelbach (Hg.) *Naturalismus*. Frankfurt/M.: Suhrkamp 2000, 128–143. Deutsche Fassung von „The Real Reason for the Standard View", in: A. Meijers (Hg.) *Explaining Beliefs: Lynne Rudder Baker and Her Critics*. Stanford: CSLI Publications 2001, 51–67.

[1] Wenn man eine in der Literatur häufig zu findende terminologische Unterscheidung aufnimmt, könnte man sagen, daß der *reduktive* Physikalist die Thesen ($PH_1$) und ($PH_2$) – und damit auch die These ($PH_3$) – vertritt, während sich der *nicht-reduktive* Physikalist nur die Thesen ($PH_1$) und ($PH_3$) zu eigen macht. Daß ein so charakterisierter nicht-reduktiver Physikalismus keine ausreichend starke physikalistische Position darstellt, ergibt sich aus der Tatsache, daß auch Vertreter dieser Position sich nicht der Frage entziehen können, wie sie es mit der These ($PH_2$) halten wollen. (Vgl. bes. Brian McLaughlin, „philosophy of mind", in: Robert Audi (Hg.), *The Cambridge Dictionary of Philo-*

dem Adjektiv ‚physisch' in den Thesen ($PH_1$) und ($PH_2$) zu verstehen ist. Was sind physische Dinge? Und was sind physische Eigenschaften?

Bleiben wir zunächst bei der ersten Frage. Klare Beispiele für physische Dinge sind: Protonen, Zuckermoleküle, Steine, Sterne, aber auch Wasserhähne, Besen und Plattenspieler. Nichtphysische Dinge sind dagegen: Gott, die Engel, Cartesische Seelen, der *élan vital*, aber auch Mengen, Zahlen und Propositionen. Gibt es ein klares Merkmal, daß es gestattet, nichtphysische Dinge eindeutig von den physischen abzugrenzen?

Eine Antwort auf diese Frage zu geben, ist in der Philosophie des öfteren versucht worden. Descartes etwa kennt zwei Arten von Substanzen (Dingen): physische Dinge (*res extensae*) und denkende Dinge (*res cogitantes*). Die einzige wesentliche Eigenschaft physischer Dinge ist ihre Ausdehnung (*extensio*); die einzige wesentliche Eigenschaft denkender Dinge ist das Denken oder Bewußtsein (*cogitatio*). Nach Descartes sind physische und denkende Dinge also säuberlich voneinander getrennt. Die ersteren befinden sich in Raum und Zeit und sind unfähig zu denken; die letzteren dagegen denken (ständig), haben aber weder einen Ort im Raum noch eine räumliche Ausdehnung.

Aus moderner Sicht ist diese Zweiteilung Descartes' jedoch unbefriedigend – unter anderem deshalb, weil in ihr kein Platz bleibt für abstrakte Dinge wie Mengen, Zahlen oder Propositionen. Wenn man in einem modernen Lexikon nachschlägt, welche Charakteristika diese dritte mögliche Art von Dingen auszeichnet, stößt man auf Listen wie diese: Abstrakte Dinge sind nicht wahrnehmbar, man kann nicht auf sie zeigen, sie haben keine (physischen) Ursachen und Wirkungen, und sie haben keinen Ort in Raum und Zeit.[2] Einige dieser Charakteristika treffen allerdings nicht nur auf abstrakte Dinge zu. Welche Dinge wahrnehmbar sind und auf welche Dinge man zeigen kann, hängt nicht nur von ihrer Art, sondern – bei physischen Dingen – auch von unserem Wahrnehmungsapparat und deshalb un-

---

*sophy*, Cambridge 1995, S. 603; sowie Ansgar Beckermann, *Analytische Einführung in die Philosophie des Geistes*, Berlin/New York 1998, Kapitel 6.) Im folgenden wird sich aber zeigen, daß jemand, der die These ($PH_2$) ablehnt, nicht wirklich als Physikalist gelten kann.

[2] A. D. Oliver, „abstract entities", in: Honderich (Hg.), *The Oxford Companion to Philosophy*, Oxford 1995, S. 3. In diesem Zusammenhang ist es vielleicht sinnvoll, darauf hinzuweisen, daß Anti-Physikalisten offenbar ganz verschiedene Positionen einnehmen können: Sie können wie Descartes oder die Vitalisten die These vertreten, daß es neben den physischen auch nichtabstrakte nichtphysische Dinge gibt, die den Lauf der Welt mit beeinflussen. Sie können aber auch der Auffassung sein, daß es neben den physischen auch abstrakte Gegenstände gibt. Im Streit um den Physikalismus spielt die Existenz abstrakter Gegenstände erstaunlicherweise jedoch häufig keine besondere Rolle.

ter anderem von der Dimension dieser Dinge ab. Auch Elektronen sind nicht wahrnehmbar; und auf ein Positron zu zeigen, dürfte ebenfalls recht schwer sein. Bleiben also nur die beiden Hauptcharakteristika nichtphysischer Dinge:

– Nichtphysische Dinge haben keinen Ort im Raum und keine Ausdehnung.
– Nichtphysische Dinge haben keine physischen Ursachen und Wirkungen.

Auch diese beiden Merkmale führen jedoch zu unbefriedigenden Ergebnissen:

– Wenn es zu den charakteristischen Merkmalen der nichtphysischen Dinge gehört, keine physischen Ursachen und Wirkungen zu haben, dann zählen Gott, Engel, Cartesische Seelen und der *élan vital* (so wie diese Dinge normalerweise verstanden werden) *nicht* zu den nichtphysischen Dingen.
– Wenn das entscheidende Merkmal des Nichtphysischen ist, keinen Ort im Raum zu haben, dann ist der *élan vital kein* nichtphysisches Ding.
– Und auch wenn beide Merkmale zusammen entscheidend sein sollen, wäre wiederum zumindest der *élan vital kein* nichtphysisches Ding. Außerdem kann man sich eine Menge anderer problematischer Fälle zumindest vorstellen: Astralleiber[3] oder Gespenster wie den Geist in Aladins Wunderlampe.[4]

Diese Probleme sprechen meines Erachtens dafür, als Antwort auf die Frage, was physische Dinge sind, eine radikalere Lösung ins Auge zu fassen – eine Lösung, die auf dem Grundsatz der antiken Atomisten beruht: „Letzten Endes gibt es nur Atome und das Leere." Physisch ist alles, was materiell ist. Und materiell ist alles, was aus den kleinsten Bausteinen der Materie aufgebaut ist – den letzten Elementarteilchen. So verstanden gibt es zwei Arten von physischen Dingen: erstens die von der Physik postulierten

---

[3] „Astralleib oder Ätherleib, in unterschiedlichen (religiösen, philosophischen u. a.) Weltdeutungssystemen die Gestalt der zu den Sternen entrückten Seelen; in der Anthroposophie der ätherisch gedachte Träger des Lebens *im* Körper des Menschen; im Okkultismus ein dem irdischen Leib innewohnender *übersinnlicher* Zweitkörper." (Meyers Lexikonverlag – Hervorhebung vom Verf.)

[4] Wenn Sätze sinnvoll sind wie „Nachdem seine Seele ihn verlassen hatte, schwebte sie noch eine Zeitlang über seinem Körper", müßten nach diesem Kriterium sogar Seelen aus dem Kreis der nichtphysischen Dinge ausgeschlossen werden. Aus dieser Überlegung ergibt sich, daß auch Kims Definition „Alles, was zumindest eine physische Eigenschaft hat, ist ein physisches Ding" das Problem nicht löst. Vgl. Jaegwon Kim, *Philosophy of Mind*, Boulder, Col. 1996, S. 11.

Basisentitäten – die letzten Elementarteilchen – und zweitens alles, was aus diesen Elementarteilchen (und aus nichts sonst) aufgebaut ist: Atomkerne, Atome und Moleküle sowie alle Dinge, die nur aus Atomen und Molekülen bestehen (Regentropfen, Steine und Blumen, aber auch Transistoren, Autos und Computer). Mein Vorschlag ist also, die These ($PH_1$) so zu verstehen:

($PH_1'$) Alle Dinge, die es gibt, sind Elementarteilchen oder Dinge, die vollständig aus Elementarteilchen aufgebaut sind.[5]

Damit kommen wir zur zweiten Frage: Was sind physische Eigenschaften? In seiner kurzen Charakterisierung des Physikalismus schreibt Wayne Davis:

> *Physicalism*. The doctrine that everything is physical. […] Physicalists hold that the real world contains nothing but matter and energy, and that objects have only physical properties, such as spatio-temporal position, mass, size, shape, motion, hardness, electrical charge, magnetism, and gravity.[6]

Sicher wird kaum jemand bestreiten, daß die von Davis angeführten Eigenschaften physische Eigenschaften sind; aber seine Liste ist sicher nicht vollständig. Wenn Gravitation zu den physischen Eigenschaften gehört, dann auch die elektromagnetische, die schwache und die starke Wechselwirkung; wenn Härte dazugehört, dann auch Plastizität usw. Auch hier stellt sich also die Frage: Gibt es ein klares Kriterium, anhand dessen man physische von nichtphysischen Eigenschaften unterscheiden kann?

Bei der Beantwortung dieser Frage scheint es mir sinnvoll, die schon getroffene Unterscheidung zwischen Elementarteilchen auf der einen und aus diesen aufgebauten komplexen physischen Dingen auf der anderen Seite noch einmal aufzugreifen. Und zwar aus zwei Gründen. Erstens, weil komplexe physische Dinge Eigenschaften haben, die Elementarteilchen nicht haben können. (Zu diesen sogenannten systemischen Eigenschaften gehören zum Beispiel die Aggregatzustände. Kein Elementarteilchen – ja nicht einmal ein einzelnes Atom und Molekül – kann gasförmig, flüssig oder fest sein.) Und zweitens, weil es so aussieht, als sei die Anzahl der physischen Eigenschaften, die Elementarteilchen haben können, relativ überschaubar, während die Menge der physischen Eigenschaften komplexer Dinge unbestimmt ist. Die physischen Eigenschaften von Elementarteilchen lassen sich daher in Form einer Liste angeben, was bei den physischen Eigenschaften komplexer Dinge nicht möglich ist. Deshalb schlage ich vor, auf die Frage, was physische Eigenschaften sind, eine zweiteilige Antwort zu geben:

---

[5] Zu dieser Formulierung vgl. Geoffrey Hellman and Frank Thompson, „Physicalism: Ontology, Determination, and Reduction“, in: *Journal of Philosophy* 72 (1975), S. 551–564.

[6] Wayne A. Davis, „physicalism“, in: Ted Honderich (Hg.), *The Oxford Companion to Philosophy*, Oxford 1995, S. 679.

(PE) (a) Zu den physischen Eigenschaften gehören die Basiseigenschaften *raum-zeitlicher Ort*, *Masse*, *elektrische Ladung*[7] und alle Eigenschaften, die aus diesen abgeleitet werden können (*Geschwindigkeit*, *Beschleunigung*, *etc.*).

(b) Die Eigenschaften komplexer Dinge sind physische Eigenschaften, wenn sie auf die physischen Eigenschaften ihrer Teile und auf deren räumliche Anordnung reduziert werden können.

Auf der Grundlage dieser Definition und der vorangegangenen Überlegungen kann die These ($PH_2$) so präzisiert werden:

($PH_2'$) (a) Elementarteilchen haben nur physische Basiseigenschaften.

(b) Alle Eigenschaften komplexer Dinge können auf die physischen Eigenschaften ihrer Teile und auf deren räumliche Anordnung reduziert werden.

## 2

Auch bei dieser Präzisierung bleibt jedoch noch eine Frage offen: Was heißt es, daß eine Eigenschaft *F* eines komplexen physikalischen Gegenstandes (eines Systems) auf die physischen Eigenschaften seiner Teile und auf deren räumliche Anordnung reduziert werden kann?[8]

Für viele Autoren gibt es auf diese Frage nur zwei mögliche Antworten: den Semantischen Physikalismus und die Identitätstheorie. Beiden Positionen zufolge ist die These ($PH_2'$) (b) jedoch nicht haltbar. Und deshalb sind diese Autoren der Auffassung, daß ($PH_2'$) (b) entweder falsch oder zumindest falsch formuliert ist. Dies ist jedoch nicht zwingend. Denn es gibt eine überzeugende Alternative zum Semantischen Physikalismus und zur Identitätstheorie – eine Alternative, die auf C. D. Broads Unterscheidung zwischen *mechanisch erklärbaren* und *emergenten* Eigenschaften zurückgeht.

Broad war dem ontologischen Physikalismus durchaus zugeneigt, auch wenn er nicht alle Thesen dieser Position teilte. Er war ein Anhänger der These ($PH_1$), der zufolge alle Dinge, die es gibt, aus physischen Teilen und nur aus solchen Teilen bestehen. Und er war der Meinung, daß alle Systemeigenschaften eine physische Basis haben. Damit ist folgendes gemeint. Ein komplexes System *S*, das eine Eigenschaft *F* besitzt, besteht aufgrund der These ($PH_1$) aus physischen Bestandteilen $C_1$, …, $C_n$ die auf die Weise *R*

---

[7] Diese Liste ist nicht als vollständige Aufzählung gemeint; falls die Physik weitere Basiseigenschaften entdeckt, müßten diese ebenfalls in die Bedingung (PE) (a) aufgenommen werden.

[8] Zur folgenden Argumentation vgl. Ansgar Beckermann, „Eigenschafts-Physikalismus“, in: *Zeitschrift für philosophische Forschung* 50 (1996), S. 3–25.

räumlich angeordnet sind; dieses System besitzt also die Mikrostruktur [$C_1$, …, $C_n$; $R$]. Broad war nun der Überzeugung, daß es unmöglich ist, daß sich zwei Systeme mit derselben Mikrostruktur in ihren Eigenschaften unterscheiden. Mit anderen Worten, Broad zufolge gilt der Grundsatz:

(*) Wenn *ein* System mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] die Eigenschaft $F$ besitzt, dann gilt dies für alle Systeme mit dieser Mikrostruktur, das heißt, dann ist der Satz (i) „Für alle $x$: wenn $x$ die Mikrostruktur [$C_1$, …, $C_n$; $R$] hat, dann hat $x$ die Eigenschaft $F$" ein wahres Naturgesetz.[9]

Jede Mikrostruktur, die den Satz (i) erfüllt, kann man eine mikrostrukturelle Basis der Systemeigenschaft $F$ nennen. Offenbar gibt es nach Broad für jede Systemeigenschaft $F$ eine mikrostrukturelle Basis. Denn immer wenn ein System die Eigenschaft $F$ hat, hat es eine bestimmte Mikrostruktur, und wegen des Grundsatzes (*) ist diese Mikrostruktur eine mikrostrukturelle Basis für $F$.

Die These, daß jede Systemeigenschaft eine mikrostrukturelle Basis besitzt, ist jedoch nicht identisch mit der These ($PH_2'$) (b). Denn Broad zufolge muß man, wie schon gesagt, zwischen *mechanisch erklärbaren* und *emergenten* Systemeigenschaften unterscheiden. Diese beiden Begriffe definiert Broad in etwa so:[10]

(ME) Eine Eigenschaft $F$ eines komplexen Systems mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] ist genau dann *mechanisch erklärbar*, wenn
  (a) der Satz „Für alle $x$: wenn $x$ die Mikrostruktur [$C_1$, …, $C_n$; $R$] hat, dann hat $x$ die Eigenschaft $F$" ein wahres Naturgesetz ist und wenn
  (b) $F$ (wenigstens im Prinzip) aus der vollständigen Kenntnis all der Eigenschaften deduziert werden kann, die die Komponenten $C_1$, …, $C_n$ isoliert oder in anderen Anordnungen besitzen.

(E) Eine Eigenschaft $F$ eines komplexen Systems mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] ist genau dann *emergent*, wenn
  (a) auf der einen Seite der Satz „Für alle $x$: wenn $x$ die Mikrostruktur [$C_1$, …, $C_n$; $R$] hat, dann hat $x$ die Eigenschaft $F$" ein wahres Naturgesetz ist,
  (b) wenn auf der anderen Seite $F$ aber nicht einmal im Prinzip aus der vollständigen Kenntnis all der Eigenschaften deduziert wer-

[9] Broad war also der Meinung, daß Systemeigenschaften stark über mikrostrukturellen Eigenschaften supervenieren. Allerdings gilt dies natürlich nur für nichtrelationale Systemeigenschaften.

[10] Vgl. Charles D. Broad, *The Mind and Its Place in Nature*, London 1925, S. 61.

den kann, die die Komponenten $C_1$, …, $C_n$ isoliert oder in anderen Anordnungen besitzen.

Allen Systemeigenschaften – den emergenten ebenso wie den mechanisch erklärbaren – ist nach Broad also gemeinsam, daß sie eine mikrostrukturelle Basis besitzen. Die emergenten unterscheiden sich von den mechanisch erklärbaren Systemeigenschaften jedoch dadurch, daß man die letzteren „(wenigstens im Prinzip) aus der vollständigen Kenntnis all der Eigenschaften deduzieren kann, die die Komponenten $C_1$, …, $C_n$ isoliert oder in anderen Anordnungen besitzen“, während dies für die ersteren nicht gilt.

Es ist nicht ganz leicht zu verstehen, wie Broads komplizierte Formel „*F* kann (wenigstens im Prinzip) aus der vollständigen Kenntnis all der Eigenschaften deduziert werden, die die Komponenten $C_1$, …, $C_n$ isoliert oder in anderen Anordnungen besitzen“ genau zu verstehen ist. Mir scheint aber, daß er in etwa folgendes gemeint hat: *F* kann genau dann aus der vollständigen Kenntnis all der Eigenschaften deduziert werden, die die Komponenten $C_1$, …, $C_n$ isoliert oder in anderen Anordnungen besitzen, wenn aus den *allgemeinen*, für Gegenstände mit den *fundamentalen* Eigenschaften der Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen folgt, daß Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle für die Systemeigenschaft *F* charakteristischen Merkmale besitzen.

Insgesamt denke ich daher, daß man die beiden Definitionen (ME) und (E) präziser so formulieren kann:

(ME′) Eine Eigenschaft *F* eines komplexen Systems mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] ist genau dann *mechanisch erklärbar*, wenn
- (a) der Satz „Für alle *x*: wenn *x* die Mikrostruktur [$C_1$, …, $C_n$; *R*] hat, dann hat *x* die Eigenschaft *F*“ ein wahres Naturgesetz ist und wenn
- (b) aus den *allgemeinen*, für Gegenstände mit den *fundamentalen* Eigenschaften der Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen folgt, daß Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle für die Eigenschaft *F* charakteristischen Merkmale besitzen.

(E′) Eine Eigenschaft *F* eines komplexen Systems mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] ist genau dann *emergent*, wenn
- (a) auf der einen Seite der Satz „Für alle *x*: wenn *x* die Mikrostruktur [$C_1$, …, $C_n$; *R*] hat, dann hat *x* die Eigenschaft *F*“ ein wahres Naturgesetz ist,
- (b) wenn auf der anderen Seite aber *nicht* aus den *allgemeinen*, für Gegenstände mit den *fundamentalen* Eigenschaften der Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen folgt, daß Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle für die Eigenschaft *F* charakteristischen Merkmale besitzen.

Meiner Meinung nach ist die so präzisierte Unterscheidung zwischen emergenten und mechanisch erklärbaren Eigenschaften unter anderem deshalb von großer Bedeutung, weil in der Bedingung (b) der Definition (ME′) ein überzeugender und sehr allgemeiner *Realisierungs*- bzw. *Reduktionsbegriff* enthalten ist, den man so formulieren kann:

(R) Die Systemeigenschaft $F$ eines komplexen Systems ist genau dann durch dessen Mikrostruktur [$C_1$, …, $C_n$; $R$] *realisiert* bzw. auf diese Mikrostruktur *reduzierbar* , wenn aus den *allgemeinen*, für Gegenstände mit den *fundamentalen* Eigenschaften der Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen folgt, daß Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] alle für die Systemeigenschaft $F$ charakteristischen Merkmale besitzen.

Dieser auf Broad zurückgehende Reduktionsbegriff hat mindestens drei Vorzüge.

– Er setzt nicht voraus, daß sich Prädikate, die Systemeigenschaften ausdrücken, mit Hilfe von Ausdrücken definieren lassen, die sich auf Mikrostrukturen beziehen. Damit vermeidet er die Probleme des Semantischen Physikalismus.
– Er ist mit der Multirealisierbarkeit von Systemeigenschaften vereinbar, da ihm zufolge Eigenschaftsreduktionen auch ohne die Existenz von Brückengesetzen möglich sind. Damit vermeidet er die Probleme der Identitätstheorie.
– Er wird allen Intuitionen gerecht, die normalerweise mit der Idee von Eigenschaftsreduktionen verbunden sind.

Nehmen wir als Beispiele die Eigenschaften, flüssig bzw. durchsichtig zu sein – zwei Makroeigenschaften physischer Systeme, von denen wohl jeder annimmt, daß sie auf die Mikrostrukturen dieser Systeme reduzierbar sind. Warum ist das so? Bleiben wir zunächst bei der Eigenschaft, flüssig zu sein. Flüssigkeiten unterscheiden „sich von Gasen dadurch, daß ihr Volumen (weitgehend) druckunabhängig (inkompressibel) ist, von festen Körpern dadurch, daß ihre Form veränderlich ist und sich der Form des jeweiligen Gefäßes anpaßt.“[11] Dies liegt auf der einen Seite daran, daß bei Flüssigkeiten – anders als bei Gasen – die Molekel so dicht wie möglich ‚gepackt‘ sind. Enger ‚zusammenrücken‘ können sie nicht (oder nur bei sehr großem Kraftaufwand), weil die Abstoßungskräfte zwischen den Molekeln dies nicht zulassen. Auf der anderen Seite sind die Molekel in Flüssigkeiten aber gegeneinander verschiebbar, sie können sozusagen frei übereinanderrollen, während die Molekel fester Körper durch die Kräfte, die sie aufeinander ausüben, an ihren relativen Position festgezurrt sind. Die Molekel

[11] Art. „Flüssigkeit“, Meyers Lexikonverlag.

eines festen Körpers können sich daher nur im Verband bewegen. Der ganze Körper bewegt sich, die relative Position seiner Molekel bleibt dabei unverändert, und deshalb behält der Körper seine Form. Offenbar ist es keine Frage, daß sich die Kräfte, die Molekel unter bestimmten Bedingungen aufeinander ausüben, aus den allgemeinen für sie geltenden Naturgesetzen ergeben. Also ergibt sich aus diesen Naturgesetzen auch, ob ein Stoff unter diesen Bedingungen flüssig ist oder nicht. Er ist flüssig, wenn die anziehenden Kräfte groß genug sind, um die Molekel bis auf einen Mindestabstand zusammenrücken zu lassen, aber nicht groß genug, um sie an ihren relativen Positionen festzuzurren.

Bei der Eigenschaft, durchsichtig zu sein, liegen die Dinge ganz ähnlich. Eine Glasscheibe ist durchsichtig, da sie Licht (Photonen) des sichtbaren Spektrums gleichmäßig und fast vollständig durchläßt. Auch hier scheint klar, daß dies an der physikalischen Struktur der beteiligten Moleküle und an deren Anordnung liegt. Im Einzelfall mag es schwierig sein, zu zeigen, daß aus den allgemeinen Naturgesetzen folgt, daß Moleküle von einer bestimmten physikalischen Beschaffenheit und in einer bestimmten räumlichen Anordnung (fast) keine Photonen absorbieren. Aber die meisten von uns würde es sicher sehr wundern, wenn es nicht so wäre. Außerdem hätte es schwerwiegende theoretische Folgen, wenn es sich anders verhielte. Auf diese Folgen werde ich gleich zu sprechen kommen.

## 3

Bis jetzt haben wir uns hauptsächlich mit der Frage beschäftigt, wie die Teilthesen des ontologischen Physikalismus genau zu verstehen sind. In diesem Abschnitt soll nun das im Titel angekündigte Argument zur Sprache kommen, das für die Richtigkeit dieser Thesen spricht – das heißt genauer: für die Richtigkeit der These ($PH_2'$) (b).

Dieses Argument geht von der Frage aus, was es eigentlich bedeuten würde, wenn diese These falsch wäre. Nach den bisherigen Überlegungen besagt die These ($PH_2'$) (b), daß alle Systemeigenschaften auf die physikalischen Mikrostrukturen der betreffenden Systeme reduzierbar sind. Wenn diese These falsch wäre, würde das also heißen, daß zumindest einige Systemeigenschaften nicht auf diese Weise reduziert werden können – bzw. in der Terminologie Broads: daß zumindest einige Systemeigenschaften nicht mechanisch erklärbar, sondern emergent sind. Die Frage ist also: Was würde es bedeuten, wenn es emergente Systemeigenschaften gäbe? Was würde es zum Beispiel bedeuten, wenn die Eigenschaft, magnetisch zu sein, emergent wäre?

Vorab scheinen zwei Dinge klar zu sein. *Erstens*: Zu den charakteristischen Merkmalen der Eigenschaft, magnetisch zu sein, gehört, daß sich

magnetische Dinge (bzw. Dinge in der Umgebung magnetischer Dinge) auf spezifische Weise verhalten:

- Magnetische Dinge ziehen Eisenfeilspäne in ihrer Umgebung an.
- Eine Kompaßnadel in der Nähe eines magnetischen Dings zeigt in dessen Richtung.
- Magnetische Dinge induzieren einen Strom in Kreisleitern, durch die sie geführt werden.
- Magnetische Dinge magnetisieren nichtmagnetische Eisenstücke in ihrer Umgebung. Etc.

*Zweitens*: Die spezifischen Verhaltensweisen, die für magnetische Dinge charakteristisch sind, betreffen nicht nur *makroskopische* Dinge, sondern auch deren *mikroskopische* Teile.

- Wenn sich eine Kompaßnadel in der Nähe eines magnetischen Dings in dessen Richtung dreht, dann deshalb, weil *alle Moleküle und Atome*, aus denen die Kompaßnadel besteht, entsprechende Bewegungen ausführen.
- Wenn in einer Spule, durch die ein magnetischer Gegenstand geführt wird, ein Strom fließt, dann deshalb, weil sich *die Elektronen* in dieser Spule auf spezifische Weise bewegen.

Magnetische Dinge bewirken makroskopische Verhaltensweisen also, indem sie ein entsprechendes Verhalten der mikroskopischen Teile der jeweiligen Gegenstände hervorrufen.

Was folgt aus diesen beiden Punkten, wenn wir annehmen, die Eigenschaft, magnetisch zu sein, sei emergent? Erstens natürlich, daß die Verhaltensweisen, die für magnetische Dinge charakteristisch sind, *nicht* auf die allgemeinen Naturgesetze zurückgeführt werden können, die für die physischen Teile dieser Dinge gelten. Mit anderen Worten: Wenn die Eigenschaft, magnetisch zu sein, emergent ist, ergibt sich weder die Tatsache, daß in einer Spule, durch die ein magnetischer Gegenstand *S* geführt wird, ein Strom fließt, noch die Tatsache, daß sich eine Kompaßnadel in die Richtung von *S* dreht, aus den Naturgesetzen, auf denen das Verhalten der physischen Komponenten von *S* im allgemeinen beruht.

Doch damit noch nicht genug. Da das Fließen des Stromes in der Spule auf der Bewegung bestimmter Elektronen beruht und da sich das Drehen der Kompaßnadel aus den Bewegungen der Atome und Moleküle ergibt, aus denen diese Nadel besteht, ergibt sich die weitere Konsequenz: Falls die Eigenschaft, magnetisch zu sein, emergent ist, ergeben sich nicht einmal die Bewegungen der Elektronen in der Spule bzw. die Bewegungen der Atome und Moleküle, aus denen die Kompaßnadel besteht, aus den für die physischen Komponenten von *S* geltenden Naturgesetzen.

Die äußerst unliebsame Konsequenz wäre also: Wenn die Eigenschaft, magnetisch zu sein, emergent wäre, wären die grundlegenden Gesetze der Elementarteilchenphysik auf beunruhigende Weise *unvollständig*. In jedem Fall, in dem die Bewegungen der Elektronen in einer Spule dadurch bewirkt werden, daß ein magnetischer Gegenstand durch diese Spule geführt wird, und in jedem Fall, in dem sich die Atome und Moleküle, aus denen eine Kompaßnadel besteht, deshalb in Bewegung setzen, weil sich diese Nadel auf einen in der Nähe befindlichen magnetischen Gegenstand hin ausrichtet, ließen sich diese Bewegungen *nicht* auf die grundlegenden Gesetze der Elementarteilchenphysik zurückführen. Da alle Bewegungsveränderungen letzten Endes durch entsprechende Kräfte hervorgerufen werden, kann man dies auch so ausdrücken: Wenn die Eigenschaft, magnetisch zu sein, emergent wäre, würde das Verhalten der Elektronen in einer Spule und das Verhalten der Atome und Moleküle einer Kompaßnadel zumindest in manchen Fällen durch Kräfte bestimmt, die sich nicht aus den grundlegenden Gesetzen der Elementarteilchenphysik ableiten lassen.

Und dieses Ergebnis läßt sich offenbar verallgemeinern: Jede emergente Eigenschaft *F*, die zumindest zum Teil dadurch charakterisiert ist, daß sich Gegenstände, die diese Eigenschaft besitzen, auf eine bestimmte Art und Weise verhalten bzw. daß Gegenstände mit dieser Eigenschaft das Verhalten anderer Gegenständen kausal beeinflussen, führt zu einer Lücke in der Elementarteilchenphysik. Denn daß *F* emergent ist, impliziert, daß das Verhalten der physischen Komponenten der Gegenstände, die *F* besitzen, bzw. der Gegenstände, die mit solchen Gegenständen interagieren, zumindest in manchen Fällen durch Kräfte bestimmt wird, die sich nicht aus den Gesetzen der Elementarteilchenphysik ergeben. Zumindest gilt dies dann, wenn das Oberflächenverhalten, das durch *F* verursacht wird, unmittelbar mit dem Verhalten der physischen Komponenten der beteiligten Gegenstände zusammenhängt. Falls es emergente Eigenschaften gibt, ist die Elementarteilchenphysik also unvollständig. In diesem Fall läßt sich nicht alles, was auf der Ebene der Elementarteilchen passiert, mit ihren Gesetzen erklären.

Allerdings gibt es vielleicht doch noch einen Weg, diese unliebsame Konsequenz zu vermeiden. Aufgrund der Broadschen Definitionen haben nämlich, wie wir schon gesehen haben, auch alle emergenten Eigenschaften eine mikrostrukturelle Basis. Das heißt, nach Broad gibt es für jede emergente Eigenschaft *F* eine Menge ***M*** von Mikrostrukturen, für die gilt:

1. Ein System *x* hat *F* nur dann, wenn es eine der Mikrostrukturen besitzt, die zu ***M*** gehören;
2. Für alle Elemente $M_i$ von ***M*** gilt: Wenn *x* die Mikrostruktur $M_i$ besitzt, dann hat *x F*.

Auch wenn die Eigenschaft, magnetisch zu sein, emergent ist, kann das System *S* diese Eigenschaft daher nur besitzen, wenn es eine Mikrostruktur [$C_1$, …, $C_n$; *R*] besitzt, für die der Satz „Für alle *x*: wenn *x* die Mikrostruktur [$C_1$, …, $C_n$; *R*] hat, dann ist *x* magnetisch" ein wahres Naturgesetz darstellt.

Wenn das so ist, ist es jedoch nicht nötig, zur Erklärung der Bewegung der Elektronen in der Spule, durch die *S* geführt wird, und der Bewegungen der Atome und Moleküle der Kompaßnadel in der Nähe von *S* die Mikroebene zu verlassen. Denn alles, was man darauf zurückführen kann, daß *S* magnetisch ist, kann man offenbar ebenso gut erklären, indem man darauf verweist, daß *S* aus den Komponenten $C_1$, …, $C_n$ besteht, die auf die Weise *R* angeordnet sind. Mit anderen Worten: Wenn Broad recht hat, gibt es für alles, was dadurch bewirkt wird, daß ein Gegenstand eine emergente Eigenschaft hat, auch eine Erklärung auf der Mikroebene. Anders als bisher behauptet, scheint die Existenz emergenter Eigenschaften also nicht die Unvollständigkeit der Elementarteilchenphysik zu implizieren.

Mit diesem Einwand würde der entscheidende Punkt jedoch gerade verfehlt. Denn das beunruhigende Ergebnis der bisherigen Überlegungen ist nicht, daß die Existenz emergenter Eigenschaften die Existenz von Wirkungen auf der Ebene der Elementarteilchen impliziert, für die es auf dieser Ebene selbst keine Erklärungen gibt, sondern daß die Existenz emergenter Eigenschaften die Existenz von Wirkungen auf der Ebene der Elementarteilchen impliziert, die sich nicht aus den *allgemeinen Gesetzen der Elementarteilchenphysik* ergeben. Natürlich kann man dem Broadschen Ansatz zufolge die Bewegungen der Elektronen in der Spule und die Bewegungen der Atome und Moleküle der Kompaßnadel darauf zurückführen, daß *S* aus den Komponenten $C_1$, …, $C_n$ besteht, die auf die Weise *R* angeordnet sind. Wenn die Eigenschaft, magnetisch zu sein, emergent ist, kann jedoch *diese Tatsache selbst* – die Tatsache, daß die auf die Weise *R* angeordneten physischen Teile von *S* ebendiese Wirkungen haben – ihrerseits nicht aus den allgemeinen Gesetzen der Elementarteilchenphysik abgeleitet werden. Wenn die Eigenschaft, magnetisch zu sein, emergent ist, handelt es sich hier um ein theoretisch nicht erklärbares *factum brutum*. Daß Komponenten der Art $C_1$, …, $C_n$, die auf die Weise *R* angeordnet sind, die genannten Wirkungen haben, ist in diesem Fall ein nicht weiter ableitbares, letztes Gesetz (in Broads Worten: „an unique and ultimate law") – ein Gesetz, von dem wir auch nur aufgrund von unmittelbarer Beobachtung wissen können, daß es besteht.

Wenn zuvor gesagt wurde, daß die Existenz emergenter Eigenschaften in gewisser Weise die Unvollständigkeit der Elementarteilchenphysik zur Folge hätte, ist damit also folgendes gemeint. Wenn es emergente Eigenschaften gäbe, dann wären die *grundlegenden Gesetze* der Elementarteil-

chenphysik *nicht allgemein*. Dann ließe sich nicht alles, was auf der Ebene der Elementarteilchen passiert, mit Hilfe *dieser* Gesetze erklären. Oder anders ausgedrückt: Dann bestünde die Elementarteilchenphysik aus einer kleinen Zahl von Grundgesetzen und einer unüberschaubaren Zahl von Ausnahmeregeln. Das wäre in etwa so, als würde die Gravitationskraft, die zwei Körper aufeinander ausüben, zwar in den meisten Fällen dem Gravitationsgesetz

$$F = \frac{m_1 . m_2}{r^2}$$

entsprechen, aber eben nicht immer – zum Beispiel weil im Fall $m_1 = 1$, $m_2 = 10$ und $r = 1$ diese Kraft nicht 10, sondern nur 7 Newton beträgt; weil im Fall $m_1 = 12$, $m_2 = 16$ und $r = 8$ diese Kraft nicht 3, sondern 4 Newton beträgt; und weil im Fall $m_1 = 45$, $m_2 = 10$ und $r = 15$ diese Kraft nicht 2, sondern 212 Newton beträgt.

Es ist natürlich nicht ausgeschlossen, daß die Dinge auf der Ebene der Elementarteilchen tatsächlich so liegen, daß Ereignisse auf dieser Ebene zwar in den meisten Fällen mit Hilfe einiger allgemeiner Grundgesetze, in einer ganzen Reihe von Einzelfällen jedoch nur mit Hilfe von Ausnahmeregeln erklärt werden können, die jeweils nur auf einen Fall zutreffen. Ich sehe aber keinen Grund für die Annahme, daß es tatsächlich so ist. Und ich denke, daß viele mit mir die Auffassung teilen, daß es höchst ungewöhnlich wäre, wenn die Elementarteilchenphysik tatsächlich in diesem Sinne ‚inhomogen' wäre. Wenn es emergente Eigenschaften gäbe, müßte dies jedoch der Fall sein. Das heißt, man hat nur die Wahl zwischen der Annahme der Existenz emergenter Eigenschaften und der Annahme, daß es sich bei der Elementarteilchenphysik um eine ‚homogene' Wissenschaft handelt, daß auf der Ebene der Elementarteilchen sozusagen alles mit rechten Dingen zugeht. Mir scheint die zweite Annahme plausibler. Das heißt, ich gehe hier davon aus, daß die folgenden beiden Prinzipien zutreffen.

1. Es gibt ein System von *allgemeinen grundlegenden* Naturgesetzen, das ausreicht, das gesamte Verhalten aller Elementarteilchen zu erklären (soweit es überhaupt erklärbar ist).
2. Dieses System *enthält keine Ausnahmegesetze*, in denen festgestellt wird, daß sich die Elementarteilchen, wenn sie in ganz bestimmte räumliche Konstellationen kommen, anders verhalten, als dies aufgrund der allgemeinen grundlegenden Naturgesetze zu erwarten wäre.

Wenn diese Prinzipien zutreffen, kann es aber keine emergenten Eigenschaften geben. Und das bedeutet auch: Wenn diese Prinzipien zutreffen, muß die These ($PH_2'$) (b) wahr sein.

# 2

# Die reduktive Erklärbarkeit des phänomenalen Bewusstseins – C.D. Broad zur Erklärungslücke[*1]

## I.

Zu Beginn des 20. Jahrhunderts war die Frage, ob Leben rein mechanisch erklärt werden könne, noch genau so heiß umstritten wie das Leib-Seele-Problem heute. Zwei Parteien standen sich unversöhnlich gegenüber. Auf der einen Seite die *Biologischen Mechanisten* mit der Auffassung, daß die für Lebewesen charakteristischen Eigenschaften (Stoffwechsel, Fortpflanzung, Wahrnehmung, zielgerichtetes Verhalten, Morphogenese) genauso mechanisch erklärt werden können wie das Verhalten einer Uhr, das sich mit physikalischer Zwangsläufigkeit aus den Eigenschaften und der Anordnung ihrer Zahnräder, Federn und Gewichte ergibt. Auf der anderen Seite die *Substanz-Vitalisten*, die die entgegengesetzte Meinung vertraten, Leben könne nur durch die Annahme einer nichtphysischen Substanz erklärt werden – einer Entelechie oder eines *élan vital*. Als Broad in den frühen zwanziger Jahren seine Überlegungen zum Begriff der Emergenz entwickelte, verfolgte er unter anderem das Ziel, Raum für eine dritte Position zwischen diesen beiden Extremen zu schaffen – eine Position, die er *Emergenten Vitalismus* nannte.

Broads erster Schritt bestand darin, darauf aufmerksam zu machen, daß das Problem des Vitalismus nur der Spezialfall eines sehr viel generelleren Problems ist – des Problems, welche Beziehung zwischen den *Makroeigenschaften* eines komplexen Systems und den *Eigenschaften und der Anordnung seiner physischen Teile* besteht.[2] Im Hinblick auf diese Frage gibt es im Prinzip nur zwei mögliche Antworten:

---

* Erstveröffentlichung in: M. Pauen und A. Stephan (Hg.) *Phänomenales Bewusstsein*. Paderborn: mentis Verlag 2002, 122–147.

1 Bei diesem Aufsatz handelt es sich um die deutsche Fassung des Artikels Beckermann 2000. Ich möchte Andreas Hüttemann danken, der mich durch sein Nachfragen dazu gebracht hat, Broads *The Mind and Its Place in Nature* noch einmal noch gründlicher durchzuarbeiten. Dank schulde ich auch Antonia Barke und Christian Nimtz für ihre hilfreichen Anmerkungen zu einer früheren Fassung dieses Aufsatzes.

2 Broad spricht statt von den Makroeigenschaften oft nur spezieller vom Makro*verhalten* komplexer Gegenstände. Dies liegt daran, daß er der Meinung war, daß nur solche Eigenschaften mechanisch erklärbar sein können, für die es

1. Die Makroeigenschaft *F* eines komplexen Systems *S* läßt sich *nicht* allein aus den Eigenschaften und der Anordnung der physischen Teile von *S* erklären; *F* kann vielmehr nur durch die Annahme erklärt werden, daß *S* eine weitere *nichtphysische* Komponente enthält, die in allen Systemen vom Typ *S* vorhanden ist und in allen anderen Systemen fehlt.

Antworten dieser Art nennt Broad *Komponententheorien*. Die andere Möglichkeit ist:

2. Die Makroeigenschaft *F* des Systems *S* läßt sich sehr wohl aus den Eigenschaften und der Anordnung seiner physischen Teile erklären.

In diesem Fall muß man Broad zufolge jedoch zwei weitere Möglichkeiten unterscheiden. Auch wenn sich die Makroeigenschaft *F* aus den Eigenschaften und der Anordnung der physischen Teile von *S* erklären läßt, kann *F* immer noch *reduktiv erklärbar*[3] oder *emergent* sein. Vertreter der Theorie der reduktiven Erklärbarkeit und Vertreter der Emergenztheorie sind sich also einig in der Ablehnung der These,

> that there need be any peculiar *component* which is present in all things that behave in a certain way and is absent from all things which do not behave in this way. [Both say] that the components may be exactly alike in both cases, and [they try] to explain the difference of behaviour wholly in terms of difference of structure. (Broad 1925, 58 f.)

Die Theorie der reduktiven Erklärbarkeit und die Emergenztheorie unterscheiden sich jedoch grundsätzlich in der Antwort auf die Frage, *auf welche Weise* das Verhalten der Komponenten die Makroeigenschaften komplexer Gegenstände erklärt.

> On [the theory of emergence] the characteristic behaviour of the whole *could* not, even in theory, be deduced from the most complete knowledge of the behaviour of its components, taken separately or in other combinations, and of their proportions and arrangements in this whole. (Broad 1925, 59)

---

eine behaviorale Analyse gibt. Von ihm so genannte ‚pure qualities', die nicht behavioral analysiert werden können, sind Broad zufolge auf jeden Fall emergent. Darauf werde ich im Abschnitt 2 zurückkommen.

[3] Broad spricht nicht von reduktiver, sondern von *mechanischer* Erklärbarkeit. Er unterscheidet jedoch zwischen *Mechanismus* und *Reinem Mechanismus*. Der zweiten Position zufolge bedeutet ‚mechanisch erklärbar' in etwa ‚explainable just by reference to the laws of classical mechanics', der ersten Position zufolge dagegen nur ‚explainable by reference to all general chemical, physical and dynamical laws'; vgl. 1925, 46. Meistens verwendet Broad den Ausdruck ‚mechanisch erklärbar' in diesem schwächeren Sinn. Um Mißverständnissen vorzubeugen, scheint es mir deshalb besser, statt dessen den Ausdruck ‚reduktiv erklärbar' zu verwenden.

Welche Eigenschaften komplexer Systeme in diesem Sinne als emergent zu gelten haben, war schon zu Broads Zeiten äußerst umstritten. Er selbst war aber offenbar der Auffassung, daß z. B. das Verhalten chemischer Verbindungen in dem von ihm erläuterten Sinne emergent ist. Zumindest war er der Meinung,

> that, so far as we know at present, the characteristic behaviour of Common Salt cannot be deduced from the most complete knowledge of the properties of Sodium in isolation; or of Chlorine in isolation; or of other compounds of Sodium, such as Sodium Sulphate, and of other compounds of Chlorine, such as Silver Chloride. (Broad 1925, 59)

Vertreter der Theorie der reduktiven Erklärbarkeit sahen das jedoch ganz anders. Denn diese Theorie kennzeichnet Broad so:

> On [the theory of reductive explainability] the characteristic behaviour of the whole is not only completely *determined by* the nature and arrangement of its components; in addition to this it is held that the behaviour of the whole could, in theory at least, be *deduced* from a sufficient knowledge of how the components behave in isolation or in other wholes of a simpler kind. (Broad 1925, 59)

Maschinen sind Broad zufolge die besten Beispiele für komplexe Gegenstände, deren Verhalten vollständig reduktiv erklärbar ist. Bei Uhren etwa gibt es sicher keinen Grund für die Annahme, daß ihr Verhalten auf einer besonderen nichtphysischen Komponente beruht, die in Uhren und nur in Uhren vorkommt. Komponententheorien sind für die Erklärung des Verhaltens von Uhren absolut unangemessen. Es gibt aber auch keinen Grund für die Annahme, daß das Verhalten von Uhren emergent wäre. Offenbar kann man dieses Verhalten vollständig aus der spezifischen Anordnung der Federn, Zahnräder und Gewichte sowie aus den allgemeinen Gesetzen der Mechanik ableiten, die für alle materiellen Gegenstände und nicht nur für die Komponenten von Uhren gelten.

Grundsätzlich kann man den Unterschied zwischen Emergenztheorie und Theorie der reduktiven Erklärbarkeit Broad zufolge deshalb so erläutern:

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C). The [theory of reductive explainability] rejects the last clause of this assertion. (Broad 1925, 61)

Zwei Dinge sind hier entscheidend:

1. Beiden – emergenten und reduktiv erklärbaren – Eigenschaften ist *gemeinsam*, daß sie nomologisch von den jeweiligen Mikrostrukturen der entsprechenden Systeme abhängen. Wenn ein System *S* aus den Teilen $C_1, \ldots, C_n$ besteht, die in der Weise *R* angeordnet sind, kurz: wenn *S* die Mikrostruktur $[C_1, \ldots, C_n; R]$ besitzt, gilt also: Der Satz

   „Alle Systeme mit der Mikrostruktur $[C_1, \ldots, C_n; R]$ haben die Makroeigenschaft F“

   ist ein *wahres Naturgesetz* – unabhängig davon, ob es sich bei *F* um eine emergente oder um eine reduktiv erklärbare Eigenschaft handelt.[4]

2. Reduktiv erklärbare Eigenschaften können darüber hinaus (zumindest im Prinzip) aus der *vollständigen Kenntnis* all der Eigenschaften *abgeleitet* werden, die die Komponenten der entsprechenden Systeme *isoliert* oder *in anderen Anordnungen* haben; bei emergenten Eigenschaften ist dies nicht möglich.

Broads Begriffe der reduktiven Erklärbarkeit und der Emergenz kann man daher so zusammenfassen:

(RE) Die Makroeigenschaft *F* eines komplexen Systems *S* mit der Mikrostruktur $[C_1, \ldots, C_n; R]$ ist genau dann *reduktiv erklärbar*, wenn *F* (zumindest im Prinzip) aus der vollständigen Kenntnis all der Eigenschaften abgeleitet werden kann, die die Komponenten $C_1, \ldots, C_n$ isoliert oder in anderen Anordnungen besitzen.

(E) Die Makroeigenschaft *F* eines komplexen Systems *S* mit der Mikrostruktur $[C_1, \ldots, C_n; R]$ ist genau dann *emergent*, wenn folgendes gilt:

   (a) Der Satz „Alle Systeme mit der Mikrostruktur $[C_1, \ldots, C_n; R]$ haben die Eigenschaft *F*“ ist ein wahres Naturgesetz, aber

   (b) *F* kann nicht (nicht einmal im Prinzip) aus der vollständigen Kenntnis all der Eigenschaften abgeleitet werden, die die Komponenten $C_1, \ldots, C_n$ isoliert oder in anderen Anordnungen besitzen.

Worauf Broad mit diesen Definitionen hinauswill, scheint im Prinzip ziemlich klar. Aber warum wählt er die komplizierte Formulierung „from the

[4] Sowohl emergente als auch reduktiv erklärbare Eigenschaften *supervenieren* also nomologisch über mikrostrukturellen Eigenschaften. Dies ist offenbar der Grund dafür, daß Broad zufolge *beide* Arten von Eigenschaften durch Bezugnahme auf die Mikrostruktur der betreffenden Systeme erklärt werden können. Dabei geht Broad allerdings von einem recht schwachen Erklärungsbegriff aus.

most complete knowledge of the properties of [the components $C_1$, …, $C_n$] *in isolation or in other wholes*"?

Zunächst einmal war sich Broad offenbar darüber im Klaren, daß der Begriff einer emergenten Eigenschaft aus trivialen Gründen leer wäre, wenn man bei der Ableitung der Makroeigenschaft eines Systems *alle* Eigenschaften seiner Teile zuließe. Etwa 20 Jahre nach der Veröffentlichung von *The Mind and Its Place in Nature* haben Hempel und Oppenheim dieses Problem – unter Bezugnahme auf eine Bemerkung Grellings – so auf den Punkt gebracht:

> If a characteristic of a whole is counted as emergent simply if its occurrence cannot be inferred from a knowledge of all the properties of its parts, then, as Grelling has pointed out, no whole can have any emergent characteristics. Thus … the properties of hydrogen include that of forming, if suitably combined with oxygen, a compound which is liquid, transparent, etc. Hence the liquidity, transparence, etc. of water *can* be inferred from certain properties of its chemical constituents. (Hempel/Oppenheim 1948, 149)

Wenn man diese Konsequenz vermeiden will, müssen solche Ableitungen verhindert werden. Und Broads Formulierung dient genau diesem Zweck. Mit ihr will er sicherstellen, daß bei der Ableitung der Makroeigenschaft eines Systems aus den Eigenschaften seiner Teile *nicht* auf Eigenschaften wie die von Hempel und Oppenheim genannten zurückgegriffen werden darf. Hätte Broad dieses Problem aber nicht auch eleganter lösen können? Klar ist, daß bei der Ableitung der Makroeigenschaft $F$ eines Systems aus den Eigenschaften und der räumlichen Anordnung seiner Teile $C_1$, …, $C_n$ nicht auf solche ‚*ad-hoc*'-Eigenschaften der Teile Bezug genommen werden darf wie die, daß Dinge der Art $C_1$, …, $C_n$, wenn sie auf die Weise $R$ angeordnet sind, einen komplexen Gegenstand ergeben, der die Eigenschaft $F$ besitzt. Die Frage ist nur, wie dies erreicht werden kann, ohne daß zugleich Eigenschaften ausgeschlossen werden, auf die zurückzugreifen in diesem Zusammenhang legitim wäre.

Bei der Beantwortung dieser Frage ist es hilfreich, den Blick von den Eigenschaften abzuwenden und statt dessen nach den *Gesetzen* zu fragen, auf die bei der Ableitung der Makroeigenschaft eines Systems zurückgegriffen werden darf. Bei diesen Gesetzen ergibt sich das Problem einer möglichen Trivialisierung des Emergenzbegriffs nämlich in analoger Weise. Nicht nur bei reduktiv erklärbaren, auch bei emergenten Eigenschaften ist, wie wir schon gesehen hatten, das Gesetz

(*) Alle Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] haben die Makroeigenschaft $F$

ein wahres Naturgesetz. Wenn man dieses Gesetz bei der Ableitung von $F$ verwenden dürfte, gäbe es daher ebenfalls keine emergenten Eigenschaften.

D. h., Hempel und Oppenheim hätten ihren Punkt auch so formulieren können:

> It is a true law of nature that, if suitably combined with oxygen, hydrogen forms a compound which is liquid, transparent, etc. Hence the liquidity, transparence, etc. of water *can* be derived by means of the laws of nature.[5]

Broad muß somit auch die Bezugnahme auf Gesetze wie das Gesetz (*) verhindern. Und dies war ihm durchaus klar, wie sich z. B. aus der folgenden Passage ergibt, in der es noch einmal um das Verhalten von Uhren geht.

> We know perfectly well that the behaviour of a clock can be deduced from the particular arrangement of springs, wheels, pendulum, etc., in it, and from *general laws of mechanics and physics which apply just as much to material systems which are not clocks*. (Broad 1925, 60 – Herv. d. Vf.)

Ganz offensichtlich war Broad der Meinung, daß man bei dem Versuch, die Makroeigenschaften eines Systems aus den Eigenschaften und der räumlichen Anordnung seiner Teile abzuleiten, nur *allgemeine Gesetze* verwenden darf – Gesetze, die für die Teile eines komplexen Systems *völlig unabhängig von ihrer spezifischen Anordnung* gelten. Auf die Frage „Auf welche Eigenschaften darf bei einer solchen Ableitung zurückgegriffen werden?“ gibt es daher eine naheliegende Antwort: „Auf genau die Eigenschaften, die in diesen allgemeinen Gesetzen erwähnt werden.“ Wenn das so ist, könnte man Broads Klausel aber durch die folgende ersetzen:

> wenn $F$ mit Hilfe allgemeiner Naturgesetze aus den Eigenschaften der Teile $C_1, \ldots, C_n$ abgeleitet werden kann, die in diesen Gesetzen erwähnt werden.

Letzten Endes führt diese Überlegung aber zu einer noch radikaleren Vereinfachung. Denn offenkundig ist die Bezugnahme auf zulässige Eigenschaften in der verbesserten Formulierung völlig überflüssig; es reicht aus, die *Gesetze* anzuführen, die bei den ins Auge gefaßten Ableitungen verwendet werden dürfen. Meiner Meinung nach sollte man Broads Klausel deshalb so umformulieren:

> wenn $F$ aus den *allgemeinen* Naturgesetzen abgeleitet werden kann, die für Teile der Art $C_1, \ldots, C_n$ gelten.

Nachdem dieser Punkt grundsätzlich geklärt ist, bleibt jedoch noch eine interessante Detailfrage: Was steckt eigentlich dahinter, wenn Broad schreibt, daß wir nicht nur untersuchen müssen, welche Eigenschaften die Teile ei-

---

[5] Daß sich Broad über *beide* Möglichkeiten der Trivialisierung des Emergenzbegriffs im Klaren war, zeigt sich unter anderem auf den Seiten 65 f. von Broad 1925.

nes Systems *in Isolation* besitzen, sondern auch, wie sie sich *in anderen Anordnungen* verhalten?

Wir hatten gesehen, daß sich nach Broad reduktiv erklärbare von emergenten Eigenschaften dadurch unterscheiden, daß jene aus den allgemeinen Naturgesetzen abgeleitet werden können, die für Teile der Art $C_1$, …, $C_n$ gelten. Aber wie kann man diese allgemeinen Naturgesetze herausfinden? Broad war offenbar der Meinung, daß man in diesem Zusammenhang immer zwei Dinge tun muß: Erstens muß man beobachten, wie sich die Teile in Isolation verhalten; und zweitens muß man untersuchen, wie sie sich als Teile anderer Systeme verhalten. Warum sind beide Schritte nötig?

Offenbar hatte Broad unter anderem das dynamische Verhalten von Systemen im Auge, auf die unterschiedliche Kräfte einwirken.[6] Wenn man herausfinden will, welche Gesetze in Fällen dieser Art gelten, ist es sinnvoll, zunächst das Verhalten von Gegenständen zu untersuchen, auf die nur eine Kraft wirkt. Auf diese Weise kommt man zu dem zentralen zweiten Newtonschen Gesetz $F = m \cdot a$. Wenn man aber wissen will, wie sich ein Gegenstand *im allgemeinen* verhält, d. h. wie er sich verhält, wenn gleichzeitig mehrere Kräfte auf ihn einwirken, reicht die Kenntnis dieses Gesetzes allein nicht aus. Vielmehr benötigen wir für den allgemeinen Fall auch ein Gesetz, das uns sagt, wie verschiedene Kräfte zusammenwirken – das Gesetz der Vektoraddition von Kräften. Und dieser Punkt läßt sich nach Broad verallgemeinern. Grundsätzlich benötigt man immer zwei Arten von Gesetzen: (a) Gesetze, aus denen hervorgeht, wie jeder einzelne Faktor das Verhalten eines Gegenstandes beeinflußt, und (b) Gesetze, aus denen hervorgeht, welches Resultat sich ergibt, wenn verschiedene Einflußfaktoren gleichzeitig auf einen Gegenstand einwirken. Gesetze der zweiten Art nennt Broad ‚laws of composition'. Und er betont mit Nachdruck ihre Unerläßlichkeit:

> It is clear that in *no* case could the behaviour of a whole composed of certain constituents be predicted *merely* from a knowledge of the properties of these constituents, taken separately, and of their proportions and arrangements in the particular complex under consideration. Whenever this *seems* to be possible it is because we are using a suppressed premise which is so familiar that it has escaped our notice. The suppressed premise is the fact that we have examined other complexes in the past and have noted their behaviour; that we have found a general law connecting the behaviour of these wholes with that which their constituents would show in isolation; and that we are assuming that this law of composition will hold also of the particular complex whole at present under consideration. (Broad 1925, 63)

[6] Vgl. Broad 1925, 62 und 63 f.

An dieser Stelle gibt es jedoch eine Unklarheit. So wie Broad sich in dieser Passage ausdrückt, erweckt er den Anschein, als würden *laws of composition* das Verhalten eines *Systems* mit dem Verhalten seiner Teile verknüpfen.[7] In diesem Fall hätten *laws of composition* also den Status von *Brückenprinzipien*, die die Ebene der Teile mit der Ebene des Ganzen verbinden. Direkt im Anschluß an die zitierte Passage kommt Broad jedoch wieder auf das Beispiel der Erklärung des dynamischen Verhaltens von Gegenständen zurück, auf die verschiedene Kräfte wirken:

> For purely dynamical transactions this assumption is pretty well justified, because we have found a simple law of composition and have verified it very fully for wholes of very different composition, complexity, and internal structure. It is therefore not particularly rash to expect to predict the dynamical behaviour of any material complex under the action of any set of forces, however much it may differ in the details of its structure and parts from those complexes for which the assumed law of composition has actually been verified. (Broad 1925, 63 f.)

Das *law of composition*, das er hier anspricht, ist offenbar das Gesetz der Vektoraddition von Kräften.[8] Aber dieses Gesetz sagt nicht, wie sich das Verhalten eines Ganzen aus dem Verhalten seiner Teile ergibt, sondern wie sich die Teile eines Ganzen verhalten, wenn mehrere Kräfte auf sie wirken. Gesetze dieser Art sollte man daher vielleicht besser „*laws of interaction*" nennen.

Allerdings: Wie auch immer man Broad interpretiert, er scheint in beiden Lesarten Recht zu behalten. Auf der einen Seite benötigt man *laws of interaction* oder *Interaktionsgesetze*. Denn das Verhalten eines Systems kann man nur dann aus den Eigenschaften und der Anordnung seiner Teile ableiten, wenn man weiß, wie sich die Teile selbst in dieser spezifischen Anordnung verhalten, und dies kann man – im allgemeinen Fall – nur wissen, wenn man weiß, wie die verschiedenen Faktoren zusammenwirken, die das Verhalten der Teile beeinflussen. Auf der anderen Seite sind jedoch auch *laws of composition* oder *Brückenprinzipien* unerläßlich, da man das Verhalten eines Systems nicht aus dem Verhalten seiner Teile ableiten kann, wenn man nicht weiß, wie das Systemverhalten mit dem Verhalten der Teile zusammenhängt. Wenn es darum geht, das Verhalten eines Systems *S* aus den Eigenschaften seiner Teile und deren Anordnung abzuleiten, benötigt man daher in der Tat *drei* Arten von Gesetzen:[9]

---

[7] Nur der Zusatz „in isolation" ist bei dieser Lesart irritierend.

[8] Vgl. Broad 1925, 62.

[9] Auf die Tatsache, daß man bei dem Versuch, das Verhalten eines Systems *S* aus den Eigenschaften seiner Teile und deren Anordnung abzuleiten, immer drei Arten von Gesetzen benötigt, wird z. B. in Hüttemann/Terzidis 2000 auf-

a. *Einfache Gesetze*, aus denen hervorgeht, wie sich jedes Teil von *S* verhält, wenn jeweils nur ein Einflußfaktor auf es einwirkt;
b. *Interaktionsgesetze*, die besagen, wie sich die Teile von *S* verhalten, wenn verschiedene Einflußfaktoren gleichzeitig auf sie einwirken; und
c. *laws of composition* oder *Brückenprinzipien*, aus denen hervorgeht, wie sich *S* als Ganzes verhält, wenn sich seine Teile auf eine bestimmte Weise verhalten.

Es ist wichtig, hier noch einmal zu betonen, daß alle diese Gesetze grundlegende allgemeine Gesetze sein oder aus grundlegenden allgemeinen Gesetzen folgen müssen. Denn nur solche Gesetze sind bei der Ableitung des Verhaltens eines Systems aus den Eigenschaften und der Anordnung seiner Teile zulässig. Wenn man all dies berücksichtigt, scheint es aber angemessen, Broads Definition (RE) und (E) so zu präzisieren:

(RE′) Die Makroeigenschaft *F* eines komplexen Systems *S* mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] ist genau dann *reduktiv erklärbar*, wenn folgendes gilt:
  (a) Die Art und Weise, wie sich die Teile $C_1$, …, $C_n$ verhalten, wenn sie auf die Weise *R* angeordnet sind, läßt sich aus den allgemeinen einfachen Gesetzen und den allgemeinen Interaktionsgesetzen ableiten, die für diese Teile gelten; und
  (b) es gibt ein allgemeines Brückenprinzip, demzufolge *S* die Makroeigenschaft *F* hat, wenn sich seine Teile $C_1$, …, $C_n$ so verhalten, wie sie es tun, wenn sie auf die Weise *R* angeordnet sind.

(E′) Die Makroeigenschaft *F* eines komplexen Systems *S* mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] ist genau dann *emergent*, wenn folgendes gilt:
  (a) Der Satz „Alle Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] haben *F*“ ist ein wahres Naturgesetz, aber
  ($b_1$) die Art und Weise, wie sich die Teile $C_1$, …, $C_n$ verhalten, wenn sie auf die Weise *R* angeordnet sind, läßt sich *nicht* aus den allgemeinen einfachen Gesetzen und den allgemeinen Interaktionsgesetzen ableiten, die für diese Teile gelten; *oder*

---

merksam gemacht. Die Unerläßlichkeit von ‚laws of composition‘ betont McLaughlin 1992.

Daß Broad selbst nicht klar zwischen Interaktionsgesetzen und Brückenprinzipien unterscheidet, mag zumindest zum Teil daran liegen, daß es in einigen Fällen einen engen Zusammenhang zwischen diesen beiden Arten von Gesetzen gibt. Wenn wir beispielsweise wissen, welche Kräfte die Teile eines Körpers *S* auf einen benachbarten Körper *S*′ausüben, dann wissen wir – aufgrund des Gesetzes der Vektoraddition von Kräften – auch, welche Kraft *S* selbst auf *S*′ausübt. Im allgemeinen benötigen wir aber spezifische Prinzipien, die die Ebene des Ganzen mit der Ebene der Teile verbinden.

($b_2$) es gibt *kein* allgemeines Brückenprinzip, demzufolge *S* die Makroeigenschaft *F* hat, wenn sich seine Teile $C_1$, …, $C_n$ so verhalten, wie sie es tun, wenn sie auf die Weise *R* angeordnet sind.

Zwei Punkte möchte ich hier noch hervorheben. Der erste betrifft die Frage, warum sich bei Broad kein einziges Beispiel für ein Brückenprinzip findet, obwohl solche Prinzipien doch unerläßlich sind, wenn man zeigen will, daß das Verhalten eines Systems reduktiv erklärbar ist. Ich vermute, daß dies daran liegt, daß die Brückenprinzipien, die für die von Broad diskutierten Fälle relevant sind, so trivial sind, daß es Broad gar nicht in den Sinn kam, sie explizit zu erwähnen. So scheint z. B. das folgende Brückengesetz trivialerweise wahr zu sein:

($P_1$) Wenn wir wissen, wie sich alle Teile eines Systems bewegen, wissen wir auch wie sich das System selbst bewegt.

Wenn man etwa an eine Scheibe denkt, deren Teile alle mit derselben Winkelgeschwindigkeit, in derselben Richtung und in derselben Ebene um einen Punkt im Innern der Scheibe kreisen, dann ist völlig klar, daß sich die Scheibe selbst um eben diesen Punkt dreht.[10] Und wenn man an das Volumen oder die Gestalt eines Körpers denkt, dann scheint ebenfalls völlig klar, daß diese durch die Orte, an denen sich seine Teile aufhalten, bzw. durch die relativen Positionen dieser Teile bestimmt sind. Nichts könnte selbstverständlicher sein. Die für Broads Beispiele relevanten Brückenprinzipien haben also nicht nur den Charakter sehr allgemeiner Naturgesetze; es scheint sogar so zu sein, daß wir uns die Falschheit dieser Prinzipien gar nicht vorstellen können. Diese Prinzipien scheinen den Status von *a priori*-Wahrheiten zu besitzen.

Damit kommen wir zum zweiten Punkt. Broad selbst betont, daß das Gesetz „Alle Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] haben die Makroeigenschaft *F*“ bei reduktiv erklärbaren Eigenschaften einen ganz anderen Status hat als bei emergenten. Wenn *F* emergent ist, ist dieses Gesetz, wie Broad sagt, ein nicht weiter ableitbares Gesetz (‚a *unique* and *ultimate* law‘). D. h., dieses Gesetz ist (a) kein Spezialfall, der aus einem allgemeinen Gesetz durch Einsetzung bestimmter Werte für bestimmbare Variablen gewonnen werden kann. Es ist (b) kein Gesetz, das durch Kombination zweier oder mehrerer allgemeiner Gesetze gewonnen werden kann. Und was vielleicht am wichtigsten ist: Wenn *F* emergent ist, dann kann dieses Gesetz (c) *nur* dadurch *entdeckt* werden, daß man eine Reihe von Systemen mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] untersucht, daß man dabei feststellt, daß alle diese Systeme die Eigenschaft *F* haben, und daß man dieses Er-

[10] Mit Bezug auf Uhren sehen die Dinge ganz ähnlich aus.

gebnis induktiv auf alle Systeme mit dieser Mikrostruktur überträgt.[11] Bei reduktiv erklärbaren Eigenschaften liegen die Dinge dagegen ganz anders.

> In order to predict the behaviour of a clock a man need never have seen a clock in his life. Provided he is told how it is constructed, and that he has learnt from the study of *other* material systems the general rules about motion and about the mechanical properties of springs and of rigid bodies, he can foretell exactly how a system constructed like a clock must behave. (Broad 1925, 65)

Wenn die Makroeigenschaft *F* eines Systems *S* mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] reduktiv erklärbar ist, kann man also auch *ohne je ein System mit dieser Mikrostruktur untersucht zu haben* wissen, daß *S* – genauso wie alle Systeme mit dieser Mikrostruktur – *F* besitzt. In diesem Fall folgt dies einfach aus den allgemeinen für die Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen (zu denen, wie gesagt, sowohl einfache Gesetze als auch Interaktionsgesetze als auch Brückenprinzipien gehören). Bei reduktiv erklärbaren Eigenschaften ist es daher in diesem Sinne *undenkbar*, daß ein System zwar die Mikrostruktur [$C_1$, …, $C_n$; *R*], aber nicht die Eigenschaft *F* besitzt. Wenn aus den allgemeinen Naturgesetzen folgt, daß alle Systeme mit dieser Mikrostruktur *F* besitzen, dann ist es – zumindest *relativ zu diesen Naturgesetzen* – unmöglich, daß ein System die Mikrostruktur [$C_1$, …, $C_n$; *R*] hat, die Eigenschaft *F* aber nicht besitzt. Damit haben wir einen eindeutigen *Test*, um herauszufinden, ob die Makroeigenschaft *F* eines Systems reduktiv erklärbar ist. Wir müssen nur fragen, ob diese Eigenschaft *vor dem ersten Auftreten* von Systemen mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] hätte prognostiziert werden können bzw. ob es – relativ zu den grundlegenden Naturgesetzen – undenkbar ist, daß ein System mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] die Eigenschaft *F* nicht besitzt.

---

[11] Mit Bezug auf das Verhalten von Silberchlorid schreibt Broad:
„[T]he law connecting the properties of silver-chloride with those of silver and of chlorine and with the structure of the compound is, so far as we know, an *unique* and *ultimate* law. By this I mean (*a*) that it is not a special case which arises through substituting certain determinate values for determinable variables in a general law which connects the properties of *any* chemical compound with those of its separate elements and with its structure. And (*b*) that it is not a special case which arises by combining two more general laws, one of which connects the properties of *any* silver-compound with those of elementary silver, whilst the other connects the properties of *any* chlorine-compound with those of elementary chlorine. So far as we know there are no such laws. It is (*c*) a law which could have been discovered *only* by studying samples of silver-chloride itself, and which can be extended inductively *only* to other samples of the same substance.“ (Broad 1925, 64 f.)

## II.

Kommen wir noch einmal zu der Frage zurück, welche Eigenschaften Broad selbst für emergent hielt und welche Gründe er dafür hatte. Im Hinblick auf das charakteristische Verhalten chemischer Verbindungen kennen wir seine Gründe schon. Broad zufolge gibt es in diesen Fällen einfach keine geeigneten *laws of composition* oder *Brückengesetze*:

> The example of chemical compounds shows us that we have no right to expect that the same simple law of composition will hold for chemical as for dynamical transactions. … It would of course (on any view) be useless merely to study silver in isolation and chlorine in isolation; for that would tell us nothing about the law of their conjoint action. This would be equally true even if a mechanistic explanation of the chemical behaviour of compounds were possible. The essential point is that it would also be useless to study chemical compounds in general and to compare their properties with those of their elements in the hope of discovering a *general* law of composition by which the properties of *any* chemical compound could be foretold when the properties of its separate elements were known. So far as we know, there is no general law of this kind. … No doubt the properties of silver-chloride are completely *determined* by those of silver and of chlorine; in the sense that whenever you have a whole composed of these two elements in certain proportions and relations you have something with the characteristic properties of silver-chloride, and that nothing has these properties except a whole composed in this way. But the law connecting the properties of silver-chloride with those of silver and of chlorine and with the structure of the compound is, so far as we know, an *unique* and *ultimate* law. (Broad 1925, 64 f.)

Im Hinblick auf die für Lebewesen charakteristischen Eigenschaften äußert Broad sich ähnlich:

> A living body might be regarded as a compound of the second order, *i.e.*, a compound composed of compounds; …. Now it is obviously possible that, just as the characteristic behaviour of a first-order compound could not be predicted from any amount of knowledge of the properties of its elements in isolation or of the properties of other first-order compounds, so the properties of a second-order could not be predicted from any amount of knowledge about the properties of its first-order constituents taken separately or in other surroundings. … [S]o the only way to find out the characteristic behaviour of living bodies may be to study living bodies as such. (Broad 1925, 67)

Sowohl was das charakteristische Verhalten chemischer Verbindungen als auch was die typischen Eigenschaften von Lebewesen angeht, gelten Broads Überlegungen heute jedoch als weitgehend überholt. Wir wissen inzwischen, daß die elektrische Leitfähigkeit von Metallen darauf beruht, daß sich in der äußersten Schale ihrer Atome nur wenige Elektronen befinden, die leicht abgespalten werden können und daher relativ frei beweglich

sind. Wir wissen, daß sich das Metall Natrium mit Chlor verbindet, weil Chloratome ihre äußerste Elektronenschale durch die von Natriumatomen abgegebenen Elektronen vervollständigen können; dabei entstehen Natrium- und Chlorionen, die so starke Anziehungskräfte aufeinander ausüben, daß sie sich in einer Gitterstruktur anordnen. (Dies ist der Grund dafür, daß Natriumchlorid bei normaler Temperatur und normalem Druck fest ist.) Wir wissen, daß Natriumchlorid wasserlöslich ist, weil Wassermoleküle aufgrund ihrer Dipolstruktur die Natrium- und Chlorionen aus ihren Gitterpositionen herauslösen können. Und wir kennen inzwischen auch einen Großteil der chemischen Vorgänge, auf denen z. B. die Atmung, die Verdauung und die Fortpflanzung von Lebewesen beruht. Broad selbst hatte diese Entwicklung allerdings nicht ausgeschlossen. Mit Bezug auf den Vorgang der Atmung schreibt er:

> [S]ince [the process of breathing] is a movement and since the characteristic movements of some complex wholes (*e.g.*, clocks) *can* be predicted from a knowledge of their structure and of other complex wholes which are not clocks, it cannot be positively *proved* that breathing is an ‚ultimate characteristic' or that its causation is emergent and not mechanistic. Within the physical realm it always remains logically possible that the appearance of emergent laws is due to our imperfect knowledge of microscopic structure or to our mathematical incompetence. (Broad 1925, 81)

Anders liegen die Dinge allerdings bei dem, was Broad ‚trans-physikalische Prozesse' nennt:

> But this method of avoiding emergent laws is not logically possible for trans-physical processes …. (Broad 1925, 81)

Aber was sind trans-physikalische Prozesse und warum glaubt Broad, daß diese Prozesse niemals reduktiv erklärt werden können? Um dies zu verstehen, muß man genauer auf Broads Begriff der ‚reinen Eigenschaft' (‚*pure quality*') eingehen. Denn ‚trans-physikalisch' nennt Broad gerade die Gesetze, die die Mikrostruktur eines Systems mit seinen reinen Eigenschaften in Verbindung bringen, also alle Gesetze der Form „Jedes System mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] hat die reine Eigenschaft $F$".[12] Broads offizielle *Definition* für den Begriff der reinen Eigenschaft lautet:

> By calling [qualities such as red, hot, etc.] ‚pure qualities' I mean that, when we say ‚This is red', ‚This is hot' and so on, it is no part of the meaning of our predicate that ‚this' stands in such and such relation to something else. It is *logically* possible that this should be red even though ‚this' were the only thing in the world …. (Broad 1925, 52)

[12] Vgl. Broad 1925, 52.

Doch diese Definition ist eigenartig; denn für Broad sind reine Eigenschaften ganz offensichtlich gerade die Eigenschaften komplexer Dinge, die traditionell als ‚sekundäre Qualitäten' bezeichnet werden – also die Eigenschaften *Temperatur*, *Farbe*, *Geschmack*, *Geruch*.[13] Herkömmlich sind sekundäre Qualitäten aber nichts anderes als Kräfte, in uns bestimmte Empfindungen hervorzurufen. Wie kann ein Gegenstand also reine Eigenschaften – d. h. sekundäre Qualitäten – besitzen, ohne daß es wahrnehmende Wesen gibt? Diesem Problem will ich hier jedoch nicht nachgehen, sondern einfach zu der Frage zurückkehren, warum Broad glaubt, reine Eigenschaften seien notwendigerweise emergent.

Wir hatten schon gesehen, daß Broad der Meinung war, daß das charakteristische Verhalten der meisten chemischen Verbindungen emergent sei. Allerdings war er eben auch der Auffassung, daß dies nur nach dem zeitgenössischen (d. h. damaligen) Stand der Wissenschaft – „so far as we know at present" – gelte und daß es durchaus möglich sei, daß uns die Wissenschaft eines Tages eines Besseren belehre. Chemische Verbindungen sind aber nicht nur durch ihr spezifisches Verhalten, sondern auch durch ihre reinen Eigenschaften charakterisiert. Und deshalb stellt sich für Broad die Frage: Kann sich auch von diesen reinen Eigenschaften herausstellen, daß sie nicht emergent, sondern reduktiv erklärbar sind? Broads Antwort auf diese Frage war ein klares Nein.[14]

Denken wir beispielsweise an Ammoniak – ein Gas, dessen Moleküle aus drei Wasserstoff- und einem Stickstoffatom bestehen und das die Eigenschaften hat, leicht wasserlöslich zu sein und über einen stechenden Geruch zu verfügen. Möglicherweise, so Broad, wird man eines Tages die Wasserlöslichkeit und die anderen charakteristischen *Verhaltensweisen* von Ammoniak aus den Eigenschaften seiner Komponenten und deren Anordnung erklären können, mit seinem Geruch aber ist das anders. Denn nach Broad ist es *theoretisch* unmöglich, diesen Geruch reduktiv zu erklären. Warum? Broads Antwort lautet: Nicht einmal ein mathematischer Erzengel – also ein Wesen, das vollständig über alle *allgemeinen* Naturgesetze informiert ist und das auch die kompliziertesten mathematischen Berechnungen im Bruchteil einer Sekunde ausführen kann – könnte voraussagen, welchen Geruch die Verbindung aus drei Wasserstoff- und einem Stickstoffatom hat.

---

[13] Vgl. Broad 1925, 46 ff. and 79 f.; es scheint mir jedoch auch möglich, daß Broad unter reinen Eigenschaften letzten Endes das versteht, was man heute ‚Qualia' nennt.

[14] Vgl. zum folgenden Broad 1925, 71 f.

> [Even a mathematical archangel] would be totally unable to predict that a substance with [the microscopic structure of ammonia] must smell as ammonia does when it gets into the human nose. The utmost that he could predict on this subject would be that certain changes would take place in the mucous membrane, the olfactory nerves and so on. But he could not possibly know that these changes would be accompanied by the appearance of a smell in general or of the peculiar smell of ammonia in particular, unless someone told him so or he had smelled it for himself. If the existence of the so-called ‚secondary qualities' … depends on the microscopic movements and arrangements of material particles which do not have these qualities themselves, then the laws of this dependence are certainly of the emergent type. (Broad 1925, 71 f.)

Warum wäre selbst ein mathematischer Erzengel in dieser Weise beschränkt? Nach den bisherigen Überlegungen muß der Grund darin liegen, daß aus den *allgemeinen* Gesetzen, die für Wasserstoff- und Stickstoffatome gelten, einfach nicht folgt, daß eine Verbindung aus drei Wasserstoff- und einem Stickstoffatom auf die für Ammoniak charakteristische Weise riecht. Aus diesen Gesetzen (und aus den Gesetzen der Neurophysiologie) folgt bestenfalls, daß in den Riechzellen in der Nase, im *nervus olfactorius* und im Gehirn einer Person, auf deren Nasenschleimhaut Ammoniakmoleküle treffen, bestimmte elektro-chemische Veränderungen stattfinden; aus ihnen folgt aber nicht, daß diese Veränderungen mit einer bestimmten Geruchsempfindung verbunden sind. Oder anders ausgedrückt: Das Gesetz, demzufolge bestimmte Veränderungen im Nervensystem einer Person zu einer solchen Geruchsempfindung führen, ist nicht aus den *allgemeinen* Naturgesetzen ableitbar; es ist ein emergentes oder, wie Broad auch sagt, ein bereichsübergreifendes Gesetz (‚*trans-ordinal law*') – ein Gesetz, daß die Mikrostruktur eines Systems mit einer seiner nicht ableitbaren Eigenschaften verbindet.

Broads Begründung für den notwendig emergenten Charakter von reinen Eigenschaften läßt sich somit so zusammenfassen: Reine Eigenschaften sind sekundäre Qualitäten und zu den charakteristischen Merkmalen sekundärer Qualitäten gehört, daß sie in uns bestimmte Empfindungen hervorrufen. Aus den *allgemeinen* Naturgesetzen folgt jedoch nicht, daß ein System *S* mit einer bestimmten Mikrostruktur in uns eine bestimmte Empfindung hervorruft; aus diesen Gesetzen folgt bestenfalls, daß durch das von *S* reflektierte Licht oder durch die Moleküle, die *S* in die Luft abgibt, bestimmte Veränderungen in unserem Zentralnervensystem hervorgerufen werden. Der zentrale Punkt in dieser Argumentation ist also die These, daß es sich bei dem *Brückengesetz*, das bestimmte Vorgänge in unserem ZNS mit unseren Empfindungen verbindet, um ein emergentes Gesetz (‚a *unique* and *ultimate* law') handelt, das nicht auf die allgemeinen Naturgesetze zurückgeführt werden kann. Broads Hauptgrund für den emergenten Charak-

ter reiner Eigenschaften ist somit die Annahme, daß Empfindungen nicht aus dem abgeleitet werden können, was im Gehirn einer Person vorgeht. Doch damit stellt sich natürlich die Frage, welche Argumente er für diese Annahme anführt.

Ein zentraler Grund war für Broad offensichtlich, daß seiner Meinung nach Empfindungen – ebenso wie alle anderen mentalen Zustände – nicht behavioral analysiert werden können.[15] Hierin liegt seiner Meinung nach der Hauptunterschied zwischen dem Leib-Seele- und dem Vitalismus-Problem.

> The one and only kind of evidence that we ever have for believing that a thing is alive is that it behaves in certain characteristic ways. *E.g.,* it moves spontaneously, eats, drinks, digests, grows, reproduces, and so on. Now all these are just actions of one body on other bodies. There seems to be no reason whatever to suppose that ‚being alive' means any more than exhibiting these various forms of bodily behaviour. … But the position about consciousness, certainly seems to be very different. It is perfectly true that an essential part of our evidence for believing that anything but ourselves has a mind and is having such and such experiences is that it performs certain characteristic bodily movements in certain situations. … But it is plain that our observation of the behaviour of external bodies is not our only or our primary ground for asserting the existence of minds and mental processes. And it seems to me equally plain that by ‚having a mind' we do not mean simply ‚behaving in such and such ways'. (Broad 1925, 612 f.)

Die Falschheit des Behaviorismus[16] ergibt sich für Broad im wesentlichen aus zwei Überlegungen: (1) Wenigstens mir selbst schreibe ich mentale Zustände nicht aufgrund von beobachtetem Verhalten zu. Selbst wenn es zuträfe, daß sich mein Körper auf die charakteristische Weise *A* bewegt, wenn ich einen Stuhl sehe, und auf die charakteristische Weise *B*, wenn ich eine Glocke höre, wären diese Bewegungen doch niemals der Grund dafür, daß ich von mir selbst sage, daß ich einen Stuhl sehe oder eine Glocke höre.

> I often know without the least doubt that I am having the experience called ‚seeing a chair' when I am altogether uncertain whether my body is acting in any characteristic way. And again I distinguish with perfect ease between the experience called ‚seeing a chair' and the experience called ‚hearing a bell'

---

[15] Vgl. zum folgenden besonders den Abschnitt „Reductive Materialism or ‚Behaviourism'", Broad 1925, 612–624. Stephan 1993 kommt zu einer ähnlichen Analyse der Broadschen Argumentation.

[16] Broad unterscheidet zwischen ‚molarem' und ‚molekularem' Behaviorismus. Nur der ‚molare' Behaviorismus entspricht dem, was man heute ‚Behaviorismus' nennt; der ‚molekulare' Behaviorismus ist eher eine Version der psychophysischen Identitätstheorie.

> when I am quite doubtful whether my bodily behaviour, if any, on the two occasions has been alike or different. (Broad 1925, 614)

(2) Jeder gute Schauspieler kann sich genauso verhalten wie jemand, der Schmerzen oder große Freude empfindet. Aus der Tatsache, daß sich jemand auf eine bestimmte Weise verhält, kann man daher nicht mit analytischer Sicherheit darauf schließen, daß er bestimmte Empfindungen oder Wahrnehmungen tatsächlich hat. Und das bedeutet generell: Wenn sich ein Wesen *A* genauso verhält wie jemand, der wirkliche Empfindungen hat, ist es immer möglich zu fragen: „Hat *A* tatsächlich Empfindungen oder verhält es sich nur so?“ Mit Bezug auf die Frage, ob ein Wesen wirklich intelligent ist bzw. ob es wirklich einen Geist hat, formuliert Broad diesen Punkt so:

> However completely the behaviour of an external body answers to the behaviouristic tests for intelligence, it always remains a perfectly sensible question to ask: ‚Has it really got a mind, or is it merely an automaton?‘ It is quite true … that, the more nearly a body answers to the behaviouristic tests for intelligence, the harder it is for us in practice to contemplate the possibility of its having no mind. Still, the question: ‚Has it a mind?‘ is never silly in the sense that it is meaningless. … it is not like asking whether a rich man may have no wealth. (Broad 1925, 614)

Offenbar war also schon Broad ein Vertreter des ‚*absent qualia*‘- Arguments und ebenso ein Verfechter der Auffassung, philosophische Zombies seien zumindest begrifflich möglich. Allerdings: Selbst wenn Broad mit seiner Behaviorismuskritik recht hat, ist damit noch keineswegs gezeigt, daß es sich bei Gesetzen der Form „Wenn sich im Gehirn der Person *A* der neuronale Vorgang *N* abspielt, spürt *A* die Empfindung *E*“ immer um emergente Gesetze handeln muß. Warum müssen alle Gesetze, in denen die Mikrostruktur eines Systems mit einer Eigenschaft verbunden wird, die *nicht* behavioral analysierbar ist, emergent sein?

Zumindest ein Teil der Antwort auf diese Frage ergibt sich meiner Meinung nach daraus, daß Broad offenbar der Meinung war, daß sich Gesetze wie

(1) Immer wenn im ZNS einer Person die C-Fasern feuern, fühlt diese Person Schmerzen

in ihrem Status zu deutlich von den folgenden Brückenprinzipien unterscheidet:

($P_1$) Wenn man weiß, wie sich alle Teile von *S* bewegen, weiß man auch, wie sich *S* selbst bewegt.

($P_2$) Wenn man den Ort aller Teile von *S* kennt, kennt man auch das Volumen von *S*.

($P_3$) Wenn man die relativen Positionen aller Teile von *S* kennt, kennt man auch die Gestalt von *S*.

Denn von diesen Prinzipien kann man sich – in einem bestimmten Sinn – einfach nicht vorstellen, daß sie falsch sind. Bei dem Gesetz (1) kann man dies aber sehr wohl. Und genau deshalb hat (1) nicht den Status, der nötig wäre, um die Rolle eines Brückenprinzips zu spielen.

Doch Broad hat noch ein zweites sehr interessantes Argument – ein Argument, das zeigt, daß er nicht nur das *Argument der Erklärungslücke*, sondern auch das *Argument vom unvollständigen Wissen* vorausgeahnt hat:

> We have no difficulty in conceiving and adequately describing determinate possible motions which we have never witnessed and which we never shall witness. … But we could not possibly have formed the concept of such a colour as blue or such a shade as sky-blue unless we had perceived instances of it, no matter how much we had reflected on the concept of Colour in general or on the instances of other colours and shades which we *had* seen. It follows that, even when we know that a certain *kind* of secondary quality … pervades … a region when and only when such and such a *kind* of microscopic event … is going on within the region, we still could not possibly predict that such and such a determinate event of the kind … would be connected with such and such a determinate shade of colour …. The trans-physical laws are then *necessarily* of the emergent type. (Broad 1925, 80)

Entscheidend ist nach Broad, daß wir den Begriff einer bestimmten Empfindung[17] erst *bilden* können, nachdem wir zum ersten Mal diese Empfindung selbst hatten. Wenn das so ist, kann das Auftreten einer Empfindung aber nicht vorhergesagt werden, bevor sie zum ersten Mal erlebt wurde. Und allein daraus folgt schon, daß Empfindungen emergent sind. Denn für reduktiv erklärbare Eigenschaften ist, wie wir am Ende des letzten Abschnitts gesehen hatten, charakteristisch, daß sie auch schon vor ihrem ersten Auftreten vorhergesagt werden können.

Am Ende dieses Abschnitts möchte ich noch einen weiteren Punkt hervorheben. Broad war unter anderem deshalb der Auffassung, daß Empfindungen notwendig emergent sind, weil es sich bei den Gesetzen, die Hirnzustände mit Empfindungen verbinden, um nicht ableitbare Gesetze (‚*unique and ultimate laws*') handelt, die nicht den Status von Brückenprinzipien besitzen. Und auf den ersten Blick könnte es so aussehen, als habe Broad für die These vom emergenten Charakter des Verhaltens chemischer Verbindungen ganz ähnliche Gründe. Immerhin schreibt er:

---

[17] Obwohl Broad hier von sekundären Qualitäten redet, geht es, wie wir schon gesehen haben, letzten Endes immer um die Empfindungen, die durch sekundäre Qualitäten hervorgerufen werden.

> The example of chemical compounds shows us that we have no right to expect that the same simple law of composition will hold for chemical as for dynamical transactions. (Broad 1925, 64)

Daß Broad hier explizit vom ‚law of composition for dynamical transactions' spricht, macht aber deutlich, daß es in der Chemie nicht um Brückenprinzipien, sondern um Interaktionsgesetze geht. Broad zufolge ergibt sich der emergente Charakter des Verhaltens chemischer Verbindungen daraus, daß es keine allgemeinen Interaktionsgesetze gibt, die uns sagen, wie sich *Atome* – d.h. die Teile dieser Verbindungen – in beliebigen Konfigurationen verhalten. Mit Bezug auf die *Definition* (E′) resultiert der emergente Charakter des Verhaltens chemischer Verbindungen also aus der Bedingung ($b_1$). Bei Empfindungen ist dies anders, auch wenn sich Broad in diesem Punkt nicht besonders klar ausdrückt.[18] Denn er scheint geglaubt zu haben, daß wir selbst dann, wenn es allgemeine Interaktionsgesetze gäbe, die alles erklären, was im ZNS einer Person vorgeht, nicht voraussagen könnten, welche Empfindungen diese Person hat. In diesem Fall ist der entscheidende Punkt das Fehlen geeigneter Brückenprinzipien. Oder anders ausgedrückt: Der emergente Charakter von Empfindungen resultiert eher aus der Bedingung ($b_2$) als aus der Bedingung ($b_1$).

## III.

An dieser Darstellung der Überlegungen Broads zum emergenten Charakter von Empfindungen sollte klar geworden sein, daß sich schon bei Broad alle wesentlichen Elemente des *Arguments der Erklärungslücke* finden. Ein direkter Vergleich mit den Überlegungen Levines und Chalmers würde dies noch deutlicher machen. Ich will mich hier aber auf die Argumentation Levines beschränken. Betrachten wir noch einmal kurz die Grundzüge dieser Argumentation. Nehmen wir die beiden Aussagen:

(1) Schmerz ist identisch mit dem Feuern von C-Fasern.

(2) Die Temperatur eines idealen Gases ist identisch mit der mittleren kinetischen Energie seiner Moleküle.

Levine zufolge gibt es einen grundsätzlichen Unterschied zwischen diesen beiden Sätzen: Der zweite ist ‚vollständig explanatorisch', der erste nicht. Auf der einen Seite ist es in einem bestimmten epistemischen Sinn *undenkbar*, daß in einem Gas die mittlere kinetische Energie der Moleküle sagen wir $6.21 \cdot 10^{-21}$ Joule beträgt, daß dieses Gas aber nicht die entsprechende Temperatur von 300 K besitzt. Auf der anderen Seite scheint es aber *sehr*

[18] Der Grund dafür ist offenbar, daß Broad nicht klar zwischen Interaktionsgesetzen und Brückenprinzipen unterschieden hat.

*wohl denkbar*, daß ich keine Schmerzen fühle, obwohl meine C-Fasern feuern. Worauf beruht dieser Unterschied?

Levines Antwort lautet: Wenn man uns fragen würde, was wir mit dem Ausdruck ‚Temperatur' meinen, dann würden wir antworten:

(2′) Temperatur ist die Eigenschaft von Körpern, die in uns bestimmte Wärme- bzw. Kälteempfindungen hervorruft, die dazu führt, daß die Quecksilbersäule in Thermometern, die mit diesen Körpern in Berührung kommen, steigt oder fällt, die bestimmte chemische Reaktionen auslöst, und so weiter.

Mit anderen Worten: Wir würden Temperatur allein durch eine *kausale Rolle* charakterisieren. Dies würde als Antwort auf die gestellte Frage allerdings nicht ausreichen, wenn nicht noch ein zweiter Punkt hinzukäme:

> [O]ur knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat. Furthermore, antecedent to our discovery of the essential nature of heat, its causal role, captured in statements like (2′), exhausts our notion of it. Once we understand how this causal role is carried out there is nothing more we need to understand. (Levine 1983, 357)

Für den explanatorischen Charakter der Aussage (2) gibt es also zwei Gründe:

1. Unser Begriff von Temperatur erschöpft sich vollständig in einer kausalen Rolle.
2. Die Physik kann verständlich machen, daß die mittlere kinetische Energie der Moleküle eines Gases genau diese kausale Rolle spielt.

Warum ist dann aber die Aussage (1) nicht vollständig explanatorisch? Mit dem Ausdruck ‚Schmerzen' assoziieren wir doch ebenfalls eine kausale Rolle: Schmerzen werden durch die Verletzung von Gewebe verursacht, sie führen dazu, daß wir schreien oder wimmern, und sie bewirken in uns den Wunsch, den Schmerz so schnell wie möglich loszuwerden. Dies bestreitet auch Levine nicht. Und er bestreitet auch nicht, daß die Identifikation von Schmerzen mit dem Feuern von C-Fasern den Mechanismus erklärt, auf dem diese kausale Rolle beruht. Dennoch gibt es seiner Meinung nach einen entscheidenden Unterschied.

> However, there is more to our concept of pain than its causal role, there is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firing is *why pain should feel the way it does*! For there seems to be nothing about C-fiber firing which makes it naturally ‚fit' the phenomenal properties of pain, any more than it would fit some other set of phenomenal properties. Unlike its functional role, the identification of the qualitative side of pain with C-fiber firing … leaves the connection between it

> and what we identify it with completely mysterious. One might say, it makes the way pain feels into merely a brute fact. (Levine 1983, 357)

Levines erster Grund für die These, daß die Aussage (1) nicht vollständig explanatorisch ist, ist also:

1. Unser Begriff von Schmerzen erschöpft sich nicht in einer kausalen Rolle; er umfaßt auch einen qualitativen Aspekt – die Art, wie es sich anfühlt, Schmerzen zu haben.

Dies allein ist aber nicht entscheidend. Denn die Aussage (1) könnte immer noch vollständig explanatorisch sein, wenn die Neurobiologie nur verständlich machen könnte, daß sich das Feuern von C-Fasern schmerzhaft anfühlt. Levines zweiter Grund ist daher, daß genau dies nicht der Fall ist.

2. Aus den allgemeinen Gesetzen der Neurobiologie folgt nicht, daß sich das Feuern von C-Fasern auf die für Schmerzen charakteristische Weise – nämlich schmerzhaft – anfühlt.

Könnte das nicht aber daran liegen, daß die Neurobiologie noch nicht weit genug fortgeschritten ist? Könnte es nicht sein, daß die Neurobiologie eines Tages doch zeigen wird, daß sich das Feuern von C-Fasern notwendigerweise schmerzhaft anfühlt – in dem Sinne, in dem die Chemie heute schon zeigen kann, daß Natriumchlorid notwendigerweise wasserlöslich ist?

Levine meint, daß dies nicht möglich ist – aus Gründen, die er besonders in seinem Aufsatz „On Leaving Out What It's Like" (1993) erläutert. Jede Reduktion, so Levine, muß zu einer *Erklärung* des reduzierten Phänomens führen. Und daß eine solche Erklärung gelungen ist, zeigt sich in seinen Augen daran, daß es hinterher in einem epistemischen Sinn unmöglich ist, sich vorzustellen, daß das Explanans ohne das Explanandum vorliegt.

> The basic idea is that a reduction should explain what is reduced, and the way we tell whether this has been accomplished is to see whether the phenomenon to be reduced is epistemologically necessitated by the reducing phenomenon, i.e. whether we can see why, given the facts cited in the reduction, things must be the way they seem on the surface. (Levine 1993, 129)

Versuchen wir, uns am Beispiel der Makroeigenschaft, flüssig zu sein, klar zu machen, wie das gemeint ist. Flüssigkeiten unterscheiden „sich von Gasen dadurch, daß ihr Volumen (weitgehend) druckunabhängig (inkompressibel) ist, von festen Körpern dadurch, daß ihre Form veränderlich ist und sich der Form des jeweiligen Gefäßes anpaßt."[19] Dies liegt auf der einen Seite daran, daß bei Flüssigkeiten – anders als bei Gasen – die Molekel so dicht wie möglich ‚gepackt' sind. Sie können nicht (oder nur bei sehr gro-

[19] Siehe den entsprechenden Eintrag in Meyers Lexikon, elektronische Version, LexiROM 2.0.

ßem Kraftaufwand) enger ‚zusammenrücken', weil zwischen ihnen erhebliche Abstoßungskräfte bestehen. Auf der anderen Seite sind die Molekel in Flüssigkeiten aber gegeneinander verschiebbar, sie können sozusagen frei übereinander rollen. Die Molekel fester Körper sind dagegen aufgrund der Kräfte, die sie aufeinander ausüben, an ihren relativen Positionen ‚festgezurrt'. Sie können sich sozusagen nur im Verband bewegen. Der ganze Körper bewegt sich, die relativen Positionen seiner Molekel bleiben aber unverändert, und deshalb behält der Körper seine Form. Damit haben wir offensichtlich eine (wenn auch unvollständige) Aufzählung der Merkmale, durch die die Eigenschaft, flüssig zu sein, charakterisiert ist. Läßt sich die Tatsache, daß Wasser bei 20° C (und normalem Luftdruck) flüssig ist, aus den Eigenschaften seiner Moleküle ableiten? Folgt aus den allgemeinen Naturgesetzen, die für $H_2O$-Moleküle gelten, daß Wasser bei 20° C genau diese Merkmale aufweist?

Aus diesen Naturgesetzen folgt erstens,[20] daß der Abstand zwischen $H_2O$-Molekülen bei 20° C aufgrund der zwischen den Molekülen bestehenden Abstoßungskräfte nur mit großem Druck weiter verringert werden kann. Und aus ihnen folgt zweitens, daß die Anziehungskräfte zwischen den Molekülen bei 20° C nicht ausreichen, um sie an ihren relativen Positionen festzuzurren. Bei dieser Temperatur können die Moleküle ‚frei übereinander rollen'. Wenn auf alle Moleküle dieselbe Kraft wirkt, wird sich daher jedes Molekül bis zu dem Ort bewegen, an dem es sozusagen nicht mehr weiter kann.

Damit allein ist aber noch nicht gezeigt, daß *Wasser* bei 20° C alle Merkmale aufweist, die für die Eigenschaft, flüssig zu sein, charakteristisch sind. Denn bisher wissen wir nur, wie sich *die einzelnen $H_2O$-Moleküle* bei dieser Temperatur verhalten. Wir benötigen zusätzlich noch *Brückenprinzipien*,[21] aus denen hervorgeht, wie das Verhalten der gesamten Flüssigkeit mit dem Verhalten der einzelnen Moleküle zusammenhängt. Und diese Prinzipien lauten offenbar:

($P_4$) Wenn der Abstand zwischen den Molekülen eines Stoffes nur mit großem Druck verkleinert werden kann, dann läßt sich das Volumen dieses Stoffes nur mit großem Druck verringern.

($P_5$) Wenn die Moleküle eines Stoffes frei übereinander rollen können, ist die Form dieses Stoffes veränderlich und paßt sich der Form des Gefäßes an, in dem er sich befindet.

Damit ergibt sich die folgende Antwort auf die Frage, warum wir uns – nach der gegebenen Erklärung – nicht mehr vorstellen können, daß Wasser

[20] Zumindest wird dies allgemein angenommen.
[21] Vgl. Levine 1993, 131.

bei 20° C *nicht* flüssig ist. Der erste Grund dafür ist einfach, daß aus den allgemeinen Naturgesetzen folgt, daß der Abstand, den $H_2O$-Moleküle bei 20° C zueinander haben, nur mit großem Druck weiter verringert werden kann und daß die Anziehungskräfte zwischen den Molekülen bei 20° C nicht ausreichen, um sie an ihren relativen Positionen festzuzurren. Mindestens ebenso wichtig ist jedoch der zweite Grund, der sich aus dem speziellen Status der Brückenprinzipien ($P_4$) und ($P_5$) ergibt. Denn offenbar ist dieser Status dafür verantwortlich, daß wir uns *nicht* vorstellen können, daß der Abstand zwischen den Molekülen eines Stoffes nur mit großem Druck verkleinert werden kann, sich das Volumen dieses Stoffes aber schon bei geringem Druck verringert bzw. daß die Moleküle eines Stoffes zwar frei übereinander rollen können, die Form dieses Stoffes aber unveränderlich ist, so daß sie sich nicht der Form des Gefäßes anpaßt, in dem er sich befindet.

Wenn wir das Verhältnis zwischen Schmerzen und C-Fasern betrachten, liegen die Dinge Levine zufolge jedoch anders. Selbst wenn wir bis ins letzte Detail darüber informiert sind, welche neurophysiologischen Prozesse (oder welche Informationsverarbeitungsprozesse) im Gehirn ablaufen, ist es seiner Meinung nach immer noch denkbar, daß die Person, in deren Gehirn diese Prozesse ablaufen, keine Schmerzen empfindet. Worauf beruht dieser Unterschied?

Wenn wir die Erklärung, die dazu führt, daß wir uns nicht vorstellen können, daß Wasser bei 20° C nicht flüssig ist, im Detail analysieren, ergeben sich drei zentrale Punkte:

1. Die charakteristischen Merkmale der Eigenschaft, flüssig zu sein, bestehen *alle* darin, daß sich flüssige Stoffe unter bestimmten Bedingungen auf eine bestimmte Art und Weise *verhalten*.
2. Aus den allgemeinen Naturgesetzen folgt, daß zwischen $H_2O$-Molekülen bei 20° C bestimmte abstoßende und anziehende Kräfte bestehen.
3. Es gibt Brückenprinzipien, aus denen sich ergibt, daß ein Stoff, zwischen dessen Molekülen diese Kräfte bestehen, genau das Verhalten zeigt, das für die Eigenschaft, flüssig zu sein, charakteristisch ist.

Wie sieht es nun mit der vermeintlichen Erklärung von Schmerzen durch das Feuern von C-Fasern aus? Offenbar gibt es hier schon im ersten Punkt einen wesentlichen Unterschied:

Unser Begriff von Schmerzen erschöpft sich nicht in einer kausalen Rolle, und Schmerzen sind auch nicht allein durch ein bestimmtes Verhalten charakterisiert; vielmehr umfaßt unser Begriff von Schmerzen einen qualitativen Aspekt – die Art, wie es sich anfühlt, Schmerzen zu haben.

Doch dieser Punkt ist letztlich nicht entscheidend. Denn Schmerzen könnten immer noch durch das Feuern von C-Fasern erklärt werden, *wenn*

es nur Brückenprinzipien gäbe, aus denen hervorginge, daß sich das Feuern von C-Fasern auf die für Schmerzen charakteristische Weise anfühlt. Entscheidend sind daher die folgenden beiden Punkte:

2. Aus den Gesetzen der Neurobiologie folgt nur, unter welchen Bedingungen welche Neuronen mit welcher Geschwindigkeit feuern.

Und:

3. Es gibt keinerlei Brückenprinzipien, die das Feuern von Neuronen mit bestimmten Erlebnisqualitäten verbinden.

Levines Argument der Erklärungslücke beruht also genauso wie Broads Argument für den emergenten Charakter von sekundären Qualitäten und Empfindungen auf der zentralen These:

($T_l$) Es gibt keine Brückenprinzipien, die neuronale Prozesse mit Empfindungen verbinden.

Ebenso wie für Broad ergibt sich damit auch für Levine der emergente Charakter von Empfindungen eher aus der Bedingung ($b_2$) als aus der Bedingung ($b_1$) der Definition (E′). Das Problem ist nicht, daß es keine allgemeinen *Interaktionsgesetze* gibt, die uns sagen, welche Neuronen mit welcher Rate feuern, wenn sie so miteinander verbunden sind, wie das in unserem ZNS der Fall ist. Das wirkliche Problem ist, daß Sätze wie

(1) Immer wenn im ZNS einer Person die C-Fasern feuern, fühlt diese Person Schmerzen

zwar wahre Naturgesetze sein können, daß sie aber nicht den Status von Brückenprinzipien besitzen.

## IV. Literatur

Beckermann, A. (1992): „Supervenience, Emergence, and Reduction“. In: A. Beckermann, J. Kim & H. Flohr (Hg.): *Emergence or Reduction?* Berlin/New York, 94–118.

– (1996): „Eigenschaftsphysikalismus“. *Zeitschrift für Philosophische Forschung* 50, 3–25.

– (2000): „The Perennial Problem of the Reductive Explainability of Phenomenal Consciousness – C. D. Broad on the Explanatory Gap“. In: T. Metzinger (Hg.): *Neural Correlates of Consciousness*. Cambridge MA, 41–55.

– (2001): *Analytische Einführung in die Philosophie des Geistes*. 2. Aufl., Berlin/New York.

Broad, C.D. (1925): *The Mind and Its Place In Nature*. London.

Chalmers, D. (1996): *The Conscious Mind*. Oxford.

Hempel, C.G., & Oppenheim, P. (1948): „Studies in the Logic of Explanation". *Philosophy of Science* 15, 135–175.

Hüttemann, A. & O. Terzidis (2000): „Emergence in Physics". *International Studies in the Philosophy of Science* 14, 267–281.

Levine, J. (1983): „Materialism and Qualia: The Explanatory Gap". *Pacific Philosophical Quarterly* 64, 354–361.

– (1993): „On Leaving Out What It's Like". In: M. Davies & G.W. Humphreys (Hg.): *Consciousness*. Oxford, 121–136.

McLaughlin, B. (1992): „The Rise and Fall of British Emergentism". In: A. Beckermann, J. Kim & H. Flohr (Hg.): *Emergence or Reduction?* Berlin New York, 49–93.

Stephan, A. (1993): „C.D. Broads a priori-Argument für die Emergenz phänomenaler Qualitäten". In: H. Lenk & H. Poser (Hg.): *Neue Realitäten – Herausforderungen der Philosophie*. Berlin, 176–183.

# 3

## Neue Überlegungen zum Eigenschaftsphysikalismus*

Seit den Tagen von Demokrit, Platon und Aristoteles ging es in der Philosophie des Geistes immer um die Frage, ob es über das Physische hinaus einen eigenen unabhängigen und selbständigen Bereich des Mentalen gibt. Diese Frage wurde aber meist als Frage nach der Existenz unabhängiger mentaler *Substanzen* verstanden. Haben wir eine Seele oder ein Selbst? Erst in den letzten 70 – 80 Jahren ging es auch um ein anderes Thema – die Frage, ob mentale *Eigenschaften* unabhängig und selbständig sind. Viele hielten die Frage nach der Existenz einer eigenständigen Seele für entschieden; eine solche Seele gibt es nicht. Aber das Problem mentaler Eigenschaften schien schwieriger zu sein. Ist das Haben von Schmerzen oder das Nachdenken über Paris wirklich etwas Physisches? Haben wir hinreichende Gründe für die Annahme, dass mentale Eigenschaften ihrer Natur nach physisch sind? Dies ist die Frage nach der Wahrheit des Eigenschaftsphysikalismus. Sind mentale Eigenschaften unabhängig und eigenständig? Oder sind sie ihrer Natur nach physische Eigenschaften oder lassen sie sich auf physische Eigenschaften zurückführen?

In den Arbeiten, die zu diesem Thema veröffentlicht wurden, ging es jedoch nicht nur um die *Wahrheit* des Eigenschaftsphysikalismus. Ein Großteil der Debatte war vielmehr der Vorfrage gewidmet, was genau der Eigenschaftsphysikalist behaupten muss. Was muss der Fall sein, damit der Eigenschaftsphysikalismus wahr ist? Die Geschichte der Antworten auf diese Frage (zu denen der Logische Behaviorismus, die Identitätstheorie, der Funktionalismus und die Supervenienztheorie gehören) ist wieder und wieder erzählt worden. Trotzdem scheint es mir sinnvoll, diese Antworten noch einmal neu zu analysieren.

### 1. Eigenschaftsphysikalismus erfordert Identität

Die naheliegendste Antwort auf die Frage nach dem Gehalt des Eigenschaftsphysikalismus lautet: Der Eigenschaftsphysikalismus ist genau dann wahr, wenn alle mentalen Eigenschaften physische Eigenschaften bzw. mit physischen Eigenschaften identisch sind. Aber was heißt es, wenn man sagt, dass eine Eigenschaft eine physische Eigenschaft oder mit einer physischen Eigenschaft identisch ist? Was kann überhaupt mit der Aussage,

---

* Erstveröffentlichung in: M. Pauen, M. Schütte & A. Staudacher (Hg.) *Begriff, Erklärung, Bewusstsein: Neue Beiträge zum Qualia-Problem*. Paderborn: mentis 2007, 143–170. Deutsche Fassung von Beckermann (2009).

Eigenschaften seien identisch, gemeint sein? Offensichtlich sind Mark Twain und Samuel Clemens genau dann identisch, wenn die beiden Namen ‚Mark Twain' und ‚Samuel Clemens' dieselbe Person bezeichnen.[1] Für Eigenschaften wird also wohl dasselbe gelten. Eigenschaften *F* und *G* sind genau dann identisch, wenn die Prädikate ‚*F*' und ‚*G*' für dieselbe Eigenschaft stehen.[2] Die entscheidende Frage ist also, wie man das herausfinden kann.

Die Antwort auf diese Frage hängt natürlich davon ab, wie man den Begriff ‚Eigenschaft' versteht. Rudolf Carnap, dessen Aufsätze (1932) und (1932/3) zu den wichtigsten frühen Arbeiten zum Problem des Eigenschaftsphysikalismus zählen, war der Meinung, dass Eigenschaften die Sinne oder Intensionen von Prädikaten sind.[3] Aus diesem Grund stehen Carnap zufolge zwei Prädikate genau dann für dieselbe Eigenschaft, wenn sie denselben Sinn haben, wenn sie also synonym sind.[4] Für Carnap muss ein Eigenschaftsphysikalist die These vertreten, dass es zu jedem mentalen Prädikat ein synonymes Prädikat der physikalischen Sprache gibt. Oder, um es auf eine kurze Formel zu bringen, nach Carnap erfordert der Eigenschaftsphysikalismus Synonymie.

Carnap war allerdings nicht nur der Auffassung, dass dies das richtige Verständnis des Eigenschaftsphysikalismus ist. Er war auch davon überzeugt, dass der so verstandene Eigenschaftsphysikalismus wahr ist. Seiner Meinung nach ist z. B. das Prädikat ‚*x* ist jetzt aufgeregt' synonym mit dem Ausdruck ‚Der Leib des *x*, und insbesondere sein Zentralnervensystem, hat eine physikalische (Mikro-)Struktur, die dadurch gekennzeichnet ist, dass Atmungs- und Pulsfrequenz erhöht ist und sich auf gewisse Reize hin noch weiter erhöht, dass auf Fragen meist heftige und sachlich unbefriedigende Antworten gegeben werden, dass auf gewisse Reize hin erregte Bewegungen eintreten und dergl.' (1932/33, 112 ff.). Was brachte Carnap zu der Überzeugung, diese Ausdrücke seien synonym? In den frühen 30er Jahren des letzten Jahrhunderts war Carnap ein Anhänger der verifikationistischen

---

[1] In (1892) setzt sich Gottlob Frege ausführlich mit der Frage auseinander, ob Identität (Gleichheit) eine Beziehung zwischen Gegenständen oder zwischen den Namen von Gegenständen ist. Letztendlich entscheidet er sich für die erste Option. Doch das ändert nichts daran, dass wahre Identitätsaussagen der Art „*a* = *b*" uns niemals darüber *informieren*, dass der durch ‚*a*' und ‚*b*' bezeichnete Gegenstand mit sich selbst identisch ist. Das wussten wir auch schon vorher. Und was auch immer die Information ist, die uns solche Aussagen vermitteln, es bleibt wahr, dass „*a* = *b*" genau dann wahr ist, wenn ‚*a*' und ‚*b*' für dasselbe stehen.

[2] Vgl. Fußnote 5.

[3] Diese Ansicht vertritt er explizit in (1956).

[4] Vgl. auch Hempel (1949).

Theorie der Bedeutung, der zufolge zwei Prädikate genau dann synonym sind, wenn sie auf der Grundlage derselben Beobachtungen angewendet werden. Und Carnap war der Meinung, dass dies auf die beiden genannten Prädikate zutrifft. Er war sogar davon überzeugt, dass der zweite Ausdruck nichts anderes ist als eine Aufzählung alle der Beobachtungen, die unserer Anwendung des Prädikats ‚*x* ist jetzt aufgeregt' zugrunde liegen.

Diese Auffassung, die später unter dem leicht irreführenden Namen ‚Logischer Behaviorismus' bekannt wurde, ist jedoch unhaltbar. Sobald die Überzeugungskraft der verifikationistischen Theorie der Bedeutung verblasste, wurde klar, dass es unmöglich ist, z. B. den Ausdruck ‚*x* möchte ein Bier' in rein physikalischer Sprache zu analysieren. Wenn eine Person ein Bier möchte, wird sie zum Kühlschrank gehen, um sich dort ein Bier zu holen – aber nur, wenn sie *glaubt*, dass sich im Kühlschrank ein Bier befindet, dass sie nicht erschossen wird, wenn sie zum Kühlschrank geht, usw. Dass eine Person einen Regenschirm mitnimmt, mag ein Anzeichen dafür sein, dass sie glaubt, dass es regnen wird – aber nur, wenn sie den Schirm nicht benutzen will, um sich vor allzu neugierigen Blicke zu schützen. Allgemein gesprochen: Es ist unmöglich, mentale Prädikate so zu analysieren, dass nicht im Analysans andere mentale Prädikate vorkommen, die ihrerseits ebenfalls nicht zirkelfrei in physikalischer Sprache analysiert werden können.

Diese Erkenntnis bildete aber nicht das Ende, sondern sogar eher den Anfang dessen, was man heute als ‚Identitätstheorie' bezeichnet. Ende der 50er Jahre des letzten Jahrhunderts entwickelten Place und Smart die Auffassung, dass Aussagen wie „Die Temperatur eines Gases ist identisch mit der mittleren kinetischen Energie seiner Moleküle", „Blitze sind elektrische Entladungen", und „Wasser ist $H_2O$" tadellos wahre Identitätsaussagen sind, obwohl die Ausdrücke ‚Temperatur' und ‚mittlere kinetische Energie' – bzw. ‚Blitz' und ‚elektrische Entladung' oder ‚Wasser' und ‚$H_2O$' – alles andere als synonym sind (vgl. Place 1956, Smart 1959). Trotzdem, so die Vertreter der Identitätstheorie, stehen, wie die Physik uns gezeigt hat, ‚Temperatur' und ‚mittlere kinetische Energie' für dieselbe Eigenschaft.[5] Daher ist es Place und Smart zufolge zumindest möglich, dass die empirischen Wissenschaften eines Tages zum dem Ergebnis gelangen, dass auch die Ausdrücke ‚Schmerz' und ‚das Feuern von C-Fasern' für dieselbe Eigenschaft stehen, obwohl sie nicht synonym sind. Place und Smart waren also der Auffassung, dass der Eigenschaftsphysikalismus auf die These

[5] Identitätstheoretiker verstehen offenbar Aussagen über die Identität von Eigenschaften nach dem Modell von Aussagen über die Identität von Gegenständen. Was in Fußnote 1 gesagt wurde, gilt also auch für Aussagen über die Identität von Eigenschaften. Aussagen der Form „*F* ist identisch mit *G*" sind genau dann wahr, wenn ‚*F*' und ‚*G*' für dieselbe Eigenschaft stehen.

hinaus läuft: Jede mentale Eigenschaft ist mit einer physischen Eigenschaft identisch, auch wenn die entsprechenden Prädikate nicht synonym sind. Oder anders ausgedrückt: Der Eigenschaftsphysikalismus erfordert nicht Synonymie, sondern nur Identität. Wenn das so ist, stehen wir aber wieder vor der Frage: Wie finden wir heraus, dass das mentale Prädikat ‚*M*' für dieselbe Eigenschaft steht wie das physikalische Prädikat ‚*P*', wenn wir das nicht dadurch herausfinden können, dass wir prüfen, ob ‚*M*' und ‚*P*' synonym sind?

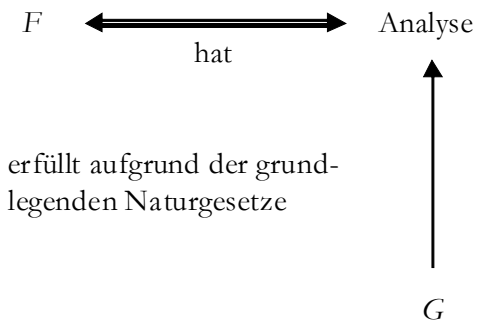
In seinem Aufsatz „Materialism and Qualia: The Explanatory Gap" (1983), in dem es ihm eigentlich um die Frage geht, ob phänomenale Zustände mit physischen Zuständen identisch sein können, gibt Joseph Levine, eher nebenbei, eine Antwort auf diese Frage, die eine ganze Kette von Diskussionen ausgelöst hat. Nach Levine ist die wahre Identitätsaussage

(1) Die Temperatur eines Gases ist identisch mit der mittleren kinetischen Energie seiner Moleküle

*vollständig explanatorisch*. Und dies beruht in seinen Augen auf zwei Tatsachen:

1. Unser Begriff von Temperatur erschöpft sich vollständig in ihrer kausalen Rolle.
2. Die Physik kann verständlich machen, dass die mittlere kinetische Energie der Moleküle eines Gases genau diese kausale Rolle spielt.

Mit anderen Worten, Levine zufolge ist die Aussage (1) vollständig explanatorisch, weil sich die Temperatur eines Gases allein unter Bezugnahme auf die mittlere kinetische Energie seiner Moleküle *reduktiv erklären* lässt. Jede reduktive Erklärung geht in zwei Schritten vor. Um zu zeigen, dass *F* allein unter Bezugnahme auf *G* reduktiv erklären lässt, muss man erstens eine Analyse von *F* geben und zweitens zeigen, dass aus den grundlegenden Naturgesetzen folgt, dass alle Gegenstände, die *G* besitzen, diese Analyse erfüllen. Das Schema reduktiver Erklärungen lässt sich also gut durch das folgende Diagramm veranschaulichen:



Entscheidend ist in diesem Zusammenhang aber, dass Levine auf den ersten Blick die Auffassung zu vertreten scheint, dass die Wahrheit von Identitätsaussagen der Form „*F* = *G*“ davon abhängt, ob es gelingt, *F* allein unter Bezug auf *G* reduktiv zu erklären. D. h., für Levine scheint eine notwendige Beziehung zwischen Identität und reduktiver Erklärbarkeit zu bestehen. Identitätsaussagen der Form „*F* = *G*“ sind nur dann wahr, wenn *F* allein unter Bezug auf *G* reduktiv erklärt werden kann.

Soweit ich sehen kann, steht – oder vielleicht sollte man besser sagen: stand – Levine mit dieser Auffassung keineswegs allein. Schon die ersten Verfechter der Identitätstheorie hatten auf die Frage, warum denn die Temperatur eines Gases mit der mittleren kinetischen Energie seiner Moleküle identisch sein soll, *unisono* die Antwort gegeben: Weil die klassische Thermodynamik auf die statistische Mechanik reduziert werden kann. Dass es sich hier um dieselbe Grundidee handelt, kann man sich leicht klar machen. Erstens ging man auf der Grundlage der seinerzeit weithin anerkannten Semantik theoretischer Terme davon aus, dass sich die Bedeutung dieser Ausdrücke implizit aus den Gesetzen ergibt, in denen sie auftreten. Die Bedeutung des Ausdrucks ‚Temperatur‘ ergibt sich demnach aus den Gesetzen der klassischen Thermodynamik; sie besteht in der kausalen Rolle, die durch diese Gesetze beschrieben wird. Und zweitens folgt aus der Reduzierbarkeit der klassischen Thermodynamik auf die statistische Mechanik, dass sich für alle Gesetze der klassischen Thermodynamik aus der statistischen Mechanik Bildgesetze (vgl. Beckermann 2001, 107 f.) ableiten lassen – Gesetze, aus denen hervorgeht, dass es auf der Ebene der statistischen Mechanik eine Eigenschaft (nämlich die mittlere kinetische Energie der Moleküle) gibt, die genau die für Temperatur charakteristische kausale Rolle einnimmt. Mit anderen Worten: Da sich die Bedeutung des Ausdrucks ‚Temperatur‘ implizit aus den Gesetzen der klassischen Thermodynamik ergibt, folgt aus der Reduzierbarkeit der klassischen Thermodynamik auf die statistische Mechanik, dass die Temperatur eines Gases allein mit Bezug auf die mittlere kinetische Energie seiner Moleküle reduktiv erklärt werden kann. Von Anfang an herrschte also die Meinung vor, dass zwischen Eigenschaftsidentität und reduktiver Erklärbarkeit ein äußerst enger Zusammenhang besteht.

## 2. Eigenschaftsphysikalismus erfordert reduktive Erklärbarkeit

Die Bedeutung reduktiver Erklärbarkeit wurde auch in einem ganz anderen Zusammenhang hervorgehoben. Zu Beginn des 20. Jahrhunderts war die Frage, ob Leben rein mechanisch erklärt werden könne, noch genau so heiß umstritten wie das Leib-Seele-Problem heute. Zwei Parteien standen sich

gegenüber. Auf der einen Seite die *Biologischen Mechanisten* mit der Auffassung, dass die für Lebewesen charakteristischen Eigenschaften (Stoffwechsel, Fortpflanzung, Wahrnehmung, zielgerichtetes Verhalten, Morphogenese) genauso mechanisch erklärt werden können wie das Verhalten einer Uhr, das sich mit physikalischer Zwangsläufigkeit aus den Eigenschaften und der Anordnung ihrer Zahnräder, Federn und Gewichte ergibt. Auf der anderen Seite die *Substanz-Vitalisten*, die die entgegengesetzte Meinung vertraten, Leben könne nur durch die Annahme einer zusätzlichen nicht-physischen Substanz erklärt werden – einer Entelechie oder eines *élan vital*. Als Broad in den frühen zwanziger Jahren seine Überlegungen zum Begriff der Emergenz entwickelte, verfolgte er unter anderem das Ziel, Raum für eine dritte Position zwischen diesen beiden Extremen zu schaffen – eine Position, die er *Emergenten Vitalismus* nannte.

Broads erster Schritt bestand darin, darauf aufmerksam zu machen, dass das Problem des Vitalismus nur der Spezialfall eines sehr viel generelleren Problems ist – des Problems, welche Beziehung zwischen den *Makroeigenschaften* eines komplexen Systems und den *Eigenschaften und der Anordnung seiner physischen Teile* besteht.[6] Im Hinblick auf diese Frage gibt es im Prinzip nur zwei mögliche Antworten: Die Makroeigenschaft *F* eines komplexen Systems *S* lässt sich oder sie lässt sich nicht allein aus den Eigenschaften und der Anordnung der physischen Teile von *S* erklären. Doch wenn *F* auf diese Weise erklärt werden kann, muss man Broad zufolge zwei weitere Möglichkeiten unterscheiden – *F* kann *mechanisch erklärbar* oder *emergent* sein. Knapp zusammengefasst, erläutert Broad selbst den Unterschied zwischen mechanischer Erklärbarkeit und Emergenz so:

> Abstrakt gesprochen behauptet die Emergenztheorie, dass es bestimmte komplexe Gegenstände gibt, die, sagen wir, aus den Komponenten A, B und C bestehen, die in der Relation R zueinander stehen; dass alle komplexen Gegenstände, die aus Komponenten der gleichen Art A, B und C bestehen, die zueinander in der gleichen Art von Relation R stehen, bestimmte charakteristische Eigenschaften besitzen; dass A, B und C in anderen Arten von komplexen Gegenständen vorkommen können, in denen die Relation nicht von der gleichen Art wie R ist; und dass die charakteristischen Eigenschaften des Ganzen R(A, B, C) nicht einmal im Prinzip aus der vollständigen Kenntnis der Eigenschaften abgeleitet werden können, die A, B und C isoliert oder in anderen komple-

[6] Broad spricht statt von den Makroeigenschaften oft nur spezieller vom Makro*verhalten* komplexer Gegenstände. Dies liegt daran, dass er der Meinung war, dass nur solche Eigenschaften mechanisch erklärbar sein können, für die es eine behaviorale Analyse gibt. Von ihm so genannte ‚pure qualities', die nicht behavioral analysiert werden können, sind Broad zufolge auf jeden Fall emergent (vgl. Beckermann 2002).

> xen Gegenständen haben, die nicht die Form R(A, B, C) besitzen. Der Mechanismus bestreitet den letzten Teil dieser Behauptung. (Broad 1925, 61)

Ich habe an anderer Stelle (Beckermann 2002) versucht, diese Passage im Detail zu analysieren und zu zeigen, dass sie letzten Endes auf die folgenden beiden Definitionen hinaus läuft.

(ME) Die Makroeigenschaft *F* eines komplexen Systems *S*, das aus den Teilen $C_1$, …, $C_n$ besteht, die auf die Weise *R* angeordnet sind, d. h. eines komplexen Systems *S* mit der Mikrostruktur [$C_1$, …, $C_n$; *R*], ist genau dann *mechanisch erklärbar*, wenn aus den allgemeinen grundlegenden für die Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen (sowie geeigneten Brückenprinzipien[7]) folgt, dass Gegenstände mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle für *F* charakteristischen Merkmale besitzen.

(E) Die Makroeigenschaft *F* eines komplexen Systems *S* mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] ist genau dann *emergent*, wenn folgendes gilt:
- (a) Es ist ein wahres (allerdings kein grundlegendes) Naturgesetz, dass alle Gegenstände mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] die Eigenschaft *F* haben; aber
- (b) es folgt *nicht* aus den allgemeinen grundlegenden für die Komponenten $C_1$, …, $C_n$ geltenden Naturgesetzen (sowie geeigneten Brückenprinzipien), dass Gegenstände mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle für *F* charakteristischen Merkmale besitzen.

---

[7] Block und Stalnaker (1999) haben argumentiert, dass die Analyse einer Eigenschaft, die reduktiv erklärt werden soll, niemals in dem Vokabular erfolgen kann, in dem die allgemeinen grundlegenden Naturgesetze formuliert sind, die man bei einer reduktiven Erklärung benötigt. Um eine solche Erklärung zu vervollständigen, müssen wir daher auf zusätzliche Brückengesetze zurückgreifen, die uns sagen, welche Makrophänomene mit welchen Mikrophänomenen *identisch* sind. Reduktive Erklärungen setzen in ihren Augen daher wahre Eigenschaftsidentitätsaussagen voraus und ersetzen sie nicht. Meiner Meinung nach benötigen wir zwar in der Tat geeignete Prinzipien, um die Kluft zwischen unterschiedlichen mereologischen Ebenen zu überbrücken. Diese Prinzipien sind aber keine *a posteriori* Identitätsaussagen. Die Prinzipien, die sich in den Wissenschaften tatsächlich finden, scheinen vielmehr unproblematische *a priori* Aussagen zu sein wie „Wenn alle Teile einer Scheibe mit derselben Winkelgeschwindigkeit um einen bestimmten Punkt kreisen, dann dreht sich die Scheibe um diesen Punkt." „Ein Gegenstand ist durchsichtig, wenn er Lichtstrahlen durchlässt." „Ein Gegenstand löst sich in einer Flüssigkeit auf, wenn beim Eintauchen in die Flüssigkeit seine Teile (Moleküle) voneinander gelöst werden und sich zwischen den Molekülen der Flüssigkeit verteilen.")

Wenn diese Interpretation richtig ist, werden zwei Dinge sofort klar. Erstens: Auch für Broad erfolgen mechanische Erklärungen in zwei Schritten. Zweitens: Diese beiden Schritte entsprechen genau denen, die wir schon bei Levine kennen gelernt hatten. Der erste Schritt besteht darin, eine Analyse von *F* zu finden, d. h. herauszufinden, welches die charakteristischen Merkmale von *F* sind.[8] Im zweiten Schritt muss dann gezeigt werden, dass aus den allgemeinen grundlegenden Naturgesetzen folgt, dass alle Gegenstände mit einer bestimmten Mikrostruktur diese für *F* charakteristischen Merkmale besitzen. Broad zufolge sollten Vertreter des Eigenschaftsphysikalismus daher behaupten, dass mentale Eigenschaften nicht emergent, sondern im Gegenteil mechanisch – oder wie wir heute sagen würde, reduktiv – erklärbar sind.

## 3. Identität und reduktive Erklärbarkeit

Broad hat nie die Auffassung vertreten, reduktive Erklärbarkeit sei eine notwendige Bedingung für Identität. Ihm ging es gar nicht um die Identität mentaler und physischer Eigenschaften. Anders Levine; zumindest auf den ersten Blick scheint er der Meinung zu sein, dass Eigenschaftsidentitätsaussagen der Form „*F* = *G*“ nur wahr sein können, wenn es möglich ist, *F* allein unter Bezug auf *G* reduktiv zu erklären. Doch dies ist schon *prima facie* unplausibel, da Identität eine *symmetrische*, reduktive Erklärbarkeit aber eine *asymmetrische* Beziehung ist.

Es ist daher kein Wunder, dass in den letzten Jahren *alle* Versuche, *Kriterien* für die Identität von Eigenschaften zu formulieren, einer grundsätzlichen Kritik unterzogen wurden. Identität, so der Kern dieser Kritik, ist eine nicht weiter analysierbare Relation. Eigenschaften sind entweder identisch oder sie sind es nicht. Auf die Frage „*Warum* sind *F* und *G* identisch?“ gibt es keine informative Antwort. Und deshalb gibt es auch keine *Kriterien*, die Eigenschaften erfüllen müssten, um identisch zu sein. Fragen kann man nur, wie man *feststellt*, ob Eigenschaften identisch sind, d. h., ob die Prädikate ‚*F*‘ und ‚*G*‘ für dieselbe Eigenschaft stehen.

---

[8] Meiner Meinung nach haben Chalmers und Jackson (2001) recht, wenn sie behaupten, dass reduktive Erklärungen im allgemeinen keine vollständige Analyse der zu erklärenden Eigenschaften voraussetzen. Es reicht aus, dass die Wahrheit des Satzes „Alle Gegenstände mit der Mikrosstruktur [$C_1$, …, $C_n$; *R*] haben die Eigenschaft *F*“ *a priori* erkannt werden kann auf der Basis der allgemeinen grundlegenden für die Komponenten $C_1$, …, $C_n$ geltenden Naturgesetze, geeigneter Brückenprinzipien und dessen, was kompetente Sprecher über die Extension von ‚*F*‘ wissen.

Diese Position hat z. B. David Papineau in seinem Aufsatz „Mind the Gap“ (1998) mit großem Nachdruck vertreten.[9] Papineau zufolge sollte sich jeder Eigenschaftsphysikalist zur Identitätstheorie bekennen. „Meine erste Aufgabe ist zu zeigen, dass Physikalismus am besten als eine These über Eigenschaftsidentität verstanden wird“ (Papineau 1998, 374). Aber, so Papineau weiter, die Identitätstheorie kann auch wahr sein, wenn sich mentale Eigenschaften nicht allein mit Bezug auf physikalische Eigenschaften reduktiv erklären lassen. Identitäten bestehen oder sie bestehen nicht. Es hat keinen Sinn zu fragen, warum zwei Dinge oder Eigenschaften identisch sind. Und deshalb spielt es für die Frage, ob *M* und *P* identisch sind, auch keine Rolle, ob wir verstehen, wie *P M* hervorbringt. Identische Eigenschaften bringen einander nicht hervor; sie sind einfach identisch. Nichts bringt sich selbst hervor.

Wie Papineau kritisieren auch Block und Stalnaker die Annahme, Physikalisten seien auf die These festgelegt, dass mentale Eigenschaften reduktiv erklärbar sind. In (1999) vertreten sie sogar die Auffassung, dies könne gar nicht sein. Denn reduktive Erklärbarkeit setze voraus, dass die zu erklärende Eigenschaft *F* so analysiert werden könne, dass in dieser Analyse nur Ausdrücke verwendet werden, die auch in den allgemeinen Naturgesetzen vorkommen. Genau dies sei im allgemeinen aber nicht möglich, und schon gar nicht bei mentalen Phänomenen. Reduktive Erklärungen mentaler Eigenschaften müssten daher in der Regel fehlschlagen. Daraus ergebe sich jedoch kein Argument gegen den Physikalismus. Denn der Physikalist sei nur auf eine Identitätsbehauptung festgelegt; und mentale Eigenschaften könnten auch dann mit physischen Eigenschaften identisch sein, wenn sie nicht reduktiv erklärt werden können. Generell sei die Rede von Identitätskriterien völlig irreführend. Statt zu fragen, welche Bedingungen erfüllt sein müssen, damit die Eigenschaften *F* und *G* identisch sind, sollten wir vielmehr fragen, wie sich Aussagen der Form „*F* = *G*“ rechtfertigen lassen.[10]

Ebenso wie Papineau vertreten Hill, McLaughlin, Block und Stalnaker also folgende Position: 1. Identität und reduktive Erklärbarkeit haben nichts miteinander zu tun. Vielmehr gilt: Wenn *F* und *G* identisch sind, dann können sie nicht aufeinander reduziert werden; nichts kann auf sich selbst reduziert werden. 2. Physikalisten sind nur auf die These festgelegt, dass mentale Eigenschaften mit physischen Eigenschaften identisch sind, und nicht auf die These, dass mentale Eigenschaften allein mit Bezug auf physische Eigenschaften reduktiv erklärt werden können.

---

[9] Allerdings war Papineau nicht der erste, der diese Auffassung vertreten hat. Vgl. z. B. Hill (1984, 1991).

[10] Vgl. zu dieser Position auch Hill (1991), Hill & McLaughlin (1999) und McLaughlin (2001).

Mit ihrer ersten These haben Hill, McLaughlin, Papineau, Block und Stalnaker in meinen Augen Recht. In frühen Debatten über die Identitätstheorie wurden zwei ganz verschiedene Ideen miteinander vermischt – die Idee der Identität und die Idee der reduktiven Erklärbarkeit. Erstens muss man feststellen, dass reduktive Erklärbarkeit mit Multirealisierbarkeit vereinbar ist; sie ist also keine hinreichende Bedingung für Identität. Aber sie ist auch keine notwendige Bedingung, wie sich am Beispiel von Wasser sehr schön zeigen lässt.

Levine hat in (1993) argumentiert, dass es, obwohl die Wahrheit von „Wasser = $H_2O$“ nicht *a priori* erkannt werden kann, *undenkbar* ist, dass $H_2O$ nicht die normalen Oberflächeneigenschaften von Wasser besitzt. Denn aus den allgemeinen grundlegenden Naturgesetzen folge, dass $H_2O$ auf Meereshöhe bei 100° C kocht, dass $H_2O$ flüssig ist, durchsichtig ist, usw. (1993, 128 f.). In seinen Augen zeigt sich daran ein wichtiger *epistemologischer* Unterschied zwischen den Aussagen „Wasser = $H_2O$“ und „Schmerz = Feuern von C-Fasern“. Denn es sei *nicht* im selben Sinne undenkbar, dass im Nervensystem einer Person die Fasern feuern, diese Person aber keine Schmerzen empfindet. Die entscheidende Frage ist aber: Ist es eine *notwendige Bedingung* für die Wahrheit von „Wasser = $H_2O$“, dass aus den allgemeinen grundlegenden Naturgesetzen folgt, dass $H_2O$ die normalen Oberflächeneigenschaften von Wasser besitzt?

Block und Stalnaker meinen, dies sei nicht so. Wenn wir herausfinden wollen, ob die Aussage „Wasser = $H_2O$“ wahr ist, stützen wir uns eher auf folgende Überlegung. Wir wissen, dass Wasser durch Erwärmen zum Kochen gebracht wird. Weiter klärt uns die Wissenschaft darüber auf, warum ein Anstieg der mittleren kinetischen Energie von $H_2O$-Molekülen zu einer bestimmten Aktivität M dieser Moleküle führt. Wenn wir annehmen, dass Wasser mit $H_2O$, Temperatur mit der mittleren kinetischen Energie der Moleküle und Kochen mit der molekularen Aktivität M identisch ist, gilt daher:

> Wir verfügen dann über eine Erklärung, wie die Erhitzung von Wasser dieses zum Kochen bringt. […] Identitäten erlauben einen Transfer von explanatorischer und kausaler Kraft, den bloße Korrelationen nicht erlauben. Annahmen wie, dass Wärme = molekulare kinetische Energie, […] usw. erlauben uns die Erklärung von Tatsachen, die wir anders nicht erklären könnten. Insofern ist unsere Schlussfolgerung, dass diese Identitätsaussagen wahr sind, durch das Prinzip des Schlusses auf die beste Erklärung gerechtfertigt. (Block & Stalnaker 1999, 23 f.)

Die Annahme, dass Wasser mit $H_2O$ und Temperatur mit der mittleren kinetischen Energie der Moleküle eines Stoffes identisch ist, führt zu einem einfacheren und kohärenteren Bild der Welt. *Dies* allein rechtfertigt diese Identitätsaussagen. Nach Block und Stalnaker müssen wir, um diese Aus-

sagen zu begründen, also nicht nachweisen, dass aus den allgemeinen grundlegenden Naturgesetzen folgt, dass $H_2O$ die normalen Oberflächeneigenschaften von Wasser besitzt.

In meinen Augen ist die Rechtfertigung der Aussage, dass Wasser $H_2O$ ist, sogar noch einfacher. Wir gehen von den beiden Hintergrundannahmen aus, dass Wasser ein chemischer Stoff ist und dass chemische Stoffe durch ihre molekulare Struktur individuiert werden. Um herauszufinden, ob Wasser $H_2O$ ist, müssen wir dann nur noch folgendes tun. Wir sammeln einige Wasserproben, bringen sie in ein chemisches Labor und bitten die Chemiker, die molekulare Struktur dieser Proben zu analysieren. Wenn wir die Antwort erhalten, dass diese Proben alle dieselbe molekulare Struktur besitzen, nämlich $H_2O$, haben wir ein Ergebnis – Wasser = $H_2O$.

Wie auch immer; dass reduktive Erklärbarkeit keine notwendige Bedingung für Identität ist, zeigt schon die folgende Überlegung. Nehmen wir an, die Chemiker teilen uns mit, dass die molekulare Struktur aller Proben $H_2O$ ist. Und nehmen wir weiter an, dass es *nicht* möglich ist zu zeigen, dass aus den allgemeinen grundlegenden Naturgesetzen folgt, dass $H_2O$ auf Meereshöhe bei 100° C kocht, dass $H_2O$ flüssig ist, durchsichtig ist, usw. Wie müssten wir diese Situation beschreiben? In meinen Augen liegt die Antwort auf der Hand. Wasser wäre nach wie vor $H_2O$; aber die Oberflächeneigenschaften von Wasser wären im Sinne Broads emergent. Das mag unwahrscheinlich erscheinen; aber es ist sicher nicht unmöglich.

Man muss also die Idee der Identität und die Idee reduktiver Erklärbarkeit im Sinne Broads tatsächlich strikt auseinander halten.[11] Und entsprechend muss man auch zwei Lesarten des Eigenschaftsphysikalismus klar voneinander unterscheiden – (a) die Auffassung, dass der Eigenschaftsphy-

[11] Es ist sehr wichtig, Broads Begriff der reduktiven Erklärung von verwandten Begriffen zu unterscheiden – z.B. dem Begriff der reduktiven Erklärung, der kürzlich von Chalmers und Jackson entwickelt wurde. Chalmers und Jackson zufolge ist ein Phänomen allein mit Bezug auf physische Phänomene reduktiv erklärbar, wenn es *a priori* aus einer vollständigen Beschreibung der physikalischen Welt folgt – einer Beschreibung, zu der auch die vollständige Physik gehört. Wie wir noch sehen werden, kann es durchaus sein, dass Identitätsaussagen wie „Wasser = $H_2O$“ im Sinne von Chalmers und Jackson reduktiv erklärbar sind. Dass die beiden Begriffe reduktiver Erklärbarkeit verschieden sind, zeigt sich schon daran, dass viele Makroeigenschaften, die in Broads Augen als emergent gelten müssen, im Sinne von Chalmers und Jackson sicher reduktiv erklärt werden können (vgl. Chalmers/Jackson, 2001, sec. 4). Der Grund dafür ist, dass Broad nur von den allgemeinen grundlegenden Naturgesetzen ausgeht, während Chalmers und Jackson eine vollständige Beschreibung der physikalischen Welt zugrundelegen – also eine Beschreibung, die auch alle Ausnahmen von den allgemeinen grundlegenden Naturgesetzen enthält.

sikalismus genau dann wahr ist, wenn alle mentalen Eigenschaften mit physischen Eigenschaften identisch sind, und (b) die Auffassung, dass der Eigenschaftsphysikalismus genau dann wahr ist, wenn alle mentalen Eigenschaften unter Bezug allein auf physische Eigenschaften im Sinne Broads reduktiv erklärt werden können. Aber ist nicht der Physikalismus generell auf die Annahme verpflichtet, dass alles in der Welt physisch ist? Warum sollten wir annehmen, dass die zweite Lesart tatsächlich eine Version des Eigenschafts*physikalismus* darstellt?

## 4. Eigenschaftsphysikalismus erfordert Supervenienz

> Eine bemerkenswerte Entwicklung in den Diskussionen des Körper-Geist-Problems der letzten ungefähr zwölf Jahre ist das Wiederaufleben des Typen-Physikalismus, der Auffassung, dass mentale Eigenschaften und Arten mit physischen Eigenschaften und Arten identisch sind. (Kim 2005, 121)

Das ist in der Tat wahr; denn seit den 60er und 70er Jahren des letzten Jahrhunderts sind eine ganze Reihe von Einwänden gegen die Identitätstheorie erhoben worden, insbesondere der Einwand der Multirealisierbarkeit mentaler Eigenschaften. Angenommen, das Prädikat ‚Schmerz' steht für eine bestimmte physische Eigenschaft, z. B. die Eigenschaft *Feuern von C-Fasern*. Dann könnte ‚Schmerz' nur auf Wesen zutreffen, deren Nervensystem C-Fasern enthält. Aber ist es wirklich plausibel anzunehmen, dass Tiere mit einem andersartigen Nervensystem oder Roboter, die Siliziumchips anstelle von Nerven enthalten, schon deshalb keine Schmerzen haben können, weil sie keine C-Fasern haben? Viel wahrscheinlicher ist doch die Annahme, dass Schmerz in verschiedenen Wesen mit ganz unterschiedlichen physischen Strukturen korreliert ist.

In Reaktion auf diesen Einwand entwickelte sich der Funktionalismus – die These, dass jeder mentale Zustand und jede mentale Eigenschaft durch eine spezifische kausale Rolle charakterisiert ist. Funktionalisten behaupten also, dass z. B. Schmerzen dadurch charakterisiert sind, dass sie (im allgemeinen) durch Gewebeverletzungen hervorgerufen werden, dass sie Schmerzäußerungen wie den Ausruf „Aua" und Verhalten zur Linderung der Schmerzen verursachen, dass sie die Konzentrationsfähigkeit beeinträchtigen usw. Für den Funktionalismus gibt es grundsätzlich zwei Lesarten.[12] Der einen Lesart zufolge bezieht sich der Ausdruck ‚Schmerz' auf die gerade geschilderte kausale Rolle. Nach einer anderen Lesart, die z. B. David Lewis vertritt, ist ‚Schmerz' eher als eine Art Kennzeichnung zu verstehen, die in jedem einzelnen Wesen für den Zustand steht, der in diesem

[12] Ein guter Überblick über die verschiedenen Versionen des Funktionalismus findet sich in Braddon-Mitchell/Jackson (1996).

Wesen die Schmerzrolle innehat oder realisiert. Beide Lesarten des Funktionalismus sind jedoch *keine* Versionen des Eigenschaftsphysikalismus. Denn keine dieser Lesarten impliziert, dass die Zustände, durch die die jeweiligen kausalen Rollen realisiert sind, physische Zustände sind.[13] Der Funktionalismus als solcher ist ontologisch neutral.[14]

Aus dem Funktionalismus entwickelte sich jedoch eine wirkliche Alternative zur Identitätstheorie – die Supervenienztheorie. Grundlage dieser Theorie ist der Gedanke, dass der Eigenschaftsphysikalismus auch dann wahr sein kann, wenn mentale Eigenschaften nicht mit physischen Eigenschaft identisch sind – vorausgesetzt, dass der Bereich des Mentalen *ontologisch* vom Bereich des Physischen *abhängt*, d. h., dass alle mentalen Tatsachen durch die physischen Tatsachen *ontologisch determiniert* sind. Mit dem Supervenienzbegriff soll diese Abhängigkeit genauer ausbuchstabiert werden.

Grundsätzlich ist Supervenienz eine Beziehung zwischen Eigenschaftsfamilien. Eine Eigenschaftsfamilie B superveniert über einer Eigenschaftsfamilie A genau dann, wenn es keine Unterschiede in den B-Eigenschaften geben kann ohne einen Unterschied in den A-Eigenschaften. Wenn $\alpha$ eine vollständige Beschreibung der Verteilung der A-Eigenschaften in einer Welt ist und $\beta$ irgendeine Aussage über die Verteilung von B-Eigenschaften in dieser Welt, dann supervenieren die B-Eigenschaften stark über den A-Eigenschaften, wenn die Aussage „Wenn $\alpha$, dann $\beta$“ für alle $\beta$ *notwendig wahr* ist.[15] Die Supervenienzversion des Eigenschaftsphysikalismus besagt also:

(SV) Wenn $\pi$ eine vollständige Beschreibung der Verteilung aller physischen Eigenschaften sowie der physikalischen Naturgesetze[16] und $\psi$ eine beliebige Aussage über die Verteilung mentaler Eigenschaft

---

[13] Das gilt sogar, wenn die kausale Rolle vollständig in physikalischer Sprache formuliert werden kann. Denn auch in diesem Fall kann der Zustand, der diese Rolle realisiert, ein nicht-physischer Zustand sein.

[14] Dies wurde schon von Putnam (1975, 436) festgestellt.

[15] Genau genommen trifft diese Formulierung nur das, was heute ‚globale Supervenienz‘ genannt wird. Auf den Unterschied zwischen globaler und nicht-globaler Supervenienz soll hier aber nicht weiter eingegangen werden.

[16] Tatsächlich sollte $\pi$ auch indexikalische physische Wahrheiten über uns selbst und eine „das ist alles“-Klausel enthalten, die feststellt, dass diese Beschreibung wirklich vollständig ist. Andernfalls würde $\pi$ nicht die mentale Tatsache implizieren, dass es keine immateriellen Geister gibt, die Schmerzen haben. Dieser Bedingung wird in Jacksons Formulierung durch die Forderung Rechnung getragen, dass alle relevanten möglichen Welten *minimale* physische Duplikate unserer Welt sein müssen. (Vgl. Chalmers/Jackson 2001)

ist, dann ist die Aussage „Wenn π, dann ψ“ für alle ψ notwendig wahr.

In jüngster Zeit hat Frank Jackson dafür die schöne Formel gefunden:

(MPD) „Jede mögliche Welt, die ein *minimales* physisches Duplikat unserer Welt ist, ist ein Duplikat *simpliciter* unserer Welt.“ (Jackson 1998, 13)

Zur Klarstellung fügt Jackson an: „[Ein] minimales physisches Duplikat unserer Welt ist eine Welt, die (a) in jeder physischen Hinsicht genauso wie unsere Welt ist (instantiierte Eigenschaft für instantiierte Eigenschaft, Gesetz für Gesetz, Relation für Relation), und (b) nichts weiter enthält (im Sinne von keine weiteren Arten oder Einzelgegenstände) als sie enthalten muss, um (a) zu erfüllen“ (ibid.).

Wenn man von dieser Supervenienztheorie ausgeht, ergibt sich sofort eine Antwort auf die Frage, warum die These, dass der Eigenschaftsphysikalismus reduktive Erklärbarkeit erfordert, überhaupt als Version des Eigenschaftsphysikalismus gelten kann. Reduktive Erklärbarkeit impliziert Supervenienz. Wenn alle mentalen Eigenschaften allein unter Bezug auf physische Eigenschaften reduktiv erklärt werden können, impliziert eine vollständige Beschreibung der Verteilung der physischen Eigenschaften (einschließlich der Beschreibung der grundlegenden Naturgesetze sowie geeigneter Brückenprinzipien) jede Aussage über die Verteilung mentaler Eigenschaften und Relationen. Allerdings, Identität impliziert ebenfalls Supervenienz. Wenn mentale Eigenschaften mit physischen Eigenschaften identisch sind, ist es offensichtlich unmöglich, dass es einen Unterschied in den mentalen Eigenschaften ohne einen Unterschied in den physischen Eigenschaft gibt. Also scheint es nur zwei Möglichkeiten zu geben – Supervenienz aufgrund von reduktiver Erklärbarkeit oder Supervenienz aufgrund von Identität. Andere Versionen der Supervenienztheorie sind meines Wissens jedenfalls nie ernsthaft vertreten worden. Insofern führt die Idee, dass Eigenschaftsphysikalismus Supervenienz erfordert, auch nicht zu einer neuen Version dieser Art des Physikalismus.

## 5. A PRIORI VS. A POSTERIORI PHYSIKALISMUS

Sowohl Vertreter der Auffassung, dass der Eigenschaftsphysikalismus Identität erfordert, als auch Anhänger der These, dass der Eigenschaftsphysikalismus reduktive Erklärbarkeit erfordert, können der Behauptung zustimmen: Wenn der Eigenschaftsphysikalismus wahr ist, dann macht das Physische das Mentale in strengsten Sinne notwendig. Allerdings gibt es Streit darüber, ob diese notwendige Abhängigkeit *a priori* oder nur *a posteriori* erkannt werden kann. „Wenn unsere Welt ohne Rest physisch ist, be-

stimmt ihre physische Natur mit Notwendigkeit im stärksten Sinne ihre mentale Natur […]. Was zur Diskussion steht, ist, ob (oder ob nicht) diese allseits anerkannte Notwendigkeit *a priori* ist, oder ob wir sie besser zur Kategorie der Notwendigkeit *a posteriori* rechnen sollten." (Jackson 2005, 252)

Unbestritten ist, dass *a priori* Physikalisten zumindest auf die folgende These festgelegt sind: Wenn wir alle physischen Fakten wüssten (einschließlich der vollständigen Physik), dann könnten wir allein auf der Grundlage dieses Wissens ohne jede weitere empirische Untersuchungen wissen, welche Dinge welche mentalen Eigenschaften haben und welche Dinge in welchen mentalen Relationen zueinander stehen. Oder, um es noch etwas präziser auszudrücken:

(*) Wenn π eine vollständige Beschreibung der physischen Welt ist (einschließlich einer vollständigen Physik) und ψ irgendeine Aussage, die eine mentale Tatsache ausdrückt, dann folgt ψ *a priori* aus π.[17]

Die Annahme, dass der Eigenschaftsphysikalismus reduktive Erklärbarkeit voraussetzt, ist offensichtlich eine Version des *a priori* Physikalismus. Wenn alle mentalen Eigenschaften allein mit Bezug auf physische Eigenschaften reduktiv erklärbar sind, können wir jede Aussage ψ aus π ableiten, wenn wir die Analyse der mentalen Prädikate kennen. Die Analyse der Prädikate einer Sprache ist aber genau das, was jeder kompetente Sprecher der Sprache kennen sollte.

Auf der anderen Seite scheint die von Hill, McLaughlin, Papineau, Block und Stalnaker vertretene Identitätsvariante des Eigenschaftsphysikalismus aber eine Version des *a posteriori* Physikalismus zu sein. Denn allgemein wird zugestanden, dass Identitätsaussagen wie „Mark Twain = Samuel Clemens" und „Wasser = $H_2O$" nur *a posteriori* als wahr erkannt werden können. Und das scheint zu implizieren, dass wir die Wahrheit dieser Aussagen nicht allein aufgrund unseres Wissens um die Bedeutung der Ausdrücke ‚Mark Twain', ‚Samuel Clemens', ‚Wasser' und ‚$H_2O$' erkennen können. Doch die Dinge sind etwas komplizierter.

Es ist sicher eine semantische Tatsache, dass die Ausdrücke ‚Mark Twain' und ‚Samuel Clemens' starre Bezeichner sind, d. h., dass sie in allen möglichen Welten dasselbe Objekt bezeichnen.[18] Und es ist auch eine se-

[17] Wieder muss π auch indexikalische physische Wahrheiten über uns selbst und eine „das ist alles"-Klausel enthalten, die feststellt, dass diese Beschreibung wirklich vollständig ist. Vielleicht ist es nötig, darauf hinzuweisen, dass der *a priori* Physikalismus im Sinne von (*) nicht behauptet, dass der Physikalismus *a priori* wahr ist.

[18] Ein starrer Bezeichner bezeichnet in jeder möglichen Welt denselben Gegenstand, sofern er in ihr überhaupt einen Gegenstand bezeichnet.

mantische Tatsache, dass ‚Mark Twain' und ‚Samuel Clemens' *de facto* dieselbe Person bezeichnen – den Schriftsteller Mark Twain. Jeder, der weiß, dass ‚Mark Twain' und ‚Samuel Clemens' starre Bezeichner sind, die tatsächlich dasselbe Objekt bezeichen, kann aber allein auf der Basis dieses Wissens erkennen, dass die Aussage „Mark Twain = Samuel Clemens" notwendig wahr ist. Und dasselbe gilt für die Aussage „Wasser = $H_2O$". Jeder, der weiß, dass die Ausdrücke ‚Wasser' und ‚$H_2O$' in unserer Welt denselben chemischen Stoff bezeichnen und dass diese Ausdrücke ebenfalls starre Bezeichner sind, die in jeder möglichen Welt denselben chemischen Stoff bezeichnen, kann allein auf der Grundlage dieses Wissens erkennen, dass „Wasser = $H_2O$" notwendig wahr ist. Dass Aussagen wie „Mark Twain = Samuel Clemens" und „Wasser = $H_2O$" nicht *a priori* als wahr erkannt werden können, beruht also nicht darauf, dass diese Aussagen durch verborgene modale Tatsachen wahr gemacht werden, deren Wahrheit wir nur *a posteriori* erkennen können. Was sie wahr macht, ist schlicht die Tatsache, dass jedes Objekt und jeder chemische Stoff in jeder möglichen Welt mit sich selbst identisch ist; und das wussten wir auch ohne empirische Untersuchungen. Dass die fraglichen Aussagen nicht *a priori* als wahr erkannt werden können, beruht vielmehr darauf, dass selbst kompetente Sprecher einer Sprache einige zentrale semantische Eigenschaften der Ausdrücke ihrer Sprache nur auf der Basis empirischer Untersuchungen herausfinden können.[19]

[19] Um Missverständnissen vorzubeugen, möchte ich versuchen, klar zu machen, wie ich die Ausdrücke ‚semantische Eigenschaft' und ‚semantische Tatsache' verstehe. Frege war der Meinung, dass jeder sprachliche Ausdruck einen Sinn und einen Bezug hat (jedenfalls sollte das in seinen Augen so sein) und dass sich Sinn und Bezug komplexer Ausdrücke in regelhafter Weise aus den Sinnen und Bezügen ihrer Teilausdrücke ergeben. Was man ein ‚Frege-Lexikon' einer Sprache nennen könnte, enthält also für jeden *nicht-komplexen* Ausdruck dieser Sprache zwei Einträge – einen, der seinen Sinn, und einen, der (für jede mögliche Welt) seinen Bezug angibt. Außerdem wäre es sinnvoll, wenn in einem solchen Frege-Lexikon alle Bezüge in kanonischer und transparenter Weise angegeben werden. D. h., wenn zwei Namen dasselbe Objekt bezeichnen, sollte dieses Objekt durch denselben Ausdruck spezifiziert werden. Wenn also z.B. die beiden Namen ‚Mark Twain' und ‚Samuel Clemens' denselben Schriftsteller bezeichnen, dann sollte dieser Bezug durch dasselbe Wort, sagen wir ‚Mark Twain', angegeben werden. Frege zufolge beziehen sich Prädikate auf Begriffe. Vertreter der Identitätstheorie dagegen würden eher sagen, dass sich Prädikate auf Eigenschaften beziehen. Doch wie dem auch sei, jedes Frege-Lexikon würde auch für jedes nicht-komplexe Prädikat einen Eintrag enthalten, in dem sein Bezug angegeben wird, d.h. der Begriff oder die Eigenschaft, für die dieses Prädikat steht. Auch dies sollte in kanonischer transparenter Weise geschehen. Wenn zwei Prädikate für denselben Begriff oder die-

Dass die kompetenten Sprecher einer Sprache keineswegs alle semantischen Eigenschaften der Ausdrücke dieser Sprache kennen, ist schon Frege aufgefallen, der betont, dass wahre Identitätsaussagen der Form „*a* = *b*“ wertvolle Erweiterungen unserer Erkenntnis darstellen können. Nach Frege kennen kompetente Sprecher zwar den Sinn, aber nicht unbedingt den Bezug der Ausdrücke ihrer Sprache.[20] Es kann also durchaus sein, dass kompetente Sprecher des Deutschen wissen, dass die Ausdrücke ‚Morgenstern' und ‚Abendstern' Himmelskörper bezeichnen, dass sie aber nicht wissen, welche Himmelskörper durch sie bezeichnet werden, und dass sie daher auch nicht wissen, dass beide Ausdrücke tatsächlich denselben Planeten bezeichnen. In ähnlicher Weise kann es sein, dass kompetente Sprecher des Deutschen durchaus wissen, dass ‚Wasser' einen chemischen Stoff bezeichnet, dass sie aber nicht wissen, welcher Stoff genau durch diesen Ausdruck bezeichnet wird, und dass sie daher auch nicht wissen, dass, wie die Dinge nun einmal sind, ‚$H_2O$' denselben Stoff bezeichnet.

Für die Identitätstheorie von Hill, McLaughlin, Papineau, Block und Stalnaker bedeutet dies Folgendes. Sie behaupten, dass der Eigenschaftsphysikalismus nur wahr sein kann, wenn alle mentalen Eigenschaften mit physischen Eigenschaften identisch sind; und sie interpretieren Eigenschaftsidentitätsaussagen nach dem Modell von Gegenstandsidentitätsaussagen. Wenn Identitätstheoretiker behaupten, dass Schmerz mit dem Feuern von C-Fasern identisch ist, dann wollen sie damit also nicht sagen, dass es zwei *verschiedene* Eigenschaften *Schmerz* und *Feuern von C-Fasern* gibt, die *de facto* identisch sind. (Verschiedene Eigenschaften sind niemals identisch.) Sie wollen vielmehr sagen, dass die Eigenschaft, für die der Ausdruck ‚Schmerz' steht, identisch ist mit der Eigenschaft, für die der Ausdruck ‚Feuern von C-Fasern' steht. Und da jedes Ding nur mit sich selbst und mit nichts anderem identisch ist, ist dies genau dann der Fall, wenn ‚Schmerz' und ‚Feuern von C-Fasern' für dieselbe Eigenschaft stehen.[21] Wenn wir also wüssten, für welche Eigenschaft(en) die Ausdrücke

---

selbe Eigenschaft stehen, dann soll ihr Bezug durch denselben Ausdruck spezifiziert werden. Semantische Tatsachen, so wie ich diesen Ausdruck verstehe, sind alle die Tatsachen, die sich allein aus den Einträgen in einem Frege-Lexikon ergeben.

[20] „Der *Sinn* eines Eigennamens wird von jedem erfaßt, der die Sprache oder das Ganze von Bezeichnungen hinreichend kennt, der er angehört […]. Zu einer allseitigen Erkenntnis der Bedeutung würde gehören, daß wir von jedem gegebenen Sinne sogleich angeben könnten, ob er zu ihr gehöre. Dahin gelangen wir nie.“ (Frege 1892, 42 – meine Hervorh.)

[21] „Identität ist absolut einfach und unproblematisch. Alles ist mit sich selbst identisch; nichts ist jemals mit etwas außer sich selbst identisch. Es gibt nie irgendein Problem damit, was dafür sorgt, dass etwas mit sich selbst identisch

‚Schmerz' und ‚Feuern von C-Fasern' stehen, könnten wir *a priori* wissen, ob die Identitätsaussage „Schmerz = Feuern von C-Fasern" wahr ist. Doch selbst kompetente Sprecher des Deutschen kennen diese semantische Tatsache nur aufgrund von empirischen Untersuchungen. Und genau deshalb ist die Wahrheit dieser Identitätsaussagen nur *a posteriori* erkennbar.

Tatsächlich sind die Dinge aber noch etwas komplizierter. Denn die Ausgangsfrage war nicht einfach, was wir *a priori* wissen können, sondern was wir *a priori* wissen *könnten*, *wenn wir über alle physikalischen Tatsachen (inklusive der vollständigen Physik) informiert wären*. Und hier scheint es nicht unplausibel anzunehmen, dass wie *unter dieser Bedingung* auch *alle semantischen Tatsachen a priori* wissen könnten. Betrachten wir noch einmal den Ausdruck ‚Wasser'. Kompetente Sprecher des Deutschen wissen, dass der chemische Stoff, den wir ‚Wasser' nennen, der Stoff ist, der sich in Flüssen und Seen befindet, der bei Regen aus den Wolken auf die Erde fällt, der aus Wasserhähnen fließt, usw. Nun ist es aber eine physikalische Tatsache, dass es sich bei diesem Stoff um $H_2O$ handelt. Aus dieser Tatsache und der Tatsache, dass ‚$H_2O$' der kanonische Name für $H_2O$ ist, können wir aber ableiten, dass ‚Wasser' und ‚$H_2O$' für denselben chemischen Stoff stehen. Chalmers und Jackson[22] gehen davon aus, dass sich dieses Ergebnis verallgemeinern lässt. Jeder Name und jedes Prädikat haben, so Chalmers und Jackson, einen deskriptiven Inhalt. Dieser deskriptive Inhalt besteht aus einer Menge von Merkmalen, von denen der kompetente Sprecher weiß, dass sie *in der wirklichen Welt* auf die durch die Namen bezeichneten Gegenstände (bzw. auf die Eigenschaften, für die die Prädikate stehen) zutreffen. Wenn es sich bei diesen Merkmalen um physische Merkmale handelt, ist es aber eine physikalische Frage, welche Gegenstände oder Eigenschaften diese Merkmale besitzen. Auf der Grundlage einer vollständigen Kenntnis der physikalischen Tatsachen müssen wir also in der Lage sein zu sagen, welche Gegenstände die Namen bezeichnen bzw. für welche Eigenschaften die Prädikate stehen. *A posteriori* Physikalisten müssen dagegen behaupten, dass wir, selbst wenn wir vollständig über alle physikalischen Tatsachen informiert wären, noch *weitere empirische Untersuchungen* anstellen müssen, um herauszufinden, ob ‚Schmerz' und ‚Feuern von C-Fasern' für dieselbe Eigenschaft stehen. Doch was für Untersuchungen könnten das sein? Unter der Voraussetzung, dass wir über alle physikalischen Tatsachen (inklusive der vollständigen Physik) informiert sind, scheint es generell möglich, *a priori* herauszufinden, ob der Satz „*F* = *G*" wahr ist, d. h., ob ‚*F*' und ‚*G*' für dieselbe Eigenschaft stehen.

---

ist; nichts kann es unterlassen, mit sich identisch zu sein. Und es gibt nie irgendein Problem damit, was dafür sorgt, dass zwei Dinge identisch sind; zwei Dinge können nie identisch sein." (Lewis 1986, 192 f.)

[22] Vgl. besonders Chalmers (2002) und Jackson (2003).

## 6. Die Semantik mentaler Prädikate und die wissenschaftliche Praxis

Unabhängig von der Frage, ob es sich bei der Identität von mentalen und physischen Eigenschaften um eine *a priori* oder eine *a posteriori* Angelegenheit handelt, können wir an dieser Stelle festhalten, dass wir im Hinblick auf die Frage, wie der Eigenschaftsphysikalismus am besten zu verstehen ist, vor der Alternative stehen: Eigenschaftsphysikalismus erfordert Identität oder Eigenschaftsphysikalismus erfordert reduktive Erklärbarkeit. Angenommen, die Vertreter der Identitätstheorie können irgendwie mit dem Einwand der Multirealisierbarkeit fertig werden,[23] welche Argumente sprechen dann noch für die eine oder die andere dieser beiden Alternativen?

Jaegwon Kim geht in (2005) ebenfalls von der Beobachtung aus, dass man im Hinblick auf das Problem des Eigenschaftsphysikalismus zwei Auffassungen unterscheiden muss. Auf der einen Seite sieht auch er Philosophen wie die Britischen Emergentisten, die der Meinung sind, dass der Eigenschaftsphysikalismus nur wahr sein kann, wenn mentale Eigenschaften allein mit Bezug auf physische Eigenschaften reduktiv erklärt werden können. Auf der anderen Seite gibt es die Identitätstheoretiker, die die Notwendigkeit reduktiver Erklärungen bezweifeln. Die Britischen Emergentisten wurden durch Fragen beunruhigt wie „*Warum* führt das Feuern von C-Fasern zu Schmerzen, und nicht zu Jucken oder Kitzeln?“ und „*Warum* und *wie* entsteht bewusste Erfahrung aus bestimmten neuronalen Zuständen?“ (vgl. etwa Kim 2005, 94). Block und Stalnaker dagegen glauben, dass solche Fragen fehlgeleitet sind. Identitäten kann man nicht erklären. Eigenschaften sind identisch oder sie sind es nicht. Es gibt keine informative Antwort auf die Frage „*Warum* sind *F* und *G* identisch?“.

> Identitäten [so argumentieren Block und Stalnaker] sollten nicht so verstanden werden als würden sie helfen, explanatorische Fragen wie ‚Warum ist Jones immer bei Bewusstsein, wenn im Gehirn Pyramidenzellenaktivität auftritt?‘ zu beantworten, sondern eher so als würden sie diese Fragen neutralisieren oder zerstreuen, d. h. als würden sie zeigen, *dass es hier gar nichts zu erklären gibt.* Warum gibt es genau dann Wasser, wenn $H_2O$ vorliegt? Warum korrelieren diese beiden miteinander? Die richtige Antwort ist hier diese: Wasser ist ein-

[23] Z. B. dadurch, dass man die Multirealisierbarkeitsthese bestreitet oder die Auffassung vertritt, mentale Eigenschaften seien mit funktionalen Eigenschaften identisch. Natürlich hat die Auffassung, dass der Eigenschaftsphysikalismus reduktive Erklärbarkeit erfordert, kein Problem mit dem Multirealisierbarkeitseinwand; denn reduktive Erklärung ist mit Multirealisierbarkeit vollständig vereinbar.

> fach $H_2O$, und es gibt hier keine Korrelation, die erklärt werden muss. (Kim 2005, 116 f.)

Kim ist allerdings der Ansicht, dass die Warum-Fragen der Britischen Emergentisten *prima facie* nicht so einfach vom Tisch gewischt werden können. Für ihn handelt es sich um ernsthafte und dringliche Fragen, deren Unangebrachtheit keineswegs auf der Hand liegt. Block und Stalnaker müssen in seinen Augen daher sehr gute Argumente haben, wenn sie uns davon überzeugen wollen, dass die Warum-Fragen der Britischen Emergentisten tatsächlich fehlgeleitet sind. Und diese Argumente haben sie seiner Meinung nach nicht. Meines Erachtens gibt es jedoch noch einen direkteren Weg, die Auffassungen von Block und Stalnaker zu widerlegen. Mentale Prädikate gehören, wie ich zeigen möchte, einfach nicht zu der Art von Prädikaten, für die die Identitätsthese überhaupt sinnvoll wäre. Entscheidend ist hier also die Semantik mentaler Prädikate.

Vor noch nicht allzu langer Zeit gehörte es, zumindest unter Philosophen, zu den fast allgemein akzeptierten Überzeugungen, dass die Bedeutung jedes Prädikats in einer Menge von notwendigen und hinreichenden Bedingungen besteht. Ein Prädikat trifft auf ein Objekt genau dann zu, wenn diese Bedingungen erfüllt sind. Aufgrund der Arbeiten von Kripke und Putnam wurde dieser Konsens jäh zerstört. Jetzt scheinen die meisten zu glauben, dass alle (also auch mentale) Prädikate eher wie Namen funktionieren – sie beziehen sich gewissermaßen auf Eigenschaften und sie treffen auf ein Objekt genau dann zu, wenn das Objekt die durch das Prädikat bezeichnete Eigenschaft hat. Empirische Untersuchungen können zur Aufklärung der Natur dieser Eigenschaft führen. Doch damit ein Prädikat auf ein Objekt zutrifft, muss das Objekt keine der Oberflächenmerkmale dieser Eigenschaft besitzen.

Erinnern wir uns noch einmal kurz an den Kern der Kripkeschen Argumentation. Kripkes Ziel war zu zeigen, dass bestimmte sprachliche Ausdrücke (insbesondere Namen und Artbegriffe) anders funktionieren als früher angenommen. Namen, so Kripke, sind starre Bezeichner, die immer, auch in modalen Kontexten, dasselbe Objekt bezeichnen. Etwas ähnliches gilt für Artbegriffe. Um dies zu zeigen, geht Kripke von Fällen folgender Art aus. Wissenschaftler laden eine Gruppe kompetenter Sprecher des Deutschen in ihr Labor ein. Die Mitglieder der Gruppe wissen, dass das übliche Gold, das sie kennen, das chemische Element mit der Ordnungszahl 79 ist. Jetzt aber weisen die Wissenschaftler auf einen Stoff, der dieselben Oberflächeneigenschaften wie Gold besitzt und sagen: „Dies ist ein Stoff, den wir in einem Meteoriten gefunden haben, der kürzlich in Sibirien auf die Erde gestürzt ist; die chemische Struktur dieses Stoffes ist *ABC*.“ Wie werden die Mitglieder der eingeladenen Gruppe darauf reagieren? Grundsätzlich gibt es zwei Möglichkeiten. Einerseits können sie sagen:

„Äußerst interessant; es gibt also einen *anderen* chemischen Stoff, der dieselben Oberflächeneigenschaften besitzt wie Gold.“ Sie können aber auch sagen: „Äußerst interessant; es gibt also neben dem chemischen Element mit der Ordnungszahl 79 noch *andere Arten von Gold*.“[24] Auf der Grundlage seiner sprachlichen Intuitionen ist sich Kripke sicher, dass die kompetenten Sprecher des Deutschen die erste Antwort geben werden. Und genau das ist der Grund für seine Überzeugung, dass ‚Gold‘ ein starrer Bezeichner ist, der in allen möglichen Welt genau das chemische Element mit der Ordnungszahl 79 bezeichnet. Wenn kompetente Sprecher des Deutschen die erste Antwort geben, zeigt das aber nicht nur, dass ‚Gold‘ ein starrer Bezeichner ist. Es zeigt auch, dass keine der bekannten Oberflächeneigenschaften von Gold für das Zutreffen von ‚Gold‘ entscheidend ist. Dieser Ausdruck bezieht sich im Deutschen auf einen bestimmten chemischen Stoff. Und chemische Stoffe werden durch ihre molekulare Struktur individuiert. Es ist daher die Aufgabe der empirischen Wissenschaften herauszufinden, auf welchen Stoff sich ‚Gold‘ bezieht. Ob ‚Gold‘ auf ein Objekt zutrifft, hängt allein von der molekularen Struktur dieses Objektes ab, nicht von seinen Oberflächeneigenschaften.

Es ist sehr wichtig, sich klar zu machen, dass das, was ich den ‚Kripke-Test‘ nennen möchte, durchaus auch anders ausgehen kann. Stellen wir uns z. B. eine Situation vor, in der Wissenschaftler sagen: „Normalerweise haben alle wasserlöslichen Stoffe die Mikrostruktur *UVW*. Aber jetzt haben wir einen Stoff mit einer anderen Mikrostruktur gefunden, der sich des ungeachtet auflöst, wenn man ihn in Wasser gibt.“ Meiner Meinung nach besteht keinerlei Zweifel, dass kompetente Sprecher des Deutschen sagen werden, dass dieser neue Stoff wasserlöslich ist, auch wenn er nicht die Mikrostruktur *UVW* besitzt. Wenn ein Wissenschaftler vorschlagen würde, Wasserlöslichkeit mit dem Besitz der Mikrostruktur *UVW* zu identifizieren, würden wir ihn wahrscheinlich für ein bisschen verrückt halten. Wir würden sagen, dass er einfach nicht verstanden hat, dass wir Dinge genau dann ‚wasserlöslich‘ nennen, wenn sie sich auflösen, wenn man sie in Wasser gibt. Anders als Gold hat Wasserlöslichkeit keine Natur, die Wissenschaftler entdecken könnten; anders als ‚Gold‘ hat ‚Wasserlöslichkeit‘ eine Analyse. Es gibt bestimmte charakteristische Oberflächeneigenschaften, die ein Objekt besitzen muss, damit der Ausdruck ‚wasserlöslich‘ auf es zutrifft.

Offenbar gibt es also zwei Arten von Prädikaten. Auf der einen Seite Prädikate (z. B. die Artbegriffe ‚Wasser‘, ‚Gold‘, usw.), die tatsächlich wie Namen funktionieren. Auf der andere Seite gibt es aber auch Prädikate (wie ‚durchsichtig‘, ‚wasserlöslich‘ und ‚giftig‘), die eine Analyse besitzen. Die-

[24] Für einige realistischere Fälle vgl. Segal (2000, ch. 5).

se Prädikate treffen auf einen Gegenstand zu, wenn er die richtigen charakteristischen Merkmale aufweist.[25]

Die These, dass es diese beiden unterschiedlichen Arten von Prädikaten gibt, im Hinblick auf die man ganz unterschiedliche Fragen stellen kann und muss,[26] wird auch durch einen genauen Blick in die wissenschaftliche Praxis bestätigt. Manchmal versuchen Wissenschaftler in der Tat, die Natur der Eigenschaft zu entdecken, für die ein bestimmtes Prädikat steht. Sie sagen uns, was Wasser ist, was Katzen sind, was Wolken sind, was Blitze sind, usw. Sie klären uns über die Natur der Dinge auf, die gewissermaßen im Sinne Lockes eine reale Wesenheit (*real essence*) besitzen. Wenn wir fragen, was Wasser ist, was Wolken sind oder was Blitze sind, erhalten wir Antworten wie „Wasser ist die chemische Verbindung $H_2O$“, „Wolken sind dichte Ansammlungen von Wassertropfen oder Eiskristallen am Himmel“ und „Blitze sind bestimmte elektrische Entladungen in der Luft“. Wenn wir diese Antworten erhalten haben, hat es keinen Sinn mehr, weiter nach dem Warum zu fragen. Man kann nicht erklären, *warum* Wasser $H_2O$ ist, *warum* Wolken dichte Ansammlungen von Wassertropfen oder Eiskristallen am Himmel sind oder *warum* Blitze bestimmte elektrische Entladungen in der Luft sind. Das ist so, und mehr ist dazu nicht zu sagen.[27]

Wissenschaftler sagen uns aber nichts über die Natur von Wasserlöslichkeit oder Durchsichtigkeit. Die Natur dieser Eigenschaften kennen wir schon, d. h., wir kennen die Analyse der entsprechenden Prädikate. Ein Objekt ist genau dann wasserlöslich, wenn es sich auflöst, wenn man es in Wasser gibt. Und ein Objekt ist genau dann durchsichtig, wenn es Lichtstrahlen (fast vollständig) durchlässt. Wasserlöslichkeit oder Durchsichtig-

---

[25] Natürlich könnte man den Unterschied, um den es mir geht, auch anders formulieren. Man könnte sagen, alle Prädikate stehen für Eigenschaften. Aber nicht alle Eigenschaften sind gleich. Einige Eigenschaften haben eine Analyse, andere nicht. Mir scheint es in diesem Zusammenhang aber sinnvoll, von Prädikaten und nicht von Eigenschaften zu reden. Ganz allgemein muss man sich darüber im Klaren sein, dass die ganze Debatte durch eine Konfusion bedroht ist – die Konfusion zwischen Prädikaten und Begriffen auf der einen und Eigenschaften auf der anderen Seite. Immer wieder hört man, dass Eigenschaften analysiert werden oder dass man versuchen sollte, die Eigenschaften eines komplexen Gegenstandes aus den Eigenschaften seiner Teile abzuleiten. Wörtlich verstanden macht das alles keinen Sinn. Man kann nur Sätze oder Propositionen ableiten, nicht Eigenschaften. Und man kann Begriffe oder die Bedeutung von Prädikaten analysieren, während es zumindest zweifelhaft ist, ob man wirklich Eigenschaften analysieren kann.

[26] Diese Idee findet sich auch in Schütte (2004).

[27] Das heißt natürlich nicht, dass diese Identitätsaussagen nicht gerechtfertigt werden können. Vgl. z. B. die oben auf Seite 57 angeführte Rechtfertigung der Aussage „Wasser ist die chemische Verbindung $H_2O$“.

keit haben keine reale, sondern nur eine nominal Wesenheit (*nominal essence*). Trotzdem können uns Wissenschaftler eine Menge über Wasserlöslichkeit oder Durchsichtigkeit erzählen. Im Hinblick auf Wasserlöslichkeit oder Durchsichtigkeit beantworten Wissenschaftler aber keine Was-, sondern Warum-Fragen. Sie können uns sagen, warum Salz wasserlöslich und Glas durchsichtig ist, d. h., sie können reduktive Erklärungen dieser Tatsachen geben. Es ist hier genau umgekehrt wie bei Wasser und Wolken. Einige Beispiele sollen zur Veranschaulichung dienen.

Kochsalz ist eine chemische Verbindung mit der Molekülstruktur NaCl. Warum ist diese Verbindung unter normalen Bedingungen fest und wasserlöslich? Nun, Natrium reagiert mit Chlor, weil Chloratome ihre äußerste Elektronenhülle mit den von den Natriumatomen abgegebenen Elektronen vervollständigen können. Bei diesem Prozess entstehen Natrium- und Chlorionen, die starke elektrische Kräfte aufeinander ausüben, was bewirkt, dass sich die Ionen in einer Gitterstruktur anordnen. Wegen der Kräfte, die zwischen den Ionen bestehen, ist Kochsalz unter normalen Bedingungen fest. Es ist aber auch wasserlöslich, da Wassermoleküle aufgrund ihrer Dipolstruktur die Natrium- und Chlorionen aus ihrer Gitterstruktur herauslösen können, so dass sie sich zwischen den Wassermolekülen verteilen.

Auch die Reinigungsfähigkeit von Seife lässt sich auf die Struktur ihrer Moleküle zurückführen. Nicht wasserlösliche Stoffe wie Öle oder Fette (und Schmutz besteht im wesentlichen aus solchen Fetten) können suspendiert und ausgeschwemmt werden, wenn sie von Molekülen umgeben werden, die am einen Ende eine lipophile langgestreckte Kohlenwasserstoffkette und am anderen eine hydrophile Alkalikarboxylatgruppe enthalten. Und genau das ist bei Seifenmolekülen der Fall.

Schließlich war die Entdeckung der Doppelhelix der entscheidende Durchbruch für das Verständnis der chemischen Prozesse, die der Biologie von Fortpflanzung und Vererbung zugrunde liegen.

Im Hinblick auf diese Beispiele sind eine Reihe von Dingen bemerkenswert. Erstens weisen sie alle die für reduktive Erklärungen charakteristische Zwei-Schritte-Struktur auf. Im ersten Schritt muss eine Analyse der Prädikate gegeben werden, um die es geht. Was heißt es, fest oder wasserlöslich zu sein oder eine Reinigungskraft zu besitzen? Der zweite Schritt besteht anschließend darin zu zeigen, dass aus den grundlegenden Naturgesetzen folgt, dass Objekte mit einer bestimmten Mikrostruktur genau diese Analyse erfüllen. Es würde nicht ausreichen, nur zu sagen, dass Stoffe, deren Moleküle sich auf eine bestimmte Weise in einer Gitterstruktur anordnen, fest und wasserlöslich sind. Vielmehr muss gezeigt werden, *dass aus den grundlegenden Naturgesetzen folgt*, dass solche Stoffe genau die Merkmale besitzen, die für Festigkeit bzw. Wasserlöslichkeit charakteristisch sind. Für die Wasserlöslichkeit von Kochsalz bedeutet das, dass man

zeigen muss, dass aus den grundlegenden Naturgesetzen folgt, dass Wassermoleküle aufgrund ihrer Dipolstruktur die Fähigkeit besitzen, Natrium- und Chlorionen aus dem Gitter, das sie gebildet haben, herauszulösen und zu bewirken, dass sie sich zwischen den Wassermolekülen verteilen.

Zweitens zeigen die angeführten Beispiele, dass reduktive Erklärungen im Sinne Broads in allen Wissenschaften – besonders aber in der Chemie und der Biologie – allgegenwärtig sind. Solche Erklärungen zu finden ist für die Wissenschaft zumindest genauso wichtig wie allgemeine Gesetze zu entdecken. (Dies ist in meinen Augen eine in der Wissenschaftstheorie viel zu wenig beachtete Tatsache.) Hinter der Suche nach reduktiven Erklärungen steckt wohl das, was ich ‚die Idee der Einheit der Welt' nennen möchte. Allgemein wird zugestanden, dass unsere Welt eine Schichtenstruktur aufweist. Atome bestehen aus Elementarteilchen, Moleküle aus Atomen, Zellen aus Molekülen, Lebewesen aus Zellen und soziale Gruppen aus Lebewesen.[28] Auf den höheren Ebenen tauchen immer wieder neue Eigenschaften auf. Atome sind weder fest noch flüssig; nur Verbände von Molekülen können diese Eigenschaften haben. Moleküle atmen und ernähren sich nicht, noch pflanzen sie sich fort; erst Lebewesen kommen diese Eigenschaften zu. Die Idee der Einheit der Welt beruht nun auf der Annahme, dass sich die Eigenschaften komplexerer Wesen grundsätzlich allein unter Bezug auf die Eigenschaften und die Anordnung ihrer Teile reduktiv erklären lassen. Die Wasserlöslichkeit von Kochsalz kann allein unter Bezugnahme auf die Eigenschaften der Atome, aus denen Salzmoleküle bestehen, erklärt werden. Die Reinigungsfähigkeit von Seife kann ebenfalls auf die Eigenschaften der Moleküle zurückgeführt werden, aus denen Seife besteht. Dass Tiere die Fähigkeit haben sich fortzupflanzen, hat eine chemisch-physiologische Erklärung, usw. Natürlich ist es eine empirische Frage, ob sich tatsächlich alle höherstufigen Eigenschaften auf diese Weise reduktiv erklären lassen. (Emergentisten bestreiten genau diese Annahme.) Die These von der Einheit der Welt lässt sich nicht *a priori* als wahr erweisen. Aber die Wissenschaft scheint das Ziel zu verfolgen, so gut es geht zu zeigen, dass diese These zumindest empirisch wahr ist.

Kommen wir auf den entscheidenden Punkt zurück. Manchmal sind Wissenschaftler zufrieden, wenn sie die Frage „*Was* ist X?" beantworten können. „Granit ist magmatisches Gestein (Plutonit) mit richtungsloskörniger Struktur; es setzt sich aus Feldspat (meist Alkalifeldspat und Plagioklas), Quarz und Glimmer (Biotit oder Muskovit) sowie kleinen Anteilen weiterer Minerale wie Hornblende, Augit, Zirkon, Apatit, Magnetit, Ilmenit und Titanit zusammen."[29] „Asthma ist eine chronische Entzündung

[28] Vgl. Kim (1998, 15).

[29] Microsoft Encarta. © 1993-2003 Microsoft Corporation. Alle Rechte vorbehalten.

der Atemwege, die durch Antikörper verursacht wird, die vom humoralen Immunsystem in Reaktion auf eingeatmete Allergene produziert werden.“ In diesen Fällen ist kein Platz für ein zusätzliches „*Warum*?“. Auf der anderen Seite gibt es aber auch Fälle, in denen es nicht nur naheliegend, sondern auch völlig legitim ist, nach dem *Warum* zu fragen. Warum ist Kochsalz wasserlöslich? Um diese Frage zu beantworten, reicht es nicht aus, einfach auf die molekulare Struktur von Kochsalz zu verweisen. Nur eine reduktive Erklärung kann diese Frage beantworten. Man muss zeigen, dass aus den grundlegenden Naturgesetzen folgt, dass sich Stoffe mit der molekularen Struktur von Kochsalz in der für wasserlösliche Stoffe charakteristischen Weise verhalten. Wenn eine solche reduktive Erklärung gegeben werden kann, ist alles in Ordnung. Dann hat sich die These von der Einheit der Welt ein weiteres Mal bewährt. Wenn nicht, gibt es gewissermaßen einen Bruch oder eine Kluft in der Natur. Die Wasserlöslichkeit von Kochsalz würde sich als *emergente* Eigenschaft erweisen – als eine wirklich neue Eigenschaft mit Wirkungen, die nicht auf die molekulare Struktur von Kochsalz zurückgeführt werden können.

Im Einklang mit der wissenschaftlichen Praxis sollten wir sagen, dass solche Brüche in der Natur zeigen würden, dass der Physikalismus falsch ist – zumindest der Physikalismus im Hinblick auf die Wasserlöslichkeit von Kochsalz. Wenn man die wissenschaftliche Praxis ernst nimmt, muss man also unterscheiden. Bei Eigenschaften, bei denen es ausreicht, Was-Fragen zu beantworten, und bei denen Warum-Fragen unabgebracht sind, ist es sinnvoll, den Eigenschaftsphysikalismus im Sinne der Identitätstheorie zu verstehen. Aber bei Eigenschaften, deren Prädikate eine Analyse besitzen und bei denen deshalb Warum-Fragen beantwortet werden müssen, muss der Eigenschaftsphysikalismus im Sinne der Theorie der reduktive Erklärbarkeit verstanden werden. Es wäre nichts gewonnen, wenn man uns nur sagen würde, Wasserlöslichkeit ist identisch mit einer bestimmten Molekülstruktur; hier können wir erst zufrieden sein, wenn außerdem gezeigt ist, dass aus den grundlegenden Naturgesetzen folgt, dass Stoffe mit dieser Molekülstruktur das für Wasserlöslichkeit charakteristische Verhalten zeigen. Auf diesem Umstand beruht die intuitive Kraft des Arguments der Erklärungslücke.

## 7. Was muss ein Eigenschaftsphysikalist behaupten?

Wenn es richtig ist, dass man zwei Arten von Prädikaten unterscheiden muss, im Hinblick auf die ganz unterschiedliche Fragen angemessen sind, stehen wir jetzt aber vor dem Problem: Zu welcher Gruppe gehören die mentalen Prädikate?

Vertreter des analytischen Funktionalismus sind eindeutig der Meinung, dass z. B. ‚Schmerz' eine Analyse besitzt. Dieses Prädikat trifft auf eine Person genau dann zu, wenn sie in einem Zustand ist, der durch Gewebeverletzungen hervorgerufen wird, Schmerzäußerungen wie den Ausruf „Aua" sowie Verhalten zur Linderung der Schmerzen verursacht, zur Beeinträchtigung der Konzentrationsfähigkeit führt usw. Qualia-Freunde beeilen sich hinzuzufügen, dass das wichtigste Merkmal von Schmerzzuständen natürlich ist, dass sie einen bestimmten qualitativen Charakter besitzen, d. h., dass es schmerzhaft ist, in diesen Zuständen zu sein. Identitätstheoretiker sind anderer Meinung. Ihres Erachtens steht ‚Schmerz' für eine Eigenschaft, von der empirische Forschung dereinst zeigen wird, dass es sich um eine physische Eigenschaft handelt. Machen wir also den Kripke-Test.

Angenommen, empirische Forschung habe gezeigt, dass Schmerz in allen bisher untersuchten Wesen mit dem Feuern von C-Fasern korreliert ist. Das Feuern von C-Fasern ist also der beste Kandidat für die physische Eigenschaft, mit der Schmerz identisch sein könnte. Stellen wir uns jetzt die folgende *kontrafaktische* Situation vor. Wissenschaftler laden eine Gruppe von kompetenten Sprechern des Deutschen in ihre Arbeitsräume ein. Die Mitglieder der Gruppe wissen, dass Schmerz bei allen bisher untersuchten Wesen mit dem Feuern von C-Faser korreliert war. Die Wissenschaftler berichten: „In jüngster Zeit haben wir eine Gruppe von Ureinwohnern einer bisher unentdeckten Insel untersucht. Mitglieder dieser Gruppe zeigen keinerlei Schmerzverhalten, wenn ihre C-Fasern feuern; und sie berichten dann auch nicht, dass sie irgendwelche schmerzhaften Empfindungen hätten. Auf der anderen Seite zeigen die Mitglieder dieser Gruppe Schmerzverhalten und berichten von schmerzhaften Empfindungen, wenn ihre D-Fasern feuern." Möglicherweise ist die Gruppe der eingeladenen kompetenten Sprecher des Deutschen für einen Moment verwirrt. Aber was werden sie sagen? Wieder gibt es zwei Möglichkeiten. Einerseits können sie sagen: „Äußerst interessant; es gibt also Menschen, die Schmerzen haben, aber keinerlei Schmerzverhalten zeigen und auch keine schmerzhaften Empfindungen haben; und darüber hinaus scheint es einen *anderen* mentalen Zustand zu geben, der bei manchen Menschen mit genau den Merkmalen verbunden ist, die bei uns für das Haben von Schmerzen charakteristisch sind." Sie können aber auch sagen: „Äußerst interessant; es gibt also Menschen, die keine Schmerzen haben, obwohl ihre C-Fasern feuern, die aber Schmerzen haben, wenn ihre D-Fasern feuern." Wenn die Gruppe der eingeladenen kompetenten Sprecher des Deutschen die Situation auf die erste Art beschreiben würde, würde das bedeuten, dass ‚Schmerz' im Deutschen ein starrer Bezeichner ist, der für eine Eigenschaft steht, deren Natur herauszufinden Aufgabe der Wissenschaft ist. Wenn sie die Situation aber auf die zweite Art beschreiben würden, würde das zeigen, dass wir das Prädikat

‚Schmerz' so verwenden, dass es auf eine Person genau dann zutrifft, wenn sie das richtige Verhalten zeigt und die richtigen Empfindungen hat. Ich bin mir ganz sicher, dass sie die Situation auf die zweite Art beschreiben würden. ‚Schmerz' gehört also zu derselben Gruppe wie ‚wasserlöslich' und ‚durchsichtig' und nicht zu derselben Gruppe wie ‚Wasser' oder ‚Asthma'.

Bemerkenswerterweise wird dieses Ergebnis auch durch Kims Beobachtung gestützt, dass es völlig legitim ist, Fragen zu stellen wie „*Warum* führt das Feuern von C-Fasern zu Schmerzen, und nicht zu Jucken oder Kitzeln?", „*Warum* und *wie* entsteht bewusste Erfahrung aus bestimmten anderen neuronalen Zuständen?". Denn wir hatten schon gesehen, dass solche Warum-Fragen genau in den Fällen angemessen sind, in denen wir es mit Prädikaten wie ‚flüssig' und ‚giftig' zu tun haben.

Wenn mentale Prädikate jedoch zur selben Gruppe gehören wie ‚wasserlöslich' und ‚durchsichtig', dann passen sie nicht zur Identitätstheorie. Identitätsaussagen sind nur bei Prädikaten sinnvoll, die wie Namen funktionieren. Nur in diesem Fall können wir nach der realen Natur der Eigenschaften fragen, für die diese Prädikate stehen. Nur in diesem Fall kann man sinnvollerweise fragen, ob diese Eigenschaften mit physischen Eigenschaften identisch sind. Wenn mentale Prädikate zur selben Gruppe gehören wie ‚wasserlöslich' und ‚durchsichtig', geht es dagegen nicht um Identität, sondern um reduktive Erklärbarkeit. Genauso wie wir legitimerweise nach einer reduktiven Erklärung dafür fragen können, dass Kochsalz oder Zucker wasserlöslich sind, können wir nach einer reduktiven Erklärung dafür fragen, dass Wesen, deren C-Fasern feuern, Schmerz verspüren. Physikalismus im Hinblick auf mentalen Eigenschaften besagt also nicht, dass mentale Eigenschaften mit physischen Eigenschaft identisch sind, er besagt, dass sich alle mentalen Eigenschaften z. B. allein unter Bezug auf die neuronale Struktur von Lebewesen reduktiv erklären lassen, dass sie also nicht im Sinne von C.D. Broad emergent sind.

*Danksagung*

Ich bin Brian McLaughlin, Christian Nimtz, Michael Schütte, Sven Walter und besonders Wolfgang Schwarz sehr dankbar für ihre äußerst hilfreichen Kommentare zu früheren Versionen dieses Artikels.

## Literatur

Beckermann, A. (2001): *Analytische Einführung in die Philosophie des Geistes*. 2. überarbeitete Auflage, Berlin/New York.

– (2002): „Die reduktive Erklärbarkeit des phänomenalen Bewusstseins – C. D. Broad zur Erklärungslücke“. In: M. Pauen & A. Stephan (Hg.) *Phänomenales Bewußtsein – Rückkehr zur Identitätstheorie?* Paderborn: mentis, 122–147; in diesem Band Beitrag 2.

– (2009): „What is Property Physicalism?“. In: B. McLaughlin, A. Beckermann & S. Walter (Hg.) *Handbook of the Philosophy of Mind*. Oxford: Oxford University Press, 152–172.

Block, N. & R. Stalnaker (1999): „Conceptual Analysis, Dualism, and the Explanatory Gap“. *The Philosophical Review* 108, 1–46. (Dt. Übersetzung in: M. Pauen, M. Schütte & A. Staudacher (Hg.) *Begriff, Erklärung, Bewusstsein: Neue Beiträge zum Qualia-Problem*. Paderborn: mentis 2007, 61–109).

Braddon-Mitchell, D. & F. Jackson (1996): *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Broad, C. D. (1925): *The Mind and Its Place In Nature*. London.

Carnap, R. (1932): „Die physikalische Sprache als Universalsprache der Wissenschaft“. *Erkenntnis* 2, 432–465.

– (1932/33): „Psychologie in physikalischer Sprache“. *Erkenntnis* 3, 107–142.

– (1956). *Meaning and Necessity*. 2nd ed. Chicago: University of Chicago Press.

Chalmers, D. (1996): *The Conscious Mind*. Oxford: Oxford University Press.

– (2002): „Does Conceivability entail Possibility?“ In: T. S. Gendler & J. Hawthorne (Hg.) *Conceivability and Possibility*. Oxford: Oxford University Press, 145–200.

Chalmers, D. & F. Jackson (2001): „Conceptual Analysis and Reductive Explanation“. *The Philosophical Review* 110, 315–61.

Frege, G. (1892): „Über Sinn und Bedeutung“. *Zeitschrift für Philosophie und Philosophische Kritik*. Wiederabgedruckt in: Gottlob Frege, *Funktion, Begriff, Bedeutung*. Hg. von G. Patzig. 6. Auflage. Göttingen: Vandenhoeck & Ruprecht.

Hempel, C. G. (1949): „The Logical Analysis of Psychology“. In: H. Feigl & W. Sellars (Hg.) *Readings in Philosophical Analysis*. New York: Appleton-Century-Crofts, 373–384.

Hill, C. (1984): „In defense of type materialism“. *Synthese* 59, 295–320.

– (1991): *Sensations*. Cambridge: Cambrdige University Press.

Hill, C. & B. McLaughlin (1999): „There are Fewer Things in Reality than are Dreamt of in Chalmers' Ontology“. *Philosophy and Phenomenological Research* 59, 445–454.

Jackson, F. (1998): *From Metaphysics to Ethics*. Oxford: Oxford University Press.

– (2003): „From $H_2O$ to Water: the Relevance to A Priori Passage“, in: H. Lillehammer & G. Rodriguez-Pereyra (Hg.) *Real Metaphysics*. London: Routledge, 84–97.

– (2005): „The Case For A Priori Physicalism“. In: C. Nimtz & A. Beckermann (Hg.) *Philosophy – Science – Scientific Philosophy.* Paderborn: mentis, 251–265.

Kim, J. (1998): *Mind in a Physical World*. Cambridge MA: MIT Press.

– (2005): *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Lewis, D. (1986): *On the Plurality of Worlds*. Oxford: Blackwell.

Levine, J. (1983): „Materialism and Qualia: The Explanatory Gap“. *Pacific Philosophical Quarterly* 64, 354–61.

– (1993): „On Leaving Out What It’s Like“. In: M. Davies & G. W. Humphreys (Hg.), *Consciousness*. Oxford: Blackwell, 121–36.

McLaughlin, B. (2001): „In Defense of New Wave Materialism: A Response to Horgan and Tienson“. In: C. Gillett & B. Loewer (Hg.) *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, 319–330.

Papineau, D. (1998): „Mind the Gap“. In: J. Tomberlin (Hg.), *Philosophical Perspectives 12: Language, Mind, and Ontology*. Oxford: Ridgeview, 373–88.

Place, U. T. (1956): „Is consciousness a brain process?“ *British Journal of Psychology* 47, 44–50.

Putnam, H. (1975): „The Nature of Mental States“. In: H. Putnam *Mind, Language, and Reality. Philosophical Papers Vol. 2*. Cambridge: Cambridge University Press, 429–440.

Schütte, M. (2004): *Reduktion ohne Erklärung*. Paderborn: mentis.

Segal, G. (2000): *A Slim Book about Narrow Content*. Cambrige MA: MIT Press.

Smart, J.J.C. (1959): „Sensations and Brain Processes“. *Philosophical Review* 58, 141–156.

# 4

# Eigenschaftsidentität und reduktive Erklärung*

1. Vor einigen Jahren haben Autoren wie Papineau, Block und Stalnaker eine generelle Kritik an der These vorgebracht, es gebe eine enge Verbindung zwischen Eigenschaftsidentitäten und reduktiver Erklärbarkeit.[1] In diesem Aufsatz versuche ich die Gründe zu erläutern, die frühe Vertreter der Identitätstheorie dazu gebracht haben, zu glauben, dass es eine solche enge Verbindung in der Tat gibt. Darüber hinaus untersuche ich die Argumente von Papineau und Block und Stalnaker. Dabei ergibt sich am Ende, dass Eigenschaftsidentität tatsächlich nichts mit reduktiver Erklärbarkeit zu tun hat, *wenn* man reduktive Erklärung im Sinn von Levine und Broad versteht. Schließlich versuche ich, Jacksons Begriff der reduktiven Erklärung zu analysieren und die Argumente kritisch zu untersuchen, die er für die Annahme vorgebracht hat, wahre Eigenschaftsidentitätsaussagen müssten in seinem Sinn reduktiv erklärbar sein.

Ende der 1950er Jahre wurde mit den beiden Aufsätzen „Is Consciousness a Brain Process?“ von U.T. Place und „Sensations and Brain Processes“ von J.J.C. Smart die Identitätstheorie aus der Taufe gehoben. Empfindungen – oder allgemeiner: mentale Zustände –, so die These, sind typidentisch mit bestimmten physischen – höchstwahrscheinlich neuronalen – Zuständen. Schmerzen z. B. sind (möglicherweise) identisch mit dem Feuern bestimmter Neuronen – z. B. dem Feuern von C-Fasern – so wie Temperatur identisch ist mit der mittleren kinetischen Energie der Moleküle (mkE), wie Wasser identisch ist mit $H_2O$ und wie Blitze identisch sind mit bestimmten elektrischen Entladungen. Und das ist so, obwohl die Ausdrücke ‚x hat Schmerzen‘ und ‚in x’s ZNS feuern die C-Fasern‘ genau so wenig synonym sind wie die Ausdrücke ‚Temperatur‘ und ‚mkE‘, ‚Wasser‘ und ‚$H_2O$‘ sowie ‚Blitz‘ und ‚elektrische Entladung einer bestimmten Art‘. Wer diese These vertritt, steht allerdings sofort vor der *epistemischen* Frage: Wie kann man *herausfinden*, dass die Eigenschaften *F* und *G* identisch sind, wenn die Prädikate ‚*F*‘ und ‚*G*‘ nicht synonym sind?

Brandt und Kim haben in ihrem Aufsatz „The Logic of the Identity Theory“ folgende Antwort gegeben: Offenbar können *F* und *G* nur identisch sein, wenn ‚*F*‘ und ‚*G*‘ nomologisch koextensional sind, d.h., wenn der Satz

---

* Deutsche Fassung von „Property Identity and Reductive Explanation“, in: S. Gozzano & C. Hill (Hg.) *New Perspectives on Type-Identity*. Cambridge: Cambridge University Press 2012, 66–87.

1 Siehe etwa Papineau (1998) und Block & Stalnaker (1999).

(1) Für alle x: *F*x genau dann, wenn *G*x

eine wahre gesetzesartige Aussage ist. In aller Regel ist es eine empirische Frage, ob das so ist. Die Wahrheit und den nomologischen Charakter von (1) können wir mit empirischen Methoden überprüfen. Aber (1) allein impliziert nicht die Wahrheit von

(2) $F = G$.

Denn (1) ist z. B. auch mit dem psycho-physischen Parallelismus vereinbar. Welche *weiteren* Gründe rechtfertigen es, ausgehend von (1) auch (2) für wahr zu halten? Brandt und Kim schreiben: „The only reason we see for taking the step is that of parsimony" (Brandt/Kim 1967, 530). Neben den empirischen Gründen, die für (1) sprechen, sind es nach Brandt und Kim also allein Überlegungen zur ontologischen Sparsamkeit, die zur Stützung von Eigenschaftsidentitätsbehauptungen herangezogen werden können.

In der frühen Diskussion der Identitätstheorie gab es aber auch noch eine andere Antwort auf die epistemische Frage – eine Antwort, die sich insbesondere am Beispiel der folgenden Identitätsaussage orientierte:

(3) Temperatur = mkE.

Was spricht für die Wahrheit von (3)? Offenbar, so eine ganze Reihe von Autoren, ganz wesentlich die Tatsache, dass sich die klassische Thermodynamik auf die statistische Mechanik reduzieren lässt. Damit war ein Zusammenhang zwischen Identitätstheorie und Theorienreduktion hergestellt, der die Diskussion maßgeblich beeinflusst hat.[2]

Theorienreduktion wurde dabei im Sinn von Nagel verstanden: Eine Theorie $T_2$ ist genau dann auf eine Theorie $T_1$ reduzierbar, wenn alle Gesetze von $T_2$ aus $T_1$ abgeleitet werden können – notfalls unter Rückgriff auf geeignete Brückengesetze. Brückengesetze werden dann benötigt, wenn die zu reduzierende Theorie $T_2$ Begriffe enthält, die in der reduzierenden Theorie $T_1$ nicht vorkommen. Dass sich die klassische Thermodynamik auf die statistische Mechanik reduzieren lässt, zeigt sich unter anderem daran, dass sich aus der statistischen Mechanik das Gesetz von Boyle und Charles

(4) $P \cdot V = N \cdot k \cdot T$.

ableiten lässt. Denn aus der statistischen Mechanik folgt direkt das Gesetz

(5) $P \cdot V = N \cdot k \cdot \frac{\overline{mv^2}}{3k}$.

---

[2] Ich denke, dass dieser Zusammenhang auch dafür verantwortlich ist, dass die Identitätstheorie allgemein als eine *reduktionistische* Position verstanden wird.

Und aus diesem Gesetz folgt (4), wenn man das folgende Brückengesetz hinzuzieht:

(6) $T = \frac{2}{3k} \cdot \frac{\overline{mv^2}}{2}$

Warum spricht die Reduzierbarkeit der klassischen Thermodynamik auf die statistische Mechanik für die Wahrheit der Identitätsaussage (3)? Von philosophischer Seite kann man hauptsächlich zwei Argumente ins Feld führen. Das erste erwähnt schon Kim in seinem Aufsatz „On the Psycho-Physical Identity Theory", wobei Kim dieses Argument Putnam zuschreibt (vgl. Kim 1966, 228 f.). Eigenschaftsidentitätsaussagen werden dadurch gerechtfertigt, dass sie Reduktionen überhaupt erst ermöglichen.

> The reduction of one scientific theory to another involves the derivation of the laws of the reduced theory from the laws of the theory to which it is reduced. If the reduction is to be genuinely inter-theoretic, the reduced theory will contain concepts not included in the vocabulary of the reducing theory, and these concepts will occur essentially in the laws of the reduced theory. Hence, if these laws are to be derived from the laws of the reducing theory in which those concepts do not occur, we shall need, as auxiliary premises of derivation, certain statements in which concepts of both theories occur. We may refer to these statements as „connecting principles." (Kim 1966, 228)

Putnam scheint, so Kim, zu argumentieren, dass gerade Identitätsaussagen wie „Gas ist eine Ansammlung von Molekülen" und „Temperatur ist mkE" in idealer Weise geeignet sind, die Rolle solcher Brückengesetze zu spielen. Diese Identitätsaussagen sind also dadurch gerechtfertigt, dass sie die Reduktion der klassischen Thermodynamik auf die statistische Mechanik ermöglichen und damit zu einer erheblichen Vereinheitlichung der physikalischen Weltsicht beitragen.[3] Gegen dieses Argument wendet Kim ein, dass etwa die Aussage (6) auch dann die Rolle eines Brückengesetzes übernehmen kann, wenn man sie nicht als Identitätsaussage, sondern nur als wahres nomologisches Bikonditional versteht.

Ein zweites philosophisches Argument lässt sich so formulieren: Wenn man davon ausgeht, dass ein Begriff wie Temperatur durch eine bestimmte kausale Rolle definiert ist und dass diese kausale Rolle genau durch die Gesetze der klassischen Thermodynamik spezifiziert wird, dann ergibt sich aus der Reduzierbarkeit der klassischen Thermodynamik auf die statistische Mechanik, dass es eine andere physikalische Größe gibt – nämlich mkE –, die genau diese kausale Rolle spielt. Da es aber unplausibel ist, anzunehmen, dass zwei verschiedene Eigenschaften dieselbe kausale Rolle

[3] Vgl. zu dieser Argumentation die Überlegungen von Block und Stalnaker (siehe unten S. 89 f.).

innehaben,[4] spricht dieses Ergebnis dafür, dass Temperatur und mkE identisch sind.

Ganz offensichtlich stützt sich Joseph Levine mit seinen Überlegungen zur Erklärungslücke genau auf dieses letzte Argument. Schon in dem Aufsatz „Materialism and Qualia: The Explanatory Gap" wird das ganz deutlich. Levine beginnt diese Überlegungen mit dem Vergleich der Identitätsaussage (3) mit der Identitätsaussage

(7) Schmerz = das Feuern von C-Fasern.

Zwischen diesen Aussagen gibt es Levine zufolge einen wichtigen Unterschied. Während es auf der einen Seite in einem gewissen Sinne *undenkbar* ist, dass in einem Gas die mittlere kinetische Energie der Moleküle einen bestimmten Wert (sagen wir, $6.21 \cdot 10^{-21}$ Joule) hat, dieses Gas aber nicht die entsprechende Temperatur von 300 K besitzt, ist es auf der anderen Seite durchaus *denkbar*, dass in meinem ZNS die C-Faser feuern, ich aber keine Schmerzen empfinde. Nach Levine liegt dies daran, dass die Aussage (3) *vollständig explanatorisch* ist, die Aussage (7) dagegen nicht. Was ist damit gemeint?

Auf diese Frage gibt Levine folgende Antwort. Wenn man uns fragen würde, was wir mit dem Ausdruck „Temperatur" meinen, würden wir antworten:

(3′) Temperatur ist die Eigenschaft von Körpern, die in uns bestimmte Wärme- bzw. Kälteempfindungen hervorruft, die dazu führt, dass die Quecksilbersäule in Thermometern, die mit diesen Körpern in Berührung kommen, steigt oder fällt, die bestimmte chemische Reaktionen auslöst, und so weiter.

Wir würden Temperatur also allein durch ihre *kausale Rolle* charakterisieren. Dies würde als Antwort auf die gestellte Frage jedoch nicht ausreichen, wenn nicht noch ein zweiter Punkt hinzukäme.

> […] our knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat. Furthermore, antecedent to our discovery of the essential nature of heat, its causal role, captured in statements like [(3′)], exhausts our notion of it. Once we understand how this causal role is carried out there is nothing more we need to understand. (Levine 1983, 357)

Der explanatorische Charakter der Aussage (3) beruht also auf zwei Tatsachen:

---

[4] Achinstein (1974) hat sogar für das Prinzip argumentiert, dass Eigenschaften genau dann identisch sind, wenn sie dieselben kausalen Eigenschaften besitzen.

1. Unser Begriff von Temperatur erschöpft sich vollständig in ihrer kausalen Rolle.
2. Physik und Chemie können verständlich machen, dass die mittlere kinetische Energie der Moleküle eines Gases genau diese kausale Rolle spielt.

Mit anderen Worten: Levine zufolge ist die Aussage (3) vollständig explanatorisch, weil aus den Gesetzen der Physik folgt, dass mkE genau die kausale Rolle spielt, durch die die Eigenschaft Temperatur charakterisiert ist. Hinter dem Ausdruck „vollständig explanatorisch" steht somit offenbar eine Theorie der reduktiven Erklärung, deren Kernidee sich so zusammenfassen lässt:

($RE_L$) Eine Eigenschaft *F* lässt sich genau dann reduktiv mit Bezug auf eine physikalisch-chemische Eigenschaft *G* erklären, wenn gilt:
(a) *F* ist charakterisiert durch eine Menge von Merkmalen $M_F$ und
(b) aus den grundlegenden Naturgesetzen lässt sich ableiten, dass alle Gegenstände, die *G* besitzen, auch alle Merkmale von $M_F$ haben.

Vor dem Hintergrund dieser Theorie lässt sich auch verstehen, warum Levine zufolge die Aussage (7) nicht vollständig explanatorisch ist. Mit dem Ausdruck „Schmerz" assoziieren wir zwar ebenfalls eine kausale Rolle: Schmerzen werden durch die Verletzung von Gewebe verursacht, sie führen dazu, dass wir schreien oder wimmern, und sie bewirken in uns den Wunsch, den Schmerz so schnell wie möglich loszuwerden. Dies bestreitet auch Levine nicht. Und er bestreitet auch nicht, dass die Identifikation von Schmerzen mit dem Feuern von C-Fasern den Mechanismus erklären würde, auf dem diese kausale Rolle beruht. Dennoch gibt es seiner Meinung nach einen entscheidenden Unterschied.

> However, there is more to our concept of pain than its causal role, there is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firing is *why pain should feel the way it does*! For there seems to be nothing about C-fiber firing which makes it naturally ‚fit' the phenomenal properties of pain, any more than it would fit some other set of phenomenal properties. Unlike its functional role, the identification of the qualitative side of pain with C-fiber firing […] leaves the connection between it and what we identify it with completely mysterious. One might say, it makes the way pain feels into merely a brute fact. (Levine 1983, 357)

Levines erster Grund für die These, dass die Aussage (7) nicht vollständig explanatorisch ist, ist also:

3. Unser Begriff von Schmerzen erschöpft sich *nicht* in einer kausalen Rolle; er umfasst auch einen qualitativen Aspekt – die Art, wie es sich anfühlt, Schmerzen zu haben.

Dies allein ist aber nicht entscheidend. Denn die Aussage (7) könnte immer noch vollständig explanatorisch sein, wenn die Neurobiologie nur verständlich machen könnte, dass sich das Feuern von C-Fasern schmerzhaft anfühlt. Levines zweiter Grund ist daher, dass genau dies nicht der Fall ist.

4. Aus den allgemeinen Gesetzen der Neurobiologie folgt nicht, dass sich das Feuern von C-Fasern auf die für Schmerzen charakteristische Weise – nämlich schmerzhaft – anfühlt.

Letzten Endes vertritt Levine also folgende Position: Wenn *F* mit Bezug auf *G* reduktiv erklärt werden kann, d. h., wenn *F* durch bestimmte – kausale und nicht kausale – Merkmale $M_F$ charakterisiert ist und wenn es eine physische Eigenschaft *G* gibt, für die aus den grundlegenden Naturgesetzen folgt, dass alle Gegenstände, die *G* besitzen, auch alle Merkmale von $M_F$ besitzen, spricht das eindeutig für die Identität von *F* und *G*. Wenn aus den grundlegenden Naturgesetzen *nicht* folgt, dass alle Gegenstände, die *G* besitzen, auch alle Merkmale von $M_F$ besitzen, bedeutet das *nicht*, dass *F* und *G nicht identisch* sind. Aber in diesem Fall ist nicht mehr zu sehen, wie wir die Identitätsaussage „*F* = *G*“ begründen können.

In seinem Aufsatz „On Leaving Out What It’s Like“ versucht Levine bemerkenswerter Weise, diese Überlegungen auch auf das Beispiel

(8) Wasser = $H_2O$

anzuwenden. Nach Levine ist (8) im folgenden Sinn genau so vollständig explanatorisch wie (3). So wie es – gegeben die fundamentalen Gesetze der Physik und Chemie – *undenkbar* ist, dass in einem Gas die mittlere kinetische Energie der Moleküle $6.21 \cdot 10^{-21}$ Joule beträgt, dieses Gas aber nicht die entsprechende Temperatur von 300 K besitzt, ist es auch *undenkbar*, dass $H_2O$ nicht die Oberflächeneigenschaften hat, die wir an Wasser feststellen können. Das ist so, weil „the chemical theory of water explains what needs to be explained“ (Levine 1993, 128).

> What is explained by the theory that water is $H_2O$? Well, as an instance of something that’s explained by the reduction of water to $H_2O$, let’s take its boiling point at sea level. The story goes something like this. Molecules of $H_2O$ move about at various speeds. Some fast-moving molecules that happen to be near the surface of the liquid have sufficient kinetic energy to escape the intermolecular attractive forces that keep the liquid intact. These molecules enter the atmosphere. That’s evaporation. The precise value of the intermolecular attractive forces of $H_2O$ molecules determines the vapour pressure of liquid masses of $H_2O$, the pressure exerted by molecules attempting to escape into saturated air. As the average kinetic energy of the

> molecules increases, so does the vapour pressure. When the vapour pressure reaches the point where it is equal to atmospheric pressure, large bubbles form within the liquid and burst forth at the liquid's surface. The water boils. I claim that given a sufficiently rich elaboration of the story above, it is inconceivable that $H_2O$ should not boil at 212° F at sea level (assuming, again, that we keep the rest of the chemical world constant). (Levine 1993, 129)

In meinen Augen wirft die These, dass diese Überlegung eine Rechtfertigung für die Identitätsaussage (8) liefert, aber mehr Fragen auf, als sie beantwortet. Auf der Welt gibt es an vielen Stellen einen – meistens flüssigen – Stoff, den wir „Wasser" nennen. Nehmen wir an, wir lassen viele Proben dieses Stoffes in einem chemischen Labor analysieren; dabei erhalten wir in allen Fällen dasselbe Ergebnis: Die Proben bestehen abgesehen von unbedeutenden Verunreinigungen aus $H_2O$. Das reicht in meinen Augen als Rechtfertigung für (8) völlig aus – vorausgesetzt, wir akzeptieren die beiden Hintergrundannahmen, dass ‚Wasser' ein Artbegriff ist, der einen chemischen Stoff bezeichnet, und dass chemische Stoffe durch ihre molekulare Struktur individuiert werden. Würde sich an dieser Rechtfertigung von (8) etwas ändern, wenn sich nicht alle, nur wenige oder sogar gar keine der Oberflächeneigenschaften von Wasser durch die chemische Theorie des Wassers erklären ließen? Würde es diese Rechtfertigung beeinträchtigen, wenn sich herausstellen sollte, dass einige dieser Oberflächeneigenschaften auf Verunreinigungen zurückgehen, die sich zufälligerweise fast überall finden – so wie sich die gelbe Farbe von Gold daraus ergibt, dass fast alle Proben von Gold Anteile von Kupfer enthalten? Würden wir dann sagen, Wasser sei gar nicht $H_2O$, sondern $H_2O$ + eine kleine Menge ABC? Nein, das würden wir nicht sagen. Auch wenn sich keine der Oberflächeneigenschaften von Wasser durch die chemische Theorie des Wassers erklären ließe, wäre Wasser unter den angegebenen Bedingungen nach wie vor $H_2O$.

Dies zeigt nicht nur, dass an Levines Thesen über den Zusammenhang von Identitätsaussagen und reduktiver Erklärbarkeit etwas nicht stimmt; es zeigt insbesondere auch, dass sich die beiden Identitätsaussagen (3) und (8) viel mehr unterscheiden, als gemeinhin angenommen wird. Kim hat dies schon in seinem Aufsatz „On the Psycho-Physical Identity Theory" gesehen. Die Identitätsaussagen (7) und (8) seien, so Kim, „disanalogous", und die Aussage (3) in ihrem Status der Aussage (7) sehr viel ähnlicher als der Aussage (8).

> „Water" and „$H_2O$" (in the sense of „substance whose molecular structure is $H_2O$") are both substantive expressions referring to physical things and not to properties, events, states, or the like. Any bit of water has a decomposition into $H_2O$ molecules; the two occupy the same spatio-temporal volume. […] In this sense, water has a decomposition into $H_2O$ molecules; gas a decomposition into molecules and atoms. […] So, water is literally made up of $H_2O$

> molecules, and a body of gas, of molecules and atoms. Temperature, however, is unlike water and gas. Temperature is not a thing that is made up of certain parts; we cannot pick out a bit of temperature or an instance of it and say that it is made up of mean kinetic energy. The domain of classical thermodynamics does not contain temperature in the way the domain of macro-chemistry contains water; rather, it contains gas, or bodies of gas, and temperature is a state variable whose values are used to characterize the thermodynamic states of a system – in other words, it is a property of the things in the domain. But it in itself is not a thing: it has no decomposition into mean kinetic energy. (Kim 1966, 230)

In der Aussage (8) geht es darum, was für eine Art Stoff Wasser ist, und die Antwort auf diese Frage wird gegeben, indem man sagt, aus welchen Molekülen Wasser besteht. Aussage (8) ähnelt insofern den Aussagen

(9) Wolken sind dichte Ansammlungen von Wassertropfen oder Eiskristallen am Himmel

und

(10) Granit ist magmatisches Gestein (Plutonit) mit richtungslos-körniger Struktur; es setzt sich aus Feldspat (meist Alkalifeldspat und Plagioklas), Quarz und Glimmer (Biotit oder Muskovit) sowie kleinen Anteilen weiterer Minerale wie Hornblende, Augit, Zirkon, Apatit, Magnetit, Ilmenit und Titanit zusammen.[5]

In den Aussagen (3) und (7) dagegen geht es um *Makro-Eigenschaften* eines Systems. Und obwohl viele Philosophen glauben, dass wir in ähnlicher Weise nach der Natur von Makro-Eigenschaften fragen können, habe ich genau wie Kim Zweifel, dass dies die richtige Art ist, das Problem zu formulieren. Eigenschaften haben keine Natur in dem Sinne, in dem chemische Stoffe eine Natur besitzen; sie sind keine Dinge, die in bestimmter Weise zusammengesetzt sind. Bei Makro-Eigenschaften ist die entscheidende Frage daher nicht die Frage nach ihrer Natur, sondern ob sie sich allein unter Bezug auf die Eigenschaften der Teile des Systems und deren Anordnung erklären lassen. Zumindest ist genau dies die Frage, um die es C.D. Broad bei Entwicklung seiner Theorie mechanischer und emergenter Erklärung ging.

2. Broad entwickelte einen Begriff der mechanischen, d. h. reduktiven, Erklärung, der dem von Levine in vielerlei Hinsicht gleicht.[6] Allerdings gibt es auch einen bemerkenswerten Unterschied. Broad interessiert sich nicht

---

[5] Place spricht in diesen Fällen vom „‚is' of composition".

[6] Eine ausführliche Analyse der Broadschen Theorie findet sich in Beckermann (2002), in diesem Band Beitrag 2.

für die *allgemeine* Frage, was es heißt, dass *F* mit Bezug auf *G* reduktiv erklärt werden kann. Ihm geht es um ein spezielleres Problem: Welche Beziehung besteht zwischen den *Makroeigenschaften* eines komplexen Systems und den *Eigenschaften und der Anordnung seiner physischen Teile*. Ist *S* ein komplexes System, das aus den Teilen *A*, *B* und *C* besteht, die auf die Weise *R* räumlich angordnet sind, und *F* eine Makroeigenschaft dieses Systems, dann behaupten Mechanisten, dass *F* zumindest im Prinzip „aus der vollständigen Kenntnis der Eigenschaften deduziert werden kann, die *A*, *B* und *C* isoliert oder in anderen komplexen Systemen besitzen, die nicht die Form *R*(*A*, *B*, *C*) haben" (Broad 1925, 61). Emergentisten bestreiten genau das, obwohl auch sie zugestehen, dass in diesem Fall der Satz „Alle Systeme mit der Form *R*(*A*, *B*, *C*) haben die Eigenschaft *F*" ein wahres Naturgesetz ist. In dem Aufsatz „Die reduktive Erklärbarkeit des phänomenalen Bewusstseins – C.D. Broad zur Erklärungslücke" (in diesem Band S. 21–45) habe ich versucht, diese Auffassung genauer zu analysieren und zu zeigen, dass sie – zumindest im Prinzip – auf die folgende Definition mechanischer Erklärbarkeit hinausläuft:

($ME_B$) Die Makroeigenschaft *F* eines komplexen Systems *S*, das aus den Teilen $C_1$, …, $C_n$ besteht, die auf die Weise *R* angeordnet sind, ist genau dann *mechanisch erklärbar*, wenn aus den für die Teile von *S* geltenden grundlegenden Naturgesetzen folgt, dass alle Systeme mit derselben Mikrostruktur alle *Merkmale* besitzen, *die für F charaktistisch sind*.

Allerdings darf man nie vergessen, dass für Broad mechanische Erklärung immer eine Zwei-Ebenen-Angelegenheit ist – auf der einen Seite gibt es die Ebene des komplexen Systems selbst und auf der anderen Seite die Ebene seiner Teile. Daher stellt sich die Frage, wie es überhaupt möglich sein soll, dass die für eine Makro-Eigenschaft *F* charakteristischen Merkmale aus den grundlegenden Naturgesetzen abgeleitet werden können, die für die Teile des Systems gelten, das diese Makro-Eigenschaft besitzt. Meiner Meinung nach besteht Broads Antwort auf diese Frage aus zwei Teilen: Damit eine Makro-Eigenschaft mechanisch erklärt werden kann, ist zunächst erforderlich, dass die Art und Weise, auf die sich die Teile $C_1$, …, $C_n$ verhalten, wenn sie auf die Weise *R* angeordnet sind, (nennen wir diese Art und Weise *V*) aus den grundlegenden allgemeinen Gesetzen folgt, die für die Teile gelten. Zweitens muss es aber zusätzlich Brückenprinzipien geben, aus denen hervorgeht, dass jedes System mit der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle Merkmale besitzt, die für die Makro-Eigenschaft charakteristisch sind, wenn sich die Teile $C_1$, …, $C_n$ auf die Weise *V* verhalten.[7]

[7] Für Beispiele siehe unten Fußnote 17. Brückenprinzipien, die verschiedene Ebenen miteinander verbinden, unterscheiden sich in vielerlei Hinsicht von

Versuchen wir, diese Idee an einem Beispiel zu erläutern. Dass die Wasserlöslichkeit von Kochsalz mechanisch erklärbar ist, lässt sich auf folgende Weise zeigen. Kochsalz besteht aus $Na^+$- und $Cl^-$-Ionen, die in einer Gitterstruktur abgeordnet sind. Aufgrund ihrer Dipolstruktur sind $H_2O$-Moleküle in der Lage, einzelne $Na^+$- und $Cl^-$-Ionen aus dieser Gitterstruktur herauszulösen und zu erreichen, dass sich diese Ionen zwischen den Wassermolekülen verteilen. All dies spielt sich noch auf der Ebene der Teile ab. Aber was haben die Dinge, die sich auf dieser Ebene abspielen, mit der Eigenschaft der Wasserlöslichkeit auf der Makro-Ebene zu tun? Diese Eigenschaft von Kochsalz wird durch das Verhalten der Ionen und Moleküle auf der Mikro-Ebene nur erklärt, wenn wir zusätzlich das folgende Brückenprinzip in Anspruch nehmen: „Wenn sich die Moleküle, aus denen ein Gegenstand besteht, voneinander lösen und zwischen den Wassermolekülen verteilen, wenn dieser Gegenstand in Wasser gegeben wird, dann löst sich dieser Gegenstand in Wasser auf.“ Broads Begriff der mechanischen Erklärung kann daher letzten Endes besser so formuliert werden:

($ME_B'$) Die Makroeigenschaft $F$ eines komplexen Systems $S$, das aus den Teilen $C_1$, …, $C_n$ besteht, die auf die Weise $R$ angeordnet sind, d. h. eines Systems mit der Mikrostruktur [$C_1$, …, $C_n$; $R$], ist genau dann *mechanisch erklärbar*, wenn Folgendes gilt:

(a) Die Art und Weise $V$, auf die sich die Teile $C_1$, …, $C_n$ verhalten, wenn sie auf die Weise $R$ angeordnet sind, lässt sich aus den grundlegenden allgemeinen Gesetzen ableiten, die für diese Teile gelten; und

(b) es gibt Brückenprinzipien, die besagen, dass Systeme mit der Mikrostruktur [$C_1$, …, $C_n$; $R$] alle *Merkmale* besitzen, *die für F charakteristisch sind*, wenn sich die Teile $C_1$, …, $C_n$ auf die Weise $V$ verhalten.

Aus dieser Formulierung folgt, dass mechanische Erklärungen auf zweierlei Weise scheitern können. Sie können scheitern, weil sich die Art und Weise, auf die sich die Teile $C_1$, …, $C_n$ verhalten, wenn sie auf die Weise $R$ angeordnet sind, nicht aus den grundlegenden allgemeinen Gesetzen ableiten lässt, die für diese Teile gelten. Und sie können scheitern, weil es keine geeigneten Brückenprinzipien gibt.

3. Es gibt also einen sehr engen Zusammenhang zwischen Levines Begriff der reduktiven Erklärung und Broads Begriff der mechanischen Erklärung. In beiden Fällen geht es um die Erklärung von (Makro-)Eigenschaften. In beiden Fällen wird vorausgesetzt, dass die zu erklärende Eigenschaft $F$

---

den Brückengesetzen, die oben in Abschnitt 1 erwähnt wurden. Dennoch verdienen auch sie es, als ‚*Brücken*prinzipien‘ bezeichnet zu werden.

durch eine Menge von Merkmalen $M_F$ charakterisiert ist. Und in beiden Fällen besteht der Kern der Erklärung darin, dass gezeigt wird, dass aus den grundlegenden Naturgesetzen folgt, dass alle Gegenstände mit der Eigenschaft *G* bzw. der Mikrostruktur [$C_1$, …, $C_n$; *R*] alle Merkmale von $M_F$ besitzen.[8] Doch zurück zur Frage nach dem *Zusammenhang* zwischen Identität und reduktiver Erklärbarkeit im Sinne Levines oder Broads. Zunächst: Broad ist an Eigenschaftsidentitäten nicht interessiert. Für ihn ist die entscheidende Frage im Vitalismusstreit nicht, ob Lebenseigenschaften mit physiko-chemischen Eigenschaften identisch sind, sondern ob man zur Erklärung dieser Eigenschaften auf eigene Komponenten zurückgreifen muss (eine Entelechie oder einen *élan vital*), ob diese Eigenschaften emergent oder ob sie mechanisch erklärbar sind. Auch wenn es um die Wasserlöslichkeit von Kochsalz geht, ist für ihn nicht die Frage, ob Wasserlöslichkeit eventuell mit einer anderen (physikalisch grundlegenderen) Eigenschaft identisch ist, sondern ob aus den allgemeinen Naturgesetzen folgt, dass $H_2O$-Moleküle in der Lage sind, die einzelnen Ionen aus einem Natrium$^+$-Chlor$^-$-Ionengitter herauszulösen, so dass sie sich zwischen den $H_2O$-Moleküle verteilen. Für Levine dagegen besteht ein Zusammenhang zwischen Eigenschafts-Identitätsaussagen und reduktiver Erklärbarkeit. Für ihn ist die Tatsache, dass *F* unter Bezug auf *G* reduktiv erklärt werden kann, ein Grund (und möglicherweise der einzige Grund), der für die Wahrheit der Identitätsaussage „*F* = *G*“ spricht.

Wir hatten schon gesehen, dass die Annahme eines engen Zusammenhangs zwischen Eigenschafts-Identitätsaussagen und reduktiver Erklärbarkeit für verschiedene Arten von Identitätsaussagen unterschiedlich plausibel ist. Die Aussage (8) „Wasser = $H_2O$“ hat mit reduktiver Erklärbarkeit offensichtlich nichts zu tun – weder inhaltlich noch epistemisch. (8) kann auch dann wahr sein, wenn sich die Oberflächeneigenschaften von Wasser nicht reduktiv erklären lassen; und die Reduzierbarheit dieser Eigenschaften ist keineswegs der einzige Grund, der uns dazu bringen könnte, (8) für wahr zu halten. Bei der Aussage (3) „Temperatur = mkE“ sind die Dinge

[8] Die diesen Erklärungsbegriffen zugrunde liegende Idee geht im Übrigen mindestens zurück bis Descartes, der z. B. im *Discours* schreibt, dass es ihm darum gehe, zu zeigen, dass sich der Prozess des Herzschlags *notwendigerweise* aus den Teilen des Herzens und deren Anordnung ergebe.
„[D]er soeben erklärte Mechanismus [ergibt] sich allein aus der Einrichtung der Organe […], die man im Herzen mit seinen Augen sehen, aus der Wärme, die man dort mit seinen Fingern spüren, und aus der Natur des Blutes, die man durch Erfahrung kennenlernen kann, und dies mit der gleichen Notwendigkeit, wie der Mechanismus einer Uhr aus der Kraft, Lage und Gestalt ihrer Gewichte und Räder folgt.“ (Descartes 1960, 81 ff.)

*prima facie* nicht so klar. Möglicherweise ist dies aber ein Hinweis darauf, dass es ein Fehler ist, (3) überhaupt als Identitätsaussage zu verstehen. Vielleicht ist der Inhalt von (3) eher: Temperatur ist in Gasen durch mkE *realisiert*. Dies würde jedenfalls dazu passen, dass die Idee der reduktiven Erklärbarkeit doch sehr viel besser zum Funktionalismus als zur Identitätstheorie passt. Levine betont ja immer wieder, dass der Begriff der Temperatur durch eine kausale Rolle charakterisiert ist – eine kausale Rolle, die offenbar implizit durch die Gesetze der klassischen Thermodynamik spezifiziert wird. Und wenn sich die Gesetze der klassischen Thermodynamik aus der statistischen Mechanik ableiten lassen, dann zeigt das, dass die Größe $\frac{2}{3k} \cdot \frac{\overline{mv^2}}{2}$ eben diese kausale Rolle innehat, mit anderen Worten, dass Temperatur in Gasen durch mkE *realisiert* wird. Außerdem ist leicht zu sehen, dass reduktive Erklärbarkeit durchaus mit multipler Realisierung vereinbar ist. Denn dass mkE in Gasen die Temperaturrolle innehat, schließt ja keineswegs aus, dass Temperatur in anderen Stoffen durch andere Eigenschaften realisiert ist,[9] sowie ja auch Wasserlöslichkeit sehr unterschiedlich realisiert sein kann.

Unabhängig von diesen Überlegungen ist auch von anderen Autoren, unter anderem von Papineau und Block und Stalnaker, die These, es gebe einen engen Zusammenhang zwischen Eigenschafts-Identitäten und reduktiver Erklärbarkeit, einer grundsätzlichen Kritik unterzogen worden. In seinem Aufsatz „Mind the Gap" bekennt sich Papineau explizit zur Identitätstheorie.

> My first task is to show that physicalism is best conceived as a thesis about property identity. (Papineau 1998, 374)

Eine mentale Eigenschaft *M* kann, so Papineau, aber sehr wohl auch dann mit einer physikalischen (oder funktionalen) Eigenschaft *P* identisch sein,

---

[9] Vgl. zu dieser Überlegung auch die Bemerkung Paul Churchlands: „Strictly speaking, however, this identity is true only for the temperature of a gas, where simple particles are free to move in ballistic fashion. In a solid, temperature is realized differently, since the interconnected molecules are confined to a variety of vibrational motions. In a plasma, temperature is something else again, since a plasma has no constituent molecules; they, and their constituent atoms, have been ripped to pieces. And even a vacuum has a so-called ‚blackbody' temperature – in the distribution of electromagnetic waves coursing through it. Here temperature has nothing to do with the kinetic energy of particles. It is plain that the physical property of temperature enjoys ‚multiple instantiations' no less than do psychological properties." (Churchland 1988, 41)

wenn *nicht* aus den grundlegenden Gesetzen der Physik folgt, dass alle Gegenstände, die die Eigenschaft *P* besitzen, die Analyse von *M* erfüllen. Identitäten bestehen oder sie bestehen nicht. Es hat keinen Sinn zu fragen, warum zwei Dinge oder Eigenschaften identisch sind. Und deshalb spielt es für die Frage, ob *M* und *P* identisch sind, auch keine Rolle, ob wir verstehen, wie *P M* hervorbringt. Identische Eigenschaften bringen einander nicht hervor, sie sind einfach identisch. Fragen kann man nur, was dafür spricht, dass *M* und *P* identisch sind. Und auf diese Frage ist nach Papineau die beste Antwort, dass *M* und *P* dieselben Ursachen und Wirkungen haben.[10]

Wie Papineau kritisieren auch Block und Stalnaker die Annahme, Physikalisten seien auf die These festgelegt, dass mentale Eigenschaften reduktiv erklärbar sind. In ihrem Aufsatz „Conceptual Analysis, Dualism, and the Explanatory Gap" vertreten sie die Auffassung, dies könne gar nicht so sein. Denn erstens setze reduktive Erklärbarkeit voraus, dass das zu erklärende Phänomen *F* so analysiert werden könne, dass in dieser Analyse nur Begriffe verwendet werden, die auch in den allgemeinen Naturgesetzen vorkommen. Genau dies sei im Allgemeinen aber nicht möglich, und schon gar nicht bei mentalen Phänomenen. Reduktive Erklärungen müssten daher in der Regel fehlschlagen. Daraus ergebe sich jedoch zweitens kein Argument gegen den Physikalismus. Denn der Physikalist sei nur auf eine Identitätsbehauptung festgelegt; und mentale Eigenschaften könnten auch dann mit physikalischen Eigenschaften identisch sein, wenn sie nicht reduktiv erklärt werden können. Natürlich müssen aber auch Block und Stalnaker eine Antwort auf die Frage geben, wie wir Identitätsbehauptungen begründen, auf welche Weise wir diese Behauptungen rechtfertigen können.

Für Block und Stalnaker werden Identitätsbehauptungen ganz allgemein durch Schlüsse auf die *beste Erklärung* gerechtfertigt. Ähnliche Auffassungen haben auch Hill und McLaughlin vertreten.[11] Für die Aussage (8) „Wasser = $H_2O$" bedeutet das konkret: Wenn wir herausfinden wollen, ob diese Aussage wahr ist, stützen wir uns auf die folgende Überlegung. Wir wissen, dass Wasser durch Erwärmen zum Kochen gebracht wird. Weiter klärt uns die Wissenschaft darüber auf, warum ein Anstieg der mittleren kinetischen Energie von $H_2O$-Molekülen zu einer bestimmten Aktivität M dieser Moleküle führt. Wenn wir annehmen, dass Wasser mit $H_2O$, Tempe-

[10] Hier sind Papineau und Levine offenbar gar nicht so weit auseinander. Denn die Reduzierbarkeit der klassischen Thermodynamik auf die statistische Mechanik zeigt ja gerade, dass die mittlere kinetische Energie der Moleküle die kausalen Eigenschaften hat, durch die die Eigenschaft Temperatur charakterisiert ist. Papineau könnte aber argumentieren, Gleichheit von Ursachen und Wirkungen ließe sich auch anders als durch reduktive Erklärung zeigen.

[11] Siehe Hill (1991), Hill & McLaughlin (1999) und Mclaughlin (2001).

ratur mit der mittleren kinetischen Energie der Moleküle und Kochen mit der molekularen Aktivität M identisch ist, gilt daher:

> Then we have an account of how heating produces boiling. If we were to accept mere correlations instead of identities, we would only have an account of how something correlated with heating causes something correlated with boiling. Further, we may wish to know how it is that increasing the molecular kinetic energy of a packet of water causes boiling. Identities allow a transfer of explanatory and causal force not allowed by mere correlations. Assuming that heat = mkE, that pressure = molecular momentum transfer, etc. allows us to explain facts that we could not otherwise explain. Thus, we are justified by the principle of inference to the best explanation in inferring that these identities are true. (Block/Stalnaker 1999, 23 f.)[12]

Die Annahme, dass Wasser mit $H_2O$ und Temperatur mit der mittleren kinetischen Energie der Moleküle eines Stoffes identisch ist, führt zu einem einfacheren und kohärenteren Bild der Welt. *Dies allein* rechtfertigt diese Identitätsaussagen. Ich hatte schon ausgeführt, dass sich in meinen Augen etwa die Aussage „Wasser = $H_2O$“ noch sehr viel einfacher begründen lässt; aber das können wir hier dahin gestellt sein lassen.

Halten wir fest, ebenso wie Papineau vertreten Block und Stalnaker die folgende Position: 1. Identität und reduktive Erklärbarkeit sind zwei verschiedene Paar Stiefel. Eigenschaften *F* und *G* können auch dann identisch sein, wenn *F* nicht reduktiv auf *G* zurückgeführt werden kann; und auch zur Begründung von Eigenschafts-Identitätsaussagen muss man nicht auf reduktive Erklärungen zurückgreifen. 2. Physikalisten sind nur auf die These festgelegt, dass mentale Eigenschaften mit physischen Eigenschaften identisch sind, und nicht auf die These, dass mentale Eigenschaften reduktiv erklärbar sind.

Mit ihrer ersten These haben Papineau, Block und Stalnaker in meinen Augen weitgehend Recht. In frühen Debatten wurden zwei ganz verschiedene Ideen miteinander vermischt – die Idee der Identität und die Idee der reduktiven oder mechanische Erklärbarkeit (im Sinne Levines und Broads). Erstens, das hatte ich schon erwähnt, ist diese Art von reduktiver Erklärbarkeit mit Multirealisierbarkeit vereinbar; sie ist also keine hinreichende Bedingung für Identität. Und sie ist, wie das Beispiel *Wasser* zeigt, auch keine notwendige Bedingung; sie ist nicht einmal notwendig zur Rechtfertigung von Identitätsaussagen.

---

[12] Mit der folgenden direkt an die zitierte Passage anschließenden Bemerkung antworten Block und Stalnaker zugleich auf Kims Einwand (vgl. oben S. 79). „If we believe that heat is correlated but not identical to molecular kinetic energy, we should regard as legitimate the question of why the correlation exits and what its mechanism is. But once we realize that heat *is* molecular kinetic energy, questions like this will be seen as wrongheaded.“

Auf der anderen Seite glaube ich aber nicht, dass Papineau, Block und Stalnaker mit ihrer zweiten These Recht haben. Wir hatten schon gesehen, dass sich die Aussage (8) in ihrem Status deutlich von der Aussage (3) unterscheidet und dass (7) eher der Aussage (3) ähnelt. Außerdem gibt es gute Gründe, zwar die Aussage (8) für eine Identitätsaussage zu halten, die Aussagen (3) und (7) aber eher als Realisierungsbehauptungen zu verstehen. So wie Broad der Auffassung war, dass es in der Vitalismusdebatte nicht um Identität, sondern um die mechanische Erklärbarkeit vitaler Eigenschaften geht, denke ich, dass es auch in der Debatte um mentale Eigenschaften nicht um Identität, sondern um reduktive Erklärbarkeit geht. Doch diesem Punkt will ich hier nicht weiter nachgehen.[13]

4. Mich interessiert an dieser Stelle vielmehr, warum es trotz der vielen Argumente, die gegen einen engen Zusammenhang von Identität und reduktiver Erklärbarkeit sprechen, in den letzten Jahren wieder vermehrt Stimmen – wie die von Frank Jackson – gegeben hat, die hier doch von einer engeren Verbindung ausgehen. Dies lässt sich in meinen Augen nur verstehen, wenn man sich klar macht, dass Jackson von einem ganzen anderen Begriff der reduktiven Erklärung ausgeht als Levine und Broad.

Jackson hat in letzter Zeit der Physikalismusdiskussion eine interessante neue Wendung gegeben. Seiner Meinung nach ist jeder Eigenschafts-Physikalist auf die folgende Behauptung festgelegt:

(MPD) „Any possible world which is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world." (Jackson 1998, 13)

Und zur Erläuterung fügt er an:

> [A] minimal physical duplicate of our world is a world that (a) is exactly like our world in every physical respect (instantiated property for instantiated property, law for law, relation for relation), and (b) contains nothing else in the sense of nothing more by way of kinds or particulars than it *must* to satisfy (a). (ebd.)

Eigenschaftphysikalisten müssen nach Jackson also mindestens behaupten:

($EP_J$) Wenn $\pi$ eine vollständige Beschreibung der physischen Welt ist[14] und $\psi$ eine beliebige wahre Aussage über Mentales, dann ist der Satz „Wenn $\pi$, dann $\psi$" mit metaphysischer Notwendigkeit wahr.

---

[13] In Beckermann (2007; in diesem Band S. 47–75) habe ich versucht, diese These ausführlich zu begründen.

[14] $\pi$ muss allerdings auch indexikalische physische Wahrheiten über uns selbst und eine „das ist alles"-Klausel enthalten, die feststellt, dass diese Beschreibung wirklich vollständig ist. Andernfalls würde $\pi$ nicht die mentale Tatsache implizieren, dass es keine immateriellen Geister gibt, die Schmerzen haben.

Jackson ist jedoch darüber hinaus der Auffassung, dass Eigenschaftsphysikalisten sogar auf die stärkere Behauptung festgelegt sind:

($EP_J'$) Wenn $\pi$ eine vollständige Beschreibung der physischen Welt ist und $\psi$ eine Aussage, die eine beliebige mentale Tatsache ausdrückt, dann folgt $\psi$ *a priori* aus $\pi$.

Jackson ist also ein Vertreter der Position, die man inzwischen ‚*a priori*-Physikalismus' nennt. Jacksons Überlegungen kann man auch so verstehen, dass in ihnen implizit ein zweiter Begriff der reduktiven Erklärung enthalten ist:

($RE_J$) Ein Phänomen ist genau dann reduktiv erklärbar, wenn es *a priori* aus $\pi$ abgeleitet werden kann.[15]

In welcher Hinsicht unterscheidet sich dieser Begriff von Levines und Broads Begriff der reduktiven Erklärbarkeit?[16]

Bei der Beantwortung dieser Frage möchte ich von einer Überlegung ausgehen, die Chalmers und Jackson in ihrem Aufsatz „Conceptual Analysis and Reductive Explanation" (2001) entwickelt haben. In Abschnitt 4 dieses Aufsatzes vertreten Chalmers und Jackson unter anderem die These, „that a macroscopic description of the world in the language of physics is implied by a microscopic description of the world in the language of physics." (Chalmers/Jackson 2001, 330 f.) Wenn $\mu$ die Konjunktion aller *mikro*physikalischer Wahrheiten über die Welt ist, dann ist in $\mu$ die *vollständige* Information über die *Struktur* und *Dynamik* der Welt auf der Mikroebene enthalten: alle Wahrheiten über die Position, die Masse und alle anderen grundlegenden Eigenschaften aller mikrophysikalischen Entitäten *zu allen Zeitpunkten*. Aus dieser Information, so Chalmers und Jackson, ergeben sich aber alle Wahrheiten über die Struktur und Dynamik der Welt auf der Makroebene – zumindest insoweit sich diese Struktur und Dynamik in Begriffen beschreiben lässt, die raumzeitliche Strukturen (Orte, Geschwindigkeiten, Gestalt usw.) und die Verteilung von Massen, Ladungen etc. betreffen.

> For example, for any given region of space at a time, the information in [$\mu$] implies information about the mass density in the region, the mass density in various subregions, the causal connections among various complex configure-

---

Dieser Bedingung wird in Jacksons Formulierung durch die Forderung Rechnung getragen, dass alle relevanten möglichen Welten *minimale* physische Duplikate unserer Welt sein müssen. (Vgl. Chalmers/Jackson 2001)

[15] Die These des *a priori* Physikalismus könnte man also auch so formulieren: Alle mentalen Phänomene sind im Sinne Jacksons reduktiv erklärbar.

[16] Der Einfachheit halber ersetze im ich im Folgenden immer Broads eigenen Ausdruck „mechanisch erklärbar" durch „reduktiv erklärbar".

> tions of matter in the region, and the extent to which the matter in the region behaves or is disposed to behave as a coherent system. This information suffices to determine which regions are occupied wholly by causally integrated systems that are disposed to behave coherently. So the information plausibly suffices for at least a geometric characterization – in terms of shape, position, mass, composition, and dynamics – of systems in the macroscopic world. (Chalmers/Jackson 2001, 330)

Außerdem gilt nach Chalmers und Jackson:

> [μ] also implies information about systems' microstructural composition, and about their distribution of systems across space and time, including the relations between systems (characterized in macrophysical terms) and about any given system's history (characterized in macrophysical terms). (Chalmers/ Jackson 2001, 331)

Ich verstehe diese Passagen so: Wenn wir über die *mikroskopische* physische Welt *vollständig* informiert sind, d. h., wenn wir von jedem mikrophysikalischen Teilchen in einer bestimmten Raum-Zeit-Region wissen, wo es sich zu welchem Zeitpunkt befindet, welche Masse und Ladung es besitzt, wie es sich im Raum bewegt, wodurch es kausal beeinflusst wird und was es selbst kausal beeinflusst usw., dann können wir aus dieser Information *a priori* ableiten, welche *makroskopischen* Gegenstände es in dieser Raum-Zeit-Region gibt, welche Gestalt diese Gegenstände haben, welche mikrostrukturelle Zusammensetzung sie besitzen, welche Masse und Ladung sie besitzen (bzw. wie Masse und Ladung in ihnen verteilt sind), wie sich diese makroskopischen Gegenstände (und ihre Teile) im Raum bewegen und wie sich ihre Gestalt, Masse und Ladung verändern.[17] Wenn es sich tatsächlich so verhält, bedeutet das aber, dass das *Verhalten* makroskopischer Gegenstände – und damit auch jede Eigenschaft, die allein durch ein bestimmtes Verhalten charakterisiert ist – in jedem Fall (im Sinne von Jackson) reduktiv erklärt werden kann. Anders als Broad annahm, kann, so verstanden,

[17] Dazu benötigt man allerdings wieder Brückenprinzipen, die die physikalische Mikro- mit der physikalischen Makroebene verbinden – Prinzipien wie „Wenn Moleküle so starke Kräfte aufeinander ausüben, dass sie ihre relativen Positionen zueinander nicht verändern und sich daher nur im Verbund bewegen, dann bilden diese Moleküle einen festen Gegenstand", „Wenn alle Teile eines Gegenstandes sich mit derselben Geschwindigkeit in dieselbe Richtung bewegen, bewegt sich der ganze Gegenstand mit derselben Geschwindigkeit in dieselbe Richtung", „Wenn alle Teile eines Diskus mit derselben Winkelgeschwindigkeit um einen Punkt im Innern des Diskus kreisen, dreht sich der Diskus um diesen Punkt", „Wenn sich die Moleküle, aus denen ein Gegenstand besteht, voneinander lösen und sich zwischen den Wassermolekülen verteilen, dann löst sich dieser Gegendstand in Wasser auf" usw. Ich persönliche denke, dass man diese Brückenprinzipen durchaus als *a priori* wahr ansehen kann.

das Verhalten von Salz gar nicht emergent sein. Damit ist klar, dass sich Broads Begriff der reduktiven Erklärbarkeit vom dem Jacksons unterscheidet. Doch worauf beruht dieser Unterschied?

Wenn wir uns auf das Verhältnis von Makro- und Mikroebene beschränken (nach Jackson können ja auch Phänomene auf der Mikroebene selbst reduktiv erklärbar sein), scheint mir der entscheidende Punkt zu sein, dass ($RE_J$) zufolge nur solche Phänomene *nicht* reduktiv erklärbar sind, die nach Broad *im zweiten Sinne* emergent sind, bei denen es also keine Brückenprinzipen gibt, die das Verhalten der Teile in geeigneter Weise mit den Eigenschaften des Ganzen verbinden. Jackson geht davon aus, dass wir über das Verhalten und die grundlegenden physikalischen Eigenschaften aller Teilchen auf der Mikroebene *vollständig* informiert sind, und fragt dann: Können wir aus dieser Information *a priori* ableiten, dass ein bestimmter Makrogegenstand *S* eine bestimmte Makroeigenschaft *F* hat? Wenn das der Fall ist, ist *F* – genauer: dieses Auftreten von *F* – reduktiv erklärbar. Für Broad reicht das nicht aus. Für Broad setzt die reduktive Erklärbarkeit eines Auftretens von *F* nicht nur voraus, dass es ein Brückenprinzip gibt, das besagt, dass *S* alle für *F* charakteristischen Merkmale besitzt, wenn sich seine Teile $C_1, \ldots, C_n$ so verhalten, wie sie dies tun, wenn sie in der Weise *R* angeordnet sind; erforderlich ist für ihn darüber hinaus, dass sich *dieses Verhalten* aus den *allgemeinen* auf der Mikroebene geltenden Gesetzen ableiten lässt. Broad zufolge müssen für die reduktive Erklärbarkeit einer Makroeigenschaft die Bedingungen (a) und (b) der Definition ($ME_B$) *beide* erfüllt sein, nach Jackson nur die Bedingung (b). Reduktive Erklärbarkeit im Sinne Broads impliziert also reduktive Erklärbarkeit im Sinne Jacksons, aber nicht umgekehrt.

Welche Beziehung besteht zwischen Identität und reduktiver Erklärbarkeit im Sinne Jacksons? Wir hatten schon gesehen, das viele Autoren wie Papineau, Block und Stalnaker in den letzten Jahren mit Vehemenz die Auffassung vertreten haben, mentale Eigenschaften könnten auch dann mit physischen Eigenschaften identisch sein, wenn sich das Mentale nicht reduktiv erklären lasse. Alle diese Autoren vertreten die folgenden beiden Thesen: 1. Physikalisten sind nur auf die These festgelegt, dass mentale Eigenschaften mit physischen Eigenschaften identisch sind. 2. Ob mentale Eigenschaften mit physischen Eigenschaften identisch sind, lässt sich einzig und allein *a posteriori* feststellen.

Zur Stützung dieser Thesen beziehen sich die genannten Autoren unter anderem auf Kripkes Überlegungen zur Semantik von Artbegriffen. Lange Zeit gehörte es, zumindest unter Philosophen, zu den allgemein akzeptierten Überzeugungen, dass die Bedeutung eines Prädikats in einer Menge von einzeln notwendigen und zusammen hinreichenden Merkmalen besteht. Dieser Auffassung zufolge trifft ein Prädikat auf einen Gegenstand

genau dann zu, wenn er diese Merkmale besitzt (oder doch zumindest die meisten von ihnen). Aufgrund der Arbeiten von Kripke und Putnam wurde dieser Konsens jäh zerstört. Heute scheinen viele zu glauben, dass alle (also auch mentale) Prädikate eher wie Namen funktionieren – sie bezeichnen Arten oder Eigenschaften und sie treffen auf einen Gegenstand genau dann zu, wenn er zu der durch das Prädikat bezeichneten Art gehört bzw. die durch das Prädikat bezeichnete Eigenschaft besitzt. Der Ausdruck ‚Gold' z. B. bezieht sich dieser Auffassung zufolge auf das chemische Element mit der Ordnungszahl 79; er trifft auf einen Gegenstand *a* also genau dann zu, wenn dieser Gegenstand (fast) ausschließlich aus Atomen besteht, die genau 79 Protonen enthalten – völlig unabhängig von den Oberflächeneigenschaften von *a*. Wenn *F* ein Ausdruck für eine natürliche Art ist, dann gibt es für *F* keine Analyse, d. h. dann gibt es keine Menge von charakteristischen Merkmalen $M_F$, für die gilt, dass jeder kompetente Sprecher weiß, dass *F* auf einen Gegenstand *a* genau dann zutrifft, wenn *a* alle (oder zumindest die meisten) Merkmale von $M_F$ besitzt.

Offensichtlich passen diese Überlegungen zur Semantik der Begriffe für natürliche Arten nicht gut zu reduktiven Erklärungen im Sinne Levines und Broads. Wenn es für Gold keine charakteristischen Merkmale gibt, kann die Eigenschaft, aus Gold zu sein, nicht reduktiv erklärt werden. Und das bedeutet, dass es für die Wahrheit von Aussagen wie „Gold = das chemische Element mit der Ordnungszahl 79" oder „Wasser = $H_2O$" nicht erforderlich sein kann, dass die Tatsache, dass etwas Gold oder Wasser ist, im Sinne von Levine und Broad reduktiv erklärbar ist. Wenn Levine darauf verweist, dass aus den allgemeinen grundlegenden Naturgesetzen folgt, dass $H_2O$ auf Meereshöhe bei 100° C kocht, dass $H_2O$ flüssig ist, durchsichtig ist usw., hat das – auch das hatte ich schon erwähnt – mit der Wahrheit der Aussage „Wasser = $H_2O$" nicht das Geringste zu tun. Denn bei diesen Oberflächeneigenschaften handelt es sich nicht um Merkmale im traditionellen Sinn. Aber wenn das so ist, wie kann Jackson dann die beiden Thesen verteidigen: (i) Wenn mentale Eigenschaften mit physischen Eigenschaften identisch sind, dann muss sich dies reduktiv erklären lassen. (ii) Wenn der Physikalismus wahr ist, dann gilt: Wenn $\pi$ eine vollständige Beschreibung der physischen Welt ist und $\psi$ eine Aussage, die eine beliebige mentale Tatsache ausdrückt, dann folgt $\psi$ *a priori* aus $\pi$.

Ganz wichtig ist hier zu sehen, dass Jackson gar nicht bestreitet, dass Aussagen wie „Wasser = $H_2O$" oder „Schmerz = das Feuern von C-Fasern" nur *a posteriori* als wahr erkannt werden können, dass es also empirischer Forschung bedarf, um die Wahrheit dieser Aussagen festzustellen. Seine Frage ist nicht, was wir *a priori* wissen können, sondern was wir *a priori* aus der *vollständigen* Kenntnis aller physikalischen Tatsachen (inklusive der vollständigen Physik) *ableiten* können. Insofern behauptet er auch

nicht, dass Aussagen wie „Wasser = $H_2O$“ oder „Schmerz = das Feuern von C-Fasern“, wenn sie denn wahr sind, in irgendeinem Sinne *a priori* wahr sind, sondern nur dass sie in diesem Fall *a priori* aus π ableitbar sind. Bevor wir auf die Frage eingehen können, warum das Jackson zufolge so ist, noch eine Bemerkung. Anders als Levine und Broad geht es Jackson nicht darum, eine einzelne Eigenschaft *F* auf eine Eigenschaft *G* oder eine Mikrostruktur zu reduzieren. Ihm geht es um die Reduktion ganzer Identitätsaussagen „*F* = *G*“, d. h. darum zu zeigen, dass sich diese Identitätsaussagen *a priori* aus π ableiten lassen. Warum ist das Jackson zufolgt der Fall?

Jackson ist sich mit seinen Gegnern einig, dass es für Ausdrücke wie ‚Gold‘, ‚Wasser‘ oder ‚Schmerzen‘ keine Analyse im traditionellen Sinn gibt. Trotzdem meint er, dass kompetente Sprecher des Deutschen – als kompetente Sprecher – etwas über den chemischen Stoff wissen, den wir ‚Wasser‘ nennen. Sie wissen, dass es sich dabei um den Stoff handelt, der sich in der *aktuellen* Welt in Flüssen und Seen befindet, der bei Regen aus den Wolken auf die Erde fällt, der aus Wasserhähnen fließt, usw. Nun ist es aber eine – in π enthaltene – physikalische Tatsache, dass es sich in der aktuellen Welt bei diesem Stoff um $H_2O$ handelt. Hieraus und aus dem Wissen, das wir als kompetente Sprecher von Wasser haben, können wir den Satz „Wasser = $H_2O$“ ableiten. Selbst wenn ‚Temperatur (in Gasen)‘ nicht durch eine charakteristische kausale Rolle *definiert* ist, wie sie etwa in den Gesetzen der klassischen Thermodynamik ausgedrückt wird, ist Jackson der Auffassung, dass kompetente Sprecher des Deutschen wissen, dass die Temperatur von Gasen genau diese Rolle in der *aktuellen* Welt spielt. Deshalb ist in seinen Augen das folgende Argument schlüssig:

(P1) Temperature in gases = that which plays the temperature role in gases in the actual world (known to be true by competent speakers of English).
(P2) That which plays the temperature role in gases in the actual world = mean molecular kinetic energy (physical fact included in π)

Therefore:

(C) Temperature in gases = mean molecular kinetic energy. (cf. Jackson 1998, 59)

Also kann „Temperatur in Gasen = mkE“ *a priori* aus π abgeleitet werden.

Dieses Resultat lässt sich Jackson zufolge verallgemeinern.[18] Jeder Name und jedes Prädikat haben in seinen Augen einen deskriptiven Inhalt. Dieser deskriptive Inhalt besteht aus einer Menge von Merkmalen, von denen der kompetente Sprecher weiß, dass sie in der *aktuellen* Welt auf die durch die Namen bezeichneten Gegenstände (bzw. auf die Eigenschaften, für die die Prädikate stehen) zutreffen. Daher müssen wir auf der Grundlage einer vollständigen Kenntnis der physikalischen Tatsachen in der Lage sein zu

[18] Vgl. Jackson (2003), aber auch Chalmers (2002).

sagen, welche Gegenstände und welche Eigenschaften identisch sind. Wenn ein kompetenter Sprecher weiß, dass die Eigenschaft *F* in der aktuellen Welt eine Menge von Merkmalen $M_F$ besitzt, und wenn es eine physikalische Tatsache ist, dass in der aktuellen Welt nur *G* diese Merkmale besitzt, kann die Aussage „*F* = *G*“ *a priori* aus π abgeleitet werden.

Gegen diese Überlegung kann der *a posteriori* Physikalist zwei Einwände erheben. Erstens kann er bestreiten, dass alle Prädikate einen deskriptiven Inhalt im Sinne Jacksons haben. Und zweitens kann er bezweifeln, dass die empirischen Fakten, aus denen wir – zusammen mit dem, was kompetente Sprecher über Schmerz wissen – z. B. erschließen könnten, dass der Satz „Schmerz = das Feuern von C-Fasern“ wahr ist, in π enthalten sind. Ich möchte mich hier auf den zweiten Einwand beschränken, der zunächst etwas merkwürdig wirken mag. Schließlich ist auch der *a posteriori* Physikalist ein Physikalist, d. h., er behauptet, dass alle Tatsachen physische Tatsachen sind. Daher sollte er doch wohl meinen, dass π alle Tatsachen überhaupt umfasst.

Wichtig ist hier, dass π eine *Beschreibung* der Welt ist, also eine Menge von Sätzen und keine Menge von Tatsachen, und dass wir verschiedene Sätze verwenden können, um dieselbe Tatsache auszudrücken. So drückt der Satz „Mehr als die Hälfte der Erde ist von Wasser bedeckt“ dieselbe Tatsache aus wie der Satz „Mehr als die Hälfte der Erde ist von $H_2O$ bedeckt“.[19]

Die Frage ist also, welche Sätze in π enthalten sind. Offenbar ist es nicht sinnvoll zu sagen, dass π alle Sätze umfasst, die eine physische Tatsache ausdrücken. Denn dann müsste in π auch der Satz „Schmerzen werden häufig durch Verletzungen verursacht“ und sogar der Satz „Schmerz = das Feuern von C-Fasern“ enthalten sein – vorausgesetzt, dass Schmerz tatsächlich mit dem Feuern von C-Fasern identisch ist. Damit geriete Jackson aber in einen Zirkel. Also kann π nicht alle Sätze enthalten, die physikalische Tatsachen ausdrücken. Und wie ist es mit dem Alternativvorschlag, dass π genau die Sätze enthält, die allein in physikalischem Vokabular formuliert sind? Wenn wir voraussetzen, dass Wissen intensional ist, ist in diesem Fall Folgendes denkbar (selbst wenn man zugesteht, dass alle Prädikate einen deskriptiven Inhalt haben): Alles, was ein kompetenter Sprecher des Deutschen über die Eigenschaft weiß, für die der Ausdruck ‚Schmerz‘ steht, weiß er nur unter Beschreibungen, die selbst mentales Vokabular enthalten. Vielleicht weiß er nur, dass Schmerz in der aktuellen Welt die *S*-Rolle spielt, wobei der Ausdruck ‚*S*‘ seinerseits nichtphysikalisches Vokabular enthält. In diesem Fall könnte es weiter sein, dass genau das Feuern von C-Fasern in der aktuellen Welt die *S*-Rolle spielt; aber der

[19] Ich verstehe ‚Tatsache‘ hier im Sinn von Russell, nicht von Frege.

entsprechende Satz wäre nicht in π enthalten, da er nicht nur physikalisches Vokabular enthält. Die Argumentation Jacksons ist daher nur überzeugend, wenn sich Folgendes plausibel machen lässt: Wenn eine Eigenschaft *F* mit einer physikalischen Eigenschaft *G* identisch ist, dann verfügt jeder kompetente Sprecher – allein aufgrund seiner Sprachkompentenz – über Wissen bzgl. *F*, das ausschließlich in physikalischen Begriffen formuliert ist und das es – zusammen mit den in π enthaltenen Sätzen – ermöglicht, die Aussage „*F* = *G*" abzuleiten. Aber das kann man nicht ohne weiteres voraussetzen. Selbst wenn man ‚reduktive Erklärung' in Jacksons Sinn versteht, scheint es daher nicht notwendig, dass der Satz „*F* = *G*" nur wahr sein kann, wenn er reduktiv erklärbar ist, d. h. wenn er *a priori* aus π abgeleitet werden kann.

## Literatur


Achinstein, Peter (1974): „The Identity of Properties". *American Philosophical Quarterly* 11, 257–275.

Beckermann, Ansgar (2002): „Die reduktive Erklärbarkeit des phänomenalen Bewusstseins – C. D. Broad zur Erklärungslücke". In: M. Pauen & A. Stephan (Hg.) *Phänomenales Bewußtsein – Rückkehr zur Identitätstheorie?* Paderborn: mentis, 122–147. In diesem Band Beitrag 2.

Beckermann, Ansgar (2007): „Neue Überlegungen zum Eigenschaftsphysikalismus". In: M. Pauen, M. Schütte & A. Staudacher (Hg.) *Begriff, Erklärung, Bewusstsein: Neue Beiträge zum Qualia-Problem*. Paderborn: mentis, 143–170. In diesem Band Beitrag 3.

Block, Ned & Robert Stalnaker (1999): „Conceptual Analysis, Dualism, and the Explanatory Gap". *The Philosophical Review* 108, 1–46.

Brandt, Richard & Jaegwon Kim (1967): „The Logic of the Identity Theory". *The Journal of Philosophy* 64, 515–537.

Broad, Charles Dunbar (1925): *The Mind and Its Place in Nature*. London: Routledge and Kegan Paul.

Chalmers, David (2002): „Does Conceivability entail Possibility?" In: T. S. Gendler & J. Hawthorne (Hg.) *Conceivability and Possibility*. Oxford: Oxford University Press, 145–200.

Chalmers, David & Frank Jackson (2001): „Conceptual Analysis and Reductive Explanation". *Philosophical Review* 110, 315–360.

Churchland, Paul M. (1988): *Matter and Consciousness*. 2nd ed., Cambridge MA: MIT Press.

Descartes, René (1960): *Discours de la méthode. Von der Methode des richtigen Vernunftgebrauchs*. Französisch-Deutsch. Übertr. u. hrsg. v. Lüder Gäbe. Hamburg: Meiner.

Hill, Christopher (1991): *Sensations*. Cambridge: Cambridge University Press.

Hill, Christoper & Brian McLaughlin (1999): „There are Fewer Things in Reality than are Dreamt of in Chalmers' Ontology". *Philosophy and Phenomenological Research* 59, 445–54.

Jackson, Frank (1998): *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jackson, Frank (2003): „From $H_2O$ to Water: the Relevance to A Priori Passage". In: H. Lillehammer & G. Rodriguez-Pereyra (Hg.) *Real Metaphysics*. London: Routledge, 84–97.

Kim, Jaegwon (1966): „On the Psycho-Physical Identity Theory". *American Philosophical Quarterly* 3, 227–235.

Levine, Joseph (1983): „Materialism and Qualia: The Explanatory Gap". *Pacific Philosophical Quarterly* 64, 354–361.

Levine, Joseph (1993): „On Leaving Out What It's Like". In: M. Davies & G. W. Humphreys (Hg.) *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell, 121–136.

McLaughlin, Brian (2001): „In Defense of New Wave Materialism: A Response to Horgan and Tienson". In: C. Gillett & B. Loewer (Hg.) *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, 319–330.

Papineau, David (1998): „Mind the Gap". In: J. Tomberlin (Hg.) *Philosophical Perspectives 12*: *Language, Mind, and Ontology*. Oxford: Basil Blackwell, 373–388.

Place, Ullin T. (1956): „Is consciousness a brain process?" *British Journal of Psychology* 47, 44–50.

Smart, John J. C. (1959): „Sensations and Brain Processes". *Philosophical Review* 58, 141–156.

# Sprachverstehen und das Computermodell des Geistes

# 5

## Sprachverstehende Maschinen
## Überlegungen zu John Searles Thesen zur Künstlichen Intelligenz*

ABSTRACT. In this paper the author tries to disentangle some of the problems tied up in John Searle's famous Chinese-room-argument. In a first step to answer the question what it would be for a system to have not only syntax, but also semantics the author gives a brief account of the functioning of the language understanding systems (LUS) so far developed in the framework of AI research thereby making clear that systems like Winograd's SHRDLU are indeed doing little more than mere number crunching. But things would be entirely different, the author argues, if the database of a LUS were built up by the system itself via some perceptual component – at least, if this perceptual component had the capacity to distinguish objects having a certain property F from objects which do not. For in this case the system could store an internal representation of the fact that the object has the property F in its database if and only if the object in fact has that property. And this would be a good basis for calling such a system a *genuine* LUS. But Searle has objected to a very similar account of J. Fodor that nothing could be further from true language understanding. The reason for this complaint seems to be that Searle holds the view that a true LUS must e.g., *know* that the word „hamburgers" refers to hamburgers and that he moreover claims that this knowledge must be *explicit* or that the system must be aware of the reference of „hamburgers" to hamburgers. The author argues that this is asking too much. For it seems plausible to say that a system is able to understand e.g., the word „hamburger" even if it has only *implicit* knowledge of the fact that „hamburger" refers to hamburgers in the sense that it has the capacity to tell hamburgers from non hamburgers and the capacity to bring the word „hamburger" together just with objects of the former kind.

### 1.

John Searle ist häufig so verstanden worden, als habe er in seinem Aufsatz „Minds, Brains, and Programs" und in späteren Arbeiten die These vertreten wollen, keine Maschine – welcher Art auch immer – sei im Wortsinn imstande, Sprache zu verstehen oder andere kognitive Zustände anzunehmen. Doch das ist nicht der Fall. Denn in dem genannten Aufsatz schreibt er ausdrücklich auch:

> I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. (MBP 422)

Es geht Searle also nicht darum zu behaupten, daß überhaupt keine Maschine Sprache verstehen oder kognitive Zustände haben kann. Und es geht ihm daher auch nicht darum zu behaupten, wir selbst, die wir dies alles können, seien keine Maschinen – ganz im Gegenteil. Seine These ist viel-

* Erstveröffentlichung in: *Erkenntnis* 28 (1988), 65–85.

mehr, daß *bestimmte* Maschinen – insbesondere Computer – keine Sprache verstehen können und daß wir selbst daher keine Computer sein können. D. h. genauer kann man Searles These so formulieren:

SEARLE:
Keine Maschine, deren Verhalten allein durch formale Veränderungen formal definierter Elemente bestimmt ist, d. h. keine Maschine, deren Verhalten als die Instantiierung eines Computerprogramms definiert ist, ist im Wortsinn imstande, Sprache zu verstehen oder andere kognitive Zustände zu haben.

Searles These ist also eingeschränkter, als viele angenommen haben. Aber sie ist deshalb nicht weniger brisant. Denn sie stellt – und das ist ja auch Searles Absicht – das gesamte Programm der Künstlichen Intelligenz-Forschung infrage. Zumindest gilt das, wenn man dieses Programm im Sinne der – wie Searle sagt – starken KI versteht, d. h. wenn man mit diesem Programm die These verbindet:

STARKE KI:
a) Es gibt Maschinen, deren Verhalten als die Instantiierung eines Computerprogramms definiert ist, die im Wortsinn Sprache verstehen oder andere kognitive Zustände annehmen können.
b) Die Programme, die das Verhalten dieser Maschinen bestimmen, *erklären* auch die menschliche Fähigkeit, Sprache zu verstehen, bzw. erklären auch, was es für Menschen bedeutet, bestimmte kognitive Zustände anzunehmen.

## 2.

Das Hauptargument, das Searle für seine These anführt, besteht in einem Gedankenexperiment,[1] das unter dem Namen *Chinese-Room-Experiment* bekannt geworden ist. Mit diesem Gedankenexperiment möchte er zeigen, daß Situationen vorstellbar sind, in denen ein Mensch z. B. in einer Dialog-

[1] Schon an dieser Stelle ist es wichtig, zu betonen, daß das Chinese-Room-Experiment nicht wirklich ein Argument ist, sondern eine Art Gedankenexperiment, in dem an die Intuitionen der Leser oder Hörer appelliert wird. Searle bittet den Leser, sich gewisse Situationen vorzustellen und dann zu sagen, ob in diesen Situationen bestimmte Zuschreibungen berechtigt sind oder nicht. Bei solch einem Gedankenexperiment ist es daher von entscheidender Bedeutung, daß die betreffenden Situation so klar wie möglich geschildert und nicht durch ungenaue Darstellungen verfälscht werden. Dieser Punkt scheint mir gerade im Hinblick auf Searles Erwiderungen auf den System- und den Roboter-Einwand, auf die ich im Schluß des Aufsatzes zu sprechen kommen werde, von großer Wichtigkeit zu sein.

situation genau das leistet, was ein Computer leisten kann, ohne dabei jedoch auch nur das Geringste von der Sprache zu verstehen, in der „der Dialog geführt wird“. Im einzelnen orientiert sich Searle bei seinem Gedankenexperiment an einem Programm von Roger Schank,[2] bei dem es im Kern darum geht, daß ein Computer, dem zuvor der Normalablauf bestimmter Situationen (z. B. eines Restaurantbesuchs) in Form von „scripts“ eingegeben wurde, Geschichten über konkrete einzelne Situationen dieser Art verstehen und Fragen im Hinblick auf diese Geschichten korrekt beantworten soll. Searle stellt uns nun folgende Parallelsituation vor:

> [S]tellen Sie sich vor, Sie wären in ein Zimmer eingesperrt, in dem mehrere Körbe mit chinesischen Symbolen stehen. Und stellen Sie sich vor, daß Sie … kein Wort Chinesisch verstehen, daß Ihnen allerdings ein auf Deutsch abgefaßtes Regelwerk für die Handhabung dieser chinesischen Symbole gegeben worden wäre. Die Regeln geben rein formal – nur mit Rückgriff auf die Syntax und nicht auf die Semantik der Symbole – an, was mit den Symbolen gemacht werden soll. Eine solche Regel mag lauten: ‚Nimm ein Kritzel-Kratzel-Zeichen aus Korb 1 und lege es neben ein Schnörkel-Schnarkel-Zeichen aus Korb 2‘. Nehmen wir nun an, daß irgendwelche anderen chinesischen Symbole in das Zimmer gereicht werden, und daß Ihnen noch zusätzliche Regeln dafür gegeben werden, welche chinesischen Symbole jeweils aus dem Zimmer herauszureichen sind. Die hereingereichten Symbole werden von den Leuten draußen ‚Fragen‘ genannt, und die Symbole, die Sie dann aus dem Zimmer herausreichen, ‚Antworten‘ – aber dies geschieht ohne Ihr Wissen. Nehmen wir außerdem an, daß die Programme so trefflich und Ihre Ausführung so brav sind, daß Ihre Antworten sich schon bald nicht mehr von denen eines chinesischen Muttersprachlers unterscheiden lassen. (GHW 31)

Der Punkt dieser Geschichte liegt auf der Hand. Der in das Zimmer Eingesperrte wird als Reaktion auf die ihm in schriftlicher Form gestellten chinesischen Fragen, die ihm selbst aber nur als Folgen von durch ihre graphische Form charakterisierten Symbolen erscheinen, Folgen von solchen Symbolen zurückgeben, die von den Fragestellern als vernünftige Antworten auf die von ihnen gestellten Fragen aufgefaßt werden können. Doch das bedeutet nicht, daß der Eingesperrte selbst deshalb verstehen müßte, was die Fragen und was seine „Antworten“ bedeuten. Der ganze Witz der geschilderten Situation liegt ja gerade darin, daß er voraussetzungsgemäß die genannten Leistungen erbringt, ohne auch nur ein Wort Chinesisch zu verstehen. Und daher, schließt Searle, verstehen auch Computer nichts von den Sprachen, in denen sie „Dialoge“ führen können. Denn, so lautet sein Argument:

---

[2] Searle selbst betont aber, seine Überlegungen träfen auch auf andere Systeme dieser Art zu, z. B. auf das System ELIZA von Weizenbaum und das System SHRDLU von Terry Winograd.

> Wenn … die Ausführung eines passenden Computerprogramms *in Ihrem Fall* nicht ausreicht, um Chinesisch zu verstehen, dann reicht das auch bei *keinem anderen digitalen Computer* aus … Wenn Sie kein Chinesisch verstehen, dann könnte auch kein anderer Computer Chinesisch verstehen; denn kraft seiner Ausführung eines Programms hat kein digitaler Computer irgendetwas, das Sie nicht haben. Der Computer hat – genau wie Sie – nichts außer einem formalen Programm für die Handhabung uninterpretierter chinesischer Symbole. (GHW 31 f.)

Obwohl diese Argumentation Searles auf sehr heftige Kritik gestoßen ist, muß man doch sagen, daß Searle mit ihr einen wichtigen Punkt getroffen hat. Denn im Kern läuft diese Argumentation darauf hinaus, daß er noch einmal mit allem Nachdruck und sehr anschaulich auf die Tatsache hinweist, daß Computer *ihrer Natur nach* weder Rechenmaschinen noch informationsverarbeitende Maschinen, sondern einfach symbolmanipulierende Maschinen sind. Oder, um auch noch den Gebrauch des Wortes „Symbol“, das ja ebenfalls auf bedeutungstragende Entitäten verweist, zu vermeiden: daß Computer ihrer Natur nach *musterverarbeitende* Maschinen sind. Denn in Computern geschieht letzten Endes nichts anderes, als daß Muster von 8, 16 oder mehr Bits, die sich in verschiedenen Registern oder an verschiedenen Speicherplätzen befinden, nach bestimmten Regeln hin- und hergeschoben oder verändert werden. Dabei orientieren sich diese Regeln selbst nur an der äußeren Gestalt der einzelnen Bitmuster und nicht etwa an einem möglichen Bedeutungsinhalt, den diese Muster haben könnten.

Wenn man diese Tatsache mit der in der Sprachwissenschaft und Sprachphilosophie geläufigen Unterscheidung von Syntax und Semantik in Zusammenhang bringt, scheint daher klar, daß Computer eingegebene sprachliche Äußerungen immer nur syntaktisch, nicht aber semantisch behandeln können. Denn während sich die Syntax herkömmlichen Definitionen zufolge nur auf die äußere Form, d. h. insbesondere auf die Wohlgeformtheit sprachlicher Ausdrücke bezieht, geht es bei der Semantik um die Bedeutung dieser Ausdrücke. Und diese Bedeutung kann, wie viele meinen, nicht auf die äußere Form sprachlicher Ausdrücke reduziert werden. Die Erfassung von Bedeutungen scheint damit etwas zu sein, was grundsätzlich außerhalb der Reichweite von Maschinen liegt, die im Prinzip nur durch ihre äußere Form charakterisierte Muster manipulieren können.

Diese Argumentation ist jedoch nicht ganz so unproblematisch, wie sie auf den ersten Blick aussehen mag. Ihr Hauptproblem liegt meiner Meinung nach darin, daß sie auf einem ziemlich unklaren Semantikkonzept beruht, in dem Semantik letztlich nur negativ in Abgrenzung gegen die Syntax definiert wird. Allerdings ist es nicht leicht, dieses Semantikkonzept durch ein genaueres und besser durchformuliertes zu ersetzen. Deshalb möchte ich hier einen anderen Weg gehen und zunächst untersuchen, wie

die von Searle kritisierten natürlichsprachlichen Computersysteme im einzelnen arbeiten, um dann möglicherweise auf diesem Weg einer Antwort auf die Frage, ob solche Systeme tatsächlich Sprache verstehen können oder nicht, etwas näher zu kommen.

## 3.

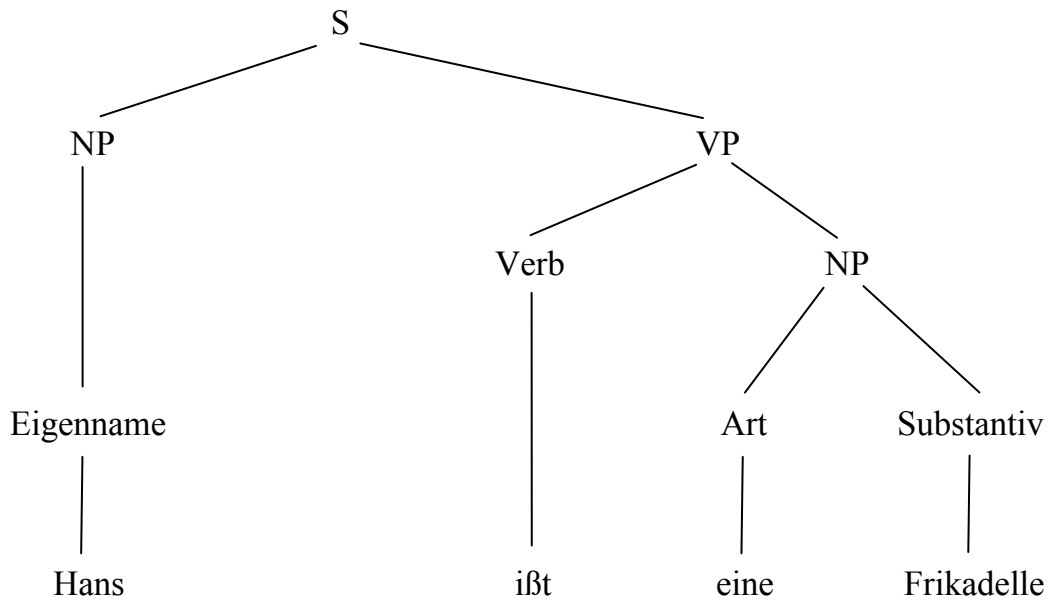
In sehr groben Zügen kann man die Arbeitsweise neuerer natürlichsprachlicher Systeme (NSS) so zusammenfassen.

In einem ersten Schritt wird jeder eingegebene Satz einer morphologisch-lexikalischen Analyse unterzogen. Als Ergebnis dieser Analyse werden jedem Wort dieses Satzes seine lexikalische Kategorie (Substantiv, Verb, Präposition, Determinator, usw.) sowie, falls sinnvoll, einige syntaktische Merkmale (Singular, Plural, definit, indefinit, usw.) zugeordnet. Als Ergebnis ergibt sich etwa für den Eingabesatz „Hans ißt eine Frikadelle" die folgende Aufstellung.[3]

| Wort | Lexikalische Kategorie | Merkmale |
|---|---|---|
| Hans | Eigenname | |
| ißt | Verb | 3. Pers., Sing., Präsens |
| eine | Artikel | indefinit, Sing., weibl. |
| Frikadelle | Substantiv | Sing., weibl. |

Im zweiten Schritt wird jeder Eingabesatz unter Berücksichtigung der Ergebnisse der morphologisch-lexikalischen Analyse mit Hilfe einer Parsing-Komponente einer syntaktischen Analyse unterzogen. Dabei wird zunächst geprüft, ob der Satz syntaktisch korrekt ist. Falls nicht, wird er zurückgewiesen. Falls er jedoch korrekt ist, wird ihm als Ergebnis der syntaktischen Analyse seine syntaktische Struktur zugeordnet, z. B. in der Form eines Strukturbaumes – für den angegebenen Beispielsatz etwa der im oberen Teil der Abb. 1 gezeigte Strukturbaum. Dieser Strukturbaum kann auch – was für die Verarbeitung durch einen Computer natürlicher ist – in der Form der im unteren Teil der Abb. 1 angeführten Liste dargestellt werden.

[3] Diese Aufstellung ist sehr verkürzt. Sie berücksichtigt z. B. nicht, daß etwa für die Wörter „eine" und „Frikadelle" die Kasus mit angegeben werden sollten und daß dementsprechend für diese Wörter verschiedene Möglichkeiten berücksichtigt werden müssen. Für diesen Zusammenhang reicht jedoch die angeführte vereinfachte Aufstellung zur Illustration aus.




```
(S  (NP  (Eigenname Hans))
    (VP  (Verb ißt)
         (NP  (Art eine)  (Substantiv  Frikadelle))))
```

Abbildung 1

In der Praxis kann es darüberhinaus sinnvoll sein, in die durch eine solche Liste gegebene Beschreibung der syntaktischen Struktur eines Satzes noch weitere Informationen mitaufzunehmen und z. B. am Hauptknoten S zu notieren, ob es sich bei dem analysierten Satz um eine Behauptung, einen Fragesatz oder einen Befehl handelt.

Diese wenigen Sätze sollen für die Darstellung der ersten beiden Schritte der Sprachverarbeitung in NSS ausreichen. Für die Ausgangsfrage ist nämlich der dritte Schritt wichtiger. Denn dieser Schritt wird herkömmlich als „semantische Analyse“ bezeichnet. In ihm scheint es also tatsächlich um die Bedeutung der eingegebenen Sätze zu gehen. Es lohnt sich daher, genauer zu untersuchen, was bei der semantischen Analyse von Sätzen konkret geschieht. Der unmittelbare Zweck dieser Analyse ist die Überführung des eingegebenen Satzes in eine *interne Repräsentation*, wenn man so will, also die Übersetzung des eingegebenen Satzes in einen entsprechenden Satz einer internen Sprache des Computers. Es gibt inzwischen sehr verschiedene Methoden der internen Repräsentation. Als Beispiel soll hier die von Winograd in seinem System SHRDLU eingesetzte Methode dienen, als semantische Repräsentationssprache Ausdrücke der KI-Programmiersprache PLANNER zu verwenden. Bei dieser Methode ergibt sich etwa für den Eingabesatz

(*) Welche Pyramiden werden von einem Block gestützt?

als semantische Repräsentation der PLANNER-Ausdruck

```
(**)  (FIND ALL  ?X1  (X1)
            (GOAL  (ISA  ?X1  PYRAMIDE))
            (FINDE  1  ?X2  (X2)
                (GOAL  (ISA  ?X2  BLOCK))
                (GOAL  (STUETZT  ?X2  ?X1)))).
```

Ich kann hier nicht darauf eingehen, wie der Eingabesatz im einzelnen in diesen PLANNER-Ausdruck überführt wird. Das ist aber auch nicht unbedingt erforderlich. Denn für diesen Zusammenhang reicht es aus zu wissen, daß alle Informationen, die dafür nötig sind, aus der Beschreibung der syntaktischen Struktur dieses Satzes, aus den Lexikoneinträgen für die in diesem Satz vorkommenden Wörter und aus einigen anderen im System vorhandenen „Wissenskomponenten“ gewonnen werden können. Die entscheidende Frage ist vielmehr, wozu diese Übersetzung in eine interne Repräsentation dient und was sie mit einem wirklichen Verstehen der Bedeutung des eingegebenen Satzes zu tun hat.

Der Grund für die Übersetzung eingegebener Sätze in interne Repräsentationen liegt zunächst darin, daß das System diese internen Repräsentationen – anders als den eingegebenen Satz – in einer zusätzlichen Auswertungs-Komponente weiter verarbeiten kann. So ist z. B. der PLANNER-Ausdruck (**) selbst eine PLANNER-Funktion, die als Teil eines PLANNER-Programms die internen Namen aller Pyramiden liefert, die tatsächlich von einem Block gestützt werden. Dabei ist allerdings die Existenz einer Datenbasis vorausgesetzt, in der das System sein „Wissen“ über die „äußere Welt“ speichert. Dies muß vielleicht noch etwas erläutert werden.

Das System SHRDLU von Terry Winograd ist so eingerichtet, daß es – auf Anweisung eines menschlichen Partners – Manipulationen in einer fiktiven

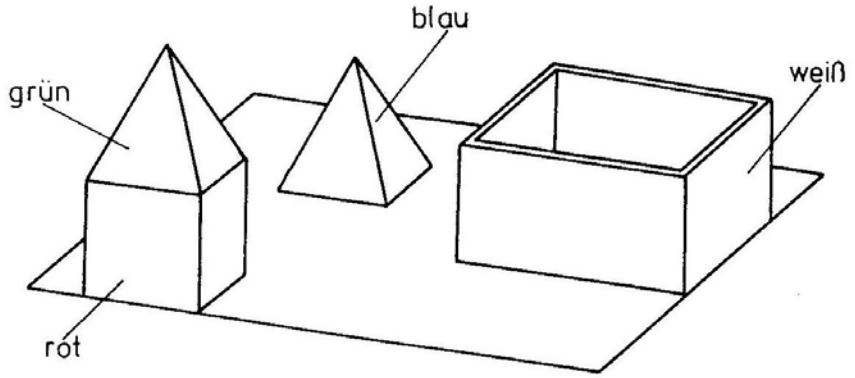



Abbildung 2

Blockwelt ausführt und darüberhinaus von diesem Partner gestellte Fragen über diese Blockwelt und über die eigenen „Handlungen“ des Systems beantwortet. Die fiktive Blockwelt selbst besteht aus einer Platte, auf der (zum Teil auch übereinander) einige Blöcke und Pyramiden sowie eine größere Box stehen. Die Abb. 2 zeigt einen möglichen Zustand dieser Blockwelt.

Eine solche Abbildung ist jedoch nur eine visuelle Hilfe für das menschliche Verständnis. Intern wird der in der Abb. 2 gezeigte Zustand eher durch eine Reihe von Ausdrücken (Listen) repräsentiert sein, z. B. durch die Ausdrücke:

```
(IST-EIN    BLOCK1    BLOCK)
(ORT        BLOCK1    (2, 1, 0))
(FARBE      BLOCK1    ROT)
(STUETZT    BLOCK1    BLOCK3)
(IST-EIN    BLOCK2    PYRAMIDE)
(ORT        BLOCK2    (3, 7, 0))
(FARBE      BLOCK2    BLAU)
(IST-EIN    BLOCK3    PYRAMIDE)
(ORT        BLOCK3    (2, 1, 2))
(FARBE      BLOCK3    GRÜN)
(IST-EIN    BLOCK4    BOX)
(ORT        BLOCK4    (8, 6, 0))
(FARBE      BLOCK4    WEISS)
```

Diese Menge von Ausdrücken bildet die Datenbasis, in der das „Wissen“ des Systems um den gegenwärtigen Zustand der Blockwelt gespeichert ist. Und auf diese Datenbasis angewandt liefert der PLANNER-Ausdruck (**) z. B. den Wert „BLOCK3“ als internen Namen für die grüne Pyramide. Dieser Name kann dann weiter verwendet werden – etwa zur Generierung einer Antwort auf die eingegebene Frage.

In gewisser Weise sieht es also so aus, als käme die Übersetzung des Eingabesatzes (*) in den PLANNER-Ausdruck (**) tatsächlich dem Verstehen des Satzes (*) gleich. Denn dieser PLANNER-Ausdruck liefert als Teil eines PLANNER-Programms genau die richtige Antwort auf die mit dem Satz (*) gestellte Frage. Beim zweiten Hinsehen erhärtet sich jedoch die Vermutung, daß Searle mit seiner Kritik auch im Hinblick auf die in diesem Abschnitt geschilderten NSS recht hat. Denn in Wirklichkeit gibt es weder eine Blockwelt noch eine grüne Pyramide noch einen roten Block. Und in Wirklichkeit wird auch die grüne Pyramide nicht durch einen Block gestützt. Die anschauliche Darstellung von Blockweltzuständen in Abbildungen wie der Abb. 2 verführt immer wieder dazu, zu vergessen, daß diese Welt nur fiktiv ist. Was es wirklich gibt, sind nur die Datenstrukturen im Computer. Und diese Datenstrukturen ihrerseits sind nichts anderes als an

bestimmten Speicherplätzen befindliche Bit-Muster. Wenn ein NSS, das nach den gerade geschilderten Strategien arbeitet, auf die Frage „Welche Pyramiden werden von einem Block gestützt?“ antwortet „Die grüne Pyramide wird von einem Block gestützt“, dann geschieht also genau das, was Searle mit seinem Chinese-Room-Experiment veranschaulichen wollte. Die Maschine analysiert und verändert eingegebene Zeichenfolgen und Bit-Muster nach Regeln, die sich allein an der physischen Gestalt dieser Muster orientieren. Und am Ende all ihrer Arbeit gibt sie eine Zeichenfolge aus, die sie allein aufgrund der äußeren Gestalt der Zeichen zusammengestellt hat. Es gibt also keinen Grund anzunehmen, die Maschine verstünde tatsächlich, was man ihr eingibt und was sie selbst ausgibt. Denn die Maschine weiß offenbar weder, was eine Pyramide ist, noch, was grün ist, noch, wann etwas von etwas gestützt wird.

4.

Ich denke, die Dinge lägen jedoch ganz anders, wenn sich die Datenbasis des im letzten Abschnitt besprochenen Systems nicht auf eine fiktive Blockwelt bezöge, sondern auf die wirkliche Umgebung des Systems, d. h. wenn das System eine Wahrnehmungskomponente enthielte, die ihrerseits diese Datenbasis als Modell der das System umgebenden Umwelt erst aufbauen würde. Dies ist durchaus keine utopische Idee. Denn nach den bisherigen Ergebnissen der KI-Forschung spricht nichts gegen die Möglichkeit visueller Systeme, die zumindest bei relativ einfachen Szenen in der Lage sind, die zu diesen Szenen gehörenden Objekte zu *identifizieren*, diese Objekte ihrer geometrischen Gestalt nach zu *klassifizieren* und die zwischen den Objekten bestehenden *räumlichen Beziehungen zu erkennen*.

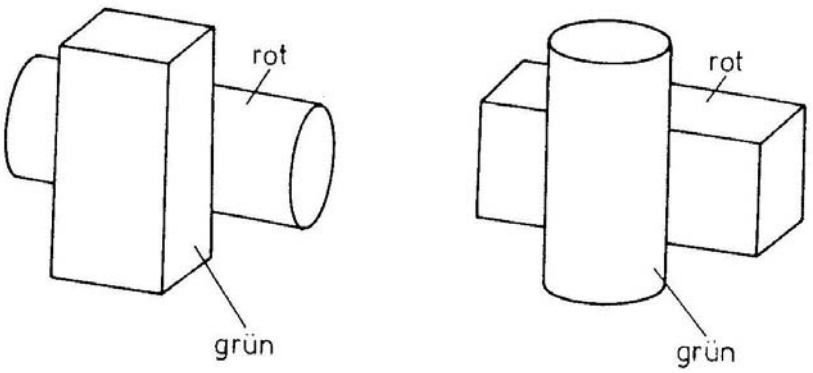



Abbildung 3

Ein solches System könnte also, wenn seine Kamera auf die in der Abb. 3 gezeigte Szene gerichtet wäre, erkennen, daß an dieser Szene vier Objekte beteiligt sind, es könnte diesen Objekten interne Namen geben (etwa OBJEKT1, …, OBJEKT4), es könnte erkennen, daß es sich bei den Objekten 1 und 4 um Quader handelt und bei den Objekten 2 und 3 um Zylinder, es könnte erkennen, daß der Zylinder 2 liegt, während der Zylinder 3 steht, es könnte erkennen, daß sich das Objekt 1 vor dem Objekt 2 befindet und das Objekt 3 vor dem Objekt 4, es könnte möglicherweise auch die Farbe der Objekte erkennen, d. h. es würde alle diese Sachverhalte in internen „Aussagen" festhalten, z. B. in der folgende Reihe von Ausdrücken:

(IST-EIN OBJEKT1 QUADER)
(FARBE OBJEKT1 GRÜN)
(VOR OBJEKT1 OBJEKT2)
(IST-EIN OBJEKT2 ZYLINDER)
(FARBE OBJEKT2 ROT)
(LIEGT OBJEKT2)
(IST-EIN OBJEKT3 ZYLINDER)
(FARBE OBJEKT3 GRÜN)
(VOR OBJEKT3 OBJEKT4)
(IST-EIN OBJEKT4 QUADER)
(FARBE OBJEKT4 ROT)

Nehmen wir nun an, daß sich ein System der gerade beschriebenen Art in einer Situation befindet, in der seine Kamera auf die in der Abb. 3 gezeigte Szene gerichtet ist, daß das System daraufhin die gerade beschriebene Datenbasis aufbaut und daß es daher, wenn man ihm die Frage stellt „Welcher Zylinder steht vor einem Quader?" antwortet „Der grüne Zylinder steht vor einem Quader". Kann man dann auch von diesem System noch sagen, es verstünde nicht wirklich, was man es fragt und was es selbst sagt?

Offenbar gehören zum Sprachverstehen sehr verschiedene Aspekte – so z. B. das Verstehen verschiedener Sprechakttypen. Aber eine zentrale These der Theorie des Sprachverstehens ist wohl die Auffassung: Die Bedeutung eines Satzes verstehen, heißt wissen, unter welchen Bedingungen dieser Satz wahr ist. D. h. die Bedeutung eines Satzes verstehen, heißt wissen, welche Wahrheitsbedingungen dieser Satz hat. Wenn man von diesem Grundsatz ausgeht, stellt sich die Frage jedoch so: Kann man von einem System wie dem geschilderten mit Recht sagen, daß es die Wahrheitsbedingungen von Sätzen kennt? Und wenn man auf diese Frage eine Antwort geben will, dann muß man sich darüber im klaren sein, was es heißt, die Wahrheitsbedingungen eines Satzes zu kennen.

Meiner Meinung nach kann man das Wissen um die Wahrheitsbedingungen von Sätzen jedoch einfach mit einer bestimmten Diskriminierungsfähigkeit identifizieren, nämlich mit der Fähigkeit, Situationen, in denen

ein Satz wahr ist, von Situation zu unterscheiden, in denen das nicht der Fall ist. Somit ist die Frage: Was ist erforderlich dafür, daß ein System S z. B. über die Fähigkeit verfügt, Situationen, in denen der Satz „Der grüne Zylinder steht vor einem Quader" wahr ist, von solchen Situationen zu unterscheiden, in denen das nicht der Fall ist? Soweit ich sehen kann, gehört zum Besitz dieser Fähigkeit unter anderem, daß das System S in dem Sinne über die in diesem Satz vorkommenden Begriffe „Zylinder", „Quader" und „vor" verfügt, daß es Situationen, in denen diese Begriffe zutreffen, von Situationen unterscheiden kann, in denen das nicht der Fall ist. Um Situationen, in denen der Satz wahr ist, identifizieren zu können, muß S also unter anderem in diesem Sinne über den Begriff „vor" verfügen. Ist das für das gerade geschilderte System der Fall?

Die Antwort auf diese Frage hängt davon ab, wie die Wahrnehmungs-Komponente dieses Systems im einzelnen arbeitet. Denn es könnte z. B. sein, daß diese Komponente einen Gegenstand A dann und nur dann als vor einem anderen Gegenstand B stehend klassifiziert, wenn A einen Teil von B verdeckt, bzw. genauer: wenn eine Ansicht der in der Abb. 4a gezeigten Art vorliegt.

In diesem Fall könnte man aber wohl nicht sagen, daß das System tatsächlich über den Begriff „vor" verfügt. Denn Bilder dieser Art können auf sehr verschiedene Weise zustande kommen. Es kann sein, daß A sich tatsächlich vor B befindet, es kann aber auch sein, daß dieses Bild von einem einzigen Objekt stammt mit drei nebeneinander angeordneten Teilen, von denen der mittlere etwas zurückgesetzt ist, und es kann auch sein, daß wir es tatsächlich nur mit einem zweidimensionalen Objekt, d. h. wirklich nur mit einem Bild zu tun haben. Von oben würden die drei geschilderten Situationen in etwa so aussehen, wie es die Abb. 4b zeigt.

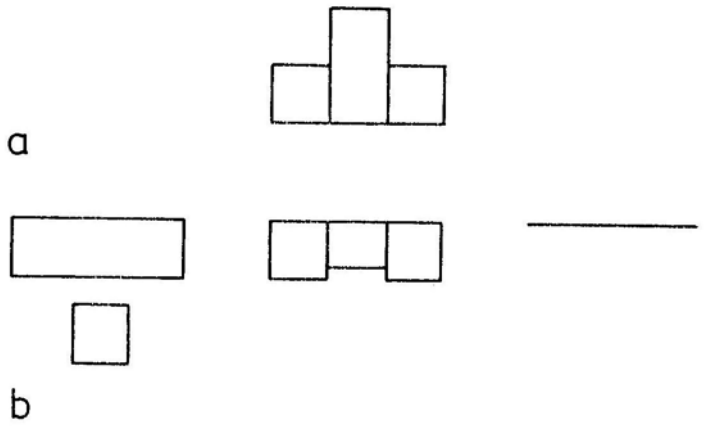



Abbildung 4

Es gibt natürlich Möglichkeiten, diese drei Situationen voneinander zu unterscheiden. Beim binocularen Sehen etwa hilft die Parallaxe bei der Bestimmung von Entfernungen, auch Texturunterschiede können entsprechende Hinweise geben. Besonders verläßlich sind jedoch die Hinweise, die sich daraus ergeben, wie sich die Ansicht einer Szene verändert, wenn sich die Objekte in dieser Szene bewegen oder wenn sich der Beobachter selbst bewegt. In den drei geschilderten Situationen z. B. würde sich die Ansicht für einen Beobachter, der sich im Uhrzeigersinn um die Szenen herumbewegt, sehr verschieden entwickeln – in etwa so, wie es in den verschiedenen Bildern der Abb. 5 gezeigt ist.

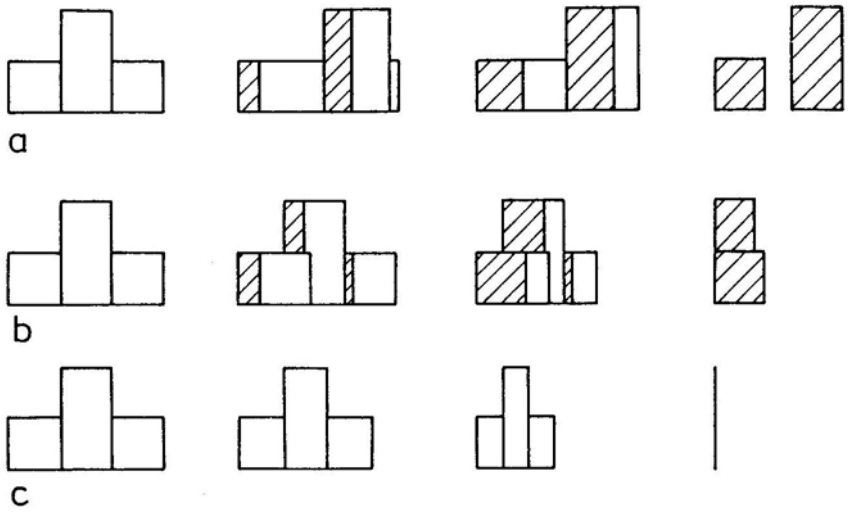



Abbildung 5

Meines Wissens ist es nicht unmöglich, ein visuelles System so einzurichten, daß es in der Lage ist, bei einer kontinuierlichen Veränderung der Ansicht einer Szene, wie sie durch normale Bewegungen hervorgerufen wird, die zu dieser Szene gehörenden Objekte zu fixieren und daher nicht nur statische Bilder, sondern auch Sequenzen aufeinanderfolgender Bilder zur Analyse der betreffenden Szene zu verwenden. Ein solches System könnte dann aber auch so eingerichtet werden, daß es vorläufige Meinungen immer wieder überprüft und gegebenfalls verändert. D. h. es scheint mir nicht unmöglich, ein System so einzurichten, daß es, auch wenn es beim ersten Anblick einer Szene zu der Überzeugung gekommen ist, daß zu dieser Szene zwei Objekte gehören, von denen sich das eine vor dem anderen befindet, diese Überzeugung revidiert, wenn es sich selbst um die Szene herumbewegt und sich dabei seine Ansicht der Szene nicht im Sinne der Folge 5a, sondern im Sinne der Folge 5b verändert. Wenn jedoch ein System in der Lage ist, auf diese Weise eine zunächst getroffene Fehlein-

schätzung zu korrigieren, d. h. die von ihm erzeugte Datenbasis entsprechend zu verändern, dann ist dies zumindest ein wichtiger Schritt hin zu der Fähigkeit, Situationen, in denen ein Gegenstand vor einem anderen steht, von Situationen zu unterscheiden, in denen das nicht so ist, und dann scheint mir nichts mehr dagegen zu sprechen, daß ein solches System tatsächlich über den Begriff „vor" verfügt, so wie wir ihn verwenden.

Wenn das aber der Fall ist und wenn ein System in ähnlicher Weise auch über die Begriffe „Zylinder", „Quader" und über die Farbbegriffe verfügt, dann kann das System auch so eingerichtet werden, daß es Situationen, in denen der Satz „Der grüne Zylinder steht vor einem Quader" wahr ist, von Situationen unterscheiden kann, in denen das nicht der Fall ist, d. h. so, daß es die Wahrheitsbedingungen dieses Satzes kennt. Und in diesem Fall scheint mir dann nichts mehr gegen die Annahme zu sprechen, daß ein solches System den gerade noch einmal angeführten Satz tatsächlich im Wortsinn *verstehen* kann.

Gegen diese Auffassung könnte man versucht sein einzuwenden, daß Systeme der gerade geschilderten Art zwar die *Extension* von Begriffen wie „vor", „Zylinder", „Quader" usw. verstehen können, daß es ihnen aber unmöglich ist, auch die *Intension* dieser Begriffe zu verstehen, daß also allen diesen Systemen *Bedeutungsverstehen* nur im Hinblick auf Extensionen und nicht im Hinblick auf Intensionen zukommt. Ein solcher Einwand würde jedoch auf einem grundlegenden Mißverständnis beruhen. Denn wenn man der herkömmlichen intensionalen Semantik folgt, dann kann die Intension eines Begriffs als eine Funktion aufgefaßt werden, die jeder möglichen Welt die entsprechende Extension dieses Begriffs zuordnet, d. h. die Menge all der Gegenstände, die in dieser möglichen Welt unter diesen Begriff fallen. Dies mag abstrakt klingen; aber es ist in diesem Zusammenhang durchaus von Bedeutung. Denn eine Funktion ist eine Zuordnung, die jedem Gegenstand aus dem Definitionsbereich der Funktion einen bestimmten Wert zuordnet. Eine Funktion kann somit durch jeden Mechanismus realisiert werden, der, angewandt auf einen Gegenstand aus dem Definitionsbereich der Funktion, den entsprechenden Wert dieses Gegenstandes erzeugt. Nach diesem Prinzip funktionieren z. B. Taschenrechner, die, wenn man zwei Zahlen eingibt und die „+"-Taste drückt, als Ergebnis die Summe dieser beiden Zahlen ausgeben. Für die hier diskutierte Frage bedeutet das, daß man jeden Mechanismus, der, wenn man ihm einen beliebigen Gegenstand vorlegt, genau dann z.B den Wert „1" ausgibt, wenn dieser Gegenstand unter den Begriff „Quader" fällt, und genau dann den Wert „0", wenn der Gegenstand nicht unter den Begriff „Quader" fällt, als eine *Realisierung der Intension* des Begriffs „Quader" auffassen kann. Jedes System, das über einen solchen Mechanismus verfügt, verfügt damit also über eine Realisierung der Funktion, die die Intension des Begriffs

„Quader“ ausmacht. Und von jedem System, das über einen solchen Mechanismus verfügt, kann man daher eher sagen, daß es die Intension, als daß es die Extension des Begriffs „Quader“ versteht.

Es ist klar, daß mancher Vertreter des zuvor angeführten Einwandes bei der Intension eines Begriffs weniger an eine bestimmte Funktion denkt, als vielmehr an die Beziehungen, in denen dieser Begriff zu anderen Begriffen steht, oder an das Wortfeld dieses Begriffes. Dazu ist zweierlei zu sagen. Erstens scheint mir der Begriff der Intension, wie er in der intensionalen Semantik verstanden wird, der grundlegendere und systematisch wichtigere Begriff zu sein. Und zweitens ist es kein Problem, Systeme der in diesem Abschnitt geschilderten Art – z. B. durch Implementation semantischer Netze – so zu erweitern, daß sie auch wissen, in welchen Beziehungen ein bestimmter Begriff zu anderen Begriffen steht.

## 5.

Wenn die bisherigen Überlegungen zutreffen, dann ist es offenbar nicht unmöglich, daß bestimmte Computer bzw. bestimmte Maschinen, deren Kern ein Computer bildet, doch im Wortsinn Sprache verstehen. Wie verhält sich dieses Ergebnis zu den Überlegungen Searles? Bedeutet es, daß Searle mit seiner These unrecht hat, daß Maschinen, deren Verhalten allein durch ein formales Programm bestimmt ist, keine Sprache verstehen können? So einfach ist es sicher nicht. Denn das Verhalten von Systemen, wie sie im letzten Abschnitt angedeutet wurden, ist offenbar *nicht nur* durch formale Programme, sondern auch dadurch bestimmt, daß diese Systeme über ihre visuellen Komponenten in bestimmten *kausalen* Beziehungen zu ihrer Umwelt stehen. Die These, daß solche Systeme möglicherweise in der Lage sind, Sprache zu verstehen, entspricht also ziemlich genau der Auffassung, die J. Fodor in seinen kommentierenden Bemerkungen zu Searles Aufsatz so formuliert hat:

> Searle is certainly right that instantiating the same program that the brain does is not, in and of itself, a sufficient condition for having those propositional attitudes characteristic of the organism that has the brain. … However, Searles treatment of the ‚robot reply‘ is quite unconvincing. Given that there are *the right kinds of causal linkages* between the symbols that the device manipulates and things in the world – including the afferent and efferent transducers of the device – it is quite unclear that intuition rejects ascribing propositional attitudes to it. (SB 431 – Hervorh. vom Verf.)

Aber obwohl Searle Fodors Zugeständnis, „daß die Instantiierung eines Programms keine hinreichende Bedingung für Intentionalität darstellt“, offenbar mit großer Freude vermerkt, hält er Fodors Gegenvorschlag immer noch für völlig unzureichend. Auch „geeignete“ kausale Verbindungen ei-

nes Computers mit der ihn umgebenden Welt befähigen diesen nicht, wirklich Sprache zu verstehen.

> [N]o matter what outside causal impacts there are on the formal tokens, these are not by themselves sufficient to give the tokens any intentional content. No matter what caused the tokens, the agent still doesn't understand Chinese. Let the egg foo yung symbol be causally connected to egg foo yung in any way you like, that connection by itself will never enable the agent to interpret the symbol as meaning egg foo yung. To do that he would have to have, for example, some *awareness* of the causal relation between the symbol and the referent; but now we are no longer explaining intentionality in terms of symbols and causes but in terms of symbols, causes, and intentionality, and we have abandoned both strong AI and the robot reply. (MBP 454)

Was ist der Grund für diese immer noch ablehnende Haltung Searles? Welche Argumente hat er für seine Auffassung, daß auch Systeme der zuvor geschilderten Art nicht wirklich in der Lage sind, Sprache zu verstehen? Einen ersten Hinweis kann man in den Erwiderungen finden, mit denen Searle auf zwei sehr interessante Einwände gegen seine Thesen reagiert hat: den System-Einwand und den Roboter-Einwand. Den System-Einwand formuliert Searle selbst so:

> While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has ‚data banks' of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part. (MBP 419)

Auf diesen Einwand erwidert Searle damit, daß er die Ausgangssituation etwas modifiziert. Man könne durchaus annehmen, so schreibt er, daß die im Zimmer eingesperrte Person alle Elemente des gerade noch einmal geschilderten Systems *internalisiert*. Sie lernt die Regeln des Regelbuchs auswendig; ebenso alles, was in den beiden Körben auf den verschiedenen Blättern mit Chinesischen Zeichen steht. Sie führt alle Berechnungen im Kopf aus. Kurz, die Person inkorporiert alles, was für das System wichtig ist. Man kann sogar annehmen, daß die Person nicht in einem Zimmer eingesperrt ist, sondern irgendwo im Freien arbeitet. Auch in diesem Fall versteht die Person jedoch kein Chinesisch; denn an den Grundzügen der Situation hat sich nichts verändert. Und das bedeutet, daß auch das Gesamtsystem kein Chinesisch versteht; denn, so wie die Situation jetzt konstruiert ist, *ist* die Person das System.

Der Roboter-Einwand stellt einen anderen Aspekt in den Vordergrund. Diesen Einwand formuliert Searle so:

> Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking – anything you like. The robot would, for example, have a television camera attached to it that enabled it to ‚see', it would have arms and legs that enabled it to ‚act', and all of this would be controlled by its computer ‚brain'. Such a robot would, unlike Schank's computer, have genuine understanding and other mental states. (MBP 420)

Auch dieser Einwand führt jedoch Searle zufolge nicht zum Ziel. Denn seiner Meinung nach ändert die Hinzufügung von „Wahrnehmungs- und Bewegungskomponenten" nichts an der Fähigkeit (oder besser: Unfähigkeit) des Systems, wirklich Sprache zu verstehen. Dies zeigt sich, so Searle, daran, daß man auf das von den Vertretern des Roboter-Einwandes ins Spiel gebrachte System dasselbe Gedankenexperiment anwenden kann. Angenommen, statt eines Computers in einem Roboter sitzt – wie in der Ausgangssituation – eine Person in einem Zimmer. Diese Person bekommt noch mehr Chinesische Schriftzeichen und noch mehr in einer für sie verständlichen Sprache abgefaßte Regeln, nach denen sie auf eingehende Symbole mit der Ausgabe von Symbolen reagiert. Weiter angenommen, einige der eingehenden Symbole kommen, ohne das die Person dies weiß, von einer Fernsehkamera und einige von ihr ausgegebene Symbole steuern die Motoren des Roboters so, daß sich seine Arme oder Beine auf eine bestimmte Weise bewegen. Offenbar ändert dies, so schreibt Searle, überhaupt nichts daran, daß die Person im Zimmer nichts anderes tut, als formale Symbole nach formalen Regeln zu manipulieren. Sie empfängt zwar „Informationen" von der Fernsehkamera, und sie gibt „Instruktionen" aus zur Bewegung der Arme und Beine des Roboters. Aber sie weiß nicht, daß sie das tut. Für sie ist die Situation allein dadurch charakterisiert, daß sie formale Symbole erhält und formale Symbole ausgibt. Und das tut sie nach rein formalen Regeln, ohne auch nur die geringste Kenntnis davon zu haben, was diese Symbole bedeuten könnten.

Mir scheint, daß Searle bei dieser Erwiderung auf den Roboter-Einwand von einem einfachen Trick Gebrauch macht. Und dieser Trick besteht darin, daß er bei der Übertragung der ursprünglichen Chinese Room-Situation auf den Roboterfall die Grenze zwischen dem Computer bzw. der Person auf der einen und den übrigen Teilen des Systems auf der anderen Seite so zieht, daß z. B. der Computer bzw. die Person als Eingabe immer nur das erhält, was die „Wahrnehmungskomponenten" als output liefern. Nur auf diese Weise kann Searle sicherstellen, daß der Computer oder die Person immer nur mit formalen Mustern und nicht direkt mit der Welt konfrontiert sind. Und auch nur auf diese Weise kann er seine Schlußfolgerung errei-

chen, daß der Computer *in* dem Roboter immer noch kein Chinesisch versteht. Doch das war gar nicht der Streitpunkt. Denn es geht ja auch nicht darum, ob das *Gehirn* einer Person eine Sprache versteht, sondern ob die ganze Person dies tut.

Aus diesem Grund wird der Trick Searles auch sofort sichtbar, wenn man den Roboter-Einwand mit dem System-Einwand verbindet, so wie die Vertreter des Roboter-Einwandes diesen Einwand offenbar auch schon von Anfang an gemeint hatten. Denn selbst Searle formuliert diesen Einwand so, daß er auf die Schlußfolgerung herausläuft, daß man von „solchen Robotern" (und nicht etwa von den in „solchen Robotern" steckenden Computern) zu recht sagen kann, daß sie tatsächlich Sprache verstehen.

Was könnte Searle auf einen solchen verbundenen Einwand erwidern? Der Strategie seiner Erwiderung auf den System-Einwand folgend müßte er behaupten, daß sich auch in diesem Fall nichts Wesentliches ändern würde, wenn die Person alle für das System (den gesamten Roboter) relevanten Teile inkorporiert. Doch das würde in diesem Fall eben auch die Inkorporierung der Fernsehkamera bedeuten und somit zur Folge haben, daß die Person als Gesamtsystem nicht mehr nur formale Symbole manipuliert, sondern auch in bestimmter Weise auf die Außenwelt reagiert, z. B. bestimmte formale Symbole überhaupt erst als Reaktion auf bestimmte von ihr wahrgenommene Situationen herstellt. Wenn jedoch die Person als Gesamtsystem etwa das Symbol „grauer Hut" immer und nur in Situationen generiert, in denen sie einen grauen Hut wahrnimmt, dann scheint es doch nicht mehr völlig unplausibel anzunehmen, daß diese Person als Gesamtsystem weiß, was das Symbol „grauer Hut" bedeutet. Fodor hat deshalb offensichtlich recht, wenn er schreibt, Searles Erwiderung auf den Roboter-Einwand sei nicht besonders überzeugend.

Searle jedoch findet seinerseits diese Auffassung Fodors wenig überzeugend. Denn seiner Meinung nach ermöglicht das bloße Bestehen von kausalen Beziehungen – welcher Art auch immer – zwischen einem Symbol und dem durch das Symbol Bezeichneten noch kein wirkliches Sprachverstehen.

> To do that [the system] would have to have, for example, some *awareness* of the causal relation between the symbol and the referent. (MBP 454)

Möglicherweise findet sich in einer eher beiläufigen Bemerkung Searles der Schlüssel für die nicht leicht zu verstehende Diskrepanz zwischen den Ansichten Searles und Fodors. Denn in seiner Erwiderung auf den System-Einwand schreibt Searle unter anderem:

> … the English subsystem *knows* that ‚hamburgers' *refers* to hamburgers …. (MBP 419 – Hervorh. vom Verf.)

Und:

But the Chinese system *knows* none of this. (ebd. – Hervorh. vom Verf.)

Die entscheidende Frage im Zusammenhang mit dieser Bemerkung scheint mir zu sein, was hier mit „wissen" gemeint sein soll. Der Hinweis auf die Notwendigkeit von „awareness" in seiner Erwiderung auf Fodor legt die Vermutung nahe, daß Searle davon ausgeht, daß zum Sprachverstehen *explizites* Wissen im Sinne von „wissen, daß" erforderlich ist. Über Wissen dieser Art verfügen die oben im Abschnitt 4. geschilderten Systeme (und an ähnliche Systeme scheint mir auch Fodor zu denken) natürlich nicht. Denn zwar wissen auch diese Systeme in einem gewissen Sinn, was das Wort „Quader" bedeutet, insofern nämlich, als sie Situationen, in denen Gegenstände die durch dieses Wort bezeichnete Eigenschaft haben, von solchen Situationen unterscheiden können, in denen das nicht der Fall ist, und als sie das Wort „Quader" daher in den verschiedensten Situationen richtig verwenden können. Aber ein solches Wissen ist offensichtlich implizit, ein „wissen, wie". Denn es besteht allein aus einer Reihe von Fähigkeiten.

Die entscheidende Frage ist somit, ob Searle recht hat, wenn er explizites Wissen um die Bedeutung sprachlicher Ausdrücke zur Voraussetzung von Sprachverstehen erklärt. Soweit ich sehen kann, führt er für diese Auffassung keine Gründe an. Und ich kann auch selbst keine systematischen Gründe erkennen. Denn Sprachverstehen ist in der Sprachphilosophie immer wieder mit der Fähigkeit, sprachliche Ausdrücke richtig zu verwenden, in Zusammenhang gebracht worden. Und diese Fähigkeit setzt offenbar nur ein implizites Wissen voraus. Mir scheint daher, daß implizites Wissen um die Bedeutung sprachlicher Ausdrücke zum Sprachverstehen ausreicht und daß man Systeme der im Abschnitt 4. behandelten Art demzufolge völlig zurecht als sprachverstehende Systeme bezeichnen kann.[4] Und dies ist, wie mir scheint, auch genau die Auffassung, die den Überlegungen Fodors zugrundeliegt.

## Literatur

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

[4] Vgl. jedoch die Überlegungen R. Cummins', der in *The Nature of Psychological Explanation* an mehreren Stellen (z. B. S. 76 ff.) auf einer ähnlichen Linie wie Searle argumentiert, daß wirkliches Verstehen voraussetzt, daß das System seine eigenen Repräsentationen versteht, daß diese Repräsentationen Repräsentationen auch für das System selbst sind.

Fodor, J.A., (SB) „Searle on what only Brains can do“. *The Behavioral and Brain Sciences* 3 (1980), 331 f.

Searle, J., (MBP) „Minds, Brains and Programs“. *The Behavioral and Brain Sciences* 3 (1980), 417–424 und 450–456.

ders., (GHW) *Geist, Hirn und Wissenschaft*. Frankfurt/M. 1986 (dt. Übersetzung von *Minds, Brains, and Science. The 1984 Reith Lectures*. London 1984).

Winograd, T., *Understanding Natural Language*. New York 1972.

ders., „Software für Sprachverarbeitung“. *Spektrum der Wissenschaft* November 1984, 88–102.

# 6

## Semantische Maschinen*

1. Daniel Dennett gehört – anders als etwa John Searle und Hubert Dreyfus – sicher nicht zu den grundsätzlichen Kritikern der KI-Forschung. In seinem Aufsatz „Three Kinds of Intentional Psychology" charakterisiert Dennett jedoch die Situation der kognitiven Psychologie auf eine Weise, die ebenfalls eine große Skepsis im Hinblick auf die Möglichkeiten der KI-Forschung zu implizieren scheint. Dennett zufolge steht die kognitive Psychologie nämlich vor einer prinzipiell unlösbaren Aufgabe – der Aufgabe zu erklären, wie eine rein syntaktische Maschine semantische Leistungen erbringen kann. Denn die Psychologie und die evolutionäre Biologie beschreiben das menschliche Gehirn als eine *semantische Maschine* (*a semantic engine*), deren Aufgabe es ist, die Bedeutung der verschiedenen eingehenden Reize zu erfassen, sie ihrem Sinne nach zu unterscheiden und dann entsprechende Handlungen zu veranlassen. Die Physiologie dagegen stellt dieses Organ als eine rein *syntaktische Maschine* (*a syntactic engine*) dar, die alle eingehenden Reize nur aufgrund ihrer strukturellen, zeitlichen und physikalischen Merkmale zu unterscheiden vermag und deren Aktivitäten allein durch diese „syntaktischen" Merkmale bestimmt wird. Nach Dennett stellt sich damit die kritische Frage:

> Now how does the brain manage to get semantics from syntax? How could any entity … get the semantics of a system from nothing but its syntax? (1981, 61)

Und Dennetts Antwort lautet:

> It couldn't. The syntax of a system doesn't determine its semantics. By what alchemy, then, does the brain extract semantically reliable results from syntactically driven operations? It cannot be designed to do an impossible task …. (ebd.)

Offenbar ist also auch Dennett der Auffassung, daß semantische Maschinen unmöglich sind.[1] Seiner Meinung nach können syntaktische Maschinen –

---

* Erstveröffentlichung in: *Intentionalität und Verstehen*, hgg. vom Forum für Philosophie Bad Homburg. Frankfurt a.M.: Suhrkamp 1990, 196–211.

[1] Dennoch unterscheidet sich die Position Dennetts natürlich erheblich von den Positionen Searles oder Dreyfus'. Denn die referierte Argumentation macht deutlich, daß Dennett auch das menschliche Gehirn für eine rein syntaktische Maschine hält. Dennett geht es also nicht um einen möglichen Unterschied zwischen Menschen und Maschinen, sondern um das ganz anders geartete Problem, welchen Platz Semantik und Intentionalität in einem Universum ha-

wie zum Beispiel Computer – bestenfalls so konstruiert werden, daß es so scheint, als würden sie semantische Aufgaben lösen, während sie tatsächlich doch immer syntaktische Maschinen bleiben. Diese Auffassung erläutert er am Beispiel einer Maschine, die zur Überwachung der an einem Telegraphen eingehenden Meldungen eingesetzt werden soll und deren Aufgabe es ist, aus diesen Meldungen genau alle Morddrohungen (genauer: alle in Englisch verfaßten Morddrohungen) auszusondern. Keine Maschine kann Dennett zufolge diese Aufgabe vollständig lösen. Denn dafür wäre es nötig, die semantische Kategorie „In Englisch verfaßte Morddrohung" syntaktisch zu definieren. Und das scheint ihm eine Aufgabe zu sein, die prinzipiell nicht gelöst werden kann. Deshalb wird sich die Maschine mit unvollkommenen Hilfsstrategien behelfen müssen, indem sie zum Beispiel alle eingehenden Meldungen daraufhin untersucht, ob sie die Zeichenfolgen

… I will kill you …

oder

… you … die … unless …

oder ähnliche Zeichenfolgen enthält. Auf diese Weise wird die Maschine zwar nicht alle, aber doch wenigstens eine gewisse Zahl von Morddrohungen aussondern können. Mehr ist nach Dennett jedoch auch nicht möglich. Denn:

> … if you want to get semantics out of syntax …, variations on this basic strategy are your only hope. You must put together a bag of tricks and hope nature will be kind enough to let your device get by. (ebd., 62 f.)

2. Wenn man Dennetts System zur Erkennung von Morddrohungen betrachtet, dann fällt auf, daß die simplen Strategien des *pattern matching*, mit denen dieses System arbeitet, den Strategien des Programms ELIZA sehr ähnlich sind, das Mitte der 60er Jahre von Joseph Weizenbaum als Programm zur Simulierung von Dialogen in einer nicht-direktiven Psychotherapie entwickelt wurde. ELIZA schien vielen ein wichtiger Fortschritt in der Entwicklung natürlichsprachlicher Systeme zu sein. Denn bei der Arbeit mit diesem Programm kamen sehr eindrucksvolle Dialoge zustande, die auf den ersten Blick wirklichen therapeutischen Dialogen zumindest verblüffend ähnlich waren. Der folgende Dialog etwa wird von Weizenbaum

---

ben können, in dem es nur syntaktische Mechanismen gibt. Ähnlich wie Searle scheint jedoch auch Dennett anzunehmen, daß zwischen syntaktischen und semantischen Maschinen eine unüberbrückbare Kluft besteht und daß daher syntaktische Maschinen niemals zugleich auch semantische Maschinen sein können.

selbst zur Veranschaulichung der Leistungen von ELIZA angeführt.[2] (In diesem Dialog sind die Sätze in Normalschrift die über eine Tastatur in den Computer eingegebenen Äußerungen des menschlichen Partners, der sich in der Situation einer Psychotherapie glaubt, und die Sätze in Kapitälchen die am Bildschirm ausgegebenen Antworten des Computers, der den Part des Therapeuten spielt.)

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DREPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLENCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK THAT I AM NOT VERY AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK THAT I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
My father is afraid of anybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

[2] Weizenbaum 1976, S. 3 f.

Erstaunlicher noch als dieser Dialog ist jedoch, mit welch einfachen Mitteln ELIZA seine Leistungen erzielt.[3] Vorab wird dem Programm für die verschiedenen Situationen, zu denen es in einem Dialog Antworten erzeugen kann, ein *script* eingegeben. Dieses *script* enthält eine Liste von Schlüsselwörtern und für jedes dieser Schlüsselwörter eine Reihe von sprachlichen Mustern, nach denen das Programm sucht, und zu jedem dieser Muster eine Reihe von Transformationsregeln. Zum Schlüsselwort „me" zum Beispiel gehört das Muster

(* YOU * ME),

und zu diesem Muster die Transformationsregel

(WHAT MAKES YOU THINK THAT I 3 YOU)

Dabei steht „*" für beliebigen Text und „3" in der Transformationsregel für das dritte Element der Eingabe (in diesem Fall also für alles zwischen „YOU" und „ME"). Die eingegebenen Sätze werden von links nach rechts nach Schlüsselwörtern durchsucht. Wenn mehrere Schlüsselwörter gefunden werden, werden sie in der Reihenfolge ihrer Wichtigkeit in einem Schlüsselwort-Speicher abgelegt. Im nächsten Schritt wird jeder eingegebene Satz mit den Mustern verglichen, die mit dem ersten der im Schlüsselwort-Speicher abgelegten Schlüsselwörter verbunden sind. Dabei wird zuerst nach schwierigeren Mustern gesucht. Denn für das zum Schlüsselwort „I" gehörende Muster

(* I * DEPRESSED *)

wird sich schwerer eine Entsprechung finden lassen als für das Muster

(* I *).

Wenn eine Übereinstimmung gefunden ist, generiert ELIZA mit Hilfe einer zugehörigen Transformationsregel eine Antwort, wobei eine zusätzliche Komponente dafür sorgt, daß nicht immer derselbe Antworttyp erscheint. Falls keine Übereinstimmung gefunden wird, wird das Verfahren mit dem nächsten Wort aus dem Schlüsselwort-Speicher wiederholt. Wenn der Schlüsselwort-Speicher leer ist, werden Antworten wie „PLEASE GO ON" oder „VERY INTERESTING" erzeugt. Außerdem verfügt ELIZA bei der Erzeugung von Antworten noch über einige zusätzliche Techniken: so kann das Programm zum Beispiel Schlüsselwörter in der Antwort ersetzen, Schlüsselwörter mit Klassen oder Situationen in Beziehung bringen („MOTHER" impliziert „FAMILY") oder im Laufe der „Unterhaltung" auf frühere Schlüsselwörter zurückkommen.

---

[3] Die folgende Beschreibung von ELIZA folgt im wesentlichen Barr/Feigenbaum 1981, 285–287.

Keine Frage, all das ist sehr geschickt ausgedacht, und es führt auch, wie Weizenbaums Beispiel zeigt, zu sehr eindrucksvollen Ergebnissen. Dennoch ist ELIZA aber sicher kein System, vom dem man zu Recht sagen könnte, daß es die eingegebenen Sätze versteht. Das wird besonders an den folgenden drei Punkten deutlich.

(a) ELIZA kann syntaktisch wohlgeformte nicht von syntaktisch falschen Sätzen unterscheiden. Denn wie das System auf die Eingabe

You are not very aggressive

mit

WHAT MAKES YOU THINK THAT I AM NOT VERY AGGRESSIVE

reagiert, so reagiert es auf die Eingabe

You are not aggressive very

mit

WHAT MAKES YOU THINK THAT I AM NOT AGGRESSIVE VERY.

(b) ELIZA kann semantisch unsinnige Sätze nicht von semantisch sinnvollen Sätzen unterscheiden. Das System murrt nicht, wenn man zum Beispiel den Satz eingibt

Please ask the table to pick up a brick for the baby.[4]

Und auf die Eingabe

You always aboabo me

antwortet ELIZA schlicht mit

WHY DO YOU THINK THAT I ALWAYS ABOABO YOU.

(c) ELIZA kann auf die verschiedenen Sprechakte des Partners nur stereotyp und oft unangemessen reagieren. Auf Fragen zum Beispiel kann es nur mit

WHY DO YOU ASK

oder mit einem Rückgriff auf frühere Themen antworten. Wirkliche Antworten geben kann das System nicht. Auch auf Bitten kann es nicht eingehen. Befehle nicht ausführen. Es kann keine oder kaum Schlüsse ziehen. Behauptungen sind irgendwann vergessen, usw. Ein eindrucksvolles Beispiel für dieses stereotype und meist unsinnige Verhalten ist der folgende Dialog,[5] der von einem Partner ausgelöst wurde, der einfach nur den Computer benutzen wollte.

---

[4] Dieser Beispielsatz stammt von Margareth Boden (siehe Boden 1977, 108).
[5] Dieser Dialog ist zitiert nach Boden 1977, 96.

Do you think that I can use the computer this morning?
WHY DO YOU ASK
I might be able to make some additional telcomp sales.
WHY AREN'T YOU SURE
My prospective customers have never seen a demonstration of the system.
DOES THIS MEAN VERY MUCH TO YOU
Of course it does.
YOU SEEM VERY POSITIVE ABOUT THAT
This is ridiculous.
WHAT DO YOU MEAN THIS IS RIDICULOUS
Is there any reason why I can't use the computer now?
TELL ME MORE ABOUT YOUR PROSPECTIVE CUSTOMERS
Please dial me up on 491–1850.

Natürlich kam ELIZA dieser Aufforderung nicht nach.

ELIZA ist also wirklich nur eine syntaktische Maschine. Und was noch schwerer wiegt: eine syntaktische Maschine, die eine semantische Maschine bestenfalls sehr unvollkommen imitiert. Doch dies ist natürlich noch kein Beweis für die These, daß semantische Maschinen prinzipiell unmöglich sind. Denn die Entwicklung natürlichsprachlicher Systeme (NSS) ist nicht bei ELIZA stehengeblieben. Sie hat vielmehr beträchtliche Fortschritte gemacht, so daß sich heute die Frage stellt, ob nicht wenigstens diese neueren Systeme als echte semantische Systeme betrachtet werden können.

3. Eine Antwort auf diese Frage ist sicher nur möglich, wenn klar ist, wie sich die neueren Systeme von einem System wie ELIZA unterscheiden. Ich will deshalb versuchen, die Arbeitsweise dieser Systeme hier zumindest in ihren Grundzügen kurz darzustellen.[6]

In der Regel beruht die Arbeitsweise aller neueren NSS auf der Tatsache, daß sich natürlichsprachliche Sätze mit Hilfe einer Phrasenstrukturgrammatik (PS-Grammatik) analysieren lassen, das heißt, daß man – sofern ein Satz *S* syntaktisch eindeutig ist (wovon ich im folgenden der Einfacheit halber ausgehen werde) – diesem Satz mit Hilfe eines Parsers in eindeutiger Weise eine strukturelle Beschreibung, zum Beispiel einen PS-Baum zuordnen kann. So entspricht etwa dem Satz

(1) Die grüne Pyramide steht auf dem roten Block

der in Abb. 1 gezeigte Strukturbaum.

[6] Vgl. zum folgenden Charniak/McDermott 1985, ch. 4.

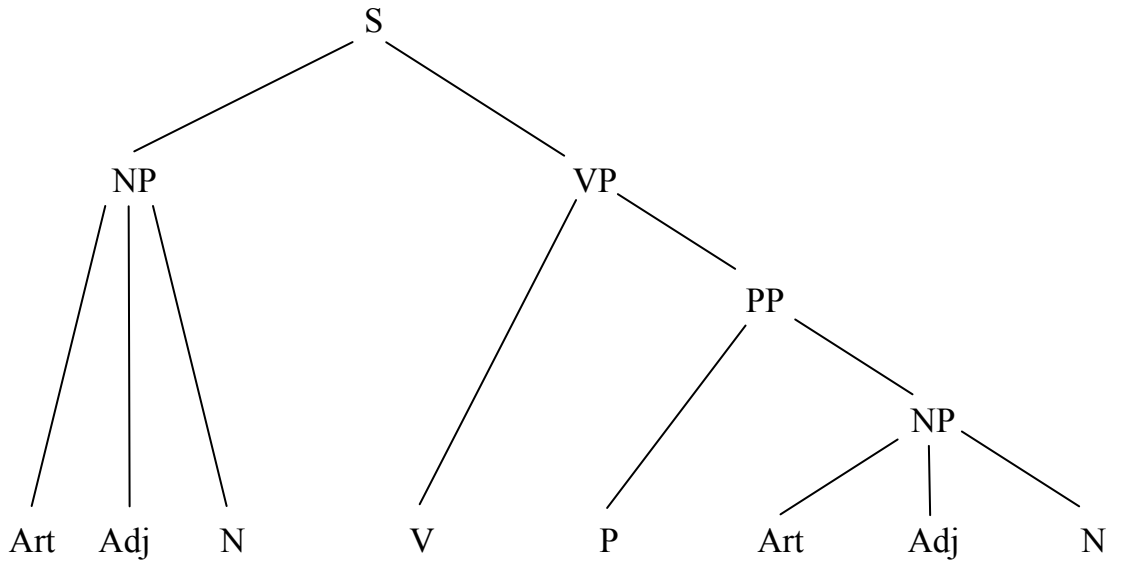



Abbildung 1

Die Frage ist nun, wie ein Programm von einem solchen Struktur-Baum ausgehend in einem Prozeß der *semantischen Analyse* eine interne Repräsentation aufbauen kann, die den Sinn dieses Satzes wiedergibt. Ich will anhand eines einfachen Beispiels versuchen, eine Antwort auf diese Frage wenigstens zu skizzieren.

Gehen wir aus von einem System *S*, das über eine interne Repräsentation der in der Abb. 2 gezeigten Blockwelt-Szene verfügt.

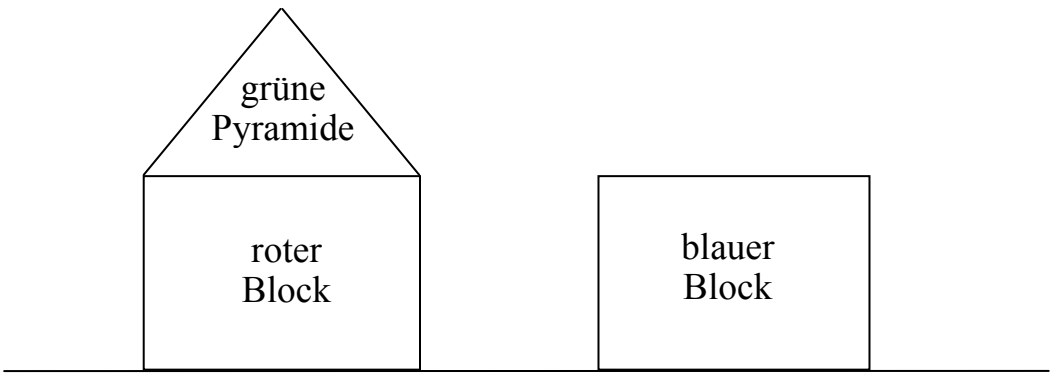



Abbildung 2 (Blockweltszene)

Eine solche Repräsentation könnte zum Beispiel die Form propositionaler Listen haben, das heißt, in dem System *S* könnten die für diese Szene wesentlichen Fakten in der folgenden Datenbasis gespeichert sein:

(IST-EIN OBJEKT-1 BLOCK)
(FARBE OBJEKT-1 ROT)
(IST-EIN OBJEKT-2 PYRAMIDE)
(FARBE OBJEKT-2 GRÜN)
(STEHT-AUF OBJEKT-2 OBJEKT-1)
(IST-EIN OBJEKT-3 BLOCK)
(FARBE OBJEKT-3 BLAU)

Wie kann für dieses System eine interne Repräsentation des Satzes (1) aussehen, von der man sinnvollerweise sagen kann, daß sie den Sinn dieses Satzes wiedergibt? Wenn wir selbst diesen Satz interpretieren, dann ist uns klar, daß in ihm von zwei Objekten die Rede ist – den Objekten, auf die sich die N-Phrasen „die grüne Pyramide“ und „dem roten Block“ beziehen – und daß der Satz besagt, daß diese beiden Objekte in einer bestimmten Relation zueinander stehen. Für die Interpretation des Satzes ist es also zunächst einmal wichtig, den Bezug der in ihm vorkommenden N-Phrasen zu klären. Fragen wir also zuerst, welche internen Repräsentationen im System *S* diese Aufgabe erfüllen können.

Am Aufbau der Datenbasis, in der die für die in der Abb. 2 gezeigten Szene wesentlichen Fakten im System *S* repräsentiert sind, wird deutlich, daß *S* für alle an dieser Szene beteiligten Objekte interne Namen verwendet: für die Pyramide den Namen „OBJEKT-2“, für den roten Block den Namen „OBJEKT-1“ und für den blauen Block den Namen „OBJEKT-3“. Um den Referenten der Phrase „die grüne Pyramide“ zu finden, benötigt *S* daher eine Funktion, die als Wert den Namen des Objekts liefert, das zugleich grün und eine Pyramide ist. Diese Aufgabe kann aber zum Beispiel durch die folgende Funktion erfüllt werden:

(2) (FINDE-WERT '?X '(UND (IST-EIN ?X PYRAMIDE) (FARBE ?X GRÜN)))

Man erhält also eine vernünftige Interpretation, wenn das System *S* bei der semantische Analyse der N-Phrase „die grüne Pyramide“ diese Funktion als interne Repräsentation zuordnet. Denn wenn *S* diese Funktion aufruft, ergibt sich als Wert der Name „OBJEKT-2“, also der interne Name des Objekts, auf das sich die interpretierte N-Phrase bezieht.

Beim Aufbau einer solchen Interpretation kann *S* zum Beispiel so vorgehen: Im internen Lexikon von *S* sind die Wörter „grün“ und „Pyramide“ mit den Klauseln

(3) (FARBE ?X GRÜN)

und

(4) (IST-EIN ?X PYRAMIDE)

assoziiert. Da „grüne Pyramide“ im PS-Baum als Adj-N Komponente analysiert wurde, deutet *S* diesen Ausdruck als Konjunktion, das heißt, ausgehend von diesen Einträgen baut die semantische Komponente von *S* als interne Repräsentation für diese beiden Wörter den Ausdruck

(5) (UND (IST-EIN ?X PYRAMIDE) (FARBE ?X GRÜN))

auf. Dies ist natürlich noch nicht die Interpretation der gesamten N-Phrase „die grüne Pyramide“. Denn es fehlt die Interpretation des bestimmten Artikels. Mit diesem Artikel ist als Bedeutung im internen Lexikon von *S* zum Beispiel der Ausdruck

(6) (FINDE-WERT (REFERENT $NP) (SINN $NP))

verbunden, der bei der Gesamtinterpretation der N-Phrase „die grüne Pyramide“ den Ausgangspunkt bildet. Dabei geht die semantische Komponente von *S* so vor, daß sie im Ausdruck (6) zunächst den Teilausdruck „(REFERENT $NP)“ durch die Variable „?X“ ersetzt und dann den Teilausdruck „(SINN $NP)“ durch das Ergebnis der Analyse des Teilausdrucks „grüne Pyramide“, das heißt also durch den Ausdruck (5). Auf diese Weise entsteht als Repräsentation der gesamten NP der Ausdruck (2).

In ähnlicher Weise wird auch die Repräsentation für den ganzen Satz (1) aufgebaut. Dabei beginnt die semantische Komponente von *S* mit dem Verb „steht“, für das sich im Lexikon zum Beispiel der Ausdruck

(7) (STEHT-AUF (REFERENT $SUBJ-NP) (REFERENT $„AUF“-PP-NP))

findet. Beim Aufbau der Repräsentation für den ganzen Satz (1) wird – ausgehend von diesem Ausdruck – im erstem Schritt die Subj-NP analysiert, wobei sich als Ergebnis der Ausdruck (2) ergibt. Durch diesen Ausdruck wird in (7) der Teilausdruck „(REFERENT $SUBJ-NP)“ ersetzt. Im zweiten Schritt wird die in die PP eingebetteten NP analysiert, wobei sich – analog zum Ausdruck (2) – als Ergebnis der Ausdruck ergibt:

(8) (FINDE-WERT ’?X ’(UND (IST-EIN ?X BLOCK) (FARBE ?X ROT)))

Durch diesen Ausdruck wird in (7) der Teilausdruck „(REFERENT $„AUF“-PP-NP)“ ersetzt, so daß sich als Ergebnis der semantischen Analyse des Satzes (1) der Ausdruck ergibt:

(9) (STEHT-AUF (FINDE-WERT ’?X ’(UND (IST-EIN ?X PYRAMIDE)
(FARBE ?X GRÜN)))
(FINDE-WERT ’?X ’(UND (IST-EIN ?X BLOCK)
(FARBE ?X ROT))))

Das heißt, da die semantische Komponente von *S* auch noch berücksichtigt, daß es sich hier um einen Behauptungssatz handelt, wird dieser Repräsenta-

tion noch ein „BEHAUPT“ vorangestellt, so daß letzten Endes die folgende Repräsentation entsteht:

(1*) (BEHAUPT ’(STEHT-AUF (FINDE-WERT ’?X ’(UND (IST-EIN ?X PYRAMIDE)
(FARBE ?X GRÜN)))
(FINDE-WERT ’?X ’(UND (IST-EIN ?X BLOCK)
(FARBE ?X ROT))))

4. Ich hatte im letzten Abschnitt den Aufbau entsprechender interner Repräsentationen als semantische Analyse bezeichnet. Aber es ist natürlich eine legitime Frage, ob dieser Sprachgebrauch angemessen ist, das heißt, ob es sich bei dem im letzten Abschnitt geschilderten Verfahren tatsächlich um eine Analyse der Bedeutung von Sätzen handelt. Und die Antwort auf diese Frage ist natürlich für unsere Ausgangsfrage von entscheidender Bedeutung, ob neuere NSS – im Gegensatz zu Systemen wie ELIZA – tatsächlich semantische Maschinen sind oder ob es sich auch bei diesen Systemen nur um bloß syntaktische Maschinen im Sinne Dennetts handelt.

In gewisser Weise sind natürlich auch die neueren NSS syntaktische Maschinen. Sie nehmen Zeichenreihen auf und transformieren sie in andere Zeichenreihen – genauso wie Searle dies in seinem Chinese-Room Beispiel eindrücklich beschrieben hat. Aber folgt aus dieser Tatsache, daß sie nicht auch semantische Maschinen sein können? Schließen sich diese beiden Charakterisierungen gegenseitig aus? Kommen wir, um einer Antwort auf diese Fragen ein Stück näher zu kommen, zunächst zurück zu Dennetts Problem, ein System zu konstruieren, das aus den eingehenden Telegraphenmeldungen genau alle Morddrohungen aussondert. Dennett war der Meinung, daß dies nicht möglich sei. Und das Hauptargument für seine These war, daß es unmöglich ist, die semantische Kategorie „In Englisch verfaßte Morddrohung“ syntaktisch zu definieren. Hinter diesem Argument scheint die generelle Überzeugung zu stehen, daß es grundsätzlich unmöglich ist, ein formales Programm zu entwickeln, das für beliebige Sätze einer natürlichen Sprache entscheidet, ob sie dieselbe Bedeutung haben oder nicht. Wenn man sich die im letzten Abschnitt geschilderte Arbeitsweise neuerer NSS ansieht, gewinnt man jedoch den Eindruck, daß diese Systeme genau dies leisten. Denn die syntaktischen und die semantischen Komponenten dieser Systeme arbeiten gerade so zusammen, daß zum Beispiel dem Satz

(10) John schlägt Jack

dieselbe interne Repräsentation zugeordnet wird wie dem entsprechenden passivischen Satz

(10′) Jack wird von John geschlagen.

Nämlich der Ausdruck:

(10*) (BEHAUPT '(UND (IST-EIN EREIGNIS-1 SCHLAGEN)
(AGENT JOHN-1) (PATIENT JACK-1)))

Analog wird dem der Aussage (1) entsprechenden Fragesatz

(11) Steht die grüne Pyramide auf dem roten Block?

eine Repräsentation zugeordnet, die sich nur im Hinblick auf die Repräsentation der Kraft des Satzes, nicht jedoch im Hinblick auf die Repräsentation des propositionalen Gehalts von der Repräsentation (1*) unterscheidet, nämlich der Ausdruck

(11*) (FRAGE '(STEHT-AUF (FINDE-WERT '?X '(UND (IST-EIN ?X PYRAMIDE)
(FARBE ?X GRÜN)))
(FINDE-WERT '?X '(UND (IST-EIN ?X BLOCK)
(FARBE ?X ROT))))

Alle diese Beispiele zeigen, daß neuere NSS im Idealfall in der Lage sind, bedeutungsgleiche von bedeutungsverschiedenen sprachlichen Ausdrücken zu unterscheiden. Denn diese Systeme arbeiten so, daß sie sprachlichen Ausdrücken dann und nur dann dieselbe interne Repräsentation zuordnen, wenn sie dieselbe Bedeutung haben.

An dieser Stelle könnte man jedoch einwenden: Möglicherweise gelingt es neueren NSS, bedeutungsgleiche von bedeutungsverschiedenen sprachlichen Ausdrücken zu unterscheiden, indem sie genau den bedeutungsgleichen Ausdrücken dieselbe interne Repräsentation zuordnen; aber daraus folgt noch nicht, daß sie die Bedeutung sprachlicher Ausdrücke auch wirklich verstehen. Denn in der Tat scheint es doch möglich zu sein, zu entscheiden, ob zwei sprachliche Ausdrücke *A* und *B* dieselbe Bedeutung haben, ohne die Bedeutung von *A* und *B* zu kennen.

Gegen diesen Einwand gibt es meiner Meinung nach zwei Erwiderungen: eine defensive und eine offensive. Die defensive knüpft an die Überlegungen Dennetts an. Dennett hatte das Problem der kognitiven Psychologie so geschildert, daß diese Wissenschaft annehme, das Gehirn könne auf die Bedeutungen eingehender Reize reagieren, während es *de facto* doch nur eine syntaktische Maschine sei. Wenn es aber syntaktische Maschinen gibt, die bedeutungsgleiche von bedeutungsverschiedenen Reizen unterscheiden können, dann gibt es hier gar kein Problem. Denn auf die Bedeutung eines Reizes reagieren können, heißt ja zunächst einmal nicht mehr als in der Lage sein, auf alle Reize mit einer bestimmten Bedeutung und nur auf diese Reize in gleicher Weise zu reagieren. Und diese Fähigkeit hängt offenbar nur von der Fähigkeit ab, bedeutungsgleiche von bedeutungsverschiedenen Reizen zu unterscheiden.

Ich halte diesen Punkt für äußerst wichtig. Aber vielleicht erscheint er manchem doch nicht ausreichend. Kommen wir also zu der offensiveren

Erwiderung. Diese Erwiderung geht von der Gegenfrage aus: Wenn die Unterscheidung bedeutungsgleicher von bedeutungsverschiedenen Ausdrücken nicht ausreicht, was mehr muß ein System können, um die Bedeutung eines Ausdrucks wirklich zu erfassen? Die naheliegende Antwort ist: es muß wissen, auf was sich die sprachlichen Ausdrücke beziehen, und das heißt wissen, welcher Gegenstand von einem definiten Gegenstandsbezeichner bezeichnet wird, wissen, auf welche Eigenschaft sich zum Beispiel das Adjektiv „grün" bezieht, wissen, unter welchen Bedingungen Sätze wahr sind, usw. Ich halte diese Antwort nicht nur für naheliegend, sondern auch für richtig. Die Frage ist nur, was hier mit „wissen" gemeint ist.

Nehmen wir als besonders zentralen Punkt das Wissen um die Wahrheitsbedingungen eines Satzes. Unter welchen Bedingungen kann man von einem System zu Recht sagen, daß es die Wahrheitsbedingungen eines Satzes kennt? Was heißt es, zu wissen, unter welchen Bedingungen ein Satz *S* wahr ist?

Meiner Meinung nach sind in diesem Zusammenhang zwei Fähigkeiten entscheidend: nämlich erstens die Fähigkeit, Situationen, in denen *S* wahr ist, von Situation zu unterscheiden, in denen das nicht der Fall ist, und zweitens die Fähigkeit, den Satz *S* systematisch mit Situationen der ersten Art in Verbindung zu bringen. Wenn das so ist, dann folgt daraus aber – und besonders der erste Punkt macht dies deutlich –, daß für die Fähigkeit, Sprache zu verstehen, nicht nur die Sprachkomponenten, sondern besonders auch die Wahrnehmungskomponenten entscheidend sind. Denn um zum Beispiel Situationen, in denen der Satz „Die grüne Pyramide steht auf dem roten Block" wahr ist, von solchen Situationen unterscheiden zu können, in denen das nicht der Fall ist, muß das System feststellen können, ob es in seiner Umgebung genau eine grüne Pyramide und genau einen roten Block gibt und ob der erste Gegenstand auf dem zweiten steht. Dazu benötigt es jedoch eine Wahrnehmungskomponente, die komplex genug ist, um folgendes leisten zu können: Sie muß erstens in der Lage sein, Einzelgegenstände in der Umgebung des Systems zu identifizieren (das heißt, sie vom Umgebungshintergrund und von anderen Einzeldingen abzugrenzen), sie muß zweitens in der Lage sein, die identifizierten Gegenstände ihrer Art und ihrer Farbe entsprechend zu klassifizieren (also zum Beispiel Pyramiden von nicht-Pyramiden und grüne von nicht-grünen Gegenständen zu unterscheiden) und sie muß drittens in der Lage sein, einfache räumliche Relationen, die zwischen den identifizierten Gegenständen in ihrer Umgebung bestehen, zu erkennen (also zum Beispiel, daß ein Gegenstand auf einem anderen steht).

Wenn wir annehmen, daß ein System mit der in der Abb. 2 gezeigten Szene konfrontiert ist, reicht es also nicht aus, daß dieses System über eine Kamera verfügt, die auf diese Szene gerichtet ist. Denn durch die Kamera

wird im System zunächst nur ein Bild der Szene erzeugt, das zum Beispiel in einer Grauwertmatrix zwischengespeichert wird. Dieser Vorgang selbst kann sicher noch nicht als Wahrnehmung bezeichnet werden. Wahrnehmung beginnt vielmehr erst, wenn durch Verarbeitung der in der Grauwertmatrix gespeicherten Daten ein internes Modell der Szene aufgebaut wird. Dabei muß die Verarbeitung, wie gerade schon erwähnt, folgendes leisten: Sie muß erstens als Ergebnis liefern, daß an der Szene drei verschiedene Gegenstände beteiligt sind (OBJEKT-1, OBJEKT-2 und OBJEKT-3), sie muß zweitens die Gegenstände der Art und der Farbe nach richtig klassifizieren und sie muß schließlich zu dem Ergebnis führen, daß der erste Gegenstand auf dem zweiten steht (aber zum Beispiel nicht der zweite auf dem dritten). Das heißt, die Verarbeitung muß letzten Endes zum Aufbau eines internen Modells dieser Szene führen, das zum Beispiel die Form der im Abschnitt 3. angeführten propositionalen Listen haben könnte.[7]

Wenn unser System über eine Wahrnehmungskomponente verfügt, die in der Lage ist, solch ein internes Modell der Umgebung des Systems aufzubauen, dann sieht man aber sofort, wie – auf dem Wege über dieses Modell – die sprachlichen und die Wahrnehmungskomponenten des Systems so zusammenarbeiten können, wie es für wirkliches Sprachverstehen nötig ist. Dies wird zum Beispiel deutlich, wenn wir noch einmal den Fragesatz

(11) Steht die grüne Pyramide auf dem roten Block?

betrachten, zu dem das System, wie schon gesagt, die interne Repräsentation

(11*) (FRAGE '(STEHT-AUF (FINDE-WERT '?X '(UND (IST-EIN ?X PYRAMIDE)
(FARBE ?X GRÜN)))
(FINDE-WERT '?X '(UND (IST-EIN ?X BLOCK)
(FARBE ?X ROT))))

aufbaut. Denn diese Repräsentation wird von dem System als Aufforderung interpretiert, auf der Grundlage der gegebenen Datenbasis den Ausdruck

(9) (STEHT-AUF (FINDE-WERT '?X '(UND (IST-EIN ?X PYRAMIDE)
(FARBE ?X GRÜN)))
(FINDE-WERT '?X '(UND (IST-EIN ?X BLOCK)
(FARBE ?X ROT))))

zu beweisen. Dazu wertet es zunächst die beiden Ausdrücke

---

[7] Wahrnehmungskomponenten, die dies leisten, konnten zwar bisher noch nicht so weit entwickelt werden wie die entsprechenden sprachlichen Komponenten. Aber die Grundzüge der Arbeitsweise visueller Systeme sind doch so weit bekannt, daß zumindest im Prinzip klar ist, wie solche Systeme konstruiert werden können. (Vgl. etwa Charniak/McDermott 1985, ch. 3)

(2) (FINDE-WERT '?X '(UND (IST-EIN ?X PYRAMIDE) (FARBE ?X GRÜN)))

und

(8) (FINDE-WERT '?X '(UND (IST-EIN ?X BLOCK) (FARBE ?X ROT))))

aus, das heißt, das System untersucht seine Datenbasis daraufhin, ob es genau ein Objekt gibt, das sowohl den Ausdruck „(IST-EIN ?X PYRAMIDE)" als auch den Ausdruck „(FARBE ?X GRÜN)" erfüllt, bzw. genau ein Objekt, das sowohl den Ausdruck „(IST-EIN ?X BLOCK)" als auch den Ausdruck „(FARBE ?X ROT)" erfüllt. Im ersten Fall findet es das OBJEKT-2 und im zweiten Fall das OBJEKT-1 und ersetzt daher im Ausdruck (9) die Teilausdrücke (2) und (8) durch die internen Namen dieser Objekte, wodurch der Ausdruck

(9′) (STEHT-AUF OBJEKT-2 OBJEKT-1)

entsteht. Im nächsten Schritt versucht das System nun, diesen Ausdruck zu beweisen, was sofort gelingt, da dieser Ausdruck Teil der Datenbasis ist. Das Ergebnis der Gesamtoperation ist also positiv. Und deshalb antwortet das System mit „Ja". Falls das Ergebnis negativ gewesen wäre, hätte es mit „Nein" geantwortet. Das Zusammenspiel von sprachlichen und Wahrnehmungskomponenten ist somit offensichtlich. Die sprachlichen Komponenten bewirken, daß das System auf die Frage (11) genau dann mit „Ja" antwortet, wenn es den Ausdruck (9′) in der Datenbasis findet. Und die Wahrnehmungskomponente bewirkt, daß dieser Ausdruck genau dann in die Datenbasis aufgenommen wird, wenn in der Szene, auf die die Kamera des Systems gerichtet ist, die grüne Pyramide tatsächlich auf dem roten Block steht. Letzten Endes antwortet das System auf die Frage (11) also genau dann mit „Ja", wenn die Wahrheitsbedingungen des propositionalen Teils dieses Satzes erfüllt sind. Das heißt, es kann nicht nur Situationen, in denen der entsprechende Aussagesatz (1) wahr ist, von Situationen unterscheiden, in denen das nicht der Fall ist, es bringt diese Situationen auch in der richtigen Weise mit dem Satz (11) in Verbindung. Damit aber sind, denke ich, alle Voraussetzungen erfüllt, um mit Recht sagen zu können, daß es diesen Satz wirklich verstanden hat. Systeme dieser Art sind also, wie mir scheint, wirklich semantische Maschinen.

## Literatur

Barr, A. & E. A. Feigenbaum, (1981) *Handbook of Artificial Intelligence* (3 Bde.). Los Altos, Cal.: William Kaufmann.

Boden, M., (1977) *Artificial Intelligence and Natural Man*. Hassocks, Sussex: Harvester Press.

Charniak, E. & D. McDermott, (1985) *Introduction to Artificial Intelligence*. Reading, Mass.: Addison-Wesley.

Dennett, D., (1981) „Three Kinds of Intentional Psychology“, in R. Healey (Hg.), *Reduction, Time and Reality*. Cambridge: Cambridge University Press, 37–61. Wiederabgedruckt in D. Dennett, *The Intentional Stance*. Cambridge MA: MIT-Press 1989, 43–68; zitiert nach dem Wiederabdruck.

Weizenbaum, J., (1976) *Computer Power and Human Reason*. San Francisco: W.H. Freeman and Company. (Dt. Übers. Frankfurt/M.: Suhrkamp 1977)

# 7

## Der Computer – ein Modell des Geistes?*

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

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

Für Descartes waren Erklärungen durch Formen oder Entelechien jedoch wissenschaftlich unbefriedigend (um das mindeste zu sagen). Und mit dieser Einschätzung hatte er sicher recht. Seiner Meinung nach waren solche Erklärungen aber auch unnötig. Denn auf der Grundlage der zu Beginn der Neuzeit entstehenden neuen Naturwissenschaft war es ihm zufolge durchaus möglich, auch die für Lebewesen charakteristischen Vorgänge rein mechanisch zu erklären. D.h., Descartes war der Meinung, daß sich diese Vorgänge mit der gleichen Notwendigkeit aus den Teilen des menschlichen und tierischen Körpers ergeben, „wie der Mechanismus einer Uhr aus der

---

* Erstveröffentlichung in: S. Krämer (Hg.) *Geist – Gehirn – Künstliche Intelligenz*: *Zeitgenössische Modelle des Denkens*. Berlin/New York: de Gruyter 1994, 71–87.

[1] Vgl. zu diesem Abschnitt Beckermann 1989.

Kraft, Lage und Gestalt ihrer Gewichte und Räder folgt" (*Discours*, 81 ff.). Und entsprechend versucht er, im *Traité de l'homme* und *La description du corps humain* für den Herzschlag, für die Ernährung, für die Wahrnehmung, für das Gedächtnis und schließlich sogar für die Fortpflanzung mechanische Erklärungen zu liefern, wobei er sich hauptsächlich an drei Modellen orientiert: am Modell der Uhr, deren Verhalten vollständig durch das mechanische Zusammenwirken ihrer Gewichte und Räder bestimmt wird; am Modell der Orgel, bei der Register und Tastenanschlag das Öffnen und Schließen der einzelnen Orgelpfeifen bewirken, und schließlich auch an den komplizierten hydraulischen Steuerungssystemen, mit denen die Gartenbaumeister seiner Zeit viele kleine Gartenfiguren zu einer Art von künstlichem Leben zu erwecken verstanden.[2]

Auch für Descartes gibt es jedoch eine *prinzipielle* Grenze für die mechanische Erklärbarkeit der Fähigkeiten von Lebewesen. Und diese Grenze liegt für ihn da, wo beim Menschen die Fähigkeiten des Denkens und Sprechens ins Spiel kommen. Im Teil V des *Discours de la méthode* erklärt Descartes ausdrücklich, daß sich Menschen seiner Meinung nach in zwei Punkten grundsätzlich von jeder Maschine, d. h. von jedem mechanischen System unterscheiden. Erstens nämlich könnten solche mechanischen Systeme „niemals Worte oder andere Zeichen dadurch gebrauchen, daß sie sie zusammenstellen, wie wir es tun, um anderen unsere Gedanken mitzuteilen". Und zweitens würden solche Systeme, auch wenn sie in einigen Punkten sehr gute Leistungen vollbrächten, „doch zweifellos bei vielem anderen versagen, wodurch offen zutage tritt, daß sie nicht aus Einsicht handeln, sondern nur aufgrund der Einrichtung ihrer Organe. Denn die Vernunft ist ein *Universalinstrument*, das bei allen Gelegenheiten zur Verfügung steht, während diese Organe für jede besondere Handlung einer besonderen Einrichtung bedürfen …" (*Discours*, S. 92 f.– Hervorh. vom Verf.).

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

Schon Descartes' unmittelbare Nachfolger sahen in dieser Position jedoch eher eine Herausforderung. War es nicht vielleicht doch möglich, den ganzen Menschen mit all seinen Fähigkeiten mechanisch zu erklären? Besonders die Philosophen der französischen Aufklärung glaubten an die

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

Möglichkeit einer positiven Antwort auf diese Frage. Der Titel des Buches *L'homme machine* (1748) von J.O. de La Mettrie spricht da eine deutliche Sprache. Alle Versuche, Descartes' Erklärungsansatz über dessen eigene Überlegungen hinauszutreiben, mußten jedoch notwendig im bloß Proklamatorischen steckenbleiben, solange es kein *Modell* gab, mit dessen Hilfe sich plausibel machen ließ, daß auch der universale Charakter der Vernunft durch ein rein mechanisches System realisiert sein kann.

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

2. Das Konzept des Computers, so wie wir ihn heute kennen und wie er inzwischen auf fast jedem Schreibtisch steht, geht auf verschiedene Wurzeln zurück. Aber die wichtigste dieser Wurzeln liegt sicher in den bahnbrechenden Arbeiten des englischen Mathematikers Alan Turing.[3] Worin bestand Turings große Leistung? Wenn man es auf einen kurzen Nenner bringen will, kann man vielleicht sagen: Erstens in dem Nachweis, daß es zu jeder arithmetischen Funktion, die überhaupt berechenbar ist, eine Maschine gibt, die diese Funktion berechnet, und zweitens in dem weit über dieses erste Ergebnis hinausgehenden Nachweis, daß es eine *universelle* Maschine gibt, die den Wert jeder beliebigen berechenbaren Funktion für jedes beliebige Argument berechnet.

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

Die Ergebnisse der logischen Grundlagenforschung sind in diesem Zusammenhang deshalb von Bedeutung, weil sie zeigen, daß man logisches Schließen – ebenso wie das numerische Rechnen – als rein formale Veränderung von (strukturierten) Zeichenreihen durchführen kann. Der formale

[3] Besonders Turing 1936/37.

Charakter der Logik war zwar schon lange vor diesen Ergebnissen bekannt. Aber Anfang der dreißiger Jahre dieses Jahrhunderts konnte Kurt Gödel zum ersten Mal zeigen,[4] daß die Prädikatenlogik 1. Stufe vollständig kalkülisierbar ist, d.h., daß es für die Prädikatenlogik 1. Stufe Kalküle K gibt, für die gilt:

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

Auf der Grundlage dieser Ergebnisse ist es nicht schwer nachzuweisen, daß es Algorithmen gibt, die für jede allgemeingültige Formel nach endlich vielen Schritten zu dem Ergebnis „allgemeingültig" führen. Im Falle nichtallgemeingültiger Formeln haben diese Algorithmen zwar den Nachteil, in einigen Fällen nie zu einem Ende zu kommen; aber trotz dieses Nachteils schien die Existenz solcher Algorithmen die Vermutung zu stützen, daß man auch das logische Schließen auf einer Maschine realisieren kann.

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

Fast zur gleichen Zeit entwickelte der Mathematiker J.A. Robinson mit seinem Resolutionsalgorithmus[7] aber eine neue und außerordentlich effektive Methode zum Beweis prädikatenlogischer Formeln. Und da sich dieser Algorithmus leicht auf einem Computer implementieren ließ, schien sich in

---

[4] Gödel 1930.
[5] Newell/Shaw/Simon 1960 und Newell/Simon 1963.
[6] Vgl. Schefe 1986, 33.
[7] Robinson 1965.

ihm der Traum eines universalen, d.h. *bereichsunabhängigen* Problemlösers nun doch zu erfüllen. Voraussetzung dafür war allerdings, daß sich zeigen ließ, daß jedes Problem auf die Ableitung einer prädikatenlogischen Formel aus einer Menge von Prämissen reduzierbar ist. Und diese Voraussetzung bildet tatsächlich die Prämisse dessen, was man das ‚Logizistische Programm' in der KI-Forschung nennen könnte. Die Gründe, die für dieses Programm sprachen, lagen auf der einen Seite natürlich in den Erfolgen, die man auf der Grundlage der genannten Prämisse erzielen konnte. Auf der anderen Seite sicher aber auch in der Einfachheit und Allgemeinheit dieser Prämisse. Denn wenn jede Problemlösung auf die Ableitung einer Formel aus einer Menge von Prämissen reduzierbar ist, dann stellt jedes System mit der Fähigkeit zum automatischen Beweisen – also z.B. jeder Computer, der nach dem Muster des Robinsonschen Resolutionsalgorithmus programmiert ist – tatsächlich eine universale Problemlösungsmaschine im Sinne Descartes' dar. Unter dieser Voraussetzung kann das Problem, zu verstehen, wie die Fähigkeit zu denken in einem rein mechanischen System realisiert sein kann, daher als zumindest im Prinzip gelöst gelten. D.h., unter dieser Voraussetzung ist zwar nicht der Computer als solcher, wohl aber *jedes System mit der Fähigkeit zum automatischen Beweisen ein mögliches Modell des Geistes*.

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

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

„Kein Problem", sagen die Konstrukteure von $R_1$. „Unser nächster Roboter muß nicht nur die beabsichtigten Wirkungen seiner Handlungen berücksichtigen, sondern auch ihre nicht beabsichtigten Nebeneffekte. D.h. er

muß in der Lage sein, alle Effekte abzuleiten, die eine Handlung in einer gegebenen Situation hervorruft“. Sie nennen ihr nächstes Modell einen „robot-deducer“, kurz $R_1D_1$, und bringen es in dieselbe Situation, in der $R_1$ vorher so schrecklich gescheitert war. Auch $R_1D_1$ kommt bald auf die Idee, daß die Handlung ZIEHE_HERAUS(WAGEN, RAUM) sein Problem lösen könnte; aber bevor er sich daran machen kann, diese Handlung auszuführen, muß er zunächst noch ableiten, welche anderen Effekte die Ausführung dieser Handlung mit sich bringen würde. Dafür benötigt er eine Menge Zeit. Und als er sich gerade anschickt, zu beweisen, daß die von ihm ins Auge gefaßte Handlung nicht die Farbe der Wände des Raumes ändern würde, ist es zu spät – wieder explodiert die Bombe.

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

Obwohl das Frame-Problem seit über 20 Jahren bekannt ist,[8] gibt es – besonders in der Diskussion zwischen KI-Forschern und Philosophen – bis heute keine Einigkeit über den genauen Gehalt dieses Problems. Es ist nicht klar, worin das Problem überhaupt besteht; es ist nicht klar, ob und, wenn ja, wie man es lösen kann; und es ist nicht klar, was aus der Existenz bzw. der Unlösbarkeit dieses Problems eigentlich folgt. Klar ist, daß das Frame-Problem nur in einem bestimmten Rahmen entsteht, nämlich dann, wenn man versucht, Programme zu schreiben, die es einem Computer oder Roboter ermöglichen, die Veränderungen zu modellieren, die bestimmte Handlungen oder Ereignisse in der Welt hervorrufen. Das Problem, über das McCarthy und Hayes 1969 zum ersten Mal in der Literatur berichteten und dem sie damals den Namen „Frame-Problem“ gaben, ergab sich jedoch nicht aus dieser Aufgabenstellung selbst, sondern aus der besonderen Art,

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

in der sie diese Aufgabe zu lösen versuchten.[9] Sie faßten die verschiedenen möglichen Zustände, die die Welt annehmen kann, als Situationen auf und deuteten Ereignisse und Handlungen als Funktionen von Situationen in Situationen. Eine Situation selbst war dabei die Menge aller der Fakten, die in dieser Situation wahr sind. Zur Berechnung der Folgesituation *s′*, die sich aus der Situation *s* ergibt, wenn in ihr das Ereignis *e* stattfindet, verwendeten McCarthy und Hayes eine Reihe von Axiomen wie z. B.

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

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

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

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

KI-Forscher bestehen häufig darauf, dies und nur dies sei das Frame-Problem, und sie betonen dann weiter, daß es für dieses Problem eine ganze Reihe von Lösungsvorschlägen gebe. Einer dieser Vorschläge ist das Programm STRIPS (Fikes und Nilsson, 1971), das zur Lösung des ursprünglichen Frame-Problems eine Strategie verwendet, die John Haugeland die „Strategie der schlafenden Hunde" genannt hat. Das Grundprinzip dieser Strategie ist ebenso einfach wie effektiv. STRIPS behandelt jede Situation als eine eigene Datenstruktur. Um herauszufinden, wie die Nachfolgesituation *s′* aussieht, die durch die Ausführung der Handlung *e* in der Ausgangssituation *s* entsteht, berechnet STRIPS die Veränderungen, die *e* in *s* bewirkt,

[9] Vgl. zum folgenden McDermott 1987.

und nimmt die entsprechenden Änderungen in der Datenstruktur vor. Alles andere wird einfach so gelassen, wie es ist. Zusätzliche Frame-Axiome, mit deren Hilfe abgeleitet werden kann, was sich nicht ändert, sind daher überflüssig.

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

Die Argumentation, die hinter dieser Auffassung steht, läßt sich kurz so zusammenfassen. Selbst wenn es – wie etwa bei dem Programm STRIPS – nicht mehr nötig ist, abzuleiten, was sich bei der Ausführung einer Handlung nicht ändert, ist das entscheidende Problem noch nicht gelöst. Denn auch die Zahl der Veränderungen, die eine Handlung bewirkt, kann schon sehr groß sein. Und wenn das so ist, dann wird nicht nur bei dem Versuch, herauszufinden, was sich bei der Ausführung einer Handlung alles *nicht* verändert (dem von Drew McDermott so genannten „inertia problem“[10]), zu viel Zeit verschwendet. Dann kostet auch schon der Versuch zu berechnen, was sich alles *verändert*, so viel Zeit, daß die Lösung eines Problems nicht im Rahmen der zur Verfügung stehenden Zeit gefunden werden kann. Intelligentes Handeln, so diese Philosophen, setzt voraus, daß ein Problem nicht nur gelöst, sondern daß es in einer angemessenen Zeit gelöst wird. Und dies wiederum ist nur möglich, wenn bei dem Versuch, das Problem zu lösen, nicht alle, sondern nur die *relevanten* Folgen einer Handlung in Betracht gezogen werden.[11]

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

> Once one has developed some suitable representation of the world about which the reasoner is expected to reason, one needs also to arrange that the system performs deductions which are appropriate for its assigned tasks and doesn’t get lost in clouds of valid, but irrelevant conclusions. (It is fairly easy to arrange that it doesn’t generate invalid conclusions.) This is variously called

[10] McDermott 1987, 117.
[11] Vgl. z. B. Pylyshyn 1987a, x.

> the theorem-proving problem, or the control problem, or the search problem, in AI. This is *not* the frame problem either. (1987, 124 – Hervorh. vom Verf.)

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

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

Was folgt aus dem Frame-Problem für die Beantwortung dieser Frage? Zum einen sicher, daß Universalität nicht alles ist. Intelligente Problemlö-

sungssysteme müssen nicht nur bereichsunabhängig, sie müssen auch schnell genug sein, d.h. sie müssen die Fähigkeit besitzen, Probleme nicht nur zu lösen, sondern in einer den Problemen angemessenen Zeit zu lösen. Kann das mit den herkömmlichen Methoden der KI-Forschung geleistet werden? Die Antwort auf diese Frage hängt natürlich davon ab, wie man die Aussichten einer weiteren Verbesserung dieser Methoden einschätzt. Aber sie hängt auch ab von einer Einschätzung der Grundlage, auf der alle diese Methoden aufbauen. Dennett hat dazu eine interessante Bemerkung gemacht:

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

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

(3) A-Stadt liegt 100 km nördlich von B-Stadt.

(4) A-Stadt liegt 200 km nordwestlich von C-Stadt.

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

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

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

genau deshalb benötigt man deduktive Verfahren, um sie explizit zu machen.

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

A-Stadt

B-Stadt

C-Stadt

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

Wenn Haugeland recht hat, dann scheint das Frame-Problem nichts weiter zu sein als ein Artefakt, das sich allein aus der Annahme ergibt, daß Repräsentationen quasi-sprachlichen Charakter haben müssen. Diese An-

[12] Haugeland 1987, 88 ff.

nahme, so Haugeland, bildet aber das Fundament der gesamten „klassischen“ KI-Forschung. Und daraus scheint zwingend zu folgen, daß Systeme, die auf den herkömmlichen Methoden der KI-Forschung beruhen, keine adäquaten Modelle des Geistes sein können, da es auf der Grundlage dieser Methoden keine Lösung für das Frame-Problem gibt.

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

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

[13] Descartes 1969, 46.

nes Programms scheitern. Denn die Erklärung physiologischer Vorgänge kann nur auf der Basis einer ausgearbeiteten Chemie erfolgen. Und die stand Descartes noch nicht zur Verfügung.

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

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

## Literatur

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Pylyshyn, Z. W. (1987a) „Preface“, in: Pylyshyn 1987, vii-xi.

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

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

Specht, R. (1966) *Descartes*. Reinbek bei Hamburg: Rowohlt.

Turing, A. (1936/37) „On Computable Numbers with an Application to the *Entscheidungsproblem*“. *Proceedings of the London Mathematical Society* 42, 230–265, und 43, 544–546.

– (1950) „Computing Machinery and Intelligence“. *Mind* 59, 433–460. Wiederabgedr. in: Boden 1990, 40–66.

# 8

## Ist eine Sprache des Geistes möglich?*

1.

Kognitionswissenschaften – in einem weiten Sinn – sind einfach alle die Wissenschaften, die sich mit der Analyse und Erklärung kognitiver Leistungen und Fähigkeiten befassen. Wenn man jedoch von *der* Kognitionswissenschaft im Singular spricht, dann ist in der Regel mehr gemeint. Für die Kognitionswissenschaft ist nicht nur ein bestimmter Forschungsgegenstand charakteristisch, sondern auch ein bestimmter Erklärungsansatz: der Informationsverarbeitungsansatz. Stillings et al. z. B. schreiben gleich auf der ersten Seite ihres 1987 erschienenen Buches *Cognitive Science – An Introduction*: „Cognitive scientists view the human mind as a complex system that receives, stores, retrieves, transforms, and transmits information." (Stillings et al. 1987: 1) Der Informationsverarbeitungsansatz führt jedoch sofort weiter zum Symbolverarbeitungsansatz. Denn offenbar kann ein System nur dann Informationen empfangen, speichern und verarbeiten, wenn es über ein System von internen Repräsentationen oder Symbolen verfügt, über eine interne Sprache, in der diese Informationen codiert sind. Zumindest ist das eine naheliegende Überlegung, die Peter Hacker so formuliert hat: „… if information is received, encoded, decoded, interpreted and provides grounds for making plans, then there must be a language or system of representation in which this is all done." (Hacker 1987: 486 f.) In der Tat ist die Annahme, daß es in kognitiven Systemen so etwas wie ein System interner Repräsentationen bzw. eine Sprache des Geistes[1] gibt, die zentrale Grundannahme vieler neuerer Arbeiten in den Bereichen der Kognitionspsychologie und der kognitiven Neurobiologie. Für diese Wissenschaften hat diese Annahme den Status einer empirischen Hypothese, d. h. für sie sind interne Repräsentationen oder Symbole theoretische Konstrukte, die deshalb postuliert werden, weil sie gut bestätigte und systematisch

---

* Erstveröffentlichung in: A. Burri (Hg.) *Sprache und Denken*. Berlin/New York: Walter de Gruyter 1997, 75–92.

[1] Der Ausdruck „Sprache des Geistes" („lingua mentis", „language of thought"), der in diesem Zusammenhang möglicherweise zum ersten Mal von G. Harman (1973) verwendet wurde, ist der Sache nach außerordentlich irreführend. Denn die Ausdrücke der Sprache des Geistes sind auch Fodor zufolge interne *physische* Systemzustände – also z. B. bestimmte neuronale Feuerungsmuster oder Bitmuster im Speicher eines Computers. Treffender wären deshalb Ausdrücke wie „Sprache des Gehirns" oder „Sprache des Computers".

besonders befriedigende Erklärungen kognitiver Leistungen ermöglichen. In der Philosophie gibt es jedoch auch Ansätze, die Annahme einer Sprache des Geistes durch sehr grundsätzliche Überlegungen zur Natur mentaler Zustände zu stützen.

Der Hauptmatador in diesem Feld ist sicher Jerry Fodor, der seine *Repräsentationale Theorie des Geistes* (RTG) über viele Jahre hinweg entwickelte,[2] bis er ihr in dem Buch *Psychosemantics* ihre sozusagen kanonische Form gegeben hat. In dieser Form umfaßt die RTG zunächst zwei Teilthesen:

> (1) Für jeden Organismus *O* und jeden Typ *A* intentionaler Zustände gibt es eine (funktionale/computationale) Relation *R*, so daß gilt:
> *O* ist genau dann in einem intentionalen Zustand des Typs *A* mit dem Inhalt *p*, wenn sich *O* in der Relation *R* zu einer mentalen Repräsentation *m* befindet und *m* die Bedeutung *p* hat.
>
> (2) Mentale Prozesse sind kausale Abfolgen einzelner mentaler Repräsentationen. („Mental processes are causal sequences of tokenings of mental representations"). (Fodor 1987: 17)

Während die zweite dieser beiden Thesen relativ klar ist, ist die erste vielleicht nicht ohne weiteres verständlich. Was also ist mit dieser These gemeint?

Wichtig ist zunächst, daß es Fodor bei dieser These um eine Antwort auf die Frage geht, wie bestimmte mentale Zustände – nämlich *intentionale* Zustände – *physisch* realisiert sein können. Intentionale Zustände – wie Wünsche, Überzeugungen, Hoffnungen und Befürchtungen – sind dadurch gekennzeichnet, daß sie *auf etwas gerichtet* sind, daß sie einen *Inhalt* haben. Man glaubt, *daß etwas der Fall ist*, man wünscht sich *einen bestimmten Gegenstand*, man hofft oder befürchtet, *daß ein bestimmtes Ereignis eintreten wird*, usw. Bei allen intentionalen Zuständen kann man also zwei Aspekte unterscheiden: die Art des Zustandes und seinen Inhalt. Mein Wunsch, ein neues Fahrrad zu erwerben, und mein Wunsch, einen alten Freund wiederzutreffen, sind intentionale Zustände derselben Art: beides sind Wünsche – allerdings Wünsche mit verschiedenen Inhalten. Meine Befürchtung, daß es heute regnen wird, und meine Überzeugung, daß es heute regnen wird, sind dagegen intentionale Zustände verschiedener Art. Aber auch sie haben etwas gemeinsam; sie haben denselben Inhalt: sie richten sich beide auf die Proposition, daß es heute regnen wird.

Fodors These ist nun, daß diesen beiden Aspekten intentionaler Zustände auch verschiedene Aspekte ihrer physischen Realisierungen entsprechen: dem Inhalt eine mentale Repräsentation und dem Zustandstyp eine be-

---

[2] Vgl. bes. Fodor 1975; 1978; 1981b; 1987.

stimmte computationale bzw. funktionale Relation. Mentale Repräsentationen sind dabei als *innere physische Strukturen* zu verstehen, die *etwas repräsentieren* und die insofern ähnlich wie die Sätze einer Sprache oder Landkarten oder die Kerben auf dem Griff des Revolvers eines Westernhelden *eine Bedeutung haben*. Konkret könnte man dabei z. B. an Listen-Strukturen denken wie sie etwa in der Programmiersprache LISP gebräuchlich sind. Daß ein System zu einer mentalen Repräsentation in einer bestimmten funktionalen oder computationalen Relation steht, bedeutet andererseits, daß diese Repräsentation in dem System (z. B. bei der Hervorbringung des Verhaltens des Systems) eine bestimmte funktionale oder computationale Rolle spielt. Hier kann man sich z. B. vorstellen, daß sich mentale Repräsentationen in verschiedenen Boxen (verschiedenen Speicherbereichen) – etwa einer *belief*- oder einer *desire*-Box – befinden und daß das System Repräsentationen, die sich in verschiedenen Boxen befinden, unterschiedlich verarbeitet. Auf jeden Fall gilt Fodor zufolge, daß ein System dann und nur dann z. B. die Überzeugung hat, daß es heute regnen wird, wenn es in dem System eine mentale Repräsentation, d. h. eine physische Struktur gibt, die die Bedeutung hat, daß es heute regnen wird, und wenn diese Repräsentation in dem System die für Überzeugungen charakteristische funktionale bzw. computationale Rolle spielt.

Mit den Thesen (1) und (2) ist Fodors RTG jedoch noch nicht vollständig charakterisiert. Denn eine zusätzliche – und wahrscheinlich sogar entscheidende – Annahme ist für ihn, daß mentale Repräsentationen eine satzartige innere Struktur besitzen. Zu den Thesen (1) und (2) kommt daher noch die eigentliche *Language of Thought*-These hinzu:

(3a) Mentale Repräsentationen sind *strukturiert*; sie haben typischerweise eine Konstituentenstruktur.

(3b) Die Teile dieser Strukturen sind „transportierbar“; dieselben Teile können in verschiedenen Repräsentationen auftreten.

(3c) Mentale Repräsentationen haben eine *kombinatorische Semantik*: Ihre Bedeutung hängt in regelhafter Weise von der Bedeutung ihrer Teile ab.

Ich kann an dieser Stelle nicht auf alle Gründe eingehen, die nach Fodor für die Richtigkeit der drei Thesen der RTG sprechen; aber ich will ein Hauptargument wenigstens andeuten.[3] Es ist schon deutlich geworden, daß Fodor diese Thesen im Zusammenhang einer Analyse intentionaler Zustände entwickelt hat. Genauer kann man sagen, daß Fodors Argumentation für diese Thesen von zwei Prämissen ausgeht:

[3] Ausführlicher werden die Fodorschen Argumente in Beckermann 1991 erläutert.

(4) Es gibt intentionale Zustände, d.h. Zustände, die zugleich kausal wirksam und semantisch evaluierbar sind.

(5) Um kausal wirksam sein zu können, müssen intentionale Zustände physisch realisiert sein.

Wenn man diese Prämissen akzeptiert, steht man jedoch sofort vor einem Problem: Intentionale Zustände haben eine ganze Reihe von Merkmalen, die einer physischen Realisierbarkeit auf den ersten Blick im Wege zu stehen scheinen. Das einzige dieser Merkmale, auf das ich hier eingehen will, besteht darin, daß Kausalbeziehungen zwischen intentionalen Zuständen häufig Rationalitätsprinzipien bzw. semantischen Relationen zwischen ihren Inhalten entsprechen. Wenn jemand *p* glaubt, dann wird er in der Regel auch alle offensichtlichen Folgerungen aus *p* glauben. Und wenn jemand *q* will und glaubt, daß *p* eine notwendige Voraussetzung zur Erreichung von *q* ist, dann wird er in der Regel auch *p* wollen.

Damit stellt sich jedoch die Frage, wie denn physische *Mechanismen* aussehen können, die solchen Kausalbeziehungen zugrundeliegen, d.h. wie physische Mechanismen aussehen können, denen ihrerseits Rationalitätsprinzipien entsprechen. Und diese Frage führt für Fodor direkt zum Symbolverarbeitungsansatz. Denn auf der einen Seite hat die Beweistheorie, auf die sich Fodor immer wieder bezieht, gezeigt, daß der Begriff der logischen Folgerung formalisiert werden kann, d.h. daß man diesen Begriff durch Bezugnahme auf geeignete Kalküle auch rein syntaktisch charakterisieren kann. Und auf der anderen Seite haben die Computerwissenschaften gezeigt, daß solche syntaktischen Umformungsprozesse mit Hilfe von Symbolverarbeitungsprozessen physisch realisiert werden können. Mit anderen Worten: Physische Mechanismen, die Rationalitätsanforderungen respektieren, lassen sich mit Hilfe von Symbolverbeitungsprozessen realisieren – allerdings nur unter der Voraussetzung, daß die zugrundeliegenden Symbole oder Repräsentationen strukturiert sind. Denn Symbolverarbeitungsprozesse bestehen gerade darin, daß sie Repräsentationen nach Regeln verändern, die auf die Struktur dieser Repräsentationen Bezug nehmen.

Fodors Hauptargument für die Thesen (1) – (3) kann man daher so zusammenfassen: Kausale Mechanismen, die Rationalitätsanforderungen respektieren, können nur auf Symbolverarbeitungsprozessen beruhen und Symbolverarbeitungsprozesse setzen strukturierte Repräsentationen voraus.

## 2.

Die RTG Fodors, die ich im letzten Abschnitt kurz dargestellt und erläutert habe, ist inzwischen von verschiedenen Autoren in sehr unterschiedlicher Weise kritisiert worden. Besonders in Oxford sind in den letzten Jahren aber vom späten Wittgenstein ausgehende kritische Überlegungen laut geworden, die nicht nur diese Theorie, sondern den gesamten Symbolverarbeitungsansatz radikal in Frage stellen.[4] Peter Hacker z. B. stellt die rhetorische Frage: „Is this [scl. the idea that there is a language in the brain] just a picturesque metaphor or helpful analogy? Or is it a symptom of widespread confusion in the presentation, description and explanation of experimental data …?“ (op. cit.: 487). Und seine Antwort lautet in der Tat, daß die Idee eines Symbolsystems im Gehirn bzw. die Idee einer Sprache des Geistes auf einer grundlegenden Begriffsverwirrung beruhe und daher wortwörtlich *sinnlos* sei. Was sind Hackers Gründe für diese niederschmetternde Diagnose?

Zunächst charakterisiert er noch einmal die Idee, die er dann attackieren will:

> The general conception at work involves the supposition that the brain has a *language* of its own, which consists of *symbols* that *represent* things. It uses the *vocabulary* of this language to *encode information* and it produces *descriptions* of what is seen …. (op. cit.: 488)

> A ‚symbolic description‘ is presumably an array of symbols which are so combined as to yield a true (or false) characterization of a certain aspect of the world. It must be cast in a certain language which has a vocabulary and grammar. (ibid.)

Aber was kann es bedeuten, daß das Gehirn über eine Sprache mit eigenem Wortschatz und eigener Grammtik verfügt? Was heißt es überhaupt, daß jemand über eine Sprache verfügt? „Someone who *has* a language has mastered a technique, acquired or possesses a skill of using symbols in accord with rules for their correct use, or – if you prefer – in accord with their meaning.“ (op. cit.: 491 f.)

Wer eine Sprache beherrscht, verfügt also über bestimmte Fähigkeiten. Er versteht in dieser Sprache gemachte Äußerungen; er kennt die Bedeutung der Wörter dieser Sprache und kann diese Wörter verwenden, um selbst die unterschiedlichsten Sprechhandlungen auszuführen: Er kann ein Taxi rufen, nach dem Weg zum Bahnhof fragen, Geschichten und Witze erzählen, Wein zum Essen bestellen, einen Freund vorstellen, eine Landschaft beschreiben, und und und. Außerdem kann er, falls er einmal nicht

---

[4] Siehe besonders Hacker 1987; aber auch den neuen Sammelband Hyman 1991.

verstanden wird, erklären, was die von ihm verwendeten Wörter bedeuten und was er mit ihnen hat sagen wollen. Alles in allem: „If [someone] understands a language he can respond in various ways to others' uses of words and sentences, as well as correcting others' errors, querying their unclarities and equivocations" (op. cit.: 492). Allein aus der Tatsache, daß Sprachbeherrschung alle diese Fähigkeiten impliziert, folgt Hacker zufolge schon zwingend, daß es im Wortsinne keinen Sinn hat, zu sagen, daß Gehirn verfüge über eine Sprache.

> Only of a creature that can perform acts of speech does it make sense to say that it has, understands, uses, a language. But it is literally unintelligible to suggest that a brain, let alone *a part of a brain*, might ask a question, have or express an intention, make a decision, describe a sunset, undertake an obligation, explain what it means, insist, assert, instruct, demand, opine, classify, and so forth. (ibid.)

Um überhaupt über eine Sprache verfügen zu können, muß man fähig sein, bestimmte Handlungen auszuführen – Handlungen, die auf einer ganz anderen Ebene liegen als die, von denen man sinnvollerweise sagen kann, daß sie von einem Gehirn oder gar von Teilen eines Gehirn ausgeführt werden könnten. Gehirne oder Gehirnteile sind daher schon aus begrifflichen Gründen keine möglichen Sprachverwender.

Aber es gibt noch mehr Gründe, die Hacker zufolge zeigen, daß die Idee einer Sprache des Gehirns immer absurder wird, je mehr wir uns über ihre Implikationen klar werden. Die Ausdrücke einer Sprache, so Hacker weiter, haben eine durch Konventionen geregelte Verwendung, und jemand, der über eine Sprache verfügt, muß die korrekte Verwendung dieser Ausdrücke kennen, d. h. er muß korrekte von unkorrekten Verwendungen unterscheiden können. Ein geregelter Sprachgebrauch in diesem Sinne, ein Sprachgebrauch, der sich an Standards der Korrektheit orientiert, kann aber nur in einer sozialen Praxis fundiert sein.

> For only where there is a practice of employing a sign can there also be an activity of matching the application of the sign against a standard of correctness. Since signs have a meaning, a use, only insofar as there is a convention, a standard of correctness for their application, there must be a *possibility* of correcting misuses by reference to the standard of correctness for the use of the expression which is embodied in an explanation of meaning. The use of language is essentially a normative activity. (op. cit.: 496)

Auch aus diesem Grund ist es Hacker zufolge völlig unmöglich, daß Gehirne oder gar Gehirnzellen über eine Sprache verfügen. Denn weder von Gehirnen noch von Gehirnzellen kann man sinnvoll sagen, daß sie Konventionen folgen. Denn Konventionen können nur da befolgt werden, wo überhaupt Konventionen existieren. Und Konventionen kann es nur da geben, wo sie in einer sozialen Gemeinschaft beim Lehren und Lernen, beim Kor-

rigieren von Fehlern und beim Erklären und Rechtfertigen von Handlungen verwendet werden.

> Only of a creature who has the *ability* to make a mistake, who can *recognize* his mistake by reference to a standard, who can *correct* his action for the *reason* that it was erroneous, only of such a creature can one say that it follows and uses conventions. (ibid.)

Aus demselben Grunde ist nach Hacker auch die Rede von cerebralen Karten völlig sinnlos. Denn auch Karten sind nur Karten *von etwas*, wenn es entsprechende Konventionen gibt. Eine bestimmte Gegend kann nur dann mit Hilfe einer Karte repräsentiert werden, wenn bei der Erstellung der Karte spezifische kartographische Konventionen befolgt wurden einschließlich bestimmter konventionell geregelter Projektionsmethoden wie etwa der Mercator-Projektion.

> So there are no representing maps without conventions of representation. There are no conventions of representation without a *use*, by intelligent, symbol-employing creatures, of the representation. And to *use* a representation correctly one must *know* the conventions of representation, understand them, be able to explain them, recognize mistakes and correct or acknowledge them when they are pointed out. Whether a certain array of lines is or is not a map is not an *intrinsic* feature of the lines, nor even a *relational* feature (that is, the *possibility* of a 1:1 mapping), but a *conventional* one (that is, the *actual* employment, by a person, of a convention of mapping). (op. cit.: 497 f.)

Es bleibt also nur die Schlußfolgerung, daß die Idee einer Sprache des Gehirns sinnlos ist. Im Gehirn kann es keine bedeutungshaltigen Symbole geben. Denn Bedeutung setzt das Bestehen von Konventionen voraus, und Konventionen implizieren die Existenz einer entsprechenden sozialen Praxis. Eine solche „soziale Praxis" ist im Hinblick auf Nervenzellen aber *begrifflich* unmöglich. Die Annahme, es gebe im Gehirn ein Symbolsystem oder eine Sprache, ist daher im Wortsinn „unbegreiflich".

## 3.

Auf den ersten Blick scheint diese Argumentation außerordentlich schlüssig. Und in der Tat bildet sie ja auch den – von vielen geteilten – Kern einer Wittgensteinianischen Bedeutungstheorie. Bei näherem Hinsehen ist sie allerdings nicht ganz so zwingend. Denn selbst die Bezugnahme auf eine soziale Praxis kann – zumindest wenn man Kripkes diesbezüglichen Überlegungen folgt[5] – den normativen Charakter von Bedeutung nicht begründen. Auf jeden Fall werden Kripkes Überlegungen von Paul Boghossian so

[5] Siehe Kripke 1982.

gedeutet.[6] Worin besteht, so fragt Boghossian, der für Bedeutung grundlegende normative Charakter? Und seine Antwort lautet:

> Suppose the expression ‚green' means *green*. It follows immediately that the expression ‚green' applies *correctly* only to *these* things (the green ones) and not to *those* (the non-greens). The fact that the expression means something implies, that is, a whole set of *normative* truths about my behaviour with that expression: namely, that my use is correct in application to certain objects and not in application to others. … meaningful expressions possess conditions of *correct* use. (1989: 513)

Genau aus dieser Tatsache ergibt sich das *skeptische Problem* für alle Bedeutungstheorien: „Having a meaning is essentially a matter of possessing a correctness condition. And the sceptical challenge is to explain how anything could possess *that*." (1989: 515) Kripkes Hauptargument gegen alle Theorien, die Bedeutung auf natürliche Eigenschaften einzelner Personen zurückführen wollen, und insbesondere gegen die dispositionale Analyse von Bedeutung lautet dementsprechend: Keine der von diesen Theorien ins Feld geführten natürlichen Eigenschaften kann die Tatsache begründen, daß mit einem Ausdruck Korrektheitsbedingungen verbunden sind; und eben deshalb sind alle diese Theorien als *Bedeutungs*theorien zum Scheitern verurteilt.

Dies ist nun der Punkt, an dem Wittgensteinianer auf einer sozialen Praxis beruhende Regeln ins Spiel bringen und argumentieren: Soweit sei zwar alles richtig; aber es zeige nur, daß Bedeutung nichts sei, was durch die Eigenschaften isolierter einzelner Personen konstituiert wird; die Bedeutung eines sprachlichen Ausdrucks ergebe sich erst aus den Regeln, auf denen der Gebrauch dieses Ausdrucks in einer Sprachgemeinschaft beruhe, und diese Regeln ergeben sich ihrerseits aus einer gemeinsamen sozialen Praxis. Aber reicht diese Antwort aus? Können Regeln und kann insbesondere eine soziale Praxis die Korrektheitsbedingungen eines sprachlichen Ausdrucks besser begründen als die natürlichen Eigenschaften von Einzelpersonen?

Daß in einer Gemeinschaft eine Regel *R* gilt, kann man im Anschluß an Hart (1961: 54 ff.) folgendermaßen erläutern:[7]

(1) Die Mitglieder der Gemeinschaft weichen selten von *R* ab;

(2) wenn ein Mitglied der Gemeinschaft von *R* abweicht, dann ist es Sanktionen seitens der anderen Mitglieder der Gemeinschaft ausgesetzt;

(3) diese Sanktionen werden im allgemeinen akzeptiert.

---

[6] Boghossian 1989.

[7] Vgl. zu dieser Formulierung auch von Savigny 1983, 34.

Wenn das so ist, dann besteht die Tatsache, daß in einer Gemeinschaft eine Regel gilt, aber auch nur in den *Dispositionen* der Mitglieder der Gemeinschaft. Und dann stellt sich natürlich die Frage, inwiefern die Dispositionen mehrerer Personen die Korrektheitsbedingungen sprachlicher Ausdrücke besser begründen können sollen als die Dispositionen einer Einzelperson.

Hierin liegt der Grund dafür, daß Kripke selbst die Bezugnahme auf die Regeln einer Sprachgemeinschaft auch nur als *skeptische* Lösung des Bedeutungsproblems akzeptiert. Eine *substantielle* Lösung ist seiner Meinung nach unmöglich. *Nichts* in der Welt kann den normativen Charakter, die Korrektheitsbedingungen sprachlicher Ausdrücke begründen. Und daher ist – in einem strikten Sinn – die Schlußfolgerung unausweichlich, daß kein sprachlicher Ausdruck die *Eigenschaft* hat, eine bestimmte Bedeutung zu haben. Es hat daher keinen Sinn zu fragen, worin diese Eigenschaft besteht. Das einzige, was wir tun können, ist zu *beschreiben*, unter welchen Bedingungen wir welchen Wörtern welche Bedeutungen *zuschreiben*, und vielleicht zu fragen, warum wir das tun.

Dabei finden wir Kripke zufolge, daß wir bei der Zuschreibung von Bedeutungen in der Tat auf die Handlungen und Dispositionen der Mitglieder von Sprachgemeinschaften Bezug nehmen. Doch daraus folgt nicht, daß es die Eigenschaft, eine bestimmte Bedeutung zu haben, wirklich gibt und daß diese Eigenschaft in diesen Handlungen und Dispositionen begründet ist.

Auf der anderen Seite meint Kripke jedoch ähnlich wie viele Wittgensteinianer, daß es einfach keinen Sinn macht, d. h. keinem verstehbaren Zweck dient, den Äußerungen einer isolierten Einzelperson Bedeutungen zuzuschreiben und daß deshalb unsere Bezugnahme auf soziale Praktiken nicht zufällig, sondern in gewisser Weise zwingend ist. Wenn sich die Frage nach der Bedeutung aber nur noch so stellt, daß es um die *Beschreibung* einer Zuschreibungspraxis und um eine *Erklärung* für diese Praxis geht, dann sind vielleicht doch noch Alternativen denkbar.

In diesem Sinn will ich im folgenden untersuchen, ob es nicht doch gute Gründe für die Praxis vieler Kognitionswissenschaftler gibt, bestimmte physische (z. B. neuronale) Strukturen als Repräsentationen aufzufassen, die eine bestimmte Bedeutung haben. Wenn sich herausstellen sollte, daß das in der Tat so ist, wäre meiner Meinung nach damit zugleich gezeigt, daß die Rede von einer Sprache des Geistes (oder Gehirns) einen vernünftigen Sinn hat – trotz aller Argumente, die Hacker und andere vorgebracht haben.

## 4.

Beginnen möchte ich jedoch mit einem Zugeständnis. Hacker hat in seinen Überlegungen sehr klar gemacht, daß es unserem normalen Gebrauch des Ausdrucks „Sprache“ zufolge eine Sprache nur geben kann, wenn es We-

sen gibt, die diese Sprache sprechen, und daß man von einem Wesen nur dann sagen kann, daß es über eine Sprache verfügt, wenn es über ein bestimmtes breit gefächertes Verhaltensrepertoire verfügt.[8] Eines seiner Argumente gegen die Idee einer Sprache des Gehirns war gerade, daß weder das Gehirn noch seine Teile über ein solches Verhaltensrepertoire verfügen *können*. Und damit hat er sicher recht.

Eine Sprache des Geistes kann es daher nur geben, wenn sie sich in bestimmter Hinsicht radikal von allen normalen Sprachen unterscheidet. Eine Sprache des Geistes ist, wenn es sie gibt, nämlich eine Sprache, die von niemandem gesprochen und auch von niemandem verstanden, ja nicht einmal gehört wird. (Wenn manche sagen, das Gehirn spreche oder verstehe diese Sprache, so ist das in der Tat nur metaphorisch zu verstehen.) Eine Sprache des Geistes ist sozusagen eine Sprache, die einfach geschieht. Token von Sätzen dieser Sprache entstehen unter bestimmten Bedingungen im Gehirn, werden dort verändert und bewirken zusammen mit anderen Satztoken bestimmte Handlungen. Die Satztoken müssen nicht geäußert werden, um zu existieren, und sie müssen nicht gehört und verstanden werden, um Wirkungen hervorzubringen. Alles das geschieht – fast möchte man sagen – wie von selbst.

Wenn das so ist, liegt aber in der Tat die Frage nahe, inwiefern man unter diesen Bedingungen überhaupt noch von einer Sprache reden kann. Diese Frage ist sicher berechtigt; und ich bin mir nicht ganz sicher, ob man sie wirklich überzeugend beantworten kann. Aber versuchsweise möchte ich von der folgenden Überlegung ausgehen: Sprache kann man zunächst einmal einfach auffassen als ein System von strukturierten Sätzen mit einer kombinatorischen Semantik. Die Sätze haben eine Bedeutung (Wahrheitsbedingungen), und diese Bedeutung hängt in regelhafter Weise ab von der Bedeutung ihrer Teilausdrücke. Man kann unterscheiden zwischen Satztypen und Satztoken, wobei Satztoken physikalische Strukturen sind, für die man angeben kann, welchen Satztyp sie realisieren. Wenn das so ist, dann kann man vielleicht aber auch sagen: Wenn es in einem System eine Menge von physikalischen Strukturen gibt, für die es ausreichende Gründe gibt, sie als Token bestimmter Satztypen mit bestimmten Bedeutungen aufzufassen, dann gibt es in diesem System eine interne Sprache der oben erläuterten Art.

---

[8] Meiner Meinung nach ist es allerdings eine interessante Frage, ob tatsächlich alle Verhaltensweisen, die Hacker anführt, notwendige Bedingungen für das Haben einer Sprache sind oder ob wir nicht auch von Wesen, die nur über einen Teil der von Hacker angeführten Fähigkeiten verfügen, sagen würden (oder sogar müßten), daß sie eine Sprache haben. Auf diese Frage kann ich hier aber leider nicht eingehen.

Vielleicht könnte man an dieser Stelle einwenden, daß der Ausdruck „Sprache“ in diesem Zusammenhang unangemessen sei und daß man deshalb eher von „Systemen interner Repräsentationen“ reden solle. Gegen diesen Vorschlag habe ich keine Bedenken. Denn Systeme interner Repräsentationen sind alles, was ein Kognitionswissenschaftler benötigt. Und ich bin sicher, daß kein Kognitionswissenschaftler je die Auffassung vertreten hat, die Sprache des Geistes sei mehr als ein System interner Repräsentationen. Auf der anderen Seite ändert diese sprachliche Klarstellung aber nichts an der argumentativen Situation. Denn Hacker macht mehr als deutlich, daß sich seine Argumente gegen die Annahme von Systemen interner *Repräsentationen* ebenso richten wie gegen die Annahme einer *Sprache* des Geistes.[9] Hacker will ganz allgemein bestreiten, daß es im Gehirn oder in Computersystemen *Bedeutung tragende Strukturen* geben könne. Ich möchte im folgenden dagegen für die These argumentieren, daß es sehr wohl gute Gründe dafür geben kann, bestimmte interne Strukturen als Systeme von Repräsentationen aufzufassen, und daß die Idee einer Sprache des Geistes in diesem Sinn daher durchaus nicht sinnlos ist. Beginnen möchte ich allerdings mit einer sehr allgemeinen wissenschaftstheoretischen Bemerkung.

## 5.

Wenn es darum geht, das Verhalten komplexer Systeme zu erklären und zu verstehen, reicht es häufig nicht aus, nur die – von Dennett[10] so genannte – *physikalische Einstellung* einzunehmen. Ein angemessenes Verständnis ergibt sich vielmehr oft erst, wenn wir auch die *funktionale Organisation* dieser Systeme verstehen. Besonders deutlich wird diese Tatsache im Bereich der Biologie, wo Erklärungen häufig nur auf der funktionalen Ebene gegeben und anatomische und physiologische Details entsprechend kaum mehr erwähnt werden. Nehmen wir irgendein Beispiel – etwa die Temperaturregelung im menschlichen Körper, die im Lehrbuch für *Biologische Psychologie* von Birbaumer und Schmidt so erklärt wird. (Ich fasse diese Darstellung hier stark zusammen.)

> Die Thermoregulation kann formal als ein kreisförmig geschlossenes Regelsystem mit negativer Rückkopplung angesehen werden. Die Körpertemperatur wird von Meßfühlern, nämlich den Thermorezeptoren überwacht, die ihre Meldungen dem zentralen Regler zuführen. Dieser stellt fest, ob die Körpertemperatur (der *Istwert*) von ihrem Sollwert abgewichen ist und verstellt ent-

[9] Vgl. z.B. Hackers These: „Nothing in the cortex constitutes a ‚symbolic representation‘ of the creature’s environment.“ (1984, 497).

[10] Die Unterscheidung zwischen physikalischer, funktionaler und intentionaler Einstellung findet sich erstmals in Dennett 1971.

> sprechend über die Aussendung von *Steuersignalen* die Stellgrößen solange, bis die Meßgrößen den Ausgleich der Abweichung signalisieren. (Birbaumer/Schmidt 1990, 117 f.)
>
> Die Körperkerntemperatur wird an verschiedenen Stellen durch temperaturempfindliche Nerven- bzw. Sinneszellen gemessen, die Körperschalentemperatur durch Thermosensoren in und unter der Haut. Der Hypothalamus, insbesondere die *Area hypothalamica posterior* wird als Integrationszentrum für die Thermoregulation angesehen. Die zentralen Effektorneurone steuern (wahrscheinlich über eine Kette von Interneuronen) die Stellglieder zur Wärmebildung und -abgabe (Wärmebildung, Isolation der Körperschale, Schweißsekretion, Verhalten). Sie erhalten ihre afferenten Zuflüsse von den äußeren und inneren Thermosensoren. Die Kältesensoren wirken direkt aktivierend auf die Effektorneurone für Wärmebildung und über Interneurone hemmend auf die Effektorneurone für die Wärmeabgabestellglieder. Die Warmsensoren sind genau umgekehrt mit den zwei verschiedenen Typen von Effektorneuronen verschaltet. (op.cit., 119 ff.)

Die fast ausschließliche Verwendung funktionalen Vokabulars ist unübersehbar. Es wird von Meßfühlern, Stellgrößen und Regelkreisen geredet ebenso wie von Integrationszentren, Thermosensoren und Effektorneuronen. Die einzigen physiologischen Begriffe scheinen anatomische Bezeichnungen wie „Area hypothalamica posterior“ zu sein. Dabei ließe sich die Geschichte auch ganz anders erzählen. Vereinfacht z. B. so: Wenn im Körper – sagen wir, aufgrund von körperlicher Arbeit – die Temperatur über einen Wert von 37°-38° Celsius steigt, bewirkt das eine erhöhte Entladungsrate bestimmter Neuronen im Körperinneren, die mit ihren Axonen bis in den Hypothalamus reichen. Im Hypothalamus wird durch die erhöhte Feuerungsrate dieser Neurone die Aktivität bestimmter sympathischer und parasympathischer Neurone gedämpft, die über Axone und neuromuskuläre Synapsen mit der glatten Muskulatur der präkapillaren Gefäße verbunden sind. Dies führt zu einer Erschlaffung dieser Muskulatur und damit zur Erweiterung der entsprechenden Gefäße.

Aber ganz unabhängig davon, daß uns diese Geschichte in ihren Einzelheiten nicht vollständig bekannt ist – sie allein würde uns auch nicht reichen. Was uns interessiert, ist nämlich die Frage, wie es unser Körper schafft, einen bestimmten *Zielzustand* zu erreichen, wie er es schafft, seine Kerntemperatur *unter sehr verschiedenen Bedingungen* relativ *konstant* zu halten. Und *das* verstehen wir erst, wenn wir sehen, daß die physiologischen Prozesse in Form eines Regelkreises zusammenwirken und daher auch mit Hilfe des entsprechenden Begriffsystems beschrieben werden können. Funktionale Begriffe kommen also besonders dann ins Spiel, wenn es nicht mehr darum geht, einzelne physische Zustände oder Aktivitäten zu erklären, sondern darum, zu verstehen, wie erfolgreiches Handeln zustande kommt, d. h. wie es ein System schafft, unter den verschiedensten Umstän-

den ein Verhalten zu produzieren, das gewissen Standards entspricht. Insgesamt gilt also:

> These 1: Das *erfolgreiche* Verhalten von Systemen können wir häufig nur dann angemessen verstehen und erklären, wenn wir bezüglich dieser Systeme von der physikalischen zur *funktionalen* Einstellung übergehen.

Zur funktionalen Analyse an dieser Stelle noch eine kurze Zusatzbemerkung: Wichtig ist in diesem Zusammenhang auch, daß Eigenschaften wie ein-Meßfühler-zu-sein oder ein-Stellglied-zu-sein keine im herkömmlichen Sinne „natürlichen" Eigenschaften sind, d.h. daß wir die entsprechenden Begriffe „Meßfühler" und „Stellglied" – anders als z.B. Begriffe wie „Pyramidenzelle" oder „neuromuskuläre Synapse" – nicht aufgrund von normalen beobachtbaren oder meßbaren neurobiologischen Merkmalen zuschreiben. Denn die Anwendbarkeit dieser Begriffe auf bestimmte neuronale Phänomene hängt davon ab, ob diese Phänomene mit anderen so zusammenwirken, daß sich ein Verschaltungsmuster ergibt, das als „kreisförmig geschlossenes Regelsystem" *interpretiert* werden kann. Um es auf eine sehr saloppe (und sicherlich auch etwas irreführende) Weise zu formulieren: Funktionale Eigenschaften finden sich nicht in der Welt; sie werden von uns in die Welt hineininterpretiert.

## 6.

Das Beispiel der Thermoregulation im menschlichen Körper ist jedoch zu unspezifisch, als daß man aus ihm etwas über den Sinn oder Unsinn der Idee einer Sprache des Geistes entnehmen könnte. Mit einem zweiten Beispiel kommen wir der Sache aber schon näher – dem Beispiel eines Schachcomputers, das auch von Dennett häufig zur Veranschaulichung herangezogen wird.[11]

Auch bei einem solchen elektronischen Gerät ist es zumindest im Prinzip möglich, jeden einzelnen Zug rein physikalisch zu erklären: Man kann feststellen, wie sich durch den Druck bestimmter Buchstaben- und Zahlentasten bestimmte Teilzustände von Siliziumchips verändern; man kann aus der Verschaltung und den Anfangszuständen dieser Chips ableiten, welche Abfolge von Zuständen sie durchlaufen, nachdem die „Enter"-Taste gedrückt wurde; und auf dieselbe Weise läßt sich schließlich ermitteln, in welchem Zustand das Gesamtgerät am Ende stehen bleibt und welche der Leuchtdioden, aus denen das Display besteht, dann leuchten bzw. nicht leuchten. Was man auf diese Weise erreichen kann, ist aber immer nur die Erklärung konkreter einzelner Endzustände auf der Grundlage des Wissens um konkrete einzelne Anfangsbedingungen. Völlig unzureichend ist dieses Verfah-

[11] Erstmals in Dennett 1971.

ren, wenn man verstehen will, aufgrund welcher Mechanismen es das Gerät schafft, immer wieder outputs zu liefern, denen Züge entsprechen, die in der jeweiligen Spielsituation plausibel oder sogar erfolgreich sind.

Ein solches Verständnis ergibt sich wiederum erst, wenn man von der physikalischen zur funktionalen Einstellung übergeht, was in diesem Fall heißt, daß man das *Programm* analysiert, das dem Verhalten des Schachcomputers zugrundeliegt. Denn erst dann ist es möglich, das, was zwischen Input und Output passiert, nicht mehr nur als eine Abfolge von Zuständen von Siliziumchips zu sehen. Erst in der funktionalen Einstellung kann man bestimmte Teilzustände dieser Chips als Repräsentationen von möglichen Konfigurationen der Figuren auf dem Schachbrett interpretieren. Und nur unter dieser Voraussetzung kann man das Geschehen zwischen Input und Output so beschreiben, wie es uns allen inzwischen geläufig ist: Das Gerät berechnet zuerst für die aktuelle Stellung die Repräsentationen aller Folgestellungen, die sich aus den für es selbst möglichen Zügen ergeben; dann zu jeder dieser Folgestellungen die Repräsentationen aller Folgestellungen, die sich aus den jeweils möglichen Zügen des Gegners ergeben; weiter zu jeder dieser Folgestellungen wieder die Repräsentationen aller Folgestellungen, die sich aus den eigenen möglichen Zügen ergeben, usw. bis zu einer durch heuristische Kriterien bestimmten Anzahl von Zügen und Gegenzügen; die einzelnen Folgestellungen werden nach vorgegebenen Kriterien bewertet; und schließlich gibt das Gerät den Zug aus, der bei (nach seinen Kriterien) optimalen Gegenzügen zu der Stellung mit der höchsten Bewertung führt.

Diese Art der Beschreibung ermöglicht nun erstmals ein Verständnis der Tatsache, daß unser Schachcomputer – in der Regel – plausible oder sogar erfolgreiche Züge macht. Denn es läßt sich zeigen, daß die Bewertungsfunktion, die der Zugauswahl zugrundeliegt, unter den jeweiligen Bedingungen tatsächlich zu plausiblen oder gar guten Stellungen führt. Wenn das Gerät zum Schluß den Zug ausgibt, der bei „optimalem“ Gegenspiel zu der am höchsten bewerteten Stellung führt, müssen seine Züge daher in der Regel recht gute Züge sein. Das gilt – wie gesagt – in der Regel; denn es gibt Stellungen, die trotz hoher Bewertung objektiv ungünstig sind, und wenn ein solcher Fall vorliegt, wählt der Computer häufig keinen besonders guten Zug. Aber das überrascht auch gar nicht. Denn wir wissen natürlich, daß der Computer manchmal Fehler macht. Für die Erklärung des Verhaltens des Computers ist die Beschreibung der Vorgänge zwischen Input und Output mit Hilfe des skizzierten Programms somit doppelt hilfreich: Sie erklärt uns, warum der Computer in der Regel gute Züge wählt, und sie erklärt uns auch, warum er manchmal furchtbare Fehler macht.

Aber zurück zum Hauptpunkt. Ich hoffe, es ist deutlich geworden, daß wir das Verhalten komplexer Systeme häufig erst dann richtig verstehen und erklären können, wenn wir von der physikalischen zur funktionalen

Einstellung übergehen, und das dies insbesondere dann der Fall ist, wenn es sich dabei um ein Verhalten handelt, das – gemessen an bestimmten Standards – als erfolgreich eingestuft werden kann. Wichtiger noch als dieser allgemeine Punkt ist jedoch ein Punkt, der sich ergibt, wenn wir genauer nachfragen, worauf wir denn eigentlich festgelegt sind, wenn wir bezüglich bestimmter Systeme die funktionale Einstellung einnehmen. Auch für diese Frage ist das Schachcomputer-Beispiel wieder außerordentlich instruktiv.

Ich hatte schon gesagt, daß im Hinblick auf einen Schachcomputer die funktionale Einstellung einzunehmen heißt, das Programm zu analysieren, daß in diesem Computer implementiert ist. Und dies wiederum bedeutet zweierlei: 1. daß man bestimmte Teilprozesse, die zwischen Input und Output ablaufen, als die Ausführung bestimmter Anweisungen auffaßt und 2. daß man rekonstruiert, wie das System die Abfolge dieser Teilprozesse organisiert. Die Ausführung einer Anweisung besteht in der Regel jedoch darin, daß eine bestimmte Datenstruktur geschaffen oder verändert wird. Und das bedeutet, daß wir bestimmte physische Prozesse nur dann als die Ausführung einer Anweisung auffassen können, wenn wir zugleich bestimmte physische Strukturen als Datenstrukturen auffassen. Für die funktionale Analyse unseres Schachcomputers heißt das konkret: Wir können das Programm, das seinem in der Regel erfolgreichen Verhalten zugrundeliegt, nur rekonstruieren, wenn wir bestimmte physische Strukturen im Inneren des Systems – die Teilzustände bestimmter Siliziumchips – als Repräsentationen von möglichen Stellungen und andere physische Strukturen dieser Art als Repräsentationen der Bewertungen von Stellungen auffassen. Wenn wir dieses Ergebnis verallgemeinern, ergibt sich die

> These 2: Die funktionale Analyse eines Systems, die allein einen Beitrag zur Beantwortung der Frage leisten kann, wie es das System schafft, sich in den unterschiedlichsten Situationen erfolgreich zu verhalten, ist in einigen Fällen nur möglich*, wenn man bestimmte physische Strukturen im Inneren des Systems als Repräsentationen auffaßt*.

An dieser Stelle könnte man – im Anschluß an die Argumente Hackers – versucht sein einzuwenden, daß die bisherige Argumentation die Tatsache völlig außer acht lasse, daß es sich bei Schachcomputern um Artefakte handelt, die von ihren Herstellern tatsächlich zu einem bestimmten Zweck programmiert wurden. Von solchen Artefakten könne man deshalb in der Tat sagen, daß in ihnen Programme ablaufen und daß es in ihnen daher auch so etwas wie Repräsentationen gebe; denn in diesem Falle gebe es jemanden – nämlich den Programmierer, der mit bestimmten physischen Strukturen bestimmte Konfigurationen von Schachfiguren repräsentieren wolle. Repräsentationen ohne eine Person, die sie verwende, seien jedoch begrifflich unmöglich.

Dieser Einwand ginge an meiner Argumentation jedoch völlig vorbei. Denn der Hauptpunkt meines Argument ist, daß wir bei manchen Systemen – *ganz unabhängig davon, wie sie entstanden sind* – annehmen müssen, daß es in ihnen Repräsentationen gibt, wenn wir verstehen wollen, wie das erfolgreiche Verhalten dieser Systeme zustandekommt. Schachcomputer müßten wir unter dieser Voraussetzung in funktionaler Einstellung also auch dann genau so beschreiben, wie ich es oben erläutert habe, wenn sie auf Bäumen wachsen würden.

In der Tat kann man an vielen Beispielen nachweisen, daß in der Neurobiologie genau diese Erklärungsstrategie verfolgt wird. Ich erinnere mich noch sehr gut an eine Diskussion, in der ich den Göttinger Physiker und Akustikexperten Manfred Schroeder einmal gefragt habe, welche neuronalen Mechanismen denn für die Lokalisation von Schallquellen verantwortlichen seien. Schroeders Antwort begann mit dem Satz: „Zunächst einmal wird im Gehirn die Kreuzkorrelation der Signale der beiden Hörnerven errechnet." Ein anderes Beispiel derselben Art findet sich in einem Artikel von J. Koenderink mit dem Titel „The Brain a Geometry Engine". Denn soweit ich diesen Artikel verstanden habe, ist seine Hauptthese, daß man die Mechanismen des visuellen Kortex am besten versteht, wenn man von der zweidimensionalen Intensitätsverteilung des auf die Retina fallenden Lichts ausgeht und die anschließende neuronale Verarbeitung so interpretiert, daß in ihr die ersten, zweiten und weitere höherstufige Ableitungen dieser Verteilung errechnet werden.

> … you may understand a large part of the structure of the front-end visual system as an embodiment of differential geometry of the visual field. … Instead of the concrete ‚edge detectors' and ‚bar detectors', one speaks of the abstract first- and second-order directional derivatives. (1990: 125)

Ich kann hier auf weitere Details nicht eingehen; aber ich denke, daß schon an diesen nur angedeuteten Beispielen klar wird, daß viele Neurobiologen bei dem Versuch, die erstaunlichen Leistungen des Gehirns zu erklären, tatsächlich die funktionale Einstellung einnehmen und daß sie dabei darüberhinaus tatsächlich Erklärungen auf der Basis der Annahme anstreben, daß im Gehirn bestimmte *Berechnungen* durchgeführt werden. Rechenprozesse aber sind Prozesse der Veränderung von Zahlzeichen; also Prozesse, bei denen Repräsentationen numerischer Werte manipuliert werden. Wenn man animmt, daß im Gehirn Rechenprozesse stattfinden, ist man damit also auch auf die Annahme festgelegt, daß es im Gehirn ein System von Repräsentationen gibt.

## 7.

Damit ist im Grunde alles Wesentliche gesagt. Genauso wie es bei Schachcomputern notwendig ist anzunehmen, daß es in ihnen Repräsentationen von Stellungen und Repräsentationen von Bewertungen gibt, wenn man verstehen will, wie diese Geräte es fertigbringen, einigermaßen erfolgreiche Züge zu produzieren, kann es im Hinblick auf andere Systeme notwendig sein anzunehmen, daß es in ihnen satzartige Repräsentationen von ihrer Umwelt oder von allgemeinen Gesetzmäßigkeiten gibt, wenn man verstehen will, worauf das erfolgreiche Verhalten dieser Systeme beruht. Dies würde z. B. für alle KI-Systeme gelten, deren Problemlösungsverhalten auf der Ableitung von Formeln aus einer Menge von Axiomen beruht. Denn das Verhalten dieser Systeme können wir nicht angemessen verstehen, wenn wir nicht einige der in ihnen ablaufenden Prozesse als Inferenzprozesse interpretieren. Und Inferenzprozesse sind Prozesse der Ableitung von satzartigen Repräsentationen aus satzartigen Repräsentationen. D. h., wir können nicht einige der in einem System ablaufenden Prozesse als Inferenzprozesse auffassen, wenn wir nicht zugleich einige der in ihm vorkommenden physischen Zustände als satzartige Repräsentationen interpretieren. Hier zeigt sich ebenso wie bei dem zuvor erläuterten Schachcomputer-Beispiel ein Primat der Interpretation von Prozessen. Bestimmte Prozesse in einem System können wir nicht angemessen verstehen, wenn wir nicht zugleich bestimmte Zustände als Repräsentationen auffassen.

Und damit ist, denke ich, auch klar, unter welchen Bedingungen wir gewissermaßen sogar gezwungen sind anzunehmen, daß es in bestimmten Systemen satzartige Repräsentationen oder Symbole gibt, d. h. daß diese Systeme eine Sprache des Geistes enthalten. Dies ist nämlich genau dann der Fall, wenn wir nur dann richtig verstehen können, wie das erfolgreiche Verhalten dieser Systeme zustandekommt, wenn wir einige der in ihnen ablaufenden physischen Prozesse als funktionale Prozesse auffassen, die ihrerseits nur als Prozesse der Erzeugung und Veränderung satzartiger Repräsentationen verstanden werden können. Noch einmal in einer These zusammengefaßt:

> These 3: Die Annahme *satzartiger Repräsentationen* ist nicht nur plausibel, sondern in gewisser Weise sogar unumgänglich, wenn wir das *erfolgreiche* Verhalten eines Systems nur durch die Annahme erklären können, daß es auf *funktionalen* Prozessen beruht, die nur als *Prozesse der Erzeugung und Veränderung satzartiger Repräsentationen* verstanden werden können.

Die Redeweise von satzartigen Repräsentationen bzw. von einer Sprache des Geistes ist also weder eine Marotte gewisser Kognitionswissenschaftler noch gar eine Marotte, die auf einer fundamentalen begrifflichen Konfusi-

on beruht. Sie ist vielmehr eine zwingende Konsequenz, die sich bei dem Versuch ergeben kann, die funktionale Architektur eines Systems zu verstehen, auf der das erfolgreiche Verhalten dieses Systems beruht.

Zum Abschluß möchte ich aber sehr nachdrücklich betonen, daß in der These 3 nur eine *Bedingung* formuliert ist. *Wenn* diese Bedingung erfüllt ist, dann kann man davon sprechen, daß es in einem System eine Sprache des Geistes gibt. Aus dieser These ergibt sich also *nicht*, daß Fodor recht hat oder daß die Kognitionswissenschaftler recht haben, die glauben, daß intelligentes Verhalten nur im Rahmen des Symbolverarbeitungsansatzes adäquat erklärt werden kann. Aber mir ging es auch nicht darum, diesen Ansatz als *sachlich angemessen* zu verteidigen, sondern nur darum, ihn vor dem Vorwurf der Begriffsverwirrung in Schutz zu nehmen.

## Literatur

Beckermann, A. (1991): „Der endgültige Todesstoß für den Repräsentationalismus? – Eine Replik auf Andreas Kemmerlings Artikel ‚Mentale Repräsentationen'". *Kognitionswissenschaft* 2, 91–98.

Birbaumer, N. & R. F: Schmidt (1990): *Biologische Psychologie*. Berlin/Heidelberg/New York.

Boghossian, P. (1989): „The Rule-Following Considerations". *Mind* 98, 507–549.

Dennett, D. (1971): „Intentional Systems". *Journal of Philosophy* 68, 87–106. Wiederabgedr. in: Dennett, D.: *Brainstorms*. Montgomery, Verm. 1978, 3–22.

Fodor, J. A. (1975): *The Language of Thought*. New York.

Fodor, J. A. (1978): „Propositional Attitudes". *The Monist* 64, 501–523. Wiederabgedr. in: Fodor 1981a, 177–203.

Fodor, J. A. (1981a): *Representations*. Cambridge, Mass.

Fodor, J. A. (1981b): „Introduction – Something on the State of the Art". In: Fodor 1981a, 1–31.

Fodor, J. A. (1987): *Psychosemantics*. Cambridge, Mass.

Hacker, P. (1987): „Languages, Minds and Brains". In: C. Blakemore & S. Greenfield (Hg.) *Mindwaves: Thoughts on Intelligence, Identity and Consciousness*. Oxford, 485–505.

Harman, G. (1973): *Thought*. Princeton, NJ.

Hart, H. L. A. (1961): *The Concept of Law*. Oxford.

Hyman, J. (Hg.) (1991): *Investigating Psychology*. London/New York.

Koenderink, J. J. (1990): „The Brain a Geometry Engine". *Psychological Research* 52, 122–127.

Kripke, S. (1982): *Wittgenstein on Rules and Privat Language.* Cambridge.

Savigny, E. von (1983): *Zum Begriff der Sprache*. Stuttgart.

Stillings, N. A. et al. (1987): *Cognitive Science – An Introduction*. Cambridge, Mass.

# Intentionalität und Qualia

# 9

## Why Tropistic Systems are not Genuine Intentional Systems [*†]

1.

If one uses a set of rather simple criteria for the ascription of intentional states one may be forced to reckon among the class of intentional systems not only men and higher animals, but also lower animals like starfish and jellyfish and perhaps even amoeba and paramecia. And, what seems worse to many, one may be forced to say that, according to these criteria, even cybernetic mechanisms like thermostats are intentional systems. Some philosophers have accepted this conclusion with a shrug, murmering „Why not?". But most of us, I think, have a strong feeling to the contrary since to speak in such an inflationary way about intentional states would deprive this kind of talk of its whole point. So we are badly in need of better criteria which are less at variance with our intuitions. What really is the difference between systems which are genuine intentional systems and systems which are not?

Since I cannot hope to answer this question in general I shall restrict myself to a minor part of the problem: the developement of a set of criteria which enable us to show that at least tropistic systems – e. g. amoeba and paramecia and thermostats – are not intentional systems. Intentional systems are, as is generally agreed, those systems which really have genuine intentional states (i. e. especially: genuine wants and beliefs) or – to put it in more cautious terms – those systems to which we can legitimately ascribe such intentional states. Hence, to claim that tropistic systems are not genuine intentional systems really is to claim that tropistic systems do not have genuine intentional states, that they do not have genuine wants and beliefs. I shall therefore try to show that this claim is indeed true, that there are good reasons for not ascribing genuine wants and beliefs to systems of this kind.

2.

---

* Erstveröffentlichung in: *Erkenntnis* 29 (1988), 125–142.

† Obviously, there is a certain resemblance between the title of this paper and the title of Fodor's recent paper „Why Paramecia Don't Have Mental Representations". This is no accident, since in this paper I address myself to almost the same questions as Fodor did in his. I shall come back to Fodor's answers at the end of section 3. An earlier draft of this paper was presented at a conference on „Aspects of Consciousness and Awareness" at the ZiF in Bielefeld. I am very much indebted to W. Ewald and K. Hillebrandt for correcting my English.

I would like to begin my considerations with a simple example. Imagine a little mechanical bug – let us call it $S_1$ – which is constructed to move towards any bright enough light in its environment. This kind of behavior is, as we shall assume, brought about in the following way: on the top of the mechanical bug $S_1$ there is a circle of 24 photo-electric cells which especially respond to light coming from the direction to which they themselves are oriented (compare figure 1). For instance, if there is a bright light at an angle of 45 degrees in front of the bug then the corresponding photo-electric cell (cell #3) will have a considerable output while all other cells remain rather „silent“. Given this situation $S_1$ will turn around (right or left, depending on whether the cell with the greatest output is on the right or on the left side of $S_1$) until the output of the front-cell (cell #0) is greatest and then move on straight forward towards the light. If the output of all cells lies under a certain threshold, $S_1$ will not move at all.

$S_1$ is what is commonly called a positively phototropic system. I shall therefore try to elucidate by means of this simple mechanical bug what I think to be the main differences between tropistic and genuine intentional systems, i. e. what in my eyes are the reasons for not ascribing wants and beliefs to things like the system $S_1$. Let us begin with the question why systems like $S_1$ are not the right kind of things to have genuine wants. As a

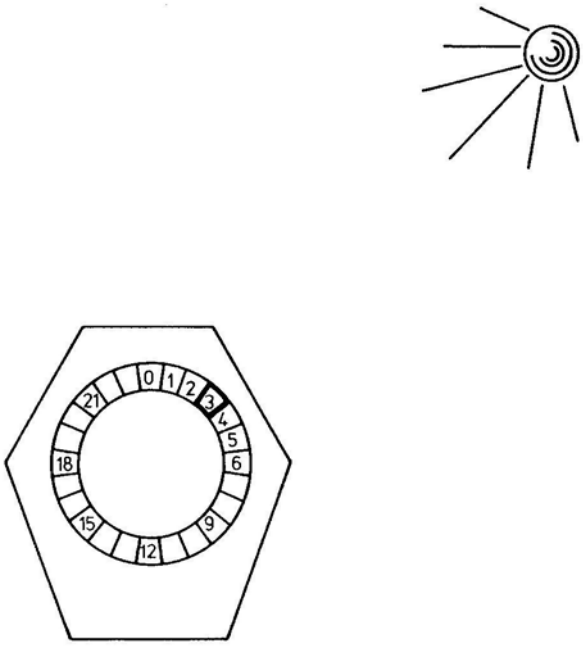



Fig. 1

first step to answering this question it is interesting to compare the system $S_1$ with objects which in a way show a quite similar behavior – though no one, I hope, is tempted to call these objects intentional systems. Think e. g. of a physical experiment which you will perhaps remember from your school days. A knitting-needle is magnetized, put through a cork and then the cork-plus-needle is placed in a little water basin so that only the north-pole of the needle stands out from the water. If now a bar-magnet is placed on the edge of the little basin the north-pole of the needle will move on a certain predictable course towards the south-pole of the bar-magnet.

One could ask whether there really is a difference in principle between the system $S_1$ and the magnetized knitting-needle since, after all, the behavior of these two objects is indeed rather similar. But in my opinion there is such a difference, and, as I see it, this difference lies in the fact that the knitting-needle is moved by *outward* forces operating on the needle while the system $S_1$ is moved from *within*. $S_1$ is, in the strict sense, an *automaton*. Systems like $S_1$ need a motor and a power supply to be able to move. This is already shown by the simple fact that they stop and stand still if their motor is broken or they have run out of fuel (or their battery is empty or whatever). In contrast to this it is characteristic of movements like the movement of the knitting-needle that they can only be prevented by the exertion of a sufficient counterforce. It is for this reason that we say that the needle is *attracted* by the bar-magnet while we wouldn't say – except metaphorically – that $S_1$ is attracted by the light, though, to be sure, even this system has an immanent tendency to approach the brightest light in its environment. Hence the difference between the knitting-needle and the system $S_1$ can be summed up like this: while the needle is pushed (or pulled) around by external forces the source of the motions of $S_1$ lies in this system itself. I think that by virtue of this property $S_1$ is a possible candidate for being an intentional system since we certainly would not ascribe wants and/or beliefs to any system which is only pushed around by outward forces. Hence, being an automaton is a necessary condition for being a possible candidate for the ascription of genuine wants and beliefs. And, therefore, the following considerations are meant to apply only to automata.

That non-automata cannot have wants and beliefs, however, characterizes intentional systems only in a negative way. What can we say more positively about the ascription of wants and/or beliefs? Or, to begin with, what can we say more positively about the ascription of wants or want-like states?

Having a certain want is – or at least implies – being disposed to behave in a certain way. Wants involve behavioral dispositions. This much is, I think, philosophical common sense. But if so, what then is the difference

between wants and want-like states on the one side and other behavioral dispositions on the other? What is it that makes wants intentional states which are to be characterized by their content? The answer is that the behavior which constitutes the realization of a non-intentional behavioral disposition – e. g. the brittleness of a windowplane or the ignitability of a match – are best characterized in purely physical terms. A plane *breaks* if it is hit by a stone, a match *ignites* if it is struck. On the other hand the behavior which constitutes the realization of an intentional behavioral disposition cannot be characterized this way. If someone wants to open the window then he is, after all, disposed to behave in a certain way. But this behavior cannot be characterized in physical terms. E. g., he is not disposed to turn left, move his legs in a physically describable way, raise his arm, etc. Rather he is disposed to do *something which will* – at least, as he believes – *have the effect that the window is open*. The behavior which constitutes the realization of an intentional behavioral disposition thus is best characterized by *the state of affairs which it will probably bring about* or, to put in another way, by the *goal-state G* which it is meant to realize. Hence, having the want that *G* implies having the disposition to do something which is apt to bring about *G* until finally *G* is achieved. So presumably we can start from the fact that any automaton *A* which has the genuine want that *G* satisfies the following two conditions:

(1) In many situations *A* exhibits a behavior that under the given circumstances is apt to bring about *G*.
(2) *A* ceases to exhibit this behavior if *G* is achieved.

These two conditions express an important point about wants, namely, that wants imply goal-directed behavior, i. e. behavior that (normally) brings about a certain goal and that comes to an end if the goal is achieved. But condition (1), as formulated above, may lead to difficulties since it does not exclude the case that *A* only under very specific conditions exhibits a behavior appropriate to bring about *G* and, at least in my opinion, we should not ascribe wants or want-like states to a system which would not under varying circumstances do what is under the respective circumstances apt to bring about *G*. Nor should we ascribe wants or want-like states to a system which exhibited the same kind of behavior under all circumstances even if this behavior under all these circumstances would have the same result. Think e. g. of a time-fuse bomb. No one would say that such a bomb wants to destroy its environment and itself at the preset time. So we had better replace condition (1) by the new condition

(1′) In a variety of different situations *A* exhibits a behavior which is apt to bring about *G* and the behavior is not the same in all situations.

Now, as to conditions (1′) and (2) it might seem that even the system $S_1$ exhibits some kind of goal-directed behavior. But with regard to this system there remains the crucial question: What exactly is the goal-state which the behavior of $S_1$ tends to realize? The description given above makes clear that $S_1$ has a tendency to move towards the brightest light in its environment, but it mentions no state which will bring the behavior of this system to an end. And so $S_1$ – like many other tropistic systems – in fact does not have a definite goal. It moves in a certain direction, but there seems to be no definite state which it is really tending to achieve. The situation would, however, be different if $S_1$ were only to turn around until the output of cell #0 is greatest without then moving straight forward. For in that case the goal-state of this slightly modified system – let us call it $S'_1$ – could be identified as „the front of $S'_1$ is facing the brightest part of its environment". Hence, the system $S'_1$ satifies both conditions (1′) and (2) and thus exhibits a kind of goal-directed behavior which is at least a necessary condition for the ascription of genuine wants. But obviously this is not yet enough for us to be able to ascribe a genuine want to this system. Hence, the question now is, what more is needed? Or, in other words: what are the reasons for not ascribing genuine wants to tropistic systems like the system $S'_1$ even though they satisfy conditions (1′) and (2)?

I see at least two points that are relevant here. First, though tropistic systems are automata – self-moving systems – they are in an important sense *passive* and not active systems. If we take a closer look at the behavior of $S_1$ or $S'_1$ we notice that these systems turn around or stand still *only if and as long as there is a corresponding causal impact from the environment on them*. If a bright enough light in the environment causes one of the photo-electric cells of $S'_1$ (except cell #0) to emit a greater output than all other cells, then $S'_1$ turns either right or left depending on which cell emits the greatest output. But this movement stops at the same moment at which the causal impact of the light on $S'_1$ is inhibited – either by dimming the light, or by turning it off, or by screening, or by whatever means. As I see it, this is a characteristic of tropistic system in general. If the object they tend to approach or to move away from is prevented from having a causal influence on the system, all behavior stops at once. This is the reason why these systems show a behavior that is in many respects indiscernible e. g. from the behavior of a magnetized knitting-needle under the causal influence of a bar-magnet, i. e. why these systems *seem* to be attracted or repelled by the objects they tend to approach or to move away from.

Imagine in contrast a very simple machine which I would like to call a random seeker. This little system also exhibits a kind of goal-directed behavior since it is designed to stop if it encounters a certain state of affairs.

But, as we shall further assume, the behavior of this system is entirely random. It moves straight on for a certain while, then changes its direction by a random angle, moves straight forward for some time, changes its direction again, etc. This little system, it seems to me, is a better candidate for the ascription of a genuine want than any tropistic system for the simple reason that *what* the system does is not entirely dependent on the varying states of its environment. It is not only an automaton, a self-moving system, but in a way also a *self-directed* system.

But let us come to the second point which is, in my view, the decisive reason for our reluctance to ascribe genuine wants to tropistic systems. As we noted before, tropistic systems exhibit in many different situations a behavior which in each of these situations is apt to bring about a certain state of affairs. But in fact they do not exhibit this behavior *because* it is apt to bring about this state of affairs. This comes out clearly if we imagine some possible situations in which e. g. the behavior of the system $S'_1$ would be absolutely inapt to achieve the goal-state of this system. Let us assume for instance that someone picks up $S'_1$ and suspends it in such a way that the wheels of the system no longer touch the floor. Then, if a bright light is put at an angle of 45 degrees in front of $S'_1$ (which causes cell #3 to have the greatest output) $S'_1$ will begin to move its wheels in a certain direction. But, evidently, this will have no effect, the position of $S'_1$ will not change a bit. Therefore, if the light is not removed or screened, the wheels of $S'_1$ will continue to move until finally $S'_1$ runs out of energy. Nothing will have been achieved as far as the goal-state of $S'_1$ is concerned. And nothing could have been achieved since in the mentioned situation none of the possible movements of $S'_1$ would have been able to bring $S'_1$ nearer to its goal-state.

Take another case. Let us assume that the system $S'_1$ is put on a disc which rotates counter-clockwise at the same speed at which $S'_1$ normally performs its turns. If we now put a light in the environment of $S'_1$ in such a way that at the moment it is put there it is located on the right side of $S'_1$, $S'_1$ will begin to turn right, again with the effect that its relative position to its environment including the light remains unchanged. But this situation is different from the situation mentioned before because in this situation $S'_1$ could do something to achieve its goal: first do nothing until the light is just in front of it and then start turning right. It's is easy to see that this doing things just in reversed order would have the desired result. But $S'_1$ does nothing like this. „Stupid" as it is it clings to doing what it always does even if this will not bring it nearer to its goal. This is speaking metaphorically. In fact this insensitivity to the special aspects of the situation is due to the already mentioned simple design of $S'_1$. Every time a certain causal impact on the system occurs it will respond with a certain kind of move-

ment – no matter whether this movement is apt to bring about one state or another. If there is an unscreened bright light on one side of or behind $S'_1$, $S'_1$ will move its wheels in one direction or the other at a preset speed and, if the bright light is just in front of it, $S'_1$ will do nothing. That this way of moving normally leads to a certain result is, as I see it, an entirely *contingent* fact. Change some features in what can be called a normal situation and the same behavior will lead to completely different results. This makes clear that, even though the behavior of tropistic systems like the system $S'_1$ is as a matter of fact apt to bring about a certain goal-state in many different situations, it is nevertheless false to say that systems of this kind exhibit such behavior *because* it is apt to bring about this state of affairs.

These considerations, in my opinion, suggest the conclusion that systems which have genuine wants differ from merely tropistic systems in the following way. Both kinds of systems exhibit in a variety of different situations a behavior which is apt to bring about a certain state of affairs. But with regard to tropistic systems this is no more than a contingent fact. Given a certain causal impact from its environment a tropistic system behaves in a certain way no matter whether this behavior will bring about a certain state of affairs or not. With regard to systems with genuine wants, however, it is not a contingent fact that they show a suitable behavior since these systems behave as they do *because* their behavior is apt to bring about a certain state of affairs. Perhaps one could even say that these systems *choose* their behavior in virtue of its being apt to bring about a certain goal-state. Hence, if an automaton *A* is a system with genuine wants it seems to satisfy not only conditions (1′) and (2) but also the additional condition

(3) The suitable behavior of *A* is due to an internal mechanism which *chooses* this behavior in virtue of its being apt to bring about *G*.

It might seem strange that in the formulation of this condition the notion of „choice“ is used though this notion, too, seems to have at most a metaphorical application in this context. But there are good reasons for using this notion here. For, if it is not a contingent fact with regard to systems with genuine wants that their behavior in varying circumstances is apt to bring about a certain goal-state, then this behavior *must be caused by an internal mechanism which ensures that this is indeed the case*. This mechanism, therefore, must effect that of all possible actions of a system that action is executed which under the given circumstances will bring about the desired goal-state – granted that there is such an action at all. Such a mechanism, I think, can well be called a choice-mechanism. Hence, there is nothing mysterious in using the notion of choice in this context.

To bring this out a little bit more I would like to sketch a possible mechanism of this kind. Imagine a system *S* which is able to exhibit three

different kinds of movement *a*, *b*, and *c*. Let us assume that this system has an internal model of its environment (I shall say something about what internal models presumably are later on) and let us further assume that it is able to produce for any given model *M* and for any of its possible movements *a*, *b* and *c* an effect-model $M^x$, which would be a model of the changed environment of *S*, which in turn would be the result of its carrying out *x*. Now, a simple choice mechanism could work as follows: In a given situation, which is represented in *S* by the model *M*, the mechanism first checks whether *M* is the model of a situation in which the goal-state *G* of *S* is realized. If so it causes *S* to do nothing, if not it produces the effect-model $M^a$. It checks whether $M^a$ is the model of a situation in which *G* is realized. If so it causes *a* to be carried out, if not it produces the effect-model $M^b$. It checks whether $M^b$ is the model of a situation in which *G* is realized. If so it causes *b* to be carried out, if not it produces the effect-model $M^c$. Finally it checks whether $M^c$ is the model of a situation in which *G* is realized. If so it causes *c* to be carried out, if not it stops and thereby causes *S* to do nothing. It is easy to see that this mechanism does what it is supposed to do. If there is a behavior which will lead to the desired goal-state *G* it will bring it about that *S* exhibits this behavior. If not it will bring it about that *S* does nothing. And even that is what one would expect a reasonable system to do which has the want that *G*.

Obviously, the mechanism just sketched is nothing else than the realization of a rudimentary form of a well known method in the field of mechanical problem-solving: the traversing of a search-tree in order to find a path to the solution of a problem. And this being so, it would not be difficult to enlarge this mechanism to enable it to check not only single actions but whole series of actions to find out whether they would lead to a given goal-state. But I do not want to labor this point further. For in the present context my aim is only to make clear that the having of genuine wants implies the capacity to make plans or at least the capacity to evalutate possible actions with regard to whether in the given circumstances they will bring about a certain (desired) state of affairs. And to make clear that tropistic systems are not the right kind of things to have genuine wants since they completely lack these capacities.

I think that the reason for drawing the dividing line just here can also be brought out by the following consideration. If we have a system *A* with an internal choice-mechanism that ensures that in any given situation *A* exhibits one of those actions that lead to a certain goal-state *G* – if there are any – then the behavior of this system can be explained by the law-like sentence

(4) In any situation *A* exhibits a behavior that brings about *G*, provided that there is such a behavior.

And this law-like sentence ascribes an intentional (want-like) state to *A* since it says in effect that *A* has a disposition which is characterized by reference to a certain state of affairs. Hence, this state has a *content*, namely, the goal-state *G*.

On the other hand the behavior e. g. of system $S'_1$ cannot be explained by the law-like sentence

(5) In any situation $S'_1$ exhibits a behavior that brings about that the front of $S'_1$ is facing the brightest part of its environment – provided that there is such a behaviour,

but rather by the three lawlike sentences

(6) a. If the brightest part of its environment is at the right side of $S'_1$, $S'_1$ moves its wheels in one direction.
b. If the brightest part of its environment is at the left side of $S'_1$, $S'_1$ moves its wheels in the other direction.
c. If the brightest part of its environment is just in front of $S'_1$, $S'_1$ does nothing,

since it is by these three statements that we are able to explain what $S'_1$ does in the two above mentioned queer situations – and in all other situations as well – and not by the statement (5). And this, I think, is the reason why, in the end, we are not entitled to ascribe to $S'_1$ the intentional (want-like) disposition that corresponds to the lawlike sentence (5).[1]

## 3.

After these remarks on why tropistic systems do not have genuine wants, let us now turn to the second part of my thesis, the claim that tropistic systems do not have genuine beliefs either. One of the key concepts which have a special significance with regard to this claim has already been mentioned – the concept of an „internal model“ or an „internal representation“. It is not the only important concept in this context. But I shall restrict my

[1] Now, this certainly is a little bit of cheating. For even if all systems *without* choice mechanisms are bound to make mistakes, systems *with* choice mechanisms also will make mistakes now and then. For absolutely fool-proof choice mechanisms, which never make any mistakes of any kind, will be at least very rare (in fact they are, as I think, impossible). So, the observable behavior of a system may not be decisive. And this leads to the conclusion that the real difference lies in the way the behavior of a system is produced – by means of an internal choice mechanism or by some other device (like a simple feedback mechanism or whatever). In this respect my account of wanting is not purely functionalistic, but has some implications as to the internal structure of a system.

considerations to the claim that tropistic systems have no internal representations of their environment.

If again we take system $S_1$ as an example of a tropistic system it might seem – at first sight – that this claim is plainly false and that on the contrary there are good reasons for attributing such internal representations to the system, since it is equipped with a subsystem (the circle of photo-electric cells on the top of $S_1$) whose states might well be interpreted in this way. For, within the limits imposed by the number and the sensitivity of the cells, each state of this circle stands in an obvious one-to-one relation to a certain state of the environment. In the same way that the above mentioned state in which cell #3 has a considerable output while all other cells remain silent corresponds to the environmental state „bright light at an angle of 45 degrees in front of $S_1$", the internal state „cell #6 high output, all other cells low output" corresponds to the environmental state „bright light exactly on the right-hand side of $S_1$", and the internal state „cell #12 high output, all other cells low output" correponds to the environmental state „bright light exactly behind $S_1$".

Moreover, the states of the circle of photo-electric cells have a feature which any system of representing states should have. *They change in rather perfect congruence with corresponding changes in the environment*. If a bright light moves steadily from the position shown in figure 1 to a position behind system $S_1$ the states of the circle change correspondingly, e. g. from the state „cell #3 high output, all other cells low output" to the state „cell #4 high output, all other cells low output" until finally state „cell #12 high output, all other cells low output" is reached. So there is a kind of isomorphism between the states of the circle and certain states of the environment. And this isomorphism also holds if system $S_1$ itself moves, since then the states of the circle also change in an appropriate way. This is certainly necessary. For what the states of the circle stand for – if they stand for anything at all – is something that can only be expressed in system-relative coordinates. Hence, summing up we can say that there exists a set of internal states of system $S_1$ such that there is an isomorphism between the elements of this set and certain states in the environment of $S_1$. That is to say, for any state *e* of these environmental states there exists a corresponding internal state *r* with the property that $S_1$ is in the state *r* if

and only if the environment is in state *e*. Why not say that the internal state *r* is a representation of the environmental state *e*?[2]

In my opinion the answer to this question is simply that we do not need to explain the behavior of system $S_1$ by reference to any internal representtations whatever. And this, in turn, is simply due to the fact that the states of $S_1$ which could possibly count as representing certain states of the environment *occur only if and as long as they are caused by these environmental states*.[3] If one wants to explain a certain piece of behavior of $S_1$ by reference to an allegedly representing state *r* of $S_1$, one can explain this piece of behavior just as well – or perhaps even better – by reference to the

---

[2] Compare e.g. F. Dretske 1981. In ch. 7 of this book Dretske defines „semantic content" in the following way:

(1) Structure *S* has the fact that *t* is *F* as its *semantic content* iff *S* carries the information that *t* is *F* in digital form. (p. 177)

And before that, in ch. 6, he writes:

(2) Structure *S* carries the information that *t* is *F in digital form* iff *S* carries no additional information about *t*, no information that is not already nested in *t's* being *F*. (p. 137)

From these two definitions one can derive, as it seems to me:

(3) Structure *S* has the fact that *t* is *F* as its *semantic content* iff it is true that *S* occurs if and only if *t* is *F*.

At least, this is possible if we interpret (2) in the following way:

(2′) Structure *S* carries the information that *t* is *F in digital form* iff the following two conditions are satisfied:

(a) if *S* is the case then *t* is *F*
(b) there is no $F' \neq F$ with: if *S* is the case then *t* is $F'$, and if *t* is $F'$ then *t* is *F*.

It must be noted here, however, that Dretske – in order to account for the possibility of false beliefs – does *not* identify beliefs with states having a certain semantic content.

[3] This already comes out if we ask again what exactly the environmental state is which e. g. the internal state „cell #6 high output, all other cells low output" stands for. In the penultimate paragraph I said that this is the state „bright light on the righthand side of $S_1$". But strictly speaking this is not true. For if we put a screen or a wall or some other nontransparent object between the light and the system $S_1$, $S_1$ will no longer be in the mentioned state. And therefore it is simply false to say that $S_1$ is in this state if and only if there is a bright light on its righthand side. What one could say at most therefore seems to be that $S_1$ is in the state „cell #6 high output, all other cells low output" if and only if there is an *unscreened* bright light on its righthand side. And the „unscreened" here is best understood figuratively as „blocking the causal impact of the light on the system $S_1$".

state which *r* is said to represent. E. g., if we want to explain a certain behavior of $S_1$ by the fact that $S_1$ is in the internal state „cell #6 high output, all other cells low output" we could as well or even better say that $S_1$ behaved the way it did because there was a bright light on the right-hand side of $S_1$.

So there is one negative lesson to be learned here about internal representations. No internal state *r* of a system *S* will count as representing an environmental state *e* if *r* occurs only if and as long as it is caused by *e*. *Internal representations must have a certain degree of causal independence from the states they represent*. More positively, this can be illustrated by a slightly modified version of $S_1$, let us call it $S_2$. Imagine the following situation. There is a bright light in the neighbourhood of $S_1$ which is not standing still, but moving slowly in a certain direction. In this situation system $S_1$ will, according to its construction, follow the light as if it were trying to reach it. But imagine further that now for a short while the light disappears behind a screen or a wall or some other kind of non-transparent object. If we assume that it is rather dark in the neighbourhood of $S_1$ while the light is behind the screen, $S_1$ clearly will stop and do nothing until the light appears again. And then it will turn to the new direction and move on.

Now let us suppose that the modified version of $S_1$, $S_2$, has a rather different construction according to which things happen as follows. If the target of $S_2$ – e. g. a light – disappears for a while behind a screen, then $S_2$ by means of some internal mechanism extrapolates the previous course of the target and by this means calculates a probable position of the target. Then $S_2$ moves as if there were a bright light at this calculated position. Obviously, $S_2$ will move the whole time as $S_1$ would move if there were no screen. And this means that $S_2$ – given that its calculations are correct – will follow its target even if it is not able to „see" it; better: even if the target has no causal impact on system $S_2$.

For this reason the behavior of $S_2$ cannot be explained by the target alone. It is true: if the calculations of $S_2$ are correct, the system of internal states, which corresponds to the various positions of the light, is the whole time in perfect accordance with the real position of the light. But this is not due to the fact that the system of these states is the whole time appropriately changing under the causal influence of the moving light. Rather the updating is – at least sometimes – accomplished by an internal mechanism of $S_2$ itself. Accordingly, it is not possible to explain the behavior of $S_2$ solely by reference to real position of the light. For while the light is behind the screen it has no causal effect on $S_2$ whatsoever. At least during this period of time $S_2$ is therefore following its own compass, i. e. at least during this time its behavior is effected solely by its own states one of which

seems really to be a representation of the actual position of the light behind the screen.

I think there are two general points that must be noted here. First, it is not possible to explain the behavior of $S_2$ entirely by reference to certain features of its environment. This corresponds to the fact that the behavior of $S_2$ is non-uniform in the sense that the system behaves differently in situations of the same kind, i. e. situations which are qualitatively the same for the system itself, which have the same causal impact on the system. If e. g. two lights with different trajectories disappear for a while we will have darkness in both cases and that means the two situations are qualitatively the same for $S_2$; but $S_2$ will behave differently in these two situations according to its different calculations of the paths of the lights. Therefore, if we want to explain the behavior of $S_2$ at all, we must take into account the internal states of the system. This, I think, really is a very general point. We must ascribe different internal states to a system *S* if its behavior is non-uniform (but not random) in the sense that the system behaves differently in situations which have the same causal impact on it. Second, the internal states which are in fact responsible for the behavior of $S_2$ belong to a system of states which – partly by internal, partly by external causes – change in such a way that these states of $S_2$ are the whole time in exact correspondence to certain states of its environment. In my opinion, these two features make system $S_2$ a very plausible candidate for the ascription of internal representations.

To put it in more general terms: seen from outside, we are entitled to ascribe internal representations to a system if the system behaves in certain situations *as if* its behavior were caused by a certain state of its environment while in fact the behavior is caused by an internal state of the system. Seen from within, we are entitled to ascribe internal representations to a system if its behavior is caused by a set of internal states which (at least normally) change in perfect accordance with corresponding changes in the system's environment and if the changes of these states of the system are not simply caused by the corresponding states of the environment themselves but are – at least in part – due to an *internal updating mechanism*. These two criteria, I think, are by and large two sides of the same coin.

These considerations bear a certain similarity to the considerations presented by J. A. Fodor in his recent paper „Why Paramecia Don't Have Mental Representations", in which he tries to develop a criterion for having mental representations and thereby also to give a necessary condition for being an intentional system. Since, according to Fodor, a system without mental representations *a fortiori* cannot be a genuine intentional system, his main suggestion in this paper is that the fundamental difference between systems with and systems without mental representations lies in the fact

that systems with mental representations are able to respond selectively to non-nomic properties of objects in their environment whereas systems without mental representations lack this ability. At least

> … *any* system that can respond selectively to non-nomic properties is, intuitively speaking, a plausible candidate for the ascription of mental representations; and any system that can't, isn't. (Fodor 1986, 11)

Obviously, this suggestion presupposes that we can make sense of the idea that there are non-nomic properties, i. e. properties which do not enter in any lawful relations. But if we grant the existence of such properties here for the sake of argument, the rationale of Fodor's proposal seems to be this. Even some tropistic systems are able to respond selectively to some specific features of their environment, say e. g. to the fact that an object *O* in their neighbourhood has a certain property *P*. Let us assume that *S* is such a system. How then has the selective behavior of *S* been brought about? In general the answer to this question will be: *O*'s having *P* causes a certain state in *S* to occur and this state in turn causes the selective behavior of *S* or, in more simple cases, *O*'s having *P* directly causes the behavior of *S*. Now, if *P* is a non-nomic property then *O*'s having *P per definitionem* does not cause anything. How then can systems react selectively to non-nomic properties at all? According to Fodor there is only one possibility: the behavior of the system has to be effected by *S's representing O as P*. In these cases it is not *O*'s having *P* that causes – directly or indirectly – the behavior of *S*, but a state of *S* which is a representation of *O*'s having *P*. Fodor's argument, therefore, can be summed up in the following way. Every system which is able to react selectively to non-nomic stimuli must possess mental representations. For non-nomic stimuli themselves cannot be causes of behavior, so there must be intermediate states, representations of the stimuli, which take over this role.

Let us take a closer look at this argument. If a system *S* responds selectively to a certain non-nomic property *P*, then this system behaves differently in situations with and situations without an object with the property *P* (short: *P*-object) in its environment. That is to say, it shows a certain behavior *B* if and only if there is a *P*-object in its neighbourhood, or in more general terms: if and only if a *P*-object is standing in a certain relation to it. But, since *P* is a non-nomic property, *S*'s doing *B* cannot be caused by the fact that there is a *P*-object in the neighbourhood of *S*. And since obviously *S*'s doing *B* cannot be caused by any other feature of its environment either (otherwise it could not count as a selective response to the property *P*), it must be caused by some internal state of *S*. Let us call this state *Z*. Now, since *S* does *B* every time it is in state *Z* and since *S* does *B* if and only if there is a *P*-object in its environment, there is after all an internal state *Z* of *S* such that *S* is in *Z* if and only if there is a *P*-object in

the environment of *S*. (If doing *B* also can be caused by another internal state *Z′* of *S* we can choose the disjunctive state „*Z* or *Z′*“ instead of *Z* alone.) But, since *P* is non-nomic not even *S*'s being in state *Z* can be caused by there being a *P*-object present. Hence, according to Fodor, we are forced to say that a certain system *S* has internal (mental) representations of its environment if, first, *S* has an internal state *Z* which corresponds perfectly to there being a *P*-object in its environment, if, second, being in state *Z* is the cause of *S*'s doing *B*, and if, finally, *S*'s being in state *Z* itself is not caused by there being a *P*-object in the environment.

So, there is a rather similar structure between Fodor's paradigm case for the ascription of internal representations and the conclusions drawn from the example of system $S'_1$. But there remains a difference in that Fodor requires the property *P* which is represented by state *Z* to be a non-nomic property. In my eyes, however, that is asking to much. For, as I see it, the kind of argument that Fodor puts forward applies already to cases in which a system behaves *as if* this behavior were caused by a certain state of the environment while *in fact* it is not caused this way even if the state of the environment does not consist in the presence of an object having a non-nomic property. That is to say, the basic criterion is not that a behavior which constitutes a selective response to a specific feature of the environment *cannot* be caused by this feature but that it is *in fact* not caused by it.

## 4.

It is time to give a short résumé. My claim was that there are good reasons for not ascribing genuine wants and beliefs to tropistic systems and hence for not reckoning these systems among the genuine intentional ones. These reasons can be summarized in the following way.

We should not ascribe genuine beliefs to tropistic systems because they have no internal representations of their environment. And they do not have such internal representations because the only feasible internal states of these systems do not have the necessary causal independence of the environment. For they occur only if and as long as they are caused by the environmental states they presumably stand for. It is this fact which also has the consequence that in explaining the behavior of tropistic systems we are not forced to take internal representations into account. For everything that can be explained by allegedly representing states can as well – or even better – be explained by the corresponding environmental states.

The reasons for not ascribing wants to tropistic systems are in a way very similar. For on the one hand it is true that tropistic systems often show a kind of goal-directed behavior. But on the other hand everything that a tropistic system does is caused by certain states of its environment. As a

consequence of this fact a tropistic system can easily be ‚tricked' by its environment into exhibiting a certain behavior even in situations in which this behavior has no chance to bring about the „desired" goal-state. Hence, tropistic systems have no genuine wants because they do not choose their behavior in virtue of its being apt to bring about a certain goal-state. Or to put it in more general terms: Because their behavior is not caused by an internal mechanism which ensures that the behavior is exhibited because under the given circumstances it is apt to lead to the „desired" result.

These formulations, however, force me to add a last note. I have deliberately tried to avoid the problems of misrepresentation and miscalculation[4] even though they have lurked around every corner. For it is certainly not true that an internal state *r* can represent an external state *e* only if *r* occurs if and only if *e* occurs. And it is certainly not true that a choice-mechanism which evaluates certain actions with respect to their being apt to bring about a certain result can be called a respective choice-mechanism only if it never makes mistakes. But a thorough discussion of these problems would have lead us to far away from my original claim, the claim that tropistic systems have no genuine wants and beliefs.

## References

Dretske, F.I. (1981) *Knowledge and the Flow of Information*. Oxford.

Fodor, J.A. (1986) „Why Paramecia Don't Have Mental Representations". *Midwest Studies in Philosophy* 10, 3–23.

[4] Except for the few remarks in note 1 above.

# 10

# Gibt es ein Problem der Intentionalität*

## 1.

Der Kern des Leib-Seele-Problems besteht darin, dass mentale Phänomene (Ereignisse, Eigenschaften, Zustände) Merkmale zu haben *scheinen*, die es auf den ersten Blick unmöglich machen, diese Phänomene in ein naturalistisches Weltbild zu integrieren – sie mit physikalischen Phänomenen zu identifizieren oder auf physikalische Phänomene zu reduzieren.[1] Heute stehen hauptsächlich zwei von in diesem Sinne kritischen Merkmalen im Mittelpunkt des Interesses.[2] Das erste ist das Merkmal intentionaler Zustände, einen *repräsentationalen* oder *semantischen Inhalt* zu besitzen. Das Problem der Naturalisierung dieser Zustände möchte ich das ‚*Problem der Intentionalität*' nennen. Das zweite kritische Merkmal ist die Eigenschaft von Wahrnehmungseindrücken und körperlichen Empfindungen, einen *qualitativen Aspekt* zu besitzen. Dieses Merkmal beruht auf der Tatsache, dass es auf eine bestimmte Weise ist oder sich auf eine bestimmte Weise anfühlt, in diesen Zuständen zu sein. Das Problem der Naturalisierung dieser Zustände wird gewöhnlich das ‚*Qualia-Problem*' genannt.

In diesem Aufsatz werde ich mich auf das Problem der Intentionalität konzentrieren und für die These argumentieren, dass dieses Problem ein *Artefakt* ist. Ich werde zu zeigen versuchen, dass intentionale Zustände gar kein mysteriöses Merkmal besitzen, das ihre Naturalisierung problematisch machen könnte. Dies *scheint* nur der Fall zu sein. Und dieser Anschein ergibt sich aus der Art und Weise, wie wir über intentionale Zustände reden; d.h. genauer: Er ergibt sich aus dem *Vokabular*, das wir verwenden, um Menschen und manchen Tieren intentionale Zustände zuzuschreiben. Meine These ist, dass dieses Vokabular *messtheoretisch* interpretiert werden sollte, analog zum Vokabular metrischer Begriffe, mit dem wir Gegenstän-

---

* Erstveröffentlichung in: U. Haas-Spohn (Hg.) *Intentionalität zwischen Subjektivität und Weltbezug*. Paderborn: mentis 2003, 19–44. Bei diesem Aufsatz handelt es sich um die deutsche Fassung von Beckermann (1996b). Bei der Übertragung vom Englischen ins Deutsche war mir Stefanie Becker eine große Hilfe.

1 Zur Frage, wie ‚Reduktion' in diesem Zusammenhang zu verstehen ist, vgl. Beckermann (1992a, 1992b, 1996a, 1997, 2001). Vgl. in diesem Band auch die Beiträge 2 und 3.

2 Eine detailliertere Darstellung der charakteristischen Merkmale intentionaler Zustände findet sich in Beckermann (2001, S. 13–17, 267–273).

den physikalische Größen wie Länge, Masse und Geschwindigkeit zuschreiben. Wenn diese Interpretation richtig ist, gibt es aber keinen Grund mehr für die Annahme, dass intentionale Zustände ein besonderes, mysteriöses Merkmal besitzen, das gewissermaßen die Grundlage für die Verwendung intentionaler Prädikate bildet. Dies anzunehmen wäre vergleichbar mit der Behauptung, dass wir physikalischen Gegenständen Größen mit Prädikaten wie ‚hat eine Masse von 2 kg' zuschreiben, weil diese Eigenschaften das mysteriöse Merkmal besitzen, einen *numerischen Inhalt* zu haben. Der messtheoretische Ansatz hat den großen Vorteil, uns die Augen dafür zu öffnen, dass wir intentionales Vokabular bei der Zuschreibung bestimmter mentaler Ereignisse nicht deshalb verwenden, *weil* diese Zustände die (mysteriöse) Eigenschaft haben, einen semantischen Inhalt zu besitzen. Vielmehr gilt im Gegenteil: Intentionale Zustände haben nur deshalb einen semantischen Inhalt (wenn es eine solche Eigenschaft überhaupt gibt), weil und insofern wir sie mit Hilfe intentionaler Prädikate zuschreiben.[3]

Auf diese These werde ich gleich zurückkommen. Zuvor jedoch noch einige allgemeine Bemerkungen über den Charakter des Problems der Intentionalität.

## 2.

In prägnanter Weise wurde dieses Problem zum ersten Mal von Franz Brentano in der folgenden berühmten Passage formuliert:

> Jedes psychische Phänomen ist durch das charakterisiert, was die Scholastiker des Mittelalters die intentionale … Inexistenz[4] eines Gegenstandes genannt haben, und das wir, obwohl mit nicht ganz unzweideutigen Ausdrücken, die Beziehung auf einen Inhalt, die Richtung auf ein Objekt …, oder die immanente Gegenständlichkeit nennen würden. In der Vorstellung ist etwas vorgestellt, in dem Urteile etwas anerkannt oder verworfen, in der Liebe geliebt, in dem Hasse gehaßt, in dem Begehren begehrt usw. Diese intentionale Inexistenz ist den psychischen Phänomenen ausschließlich eigentümlich. Kein physisches Phänomen zeigt etwas Ähnliches. Und somit können wir die psychischen Phänomene definieren, indem wir sagen, sie seien solche Phänomene, welche intentional einen Gegenstand in sich enthalten. (Brentano 1924, S. 124 f.)

Diese Formulierungen legen die Auffassung nahe, dass ein Phänomen genau dann intentional ist, wenn es in einer bestimmten Weise auf einen Gegenstand bezogen ist bzw. wenn es etwas in einer spezifischen, nicht-

---

[3] Zu diesem Thema werde ich noch mehr unten in Abschnitt 8 sagen.

[4] Der Ausdruck ‚Inexistenz' bedeutet hier etwa dasselbe wie ‚Enthaltensein'; er ist also nicht im Sinne von ‚Nichtexistenz' zu verstehen.

räumlichen Weise als Objekt in sich enthält. Heute wird das allgemein etwas anders gesehen. Heute besteht ein weitgehender Konsens darüber, dass die Intentionalität von Überzeugungen, Wünschen, etc. nicht in einer kaum verständlichen Beziehung zwischen diesen Zuständen und (möglicherweise gar nicht existierenden) Gegenständen besteht, sondern darin, dass diese Zustände semantisch bewertbar sind, dass sie Wahrheits- oder Erfüllungsbedingungen besitzen. Akzeptiert man diese Darstellung, so scheint es plausibel, Brentanos These folgendermaßen umzuformulieren:

> BRENTANOS THESE (1. Fassung)
> Intentionale Zustände zeichnen sich dadurch aus, dass sie in dem Sinn einen semantischen Inhalt haben, dass sie Wahrheits- oder Erfüllungsbedingungen besitzen. Physikalische Zustände können in diesem Sinn keinen semantischen Inhalt haben. Daher sind intentionale Zustände grundsätzlich von physischen Zuständen unterschieden.

In dieser Fassung wäre Brentanos These allerdings äußerst unplausibel. Denn warum sollte es der Fall sein, dass physikalische Zustände keinen semantischen Inhalt bzw. keine Wahrheits- oder Erfüllungsbedingungen haben können? Immerhin wird von niemandem bestritten, dass sprachliche Äußerungen einen semantischen Inhalt haben. Und sprachliche Äußerungen sind – ganz gleich, ob sie nun als Muster von Schallwellen oder als Schriftzüge auf dem Papier realisiert sind – zuerst einmal physikalische Phänomene. In dieser allgemeinen Form ist die These, dass physikalische Phänomene *per se* keine Wahrheits- oder Erfüllungsbedingungen haben können, daher sicher nicht haltbar.

Doch das ist auch gar nicht entscheidend. Die zentrale Frage in diesem Zusammenhang ist nicht, ob physikalische Phänomene einen semantischen Inhalt haben können, sondern ob die *Eigenschaft*, einen semantischen Inhalt (Wahrheits- oder Erfüllungsbedingungen) zu haben, im Rahmen eines naturalistischen Weltbilds expliziert werden kann oder ob sie diesen Rahmen sprengt. Im Hinblick auf diese Frage scheint Brentano die Auffassung vertreten zu haben, dass die Eigenschaft, einen semantischen Inhalt zu haben, über den physikalischen Bereich hinausgeht. Daher kann seine These am adäquatesten wohl so formuliert werden:

> BRENTANOS THESE (2. Fassung)
> Intentionale Zustände zeichnen sich dadurch aus, dass sie einen semantischen Inhalt (Wahrheits- oder Erfüllungsbedingungen) haben. Die *Eigenschaft, einen semantischen Inhalt zu haben*, kann jedoch nicht naturalisiert werden, d.h. sie kann nicht innerhalb eines ausschließlich physikalischen Weltbildes expliziert werden. Intentionale Zustände sind daher Zustände, die über den Bereich des Physikalischen hinausgehen.

Natürlich steckt in dieser These für jeden naturalistisch gesinnten Philosophen eine Herausforderung – zumindest wenn er oder sie nicht geneigt ist, sich dem Eliminativismus anzuschließen, um so dem ganzen Problem der Intentionalität aus dem Wege zu gehen. Intentionalität im Sinne der These Brentanos gehört sicher nicht zu den grundlegenden und irreduziblen Eigenschaften des Universums wie etwa Masse und Ladung. Folglich kann man als naturalistisch gesinnter Philosoph nur dann ein Realist bezüglich intentionaler Zustände sein, wenn man gleichzeitig Reduktionist ist.[5] Das Problem, das durch Brentanos These aufgeworfen wird, d. h. das Problem der Intentionalität, wie es gewöhnlich verstanden wird, kann daher folgendermaßen formuliert werden:

> PROBLEM DER INTENTIONALITÄT
>
> Ist es nicht doch möglich, die Eigenschaft intentionaler Zustände, einen semantischen Inhalt zu haben, zu naturalisieren? Kann die Eigenschaft, einen semantischen Inhalt zu haben, nicht doch mit einer physikalischen Eigenschaft identifiziert oder auf eine physikalische Eigenschaft reduziert werden?

## 3.

In den vergangenen Jahren haben viele analytische Philosophinnen und Philosophen versucht, dieses Problem zu lösen. Man kann fast sagen, dass dieses Thema die Debatten in der Philosophie des Geistes eine Zeit lang völlig beherrscht hat. Dennoch gibt es bis jetzt keinen Lösungsvorschlag, der auch nur von einer bescheidenen Mehrheit akzeptiert würde. Stattdessen existiert eine Vielzahl von – teilweise höchst unterschiedlichen – Ansätzen: Da ist Dretskes informationstheoretisch inspirierter Ansatz; dann gibt es Fodors Theorie, in der die Idee asymmetrisch voneinander abhängender Kausalrelationen eine zentrale Rolle spielt; und schließlich gibt es den funktional-teleologischen Ansatz, wie er besonders von Millikan und Papineau favorisiert wird.[6] Wenn man alle diese Lösungsvorschläge miteinander vergleicht, werden zwei Dinge deutlich: 1. Bei der Zuschreibung

---

[5] Vgl. Fodor (1987, S. 97).

[6] Vgl. bes. Dretske (1981, 1986), Fodor (1987, 1991), Millikan (1984, 1989) und Papineau (1985, 1988). Eine Zusammenfassung dieser Theorien findet sich auch in Beckermann (2001, S. 334–357). Auf den ersten Blick scheint der interpretationstheoretische Ansatz von Haugeland und Cummins in dieser Aufzählung zu fehlen. Dieser Ansatz ist den messtheoretischen Überlegungen, auf die ich im Folgenden noch ausführlich eingehen werde, jedoch ähnlicher als den angeführten klassischen Versuchen, Intentionalität zu naturalisieren.

semantischer Inhalte[7] spielt offenbar eine Vielzahl von Faktoren eine wichtige Rolle, wobei die Beiträge der einzelnen Faktoren nicht klar voneinander getrennt werden können. 2. Es scheint gar keine Kriterien zu geben, mit Hilfe deren wir entscheiden könnten, welche der vielen Lösungsvorschläge der richtige ist. Anders ausgedrückt: Jeder Vorschlag ist bis zu einem gewissen Grad plausibel; aber alle erwecken immer auch den Eindruck einer gewissen Beliebigkeit.

Spätestens nach dieser Diagnose sollte sich der Verdacht einstellen, dass die Frage, ob und wie die Eigenschaft, einen semantischen Inhalt zu besitzen, naturalisiert werden kann, falsch gestellt ist. Aber was genau ist falsch an dieser Frage? Wo liegt der allen genannten Ansätzen gemeinsame Fehler?

Meiner Meinung nach beruht dieser Fehler auf zwei Annahmen. Die erste Annahme ist, dass wir bei der Zuschreibung intentionaler Zustände ‚dass'-Sätze verwenden, *weil* diese Zustände selbst – ganz unabhängig davon, wie wir sie zuschreiben – die mysteriöse Eigenschaft besitzen, einen semantischen Inhalt zu haben. Und die zweite Annahme ist, dass wir in der Lage sein müssen, diese mysteriöse Eigenschaft zu naturalisieren, wenn wir im Hinblick auf intentionale Zustände Naturalisten bleiben wollen. Ich denke, dass besonders die erste Annahme auf einem grundlegenden Missverständnis beruht. Allerdings handelt es sich um ein durchaus naheliegendes Missverständnis. Wenn man etwa sagt, *S* glaubt, dass Hunde bellen, dann scheint man damit erstens zu sagen, dass *S* in einem Zustand ist, der zu einem bestimmten Typ intentionaler Zustände – zum Typ der Überzeugungen – gehört, und man scheint zweitens zu sagen, dass sich dieser Zustand von anderen Zuständen desselben Typs, d. h. von anderen Überzeugungen, durch seinen spezifischen Inhalt unterscheidet – nämlich den Inhalt, dass Hunde bellen. Wie sollte ein Zustand die Überzeugung sein, dass Hunde bellen, wenn er nicht die Eigenschaft hat, diesen Inhalt zu haben?

Aber diese Argumentation ist vorschnell: Denn aus der Tatsache, dass wir bei der *Zuschreibung* intentionaler Zustände Ausdrücke verwenden wie ‚hat die Überzeugung, dass *p*', folgt für sich genommen sicher nicht, dass die Zustände, die wir auf diese Weise zuschreiben, die mysteriöse Eigen-

---

[7] Streng genommen ist der Ausdruck ‚*Zuschreibung* von Inhalten' im Kontext der Theorien von Dretske, Fodor und Millikan nicht adäquat, weil in diesen Theorien vorausgesetzt wird, dass Inhalte von uns nicht nur *zugeschrieben* werden, sondern dass Inhalte reale intrinsische Eigenschaften realer intentionaler Zustände sind. Im Gegensatz dazu gehört der messtheoretische Ansatz, den ich im Folgenden erläutern werde, zur Gruppe von Theorien, in denen angenommen wird, dass Inhaltszuschreibungen nicht durch objektive Tatsachen festgelegt sind, sondern auf Interpretationen beruhen. (Diesen Punkt verdanke ich Peter Lanz.)

schaft aufweisen, den semantischen Inhalt *p* zu haben. In den letzten Jahren ist dieser Punkt besonders von Paul Churchland betont worden, der – ähnlich wie Field, Stalnaker, Davidson[8] und andere – die These verteidigt hat, dass das Vokabular, mit dem wir intentionale Zustände zuschreiben, analog zum Vokabular metrischer Begriffe aufgefasst werden sollte, mit dessen Hilfe wir physikalische Größen wie Länge, Gewicht oder Temperatur zuschreiben. ‚Dass'-Sätze spielen dieser Auffassung zufolge also in etwa dieselbe Rolle wie Zahlausdrücke im Vokabular metrischer Begriffe. Sicher sollte man mit dieser Analogie vorsichtig sein, da sie dazu verleiten könnte, die Unterschiede zwischen den beiden Begriffsfamilien zu verwischen. Aber im Prinzip ist sie richtig. Wenn man das Vokabular verstehen will, das wir bei der Zuschreibung intentionaler Zustände verwenden, sollte man daher am besten mit einer Analyse des Vokabulars metrischer Begriffe beginnen.

Ausdrücke wie ‚*x* hat eine Länge von *y* cm', ‚*x* hat eine Masse von *y* kg' und ‚*x* hat eine Temperatur von *y* Grad Celsius' sehen auf den ersten Blick so aus, als würden sie Relationen zwischen raumzeitlichen Gegenständen und Zahlen ausdrücken. Mit anderen Worten: Diese Ausdrücke scheinen die logische Form *xRy* zu haben, wobei *R* für ein zweistelliges Prädikat und *x* und *y* für Gegenstandsbezeichner stehen, die sich auf raumzeitliche Gegenstände bzw. auf Zahlen beziehen. Bei genauerem Hinsehen ergibt sich jedoch noch eine andere Möglichkeit. Ausdrücke wie ‚*x* hat eine Länge von *y* cm' können nämlich auch die Form $F_y x$ haben, wobei *F* für einen Operator steht, der – angewendet auf einen geeigneten Indexausdruck *y* (hier einen Ausdruck für eine positive reelle Zahl) – ein einstelliges Prädikat $F_y$ erzeugt.[9] (Solche Prädikate nenne ich im Folgenden ‚*Operator-Index-Prädikate*'.) Wenn man sagt ‚Würfel *a* hat eine Masse von 2 kg', sagt man dieser Lesart zufolge nicht, dass *a* in einer bestimmten Relation zur Zahl 2 steht, sondern dass *a* eine bestimmte monadische Eigenschaft besitzt – eine reale Eigenschaft, die kausal relevant ist. Wenn wir bei der Zuschreibung dieser Eigenschaft einen Ausdruck verwenden, der einen Ausdruck für die Zahl 2 enthält, bedeutet das also weder, dass es sich hier um eine relationa-

---

[8] Vgl. Churchland (1979, S. 100–107), Field (1980, S. 114), Stalnaker (1984, S. 9 ff.), Davidson (1974, S. 147, 1989, S. 9 ff.). Vgl. auch Dennett (1982, S. 123 ff., 1987, S. 208) und Matthews (1990). Eine Erörterung des messtheoretischen Ansatzes findet sich in Lanz (1987, S. 95–127). Field scheint einer der ersten gewesen zu sein, die behauptet haben, dass die messtheoretische Interpretation intentionaler Prädikate vielleicht eine Lösung des Brentanoschen Problems darstellen könnte (1980, S. 114).

[9] Dieser Ansatz findet sich im Detail in P. Churchland (1979, S. 100 ff.). Meines Wissens wird diese Möglichkeit zum ersten Mal von Quine erwähnt (1970, Abschnitt 2.1).

le Eigenschaft handelt, noch dass wir mit diesem Ausdruck eine Eigenschaft zuschreiben, die das mysteriöse Merkmal hat, einen bestimmten numerischen Inhalt – nämlich den Inhalt 2 – zu besitzen.[10] Wenn das so ist, ergibt sich allerdings die Frage, warum wir bei der Zuschreibung von Eigenschaften wie Masse, Länge oder Temperatur überhaupt Prädikate verwenden, die Zahlwörter als wesentlichen Bestandteil enthalten.

In diesem Zusammenhang wird üblicherweise auf die grundlegenden Überlegungen der Messtheorie verwiesen, wie sie besonders von Hempel (1952) und Suppes und Zinnes (1963) formuliert worden sind. Diese Überlegungen können am Beispiel des Begriffs der Masse wie folgt zusammengefasst werden. Dafür dass der Begriff der Masse metrisiert werden kann, sind zwei Gründe verantwortlich: 1. Auf der Menge *D* der Gegenstände, auf die wir diesen Begriff anwenden können, existiert eine zweistellige empirische Relation *H* (konkret: die Relation, die zwischen zwei Gegenständen *a* und *b* genau dann besteht, wenn *b a* überwiegt oder aufwiegt, falls *a* und *b* auf gegenüberliegende Waagschalen gelegt werden) und eine ebenfalls zweistellige additive Operation o. 2. Das empirische relationale System $<D, H, \circ>$ lässt sich strukturerhaltend auf das abstrakte relationale System $<\mathbf{R}^+, \leq, +>$ abbilden, d. h. es gibt einen Homomorphismus *m* von der Menge *D* auf die Menge der positiven reellen Zahlen mit:

(1) $aHb$ genau dann, wenn $m(a) \leq m(b)$

(2) $m(a \circ b) = m(a) + m(b)$.[11]

Das hat folgende Konsequenz: Wenn wir den Gegenständen von *D* Massen mit Hilfe von Operator-Index-Prädikaten zuschreiben, die als Indizes Ausdrücke für genau die Zahlen enthalten, die der Homomorphismus *m* diesen Gegenständen zuweist, können wir von den Prädikaten direkt able-

[10] Dass *a*'s Eigenschaft, eine Masse von 2 kg zu haben, nicht in *a*'s Relation zu der Zahl 2 besteht, ist leicht zu sehen, wenn wir bedenken, dass wir *genau dieselbe Eigenschaft* durch den Ausdruck ‚*x* hat eine Masse von 2000 g' oder ‚*x* hat eine Masse von 70,55 Unzen' zuschreiben können, obwohl die Zahlwörter, die in diesen Ausdrücken enthalten sind, doch sehr verschiedene Zahlen bezeichnen.

[11] Die Hauptthemen der Messtheorie sind das Repräsentationsproblem und das Eindeutigkeitsproblem. Das erste Problem betrifft die Frage nach der Existenz strukturerhaltender Homomorphismen, d. h. präziser, die Frage nach den Bedingungen, die ein empirisches relationales System erfüllen muss, damit man zeigen kann, dass es ein entsprechendes numerisches relationales System gibt, auf das das empirische relationale System homomorph abgebildet werden kann. Das Eindeutigkeitsproblem betrifft dann die Frage, wie viele dieser Homomorphismen unter diesen Bedingungen existieren und inwiefern sie sich voneinander unterscheiden.

sen, welche Positionen die Gegenstände im empirischen relationalen System <*D*, *H*, o> einnehmen. Wenn wir von drei Gegenständen *a, b* und *c* wissen, dass sie eine Masse von 2, 4 und 6 kg haben, wissen wir auch, dass *b* schwerer ist als *a*, dass *c* schwerer ist als *a* und dass *a* und *b* zusammen (d. h. *a* o *b*) genauso schwer sind wie *c*.[12] Der Hauptgrund für die Praxis der Zuschreibung von Masseneigenschaften durch Operator-Index-Prädikate, die durch die Anwendung von Operatoren wie ‚hat eine Masse von *y* kg' auf einen Ausdruck für eine positive reelle Zahl gebildet werden, scheint also in der Tatsache begründet zu sein, dass in diesen numerischen Ausdrücken Informationen darüber enthalten sind, welche Positionen die jeweiligen Gegenstände innerhalb der Struktur aller Gegenstände der Menge *D* einnehmen.

## 4.

Der Kerngedanke von Autoren wie Churchland, Field, Stalnaker und Davidson ist, dass Propositionen für intentionale Zustände eine ähnliche Rolle spielen wie Zahlen für physikalische Größen. Oder, um die in diesem Zusammenhang angemessenere formale Redeweise zu verwenden: dass die ‚dass'-Sätze, die wir bei der Zuschreibung von intentionalen Zuständen verwenden, eine ähnliche Rolle spielen wie die Zahlausdrücke, die bei der Zuschreibung physikalischer Größen unverzichtbar zu sein scheinen. Dies scheint jedoch vorauszusetzen, dass es im Fall intentionaler Zustände ebenfalls möglich ist, ein empirisches und ein abstraktes relationales System anzugeben (in diesem Falle offenbar ein propositionales relationales System), die homomorph aufeinander abgebildet werden können.[13] Die zen-

---

[12] Dass *a* schwerer als *b* ist, soll hier heißen, dass *bHa* und nicht *aHb* (d. h. dass *a b* überwiegt), und dass *a* genauso schwer wie *b* ist, soll heißen, dass *aHb* und *bHa*.

[13] Vgl. z. B. die folgenden zwei Passagen in Stalnaker (1984):
„What is it about such physical properties as having a certain height or weight that makes it correct to represent them as relations between the thing to which this property is ascribed and a number? The reason we can understand such properties – physical quantities – in this way is that they belong to families of properties which have a structure in common with the real numbers. Because the family of properties which are *weights* of physical objects has this structure, we can . . . use a number to pick a particular one of these properties out of the family." (1984, S. 9)
„The analogy suggests that to define a relation between a person or a physical object and a proposition is to define a class of properties with a structure that makes it possible to pick one of the properties out of the class by specifying a proposition." (1984, S. 11)

trale Frage in diesem Zusammenhang scheint daher zu sein: Wie könnten das empirische und das abstrakte relationale System in Bezug auf intentionale Zustände aussehen?

Auf den ersten Blick scheint sich die folgende Antwort anzubieten: Das empirische relationale System $<Z, C>$ besteht in diesem Fall aus einer Menge von internen Zuständen $Z$ und der Kausalrelation $C$; das abstrakte relationale System $<P, I>$ aus der Menge der Propositionen $P$ und der Relation der logischen Implikation $I$. Diese Annahme scheint sich deshalb von selbst aufzudrängen, weil sie eine einfache Antwort auf die gestellte Frage ermöglicht:

(KT) Wir schreiben die internen Zustände der Menge $Z$ mit Hilfe von Prädikaten zu, die dadurch entstehen, dass wir Operatoren wie ‚hat die Überzeugung $y$' auf ‚dass'-Sätze anwenden, weil es eine homomorphe Abbildung $f$ von der Menge $Z$ in die Menge $P$ der Propositionen gibt, für die gilt:

(i) Für beliebige Zustände $x_1, x_2 \in Z$:
$x_1$ verursacht $x_2 \Leftrightarrow f(x_1)$ impliziert logisch $f(x_2)$.

Allerdings: Bei näherem Hinsehen ist KT doch alles andere als plausibel. Erstens ist die in KT implizit enthaltene Rationalitätsannahme gleichzeitig zu stark und zu schwach.[14] Nehmen wir z. B. an, dass $f$ einem Zustand $x \in Z$ die Proposition $p$ zuordnet und dass wir $x$ daher ‚die Überzeugung, dass $p$' nennen. In diesem Fall impliziert die Bedingung (i):

(a) Für alle $y \in Z$, denen $f$ eine Proposition zuordnet, die aus $p$ logisch folgt, gilt: $x$ verursacht $y$.

(b) Für alle $y \in Z$, denen $f$ eine Proposition zuordnet, die aus $p$ nicht logisch folgt, gilt: $x$ verursacht $y$ nicht.

D. h., (i) impliziert, dass die Überzeugung, dass $p$, alle und nur die Überzeugungen, dass $q$, verursacht, für die gilt: $q$ ist eine logische Folge von $p$. Aber natürlich glauben wir keineswegs alles, was aus dem logisch folgt, was wir tatsächlich glauben. Und natürlich sollte es möglich sein, dass eine Überzeugung, dass $p$, die Überzeugung, dass $q$, auch dann verursacht, wenn $p$ nur gute Gründe für $q$ liefert, ohne $q$ logisch zu implizieren. Und weiterhin sollte es sicher auch Raum für irrationale Überzeugungen geben, d. h. Überzeugungen, die von Überzeugungen verursacht wurden, die nicht einmal gute Gründe für sie liefern.

Der zweite Mangel von KT liegt darin, dass sie alle kausalen Relationen zwischen den inneren Zuständen eines Systems und Zuständen in seiner Umgebung und auch alle kausalen Relationen zwischen den inneren Zu-

[14] Diese Kritik findet sich in ähnlicher Weise bereits in Lanz (1987, S. 117 ff.).

ständen eines Systems und seinen Handlungen gänzlich außer Acht lässt. Wenn wir einer Person *A* bestimmte intentionale Zustände zuschreiben – z. B. indem wir sagen ‚*A* hat die Überzeugung, dass es regnet' oder ‚*A* hat den Wunsch, den Rekord über 100 m Freistil zu brechen' –, dann tun wir das aber unter anderem, weil Zustände der ersten Art in einem bestimmten Sinn kausal davon abhängig sind, dass es regnet, während Zustände der zweiten Art normalerweise ein Verhalten verursachen, das *A* ihrem Ziel voraussichtlich näher bringt. Kausale Relationen zwischen den internen Zuständen einer Person und Zuständen ihrer Umgebung sowie ihrem Verhalten spielen daher bei der Wahl der ‚dass'-Sätze, mit deren Hilfe wir diese Zustände zuschreiben, ebenfalls eine wichtige Rolle.

Ein dritter Mangel ist mit dem zweiten eng verbunden. Nach KT ist der Inhalt intentionaler Zustände absolut undeterminiert. Wenn es eine Abbildung *f* gibt, die die Bedingung (i) erfüllt, dann gibt es nämlich unendlich viele verschiedene Abbildungen dieser Art. Angenommen, *f* bildet die Elemente von *Z* auf die Menge $P_f$ der Propositionen ab, die durch die ‚dass'-Sätze ‚dass $p_1$', ‚dass $p_2$', ‚dass $p_3$', etc. bezeichnet werden. Wenn man in diesen ‚dass'-Sätzen überall den Namen ‚Konrad Adenauer' durch den Namen ‚Luciano Pavarotti' ersetzt und umgekehrt und das Prädikat ‚ist Bundeskanzler' durch das Prädikat ‚ist ein Tenor' und umgekehrt, erhält man eine Menge von ‚dass'-Sätzen, die jeweils andere Propositionen bezeichnen.[15] Offenbar gelten aber zwischen diesen Propositionen dieselben logischen Beziehungen wie zwischen den Elementen von $P_f$. Wenn *f* die Bedingung (i) erfüllt, wenn wir in den ‚dass'-Sätzen, die die Elemente der Menge $P_f$ bezeichnen, alle Namen und Prädikate gleichförmig durch andere Namen und Prädikate ersetzen und wenn $f'$ schließlich die Funktion ist, die den Elementen von *Z* die Propositionen zuordnet, die durch die auf diese Weise gewonnenen ‚dass'-Sätze bezeichnet werden, dann erfüllt auch $f'$ (i).

Ein vierter und letzter Mangel von KT besteht darin, dass diese Theorie keine vernünftige Unterscheidung zwischen den unterschiedlichen *Arten* intentionaler Zustände erlaubt. Wie kann KT z. B. dem Unterschied zwischen Überzeugungen und Wünschen gerecht werden?[16] Offenbar gar

---

[15] Natürlich gilt das nur, wenn in den ‚dass'-Sätzen zumindest einer der genannten Namen oder eines der genannten Prädikate vorkommt.

[16] Auch hierzu findet sich eine ähnliche Kritik in Lanz (1987, S. 113 ff.). Nach Lanz gibt es noch einen anderen Punkt, an dem die Analogie zwischen Messtheorie und intentionaler Psychologie zusammenbricht: die Tatsache, dass die empirischen Relationen, die zwischen den Gegenständen gelten müssen, damit einige ihrer Eigenschaften messbar sind, ohne Rekurs auf Zahlen spezifiziert werden können, während die entsprechenden Relationen zwischen intentionalen Zuständen nicht ohne Rekurs auf Propositionen spezifiziert werden können (vgl. 1987, S. 107 ff.). Mir ist allerdings nicht klar, warum das so sein sollte.

nicht. Denn KT sagt uns nur etwas darüber, wie die Funktion *f* jedem Zustand von *Z* eine Proposition *p* zuordnet, sie sagt uns aber nichts darüber, ob es sich bei diesem Zustand um den Glauben, dass *p*, oder den Wunsch, dass *p*, oder einen anderen intentionalen Zustand mit dem Inhalt *p* handelt. Letzten Endes kann KT wohl nur im Hinblick auf Überzeugungen eine gewisse Plausibilität für sich in Anspruch nehmen; denn es ist zumindest nicht völlig unplausibel, anzunehmen, dass jemand, der glaubt, dass *p*, und glaubt, dass wenn *p*, dann *q*, auch glaubt, dass *q*. Eine ähnliche Annahme im Hinblick auf Wünsche wäre dagegen wenig sinnvoll.

## 5.

Ich denke, die soeben aufgezählten Probleme zeigen ziemlich deutlich, dass die erste Antwort auf die Frage „Wie könnten das empirische und das abstrakte relationale System im Falle intentionaler Zustände aussehen?“ sicher nicht haltbar ist. Aber was folgt daraus im Hinblick auf die Analogie zwischen metrischen und intentionalen Begriffen? Offenbar gibt es hier zwei Möglichkeiten. Auf der einen Seite kann man argumentieren, dass aus dem Scheitern *einer* Antwort sicher nicht folge, dass auch alle anderen Antwortversuche zum Scheitern verurteilt seien. Aus diesem Grunde sei es durchaus sinnvoll, weiter nach einer zufriedenstellenden Antwort auf die ursprüngliche Frage zu suchen.[17] Auf der anderen Seite kann man aber auch der Auffassung sein, dass sich die Schwierigkeiten, mit denen der erste Antwortversuch konfrontiert ist, in derselben – oder zumindest in sehr ähnlicher – Weise bei allen möglichen Antwortversuchen ergeben werden. Falls das so ist, wäre die weitere Suche nach einer zufriedenstellenden Lösung aber von vornherein wenig erfolgversprechend. Und wenn man dieser Diagnose zustimmt, liegt es sogar nahe, zu argumentieren, die ganze Idee einer Analogie von metrischen und intentionalen Begriffen sei verfehlt.

Meiner Meinung nach gehen diese Positionen allerdings beide von einer falschen Prämisse aus – nämlich von der Annahme, dass die Verwendung von Operator-Index-Prädikaten dann und nur dann gerechtfertigt ist, wenn es möglich ist, ein empirisches und ein abstraktes relationales System anzugeben, von denen gezeigt werden kann, dass sie homomorph aufeinander abgebildet werden können. Und diese Annahme selbst scheint auf der noch grundlegenderen Voraussetzung zu beruhen, dass wir jede Verwendung

---

Warum kann man die kausalen Relationen zwischen den internen Zuständen eines Systems nicht ohne Rekurs auf Propositionen spezifizieren? Lanz scheint offenbar zu glauben, dass man die in Frage stehenden internen Zustände nur *als intentionale Zustände* spezifizieren kann. Aber für diese Annahme gibt es, soweit ich sehen kann, keinerlei zwingende Gründe.

[17] Diese Auffassung scheint Matthews (1994) zu vertreten.

von Operator-Index-Prädikaten *rechtfertigen* müssen, d.h. dass es, wenn wir diese Prädikate bei der Zuschreibung bestimmter Eigenschaften oder Zustände verwenden, eine *Begründung in der Sache selbst* geben muss, die uns das *Recht* dazu gibt.

Mir scheint jedoch, dass gerade diese letzte Annahme völlig verfehlt ist und dass wir im Gegenteil bei der Wahl der Prädikate, mit deren Hilfe wir welche Eigenschaften auch immer zuschreiben, völlig frei sind. Wenn das so ist, ist die strittige Frage aber nicht mehr „Was kann die Zuschreibung von Eigenschaften oder Zuständen mit Hilfe von Operator-Index-Prädikaten *rechtfertigen*?", sondern nur noch die weit weniger anspruchsvolle Frage „Welche *Motive* haben wir, in dem einen Fall so und im anderen Fall anders vorzugehen?". Die Frage ist dann nur noch, welchen *Vorteil* wir davon haben, dass wir in bestimmten Fällen Operator-Index-Prädikate verwenden.

Ich will versuchen, meinen Standpunkt an einem einfachen Beispiel zu verdeutlichen. Bekanntlich verwenden wir bei der Zuschreibung von Farbeigenschaften Ausdrücke wie ‚*x* ist blau', ‚*x* ist grün', ‚*x* ist rot' etc. Wir *könnten* zu diesem Zweck allerdings ebenso gut die Ausdrücke verwenden: ‚*x* hat Farbe 1', ‚*x* hat Farbe 2', ‚*x* hat Farbe 3'.[18] Ob wir bei der Zuschreibung von Farbeigenschaften solche Operator-Index-Prädikate verwenden oder nicht, ist vollkommen uns selbst überlassen – nichts zwingt uns dazu und nichts hält uns zwingend davon ab. Normalerweise verwenden wir in diesem Fall keine Operator-Index-Prädikate – und zwar deshalb nicht, weil es keinen Vorteil mit sich bringt; aber selbst das bedeutet nicht, dass wir es nicht doch tun könnten.

Was könnte der besondere Vorteil der Verwendung von Operator-Index-Prädikaten sein? Diese Frage sollte uns dazu führen, noch einmal über die Rolle der Tatsache nachzudenken, dass wir bei metrischen Begriffen immer ein empirisches und ein abstraktes relationales System angeben können, die homomorph aufeinander abgebildet werden können. Wenn die im letzten Abschnitt skizzierten Überlegungen richtig sind, benötigen wir auch für den Gebrauch metrischer Begriffe keine Rechtfertigung. D.h., der Sinn des Nachweises, dass es für jeden dieser Begriffe ein empirisches und ein abstraktes relationales System gibt, die homomorph aufeinander abgebildet werden können, kann nicht darin bestehen, den Gebrauch metrischer Begriffe zu rechtfertigen. Aber welchen Sinn hat dieser Nachweis dann? Nun, wenn sich zeigen lässt, dass sich das empirische relationale System $<D, H, o>$ homomorph auf das abstrakte relationale System $<\mathbf{R}^+, \leq, +>$ abbilden lässt, dann folgt daraus, dass sich, wenn wir Massen mit Hilfe von Operator-Index-Prädikaten mit den entsprechenden Indexausdrücken zuschrei-

[18] In diesem Fall würden wir – in der Sprache der Messtheorie – Farben auf Nominalskalenniveau messen.

ben, aus diesen Prädikaten *mehr* ablesen lässt als nur, welche Massen die jeweiligen Gegenstände besitzen. Aus diesen Prädikaten ergibt sich – wie wir schon gesehen haben – zugleich auch, welche Positionen diese Gegenstände in der gesamten Struktur <*D*, *H*, o> einnehmen. Allgemein gesprochen zeigt daher der Nachweis, dass sich ein empirisches relationales System homomorph auf ein abstraktes relationales System abbilden lässt, welche *zusätzlichen* Informationen sich aus der Verwendung entsprechender Operator-Index-Prädikate ergeben. Und zusätzliche Informationen stellen sicher einen Vorteil dar.

Dies ist jedoch keineswegs der einzige Vorteil, der sich aus der Verwendung metrischer Begriffe ergibt. Nehmen wir als Beispiel noch einmal die physikalische Größe Masse. Wenn wir von drei Gegenständen *a, b* und *c* wissen, dass sie eine Masse von 2, 4 und 6 kg haben, dann wissen wir, wie schon erwähnt, auch, dass *b* schwerer ist als *a* und dass *a* und *b* zusammen genauso schwer sind wie *c*. Doch das ist noch nicht alles. Aus dem zweiten Newtonschen Gesetz folgt nämlich, dass wir, wenn wir die Gegenstände *a* und *b* in derselben Zeit auf dieselbe Geschwindigkeit beschleunigen wollen, für *b* die doppelte Kraft benötigen und dass, wenn auf *a* und *c* in derselben Zeit dieselbe Kraft wirkt, die Geschwindigkeit von *a* am Ende dreimal so groß ist wie die Geschwindigkeit von *c*. Mit anderen Worten: Der größte Nutzen, der sich aus der Verwendung metrischer Prädikate ergibt, beruht auf der Tatsache, dass wir *Gesetze*, die für physikalische Größen gelten, häufig in systematisch besonders zufriedenstellender Form formulieren können, wenn wir diese Größen mit Hilfe metrischer Prädikate zuschreiben. Denn in diesem Fall können wir uns bei der Formulierung dieser Gesetze auf Relationen zwischen Zahlen beziehen und gleichzeitig über Zahlen quantifizieren. Dass dies tatsächlich so ist, wird deutlich, wenn wir das zweite Newtonsche Gesetz

(1) $F = m \cdot a$

in die angemessene Operator-Index-Schreibweise überführen:

(1′) $(x)(i)(j)\,(F_i x \wedge m_j x \rightarrow a_{i/j}\, x)$.[19]

6.

Meines Erachtens liegt der Hauptnutzen, der sich aus dem Gebrauch *intentionaler Prädikate* ergibt, ebenfalls in der Tatsache, dass wir mit ihrer Hilfe bestimmte Gesetze zur Erklärung menschlichen Verhaltens auf eine Weise formulieren können, die sowohl ökonomisch als auch systematisch beson-

[19] Der Einfachheit halber habe ich hier außer Acht gelassen, dass sowohl Kräfte als auch Beschleunigungen keine skalaren Größen, sondern Vektoren sind.

ders befriedigend ist. Im Folgenden werde ich versuchen, diesen Punkt anhand eines einfachen, aber suggestiven Beispiels zu veranschaulichen. Dabei gehe ich davon aus, dass intentionale Zustände den Status theoretischer Zustände haben, die wir deshalb postulieren, weil wir nur so das Verhalten bestimmter Systeme plausibel erklären können. Bevor ich zu dem angekündigten Beispiel komme, möchte ich daher eine kurze Bemerkung zu der Frage voranschicken, unter welchen Bedingungen wir überhaupt einen Grund haben, theoretische Zustände anzunehmen.[20]

Die erste Antwort auf diese Frage lautet: Wir haben immer dann einen Grund für die Annahme theoretischer Zustände, wenn sich ein System *S* in Situationen desselben Typs systematisch *unterschiedlich* verhält. Wenn sich *S* z. B. in Situationen des Typs $\Phi$ manchmal auf die Weise *x* und manchmal auf die Weise *y* verhält, dann bietet sich dafür die einfache Erklärung an, dass *S* zwei verschiedene (theoretische/dispositionale) Zustände *A* und *B* annehmen kann, für die gilt:

(2) In Situationen des Typs $\Phi$ verhält sich *S* genau dann auf die Weise *x*, wenn *S* im Zustand *A* ist.

(3) In Situationen des Typs $\Phi$ verhält sich *S* genau dann auf die Weise *y*, wenn *S* im Zustand *B* ist.

Dies ist natürlich nur der einfachste Fall. Die Annahme theoretischer Zustände wird sehr viel fruchtbarer, wenn man (a) zusätzliche Informationen darüber hat, wann, d. h. unter welchen Bedingungen *S* in diese Zustände kommt, und wenn man (b) das Verhalten von *S* mit einer sehr kleinen Menge von theoretischen Zuständen erklären kann, weil sich Verhaltensunterschiede durch die Interaktion verschiedener Zustände erklären lassen und nicht für jede Verhaltensweise ein eigener Zustand postuliert werden muss.

Soweit die allgemeinen Bemerkungen. Nehmen wir nun folgenden Fall. *S* sei eines jener Systeme, die sich in Situationen desselben Typs nicht immer gleich verhalten. Manchmal geht *S* dorthin, wo Wasser ist, und trinkt; manchmal geht *S* dorthin, wo Bananen sind, und isst. Offenbar können diese Unterschiede im Verhalten von *S* unschwer durch die Annahme erklärt werden, dass *S* sich in zwei unterschiedlichen Zuständen *A* und *B* befinden kann, für die gilt:

(4) Wenn *S* sich im Zustand *A* befindet, sucht *S* einen Ort auf, an dem Wasser ist, und trinkt.

(5) Wenn *S* sich im Zustand *B* befindet, sucht *S* einen Ort auf, an dem Bananen sind, und isst.

---

[20] Vgl. zu den folgenden Abschnitten auch Beckermann (1986, S. 319 ff., 1992c, S. 162 ff.).

Tatsächlich, wollen wir weiter annehmen, sind die Dinge jedoch nicht ganz so einfach. In einigen Fällen, in denen wir Grund zu der Annahme haben, dass sich *S* im Zustand *A* befindet, sucht *S* nicht den Ort $x_1$ auf, an dem sich tatsächlich Wasser befindet, sondern einen Ort $x_2$, an dem gar kein Wasser vorhanden ist. Selbst in diesem Fall ist es jedoch nicht allzu schwierig, das Verhalten von *S* theoretisch zu erklären. Wir müssen lediglich annehmen, dass *S* zwei weitere Zustände *C* und *D* annehmen kann und dass das Verhalten von *S* nicht durch die Gesetze (4) und (5) erklärt wird, sondern unter anderem durch die folgenden Gesetze:

(4a) Wenn sich *S* in den Zuständen *A* und *C* befindet, sucht *S* den Ort $x_1$ auf (und versucht zu trinken).

(4b) Wenn sich *S* in den Zuständen *A* und *D* befindet, sucht *S* den Ort $x_2$ auf (und versucht zu trinken).

Vielleicht entdecken wir sogar, dass *S* den Zustand *C* genau dann annimmt, wenn die Situation, in der sich *S* befindet, den Situationen ähnelt, in denen sich tatsächlich Wasser an $x_1$ befindet, und dass Entsprechendes auch für den Zustand *D* gilt. In diesem Fall haben wir dann außer den Gesetzen (4a) und (4b) auch noch die folgenden Gesetze zur Verfügung:

(6a) Wenn sich *S* in einer Situation befindet, die den Situationen ähnelt, in denen Wasser an $x_1$ ist, dann nimmt *S* den Zustand *C* an.

(6b) Wenn sich *S* in einer Situation befindet, die den Situationen ähnelt, in denen Wasser an $x_2$ ist, dann nimmt *S* den Zustand *D* an.

Diese wirklich sehr elementaren Überlegungen zeigen bereits, dass wir bei der Erklärung des Verhaltens von *S* bei wachsender Verhaltenskomplexität sehr schnell gezwungen sind, eine ziemlich große Anzahl unterschiedlicher theoretischer Zustände zu postulieren, deren kausale Rollen nur mit Hilfe einer noch größeren Anzahl von Verhaltensgesetzen beschrieben werden können. Noch deutlicher wird das Problem, wenn wir annehmen, dass *S* im Zustand *A* nicht immer nur die Orte $x_1$ oder $x_2$ aufsucht, sondern manchmal auch die Orte $x_3$, $x_4$, $x_5$ etc., was dazu führt, dass dieses Verhalten nur durch die Annahme weiterer theoretischer Zustände *E, F, G* etc. erklärt werden kann. Und ziemlich undurchsichtig wird die Sache schließlich, wenn wir zusätzlich annehmen, dass Analoges auch für den Zustand *B* gilt, und wenn wir darüber hinaus auch noch annehmen, dass *S* nicht nur Wasser- oder Bananenorte aufsucht, sondern manchmal auch Schattenorte, Höhlenorte, Bergorte etc.

Es braucht nicht viel Phantasie, um zu erkennen, wie schnell die Anzahl der theoretischen Zustände, die wir zur Erklärung des Verhaltens von *S* benötigen, jedes handhabbare Maß übersteigt. Und wenn das Verhalten von *S* ein bestimmtes Maß an Komplexität erreicht, benötigen wir diese Vielzahl

theoretischer Zustände tatsächlich. Allerdings: Auch wenn wir das Verhalten von *S* nicht mit einer geringeren Anzahl theoretischer Zustände erklären können, so kann doch die Anzahl der *Gesetze*, in denen die kausalen Rollen dieser Zustände ausgedrückt werden, drastisch verringert werden. Denn die theoretischen Zustände, die wir postulieren mussten, um das Verhalten von *S* zu erklären, können recht einfach zu wenigen Typen zusammengefasst werden. Die Zustände *A* und *B* z. B. sind – genauso wie die Zustände, die dazu führen, dass *S* Schattenorte, Höhlenorte oder Bergorte aufsucht – dadurch charakterisiert, dass sie dazu führen, dass *S* Orte aufsucht, die ein bestimmtes Merkmal aufweisen. Alle diese Zustände können deshalb in einem bestimmten Zustandstyp zusammengefasst werden, den wir hier ‚*W*' nennen wollen: Zuständen dieses Typs ist gemeinsam, dass sie dazu führen, dass *S* Orte aufsucht, die ein bestimmtes Merkmal besitzen, sie unterscheiden sich lediglich in dem charakteristischen Merkmal, das der jeweilige Ort besitzen soll. Es bietet sich daher an, alle Zustände des Typs *W* mit Operator-Index-Prädikaten zuzuschreiben, in denen der Operator den Typ *W* ausdrückt, während die Indizes für die verschiedenen charakteristischen Merkmale stehen. Mit anderen Worten: Es bietet sich an, den Zustand *A* mit Hilfe des Prädikats ‚$W_{\text{Wasser}}$', den Zustand *B* mit Hilfe des Prädikats ‚$W_{\text{Bananen}}$' und die anderen Zustände dieses Typs mit den Prädikaten ‚$W_{\text{Schatten}}$', ‚$W_{\text{Höhle}}$' und ‚$W_{\text{Berg}}$' zuzuschreiben.

Im nächsten Schritt ist unschwer zu sehen, dass auch die Zustände *C* und *D* sowie die Zustände *E*, *F*, *G* etc. etwas gemeinsam haben. Von diesen Zuständen kann daher ebenfalls gesagt werden, dass sie zu einem Zustandstyp gehören, den wir ‚*Ü*' nennen wollen. Denn alle diese Zustände unterscheiden sich nur im Hinblick auf den Ort, den *S* aufsucht, wenn es sich im Zustand *A* – d. h. im Zustand ‚$W_{\text{Wasser}}$' – befindet. Die eigentliche Pointe wird jedoch erst deutlich, wenn wir weiter annehmen, dass für den Zustand *B* – d. h. den Zustand ‚$W_{\text{Bananen}}$'– analoge Zustände *C*′, … , *G*′ existieren, die dazu führen, dass *S* die Orte $x_1$, $x_2$, … bzw. $x_5$ aufsucht, wenn sich *S* im Zustand ‚$W_{\text{Bananen}}$' befindet, und dass für diese theoretischen Zustände *C*′, … , *G*′ die Gesetze gelten:

(5a′) Wenn sich *S* in den Zuständen *B* und *C*′ befindet, sucht *S* den Ort $x_1$ auf.

…

(5e′) Wenn sich *S* in den Zuständen *B* und *G*′ befindet, sucht *S* den Ort $x_5$ auf.

(6a′) Wenn sich *S* in einer Situation befindet, die den Situationen ähnelt, in denen sich Bananen am Ort $x_1$ befinden, dann nimmt *S* den Zustand *C*′ an.

…

(6e′) Wenn sich *S* in einer Situation befindet, die den Situationen ähnelt, in denen sich Bananen am Ort $x_5$ befinden, dann nimmt *S* den Zustand *G′* an.

Denn jetzt zeigt sich, dass die Zustände *C′*, … , *G′* offenbar ebenfalls zum Typ *Ü* gehören und dass sich die Zustände dieses Typs nicht nur in Bezug auf einen, sondern in Bezug auf zwei Parameter unterscheiden: den Ort und das charakteristische Merkmal dieses Ortes. In diesem Fall scheint es daher sinnvoll, Operator-Index-Prädikate mit zwei Indizes einzuführen und z. B. die Zustände *C*, …, *G* mit Hilfe der Prädikate ‚$Ü_{\text{Wasser, } x_1}$‘, …, ‚$Ü_{\text{Wasser, } x_5}$‘ zuzuschreiben. Entsprechendes gilt dann für die Zustände *C′*, …, *G′*, die jetzt mit Hilfe der Prädikate ‚$Ü_{\text{Bananen, } x_1}$‘, …, ‚$Ü_{\text{Bananen, } x_5}$‘ zugeschrieben werden können.[21]

Aber warum ist es naheliegend, in diesem Fall die angeführten Operator-Index-Prädikate zu verwenden? Welchen Vorteil haben diese gegenüber den Prädikaten ‚*A*‘, ‚*B*‘, ‚*C*‘, …, ‚*G*‘, ‚*C′*‘, …, ‚*G′*‘ usw.? Nun, der große Vorteil der Verwendung von Operator-Index-Prädikaten ist in diesem Fall, dass wir eine unüberschaubare Zahl von Gesetzen – nämlich die Gesetze (4a) und (4b), die Gesetze (5a) - (5e) sowie alle anderen verwandten Gesetze – *in ein einziges Gesetz* zusammenziehen können:

(4′) Für alle *x* und *y*: wenn *S* in den Zuständen $W_y$ und $Ü_{y, x}$ ist, dann sucht *S* den Ort *x* auf.

Und dasselbe gilt für die Gesetze (6a), (6b), (6a′) – (6e′) und alle verwandten Gesetze. Denn auch diese Gesetze lassen sich jetzt in einem einzigen Gesetz zusammenfassen:

(6′) Für alle *x* und *y*: Wenn sich *S* in einer Situation befindet, die den Situationen ähnelt, in denen sich *y* am Ort *x* befinden, dann nimmt *S* den Zustand $Ü_{y, x}$ an.

Der unschätzbare Wert, der sich schon in diesem einfachen Fall aus der Einführung von Operator-Index-Prädikaten ergibt, besteht also darin, dass wir eine potenziell unendliche Anzahl von Gesetzen auf eine geringe, handhabbare Anzahl reduzieren können, ohne dass wir die große Vielzahl theoretischer Zustände einschränken müssen, die wir zur Erklärung des Verhaltens von *S* tatsächlich benötigen. Denn die Zustände, die wir mit Hil-

[21] Statt zweier Indizes könnte man offenbar schon an dieser Stelle Sätze als Indexausdrücke verwenden. In diesem Fall hätten die Operator-Index-Prädikate nicht die Form $Ü_{y, x}$, sondern $Ü_{\textit{Am Ort x befindet sich y}}$. Die Einführung von Sätzen als Indexausdrücke scheint jedoch erst dann wirklich sinnvoll, wenn nicht alle Indexsätze dieselbe logische Form haben, d. h. wenn das System *S* auch Zustände annehmen kann, die sinnvollerweise mit Prädikaten wie ‚$Ü_{\text{Das Wasser ist kalt}}$‘ oder ‚$Ü_{\text{Alles, was grün ist, ist giftig}}$‘ zugeschrieben werden können.

fe der Prädikate ‚$\ddot{U}_{\text{Wasser, } x_1}$‘, ‚$\ddot{U}_{\text{Wasser, } x_2}$‘, … oder mit den Prädikaten ‚$\ddot{U}_{\text{Bananen, } x_1}$‘, ‚$\ddot{U}_{\text{Bananen, } x_2}$‘, … zuschreiben, bleiben natürlich verschieden. D. h., Operator-Index-Prädikate ermöglichen einfache Formulierungen, ohne die notwendige Komplexität zu unterdrücken.[22]

## 7.

Das Beispiel im letzten Abschnitt war natürlich so gewählt, dass es der Annahme, die Verwendung *intentionaler* Prädikate bringe dieselben Vorteile mit sich wie die Verwendung der Operator-Index-Prädikate $W_y$ und $\ddot{U}_{y, x}$, zumindest eine *prima facie* Plausibilität verleiht. Denn das Verhalten des Systems *S*, das die Einführung dieser Operator-Index-Prädikate nahe legte, hat eine große Ähnlichkeit zum Verhalten der Lebewesen, deren Verhalten wir intentional erklären, auch wenn das Letztere natürlich sehr viel komplexer ist. Mir scheint daher, dass der Grund für die Verwendung intentionaler Prädikate nicht darin zu suchen ist, dass die mit solchen Prädikaten zugeschriebenen Zustände die mysteriöse Eigenschaft besitzen, einen semantischen Inhalt zu haben. Vielmehr verwenden wir diese Prädikate, weil wir nur mit ihrer Hilfe relativ einfache und handhabbare Verhaltensgesetze formulieren können. Ohne dieses Vokabular würden wir einfach in einer Flut von Gesetzen untergehen.

Als Konsequenz ergibt sich daher, dass es das traditionell verstandene Problem der Intentionalität gar nicht gibt. Denn die traditionelle Lesart dieses Problems setzte voraus, dass intentionale Zustände ein Merkmal besitzen, dessen Naturalisierbarkeit problematisch ist – nämlich das Merkmal, einen semantischen Inhalt zu besitzen. Die Herausforderung für den naturalistisch eingestellten Philosophen bestand daher darin, für dieses Merkmal eine überzeugende naturalistische Analyse zu finden.

Sobald wir erkennen, dass wir intentionale Prädikate nicht verwenden, weil die mit ihrer Hilfe zugeschriebenen Zustände die mysteriöse Eigenschaft besitzen, einen semantischen Inhalt zu haben, sondern deshalb, weil nur diese Prädikate eine relativ einfache und handhabbare Formulierung entsprechender Verhaltensgesetze erlauben, verlieren wir, hoffe ich, das Gefühl, dass es im Zusammenhang mit intentionalen Zuständen etwas Mysteriöses gibt, das dringend einer Naturalisierung bedarf. Sobald wir den wahren Grund für die Verwendung intentionaler Prädikate verstehen, erkennen wir, dass es keinen Grund für die Annahme gibt, dass die Zustände, die wir mit diesen Prädikaten zuschreiben, sich von ‚normalen‘ theoretischen Zuständen durch einen mysteriösen semantischen Inhalt unterschei-

[22] Ich möchte nicht versäumen, darauf hinzuweisen, dass die Überlegungen in diesem Abschnitt Brian Loars Theorie in (1981) viel verdanken. Vielleicht handelt es sich hier einfach um zwei Seiten derselben Medaille.

den. So wie es ja auch keinen Grund für die Annahme gibt, dass sich die Eigenschaften, die wir durch das Prädikat ‚hat eine Masse von 2 kg' zuschreiben, von anderen ‚normalen' Eigenschaften durch einen mysteriösen numerischen Inhalt unterscheidet. Jeder, der das Letztere behauptet, würde sich einer Projektion von der *linguistischen Ebene* auf die *Ebene der Tatsachen* schuldig machen. Er würde daraus, dass wir ein Prädikat für die Zuschreibung dieser Eigenschaft verwenden, das einen Zahlausdruck enthält, schließen, dass die zugeschriebene Eigenschaft einen spezifischen numerischen Inhalt hat. Dieser Schluss ist aber offensichtlich ungültig.

Völlig analog schließen aber diejenigen, die glauben, dass es ein Problem der Intentionalität gibt, aus der Tatsache, dass wir bei der Zuschreibung intentionaler Zustände Ausdrücke verwenden, die ‚dass'-Sätze enthalten, darauf, dass die so zugeschriebenen Zustände die mysteriöse Eigenschaft besitzen, einen semantischen Inhalt zu haben. Dass auch dieser Schluss ungültig ist, zeigt nicht nur der Parallelfall metrischer Begriffe; es ergibt sich auch aus dem im letzten Abschnitt analysierten Beispiel. Denn dieses Beispiel zeigt mit aller Deutlichkeit, dass die tatsächlichen Gründe für die Verwendung intentionaler Prädikate nichts mit irgendwelchen semantischen Inhalten zu tun haben (müssen), die die Zustände, die mit Hilfe dieser Prädikate zugeschrieben werden, angeblich besitzen.

Ich denke daher, dass *es einfach kein Problem der Intentionalität gibt*, d. h. dass wir uns die Frage, wie sich die Eigenschaft, einen semantischen Inhalt zu haben, naturalisieren lässt, nicht zu stellen brauchen, weil intentionale Zustände diese Eigenschaft gar nicht besitzen. Jedenfalls nicht in dem Sinne, der traditionell unterstellt wird. Das – traditionell verstandene – Problem der Intentionalität ist ein Artefakt, das aus einer falschen Analyse der Gründe resultiert, die für unsere Verwendung intentionalen Vokabulars verantwortlich sind. Diese Analyse ist falsch, weil der Grund für die Verwendung intentionaler Prädikate eben nicht darin besteht, dass die theoretischen Zustände, die mit ihrer Hilfe zugeschrieben werden, das mysteriöse Merkmal, einen semantischen Inhalt zu haben, besitzen, sondern einzig und allein darin, dass nur die Verwendung dieser spezifischen Form von Operator-Index-Prädikaten eine einfache und überschaubare Formulierung der entsprechenden Verhaltensgesetze erlaubt.

Wenn es in diesem Zusammenhang überhaupt ein Problem gibt, so ist es nicht das Problem, zu erklären, wie Zustände mit einem semantischen Inhalt in die Welt kommen, sondern höchstens das Problem, wie das äußerst komplexe Verhalten möglich ist, für dessen Erklärung es angemessen ist, intentionale Prädikate zu verwenden.

8.

Gegen diese Schlussfolgerung könnte man allerdings den folgenden Einwand erheben.[23] Eines meiner Hauptargumente lautete: Aus der Tatsache, dass wir physikalische Größen – wie z. B. Massen von Körpern – mit Hilfe von Zahlen messen, folgt keineswegs, dass wir damit diesen Körpern einen numerischen Inhalt zuweisen. Folglich sollten wir analog auch nicht annehmen, dass die Verwendung intentionalen Vokabulars bei der Zuschreibung propositionaler Einstellungen impliziert, dass die so zugeschriebenen Zustände einen intentionalen oder repräsentationalen Inhalt besitzen. Aber was kann es heißen, dass Körper einen numerischen Inhalt haben? Vielleicht nicht mehr, als dass arithmetisches Vokabular auf Körper angewendet werden kann. So ähnlich mag die Eigenschaft, einen semantischen Inhalt zu haben, nichts weiter als die Anwendbarkeit intentionalen Vokabulars bedeuten. Unsere Überzeugungen können aber wahr oder falsch sein, unsere Wünsche erfüllt oder unerfüllt sein. Was könnte es sonst bedeuten, einen semantischen Inhalt zu haben?

Dies mag zunächst wie ein starker Einwand klingen; in meinen Augen ist es aber eher eine (fast) bedingungslose Kapitulation. Wenn man der Meinung ist, dass das Problem der Intentionalität einer der Punkte ist, die das Leib-Seele-Problem rätselhaft machen, dann heißt das, dass man annimmt, dass es im Zusammenhang mit intentionalen Zuständen *etwas sehr Spezielles* gibt – etwas, das ein *echtes* Problem für jeden darstellt, der versucht, eine naturalistische Analyse des Geistes zu finden. Philosophen, die denken, dass das Problem der Intentionalität ein *tiefes* Problem ist, müssen daher behaupten, dass die Eigenschaft, einen repräsentationalen oder semantischen Inhalt zu haben, ein rätselhaftes Merkmal einiger unserer mentalen Zustände ist – ein Merkmal, das nicht ohne große Probleme naturalisiert werden kann.

Mein Ziel war, zu zeigen, dass es keinen Grund zu der Annahme gibt, dass die Zustände, die wir mit Hilfe intentionaler Prädikate zuschreiben, tatsächlich ein solches Merkmal besitzen. Aber daraus folgt sicherlich nicht, dass intentionale Prädikate nicht auf Menschen angewendet werden können. Meine Analyse hat daher nicht die Konsequenz, dass intentionale Zustände keinen repräsentationalen oder semantischen Inhalt haben – *vorausgesetzt, dass dies nicht mehr bedeutet, als dass intentionale Prädikate zumindest manchmal auf Menschen wie dich und mich zutreffen.*

Wichtig ist, sich klar zu machen, dass wir es jetzt mit zwei sehr verschiedenen Auffassungen darüber zu tun haben, was es heißt, einen repräsentationalen oder semantischen Inhalt zu haben. Der ersten Auffassung

[23] Diesen Einwand verdanke ich Wolfgang Spohn.

zufolge ist die Eigenschaft, einen repräsentationalen oder semantischen Inhalt zu haben, ein mysteriöses und schwer fassbares Merkmal einiger unserer mentalen Zustände; der zweiten Auffassung zufolge bedeutet es nichts weiter als die Anwendbarkeit intentionalen Vokabulars. Ich habe hier versucht, die erste Ansicht zu kritisieren – die einzige Auffassung, die zu der Konklusion führt, dass das Problem der Intentionalität ein *ernstes* Problem für jeden Versuch einer naturalistischen Analyse des Geistes darstellt. Bezüglich der zweiten Auffassung sehe ich keine Schwierigkeiten, da das Problem der Intentionalität dieser Auffassung zufolge keineswegs mysteriöser ist als das ‚Problem physikalischer Größen'. Mit anderen Worten: Es gibt kein Problem der Intentionalität in dem Sinn, dass es nötig wäre, eine Naturalisierung für ein mysteriöses und schwer fassbares Merkmal zu finden. Denn es gibt keinen Grund zu der Annahme, dass die mit Hilfe intentionaler Prädikate zugeschriebenen Zustände ein solches Merkmal überhaupt besitzen. Hingegen gibt es ein Problem der Intentionalität in dem Sinne, dass es nötig ist, eine Antwort auf die Frage zu finden, unter welchen Bedingungen und aus welchen Gründen wir intentionales Vokabular bei der Erklärung des Verhaltens bestimmter Wesen verwenden. Aber nichts spricht dafür, dass es besonders schwierig oder gar unmöglich sein sollte, eine naturalistische Antwort auf *diese* Frage zu finden. Wenn mit dem Problem der Intentionalität keine weiteren Schwierigkeiten verbunden sind als mit dem Problem des numerischen Inhalts, dann können wir zumindest soviel sagen: Das Problem der Intentionalität ist kein *tiefes* philosophisches Problem.

Aus diesen Bemerkungen ergibt sich auch zumindest eine Teilantwort auf eine weitere Frage: Ist der Preis nicht zu hoch, den wir zahlen müssen, wenn wir die von mir vorgeschlagene Theorie akzeptieren? Ist diese Theorie nicht bloß eine Variante des Eliminativismus? Auf den ersten Blick könnte dies so scheinen, da eine der Hauptthesen dieser Theorie besagt, dass die Zustände, die wir mit Hilfe intentionaler Prädikate zuschreiben, nicht das – irgendwie mysteriöse – Merkmal besitzen, einen semantischen Inhalt zu haben. Wenn wir intentionale Zustände so *definieren*, dass sie genau die mentalen Zustände sind, die einen semantischen Inhalt besitzen, würde aus der Theorie daher folgen, dass solche Zustände nicht existieren. Allerdings: Wären wir tatsächlich bereit, zu sagen, dass keine Massen, keine Längen, keine Temperaturen und keine Ladungen existieren, nur weil die Eigenschaften, die wir mit metrischen Begriffen zuschreiben, nicht das mysteriöse Merkmal besitzen, einen numerischen Inhalt zu haben? In meinen Augen wäre dies völlig kontraintuitiv. Massen, Längen, Temperaturen und Ladungen *sind* die Eigenschaften, die wir mit den entsprechenden metrischen Prädikaten zuschreiben. Und dasselbe gilt für intentionale Zustände: Intentionale Zustände *sind* die Zustände, die wir mit intentionalen Prä-

dikaten zuschreiben – unabhängig davon, welche anderen charakteristischen Merkmale diese Zustände haben oder nicht haben mögen. Wenn das so ist, dann ist die Theorie, die ich vorgeschlagen habe, aber nicht eliminativistisch, da dieser Theorie zufolge die Zustände, die wir mit Hilfe intentionaler Prädikate zuschreiben, genauso real sind wie andere Zustände auch.[24]

## 9.

Als eine Art Koda möchte ich anfügen, dass die Auffassung, die ich in diesem Artikel entwickelt habe, den zusätzlichen Vorteil hat, neues Licht auf ein Problem zu werfen, das immer noch heiß diskutiert wird – das Problem von Individualismus und Anti-Individualismus. Wenn man versucht, sich über die verschiedenen Standpunkte in der Debatte um dieses Problem Klarheit zu verschaffen, ergibt sich folgendes Bild.

Burges Hauptargument[25] für seine anti-individualistische Position ist: Die Kriterien, die unserer alltagspsychologischen Praxis der Zuschreibung intentionaler Zustände zugrunde liegen, sind so geartet, dass wir unter bestimmten Bedingungen die Überzeugungen zweier Personen sogar dann mit Hilfe *unterschiedlicher Inhaltssätze* zuschreiben, wenn sich diese Personen *physikalisch nicht unterscheiden.* Dies liegt daran, dass diesen Kriterien zufolge der ‚Inhalt' von Überzeugungen nicht nur davon abhängt, was innerhalb einer Person vorgeht, sondern auch davon, in welcher Umwelt sie lebt und aufgewachsen ist, und/oder welche Sprache von der Sprachgemeinschaft gesprochen wird, zu der sie gehört. Wenn wir von $Elmar_1$ sagen, er habe die Überzeugung, dass er Arthritis in seinem Oberschenkel hat, während wir $Elmar_2$ die entsprechende Überzeugung zuschreiben, in-

---

[24] Vgl. zu dieser Stelle auch das folgende Zitat aus Davidson (1989):
„We know there is no contradiction between the temperature of the air being 32 fahrenheit or 0 celsius; there is nothing in this ‚relativism' to show that the properties being measured are not ‚real' … in the light of the considerations put forward here, this [i. e. that either of two different interpretations might correctly be applied to the same thought (or utterance) of a person] comes to no more than the recognition that more than one set of one person's utterances might be equally successful in capturing the contents of someone else's thoughts and speech. Just as numbers can capture all the empirically significant relations among weights or temperatures in infinitely many different ways, so one person's utterances can capture all the significant features of another person's thoughts and speech in different ways. This fact does not challenge the ‚reality' of the attitudes or meanings thus variously reported." (Davidson 1989, S. 16)

[25] Vgl. bes. Burge (1979, 1986).

dem wir sagen, dass er glaubt, er habe Tharthritis in seinem Oberschenkel (und all das, obwohl vorausgesetzt wurde, dass sich $Elmar_1$ und $Elmar_2$ weder hinsichtlich ihrer physikalischen Eigenschaften, noch bezüglich ihrer Geschichte unterscheiden, soweit diese Geschichte in nicht-intentionalem Vokabular erzählt werden kann), so ist dies allein auf die Tatsache zurückzuführen, dass das Wort ‚Arthritis' in der Sprachgemeinschaft von $Elmar_1$ eine andere Bedeutung hat als in der Sprachgemeinschaft, zu der $Elmar_2$ gehört.[26]

Fodors Haupteinwand gegen Burges Anti-Individualismus lautet dagegen: Burge hat zwar Recht, was die *alltagspsychologische* Praxis der Zuschreibung intentionaler Zustände angeht; aber daraus folgt nur, dass eine *wissenschaftliche Psychologie* andere Kriterien für die Zuschreibung intentionaler Zustände verwenden muss, da die Überzeugungen, die wir $Elmar_1$ und $Elmar_2$ im Alltag unter der Verwendung verschiedener Inhaltssätze zuschreiben, ‚offensichtlich' über dieselben *Kausalkräfte* verfügen; vom wissenschaftlichen Standpunkt aus müssen diese Überzeugungen daher als (typ-)identisch behandelt werden.[27]

Wenn Fodors These richtig ist, dass Zustände mit denselben Kausalkräften als typidentisch klassifiziert werden müssen, und wenn ferner auch die Annahme richtig ist, dass die Überzeugungen von $Elmar_1$ und $Elmar_2$ dieselben Kausalkräfte besitzen, ist man offenbar gezwungen, zuzugeben, dass die Glaubenszustände von $Elmar_1$ und $Elmar_2$ nicht verschieden, sondern typidentisch sind. Aber heißt das auch, dass wir bei der Zuschreibung dieser Überzeugungen keine verschiedenen ‚dass'-Sätze verwenden dürfen? Wenn die messtheoretische Interpretation intentionaler Prädikate korrekt ist, lautet die Antwort auf diese Frage: Keineswegs. Denn dieser Interpretation zufolge ist es durchaus möglich, typidentische Zustände mit Hilfe verschiedener intentionaler Prädikate zuzuschreiben. Wenn wir $Elmar_1$ eine bestimmte Überzeugung zuschreiben, indem wir sagen, er glaubt, dass er Arthritis in seinem Oberschenkel hat, dann können wir durchaus $Elmar_2$ eine *typidentische Überzeugung* zuschreiben, indem wir sagen, er glaubt, dass er Tharthritis in seinem Oberschenkel hat – und zwar in demselben Sinn, in dem wir einem Würfel *a dieselbe Masseeigenschaft* zuschreiben können, indem wir sagen ‚*a* hat eine Masse von 2 kg' und ‚*a* hat eine Masse von 2000 g'. Der Grund für die Wahl verschiedener ‚dass'-Sätze könnte

[26] Zu diesem Beispiel vgl. Burge (1979, S. 77 ff.).

[27] Vgl. zu diesem Thema besonders Fodor (1987, Kap. 2). In (1986) hat Burge zu zeigen versucht, dass auch die wissenschaftliche Psychologie – im Gegensatz zu dem, was man nach Fodors Argumentation erwarten dürfte – *de facto* anti-individualistische Kriterien für die Zuschreibung intentionaler Zustände verwendet. Für eine ausführlichere Diskussion von Burges Anti-Individualismus sowie Fodors Einwänden vgl. Beckermann (2001, S. 363–381).

lediglich sein, dass wir verschiedene Maßstäbe anwenden. Im ersten Fall verwenden wir die Sprachgemeinschaft von $Elmar_1$ als Maßstab und im zweiten Fall die Sprachgemeinschaft von $Elmar_2$. So betrachtet scheint ein großer Vorteil der messtheoretischen Interpretation intentionaler Prädikate zu sein, dass diese Interpretation sowohl den Argumenten Fodors also auch den Argumenten Burges gerecht wird und dass sie dabei eine – in meinen Augen – sehr befriedigende Lösung der Streitfrage um die individualistische oder anti-individualistische Individuation intentionaler Zustände liefert.[28]

## Literatur

Beckermann, A. (1986): „Dennetts Stellung zum Funktionalismus". *Erkenntnis* 24, 309–341.

Beckermann, A. (1992a): „Introduction – Reductive and Nonreductive Physicalism". In: A. Beckermann, H. Flohr & J. Kim (1992), 1–21.

Beckermann, A. (1992b): „Supervenience, Emergence, and Reduction". In: A. Beckermann, H. Flohr & J. Kim (1992), 94–118.

[28] Eine verwandte Überlegung findet sich auch in Loar (1988). Dort zeigt Loar mit Hilfe einiger sehr hilfreicher Beispiele, dass es einerseits Fälle gibt, in denen wir dieselben Inhaltssätze für die Zuschreibung von Überzeugungen verwenden, obwohl die zugeschriebenen Überzeugungen psychologisch betrachtet sehr verschieden sind, und dass es andererseits Fälle gibt, in denen wir verschiedene Inhaltssätze zur Zuschreibung von Überzeugungen verwenden, die – psychologisch betrachtet – ununterscheidbar sind. Loar schließt daraus, dass man zwischen einem ‚sozialen' und einem ‚psychologischen' Inhalt unterscheiden müsse. Mit anderen Worten: Wie Fodor gelangt er zu der Ansicht, dass es außer dem ‚weiten', sozialen Inhalt, auch noch einen ‚engen', psychologischen Inhalt geben muss. Meiner Meinung nach ist diese Schlussfolgerung allerdings nicht gerechtfertigt, weil sie auf der Voraussetzung beruht, dass zwei (typ-) identische intentionale Zustände, die wir alltagspsychologisch mit verschiedenen Inhaltssätzen zuschreiben, doch in gewisser Weise denselben Inhalt haben müssen – einfach weil es sich hier um intentionale Zustände handelt. Dem messtheoretischen Ansatz zufolge haben die mit intentionalen Prädikaten zugeschriebenen Zustände aber selbst gar keinen Inhalt. Daher können (typ-)identische Zustände *a forteriori* nicht dadurch charakterisiert sein, dass sie in der einen oder anderen Hinsicht denselben Inhalt haben. Die messtheoretische Interpretation erlaubt uns daher, die Streitfrage bezüglich der individualistischen oder anti-individualistischen Individuation intentionaler Zustände für beide Parteien zufriedenstellend zu lösen, ohne ‚enge' Inhalte ins Spiel zu bringen. In meinen Augen ist dies ein weiteres sehr zufriedenstellendes Ergebnis.

Beckermann, A. (1992c): „Wie real sind intentionale Zustände? Dennett zwischen Fodor und den Churchlands“. In: H. J. Sandkühler (Hg.) *Wirklichkeit und Wissen. Wirklichkeitskonzeptionen in Philosophie und Wissenschaften*. Peter Lang, Frankfurt/M., 151–176.

Beckermann, A. (1992d): „Das Problem der Intentionalität – Naturalistische Lösung oder meßtheoretische Auflösung?“ *Ethik und Sozialwissenschaft* 3, 433–447, 502–512, 520–522.

Beckermann, A. (1996a): „Eigenschafts-Physikalismus“. *Zeitschrift für philosophische Forschung* 50, 3–25.

Beckermann, A. (1996b): „Is There a Problem about Intentionality?“ *Erkenntnis* 45, 1–23.

Beckermann, A. (1997): „Property Physicalism, Reduction and Realization“. In: M. Carrier & P. Machamer (Hg.) *Mindscapes. Philosophy, Science, and the Mind*. Universitätsverlag, Konstanz/Pittsburgh University Press, Pittsburgh, 303–321.

Beckermann, A. (2001): *Analytische Einführung in die Philosophie des Geistes*. 2., überarbeitete Aufl., Walter de Gruyter, Berlin/New York.

Beckermann, A., H. Flohr & J. Kim (Hg.) (1992): *Emergence or Reduction? – Essays on the Prospects of Nonreductive Physicalism*. Walter de Gruyter, Berlin/New York.

Brentano, F. (1924): *Psychologie vom empirischen Standpunkt*. Hrsg. von O. von Kraus, Meiner Verlag, Leipzig.

Burge, T. (1979): „Individualism and the Mental“. *Midwest Studies in Philosophy* 4, 73–121.

Burge, T. (1986): „Individualism and Psychology“. *Philosophical Review* 95, 3–45.

Churchland, P. M. (1979): *Scientific Realism and the Plasticity of Mind*. Cambridge University Press, Cambridge.

Cummins, R. (1989): *Meaning and Mental Representation*. MIT Press, Cambridge MA.

Davidson, D. (1974): „Belief and the Basis of Meaning“. *Synthese* 27, 309–323. Wiederabgedruckt in: D. Davidson, *Inquiries into Truth and Meaning*. Clarendon Press, Oxford 1984, 141–154.

Davidson, D. (1989): „What is Present to the Mind?“ *Grazer Philosophische Studien* 36, 3–18.

Dennett, D. (1982): „Beyond Belief“. In: A. Woodfield (Hg.) *Thought and Object*. Clarendon Press, Oxford, 1–95. Wiederabgedruckt in D. Dennett (1987a), 117–202.

Dennett, D. (1987a): *The Intentional Stance*. MIT Press, Cambridge MA.

Dennett, D. (1987b): „About Aboutness“. In: D. Dennett (1987a), 203–211.

Dretske, F. (1981): *Knowledge and the Flow of Information*. Blackwell, Oxford.

Dretske, F. (1986): „Misrepresentation“. In: R. J. Bogdan (Hg.) *Belief – Form, Content, and Function*. Clarendon Press, Oxford, 17–36.

Field, H. (1980): „Postscript to ‚Mental Representation‘“. In: N. Block (Hg.) *Readings in the Philosophy of Psychology*. Vol. 2. MIT Press, Cambridge MA, 112–114.

Fodor, J. (1987): *Psychosemantics*. MIT Press, Cambridge MA.

Fodor, J. (1991): *A Theory of Content and Other Essays*. MIT Press, Cambridge MA.

Haugeland, J. (1981): „Semantic Engines“. In: J. Haugeland (Hg.), *Mind Design*. MIT Press, Cambridge MA, 1–34.

Haugeland, J. (1985): *Artificial Intelligence*. MIT Press, Cambridge MA.

Hempel, C. G. (1952): *Fundamentals of Concept Formation in Empirical Science*. Chicago.

Lanz, P. (1987): *Menschliches Handeln zwischen Kausalität und Rationalität*. Athenäum, Frankfurt/M.

Loar, B. (1981): *Mind and Meaning*. Cambridge University Press, Cambridge.

Loar, B. (1988): „Social Content and Psychological Content“. In: R. H. Grimm & D. D. Merrill (Hg.) *Contents of Thought.* University of Arizona Press, Tucson, 99–110.

Matthews, R. (1994): „The Measure of Mind“. *Mind* 103, 131–146.

Millikan, R. (1984): *Language, Thought, and Other Biological Categories*. MIT Press, Cambridge MA.

Millikan, R. (1989): „Biosemantics“. *Journal of Philosophy* 86, 281–97.

Quine, W. V. O. (1970): *Philosophy of Logic*. Prentice-Hall, Englewood Cliffs, NJ.

Papineau, D. (1985): „Representation and Explanation“. *Philosophy of Science* 51, 550–572.

Papineau, D. (1988): *Reality and Representation*. Basil Blackwell, Oxford.

Stalnaker, R. (1984): *Inquiry*. MIT Press, Cambridge MA.

Suppes, P. & J. Zinnes (1963): „Basic Measurement Theory“. In: R. D. Luce et al. (Hg.), *Handbook of Mathematical Psychology*. Bd. 1. New York.

# 11

## Visuelle Informationsverarbeitung und phänomenales Bewußtsein[*1]

1. Soweit es um ein angemessenes Verständnis phänomenalen Bewußtseins geht, sind am Informationsverarbeitungsansatz orientierte repräsentationalistische Theorien des Geistes – ebenso wie entsprechende neurobiologische oder funktionalistische Theorien – mit einer Reihe von Argumenten konfrontiert, die auf „inverted-“ oder „absent-qualia“-Überlegungen beruhen. Das Grundmuster dieser Überlegungen lautet: Selbst wenn wir vollständig über die neuronalen oder funktionalen oder repräsentationalen Zustände informiert wären, die dem Auftreten phänomenalen Bewußtseins zugrunde liegen, wäre es immer noch *vorstellbar*, daß diese neuronalen Zustände (oder Zustände mit derselben kausalen Rolle bzw. derselben repräsentationalen Funktion) auftreten, ohne überhaupt einen phänomenalen Gehalt zu besitzen, bzw. daß diese Zustände mit phänomenalen Gehalten einhergehen, die sich von den üblichen erheblich unterscheiden.

Auf den ersten Blick scheinen diese Argumente durchaus eine gewisse Plausibilität zu besitzen. Im Falle repräsentationalistischer Theorien beruht diese Plausibilität jedoch weitgehend darauf, daß man sich auf die repräsentationalen Zustände selbst beschränkt und die spezifische Weise des Zustandekommens der entsprechenden Repräsentationen völlig außer acht läßt. Was ich damit meine, möchte ich am Fall der visuellen Wahrnehmung verdeutlichen.

Nehmen wir an, Harvey sieht, daß vor ihm auf dem Tisch ein Glas steht. Wie würde dieser Zustand von einer repräsentationalistischen Theorie analysiert? Nun, in repräsentationalistischen Theorien finden sich in der Regel zwar Analysen für mentale Zustandstypen wie Überzeugungen und Wünsche; Wahrnehmungsprozesse werden im allgemeinen meines Wissens jedoch nicht thematisiert. Es scheint allerdings nicht unplausibel anzunehmen, daß für einen Repräsentationalisten Wahrnehmungsprozesse in erster Linie Prozesse des Erwerbs von Überzeugungen sind; für ihn werden also die Meinungen im Vordergrund stehen, in denen Wahrnehmungsprozesse resultieren. Und für diesen Typ mentaler Zustände hat er eine Analyse. Wenn Harvey sieht, daß vor ihm auf dem Tisch ein Glas steht, dann besteht dieser Zustand also unter anderem darin, daß Harvey in einer bestimmten funktionalen/computationalen Relation *R* zu einer mentalen Repräsentation

---

* Erstveröffentlichung in: T. Metzinger (Hg.) *Bewußtsein*. 2. Aufl., Paderborn: Schöningh 1996, 663–679.

1 Ich möchte Achim Stephan und Antonia Barke für ihre hilfreichen Kommentare zu einer früheren Version dieses Aufsatzes danken.

*mr* steht, die den Inhalt hat, daß auf dem Tisch direkt vor Harvey ein Glas steht. Oder salopp ausgedrückt, daß sich die mentale Repräsentation *mr* in Harveys belief-box befindet. Um Wahrnehmungsüberzeugungen von anderen Überzeugungen zu unterscheiden, wird man zusätzlich allerdings noch etwas darüber sagen müssen, wie *mr* in Harveys belief-box gelangt – z. B. daß *mr* mehr oder weniger direkt durch die Tatsache, die ihren Inhalt ausmacht, verursacht wird und daß darüber hinaus in diesem Verursachungsprozeß Harveys Augen und das von den beteiligten Objekten reflektierte Licht eine zentrale Rolle spielen.

Eine repräsentationalistische Analyse des Zustands, daß Harvey vor sich auf dem Tisch ein Glas sieht, könnte also so aussehen:

(a) In Harveys belief-box befindet sich eine mentale Repräsentation *mr* mit dem Inhalt, daß auf dem Tisch direkt vor Harvey ein Glas steht.
(b) Daß sich diese mentale Repräsentation jetzt in Harveys belief-box befindet, wird u. a. direkt durch die Tatsache verursacht, die ihren Inhalt ausmacht, und zwar auf eine Weise, bei der Harveys Augen und das von den beteiligten Objekten reflektierte Licht eine zentrale Rolle spielen.

Wenn man von dieser Analyse ausgeht, steht man jedoch sofort vor den oben angesprochenen Problemen. Wenn alles, was sich darüber sagen läßt, daß Harvey vor sich auf dem Tisch ein Glas sieht, in den beiden Punkten (a) und (b) zusammengefaßt ist, dann scheint es durchaus *vorstellbar*, daß dieser Zustand überhaupt keinen phänomenalen Gehalt besitzt bzw. daß er mit einem ganz anderen als dem üblichen phänomenalen Gehalt verbunden ist. Dies liegt meiner Meinung nach jedoch nur daran, daß im Punkt (b) zu wenig über die *Weise* gesagt wird, in der Harveys Wahrnehmungsüberzeugung zustandekommt. Mit anderen Worten: Der phänomenale Gehalt von Wahrnehmungszuständen kann im Rahmen einer repräsentationalen Theorie des Geistes nur dann angemessen erklärt werden, wenn man sich nicht nur auf mentale Repräsentationen konzentriert, sondern auch die Art und Weise genau analysiert, wie diese Repräsentationen zustande kommen. Um diese These zu illustrieren, möchte ich im folgenden die Grundzüge visueller Informationsverarbeitung skizzieren, so wie sie im Augenblick von den Kognitionswissenschaften bzw. der KI-Forschung gesehen werden.

2. Visuelle Informationsverarbeitung beginnt mit Netzhautbildern (genauer: Verteilungen der Feuerungsraten der Photosensoren in der Netzhaut) bzw. im Fall künstlicher Systeme mit von einer Fernsehkamera erzeugten Rohbildern, die in Pixelmatrizen kodiert werden. Der Gesichtssinn ist jedoch ein Fernsinn, der uns über distale Reize informieren soll. Netzhautbilder bzw. Rohbilder sind daher nur interessant, wenn aus ihnen Informationen über die Umweltszene gewonnen werden können, durch die sie hervorgeru-

fen wurden. Am Ende des Verarbeitungsprozesses muß daher eine Beschreibung bzw. Repräsentation dieser Umweltszene stehen. Kurz gesagt: Visuelle Informationsverarbeitung beginnt mit dem Netzhautbild bzw. einem Rohbild und sie endet mit einer Repräsentation der Umweltszene, durch die das Netzhautbild bzw. Rohbild hervorgerufen wurde.

Wenn man den Gesamtprozeß des Sehens betrachtet, lassen sich also zumindest drei Komponenten unterscheiden:[2]

(a) physikalische Objekte in einer Szene
(b) Bilder der Szene als Eingabe
(c) eine Beschreibung oder Repräsentation der Szene als Ausgabe

Diese Komponenten wirken folgendermaßen zusammen. Die in einer (im allgemeinen dreidimensionalen) Szene angeordneten physikalischen Objekte erzeugen unter normalen Beleuchtungsverhältnissen zuerst im Eingabemedium des betreffenden Systems ein Bild, das neuronal oder elektronisch codiert wird. (Bilder in dem hier relevanten Sinn sind also nicht mehr und nicht weniger als zweidimensionale Projektionen dreidimensionaler Szenen.) Die Aufgabe visueller Informationsverarbeitung ist es, im zweiten Schritt aus diesem Bild[3] – sozusagen auf dem Weg einer „inversen Optik" – die Szene zu rekonstruieren, die zu diesem Bild geführt hat. Der Prozeß visueller Informationsverarbeitung muß somit zu einer Ausgabe führen, aus der hervorgeht, „wo sich was" in dieser Szene befindet.

> Das „wo" bezieht sich auf räumlich-zeitliche Informationen, also auf die Rekonstruktion der Szenengeometrie. Das „was" impliziert eine Deutung des Szeneninhalts, also insbesondere die Erkennung von Objekten. (Neumann 1993: 566–567)

Visuelle Informationsverarbeitung besteht also in der Rekonstruktion einer Szene aus einem von dieser Szene verursachten Bild. Oder präziser: Aus der (Re-)Konstruktion einer Repräsentation einer Szene aus einer Kodierung eines von dieser Szene verursachten Bildes.

Heute wird im allgemeinen angenommen, daß sich innerhalb dieses Rekonstruktionsprozesses vier Stufen unterscheiden lassen,[4] die zumindest

---

[2] Vgl. Neumann 1993: 566.

[3] Im allgemeinen reicht ein einzelnes Bild allerdings nichts aus, um diese Rekonstruktion vornehmen zu können. In der Praxis geht man daher nicht von Einzelbildern, sondern von Bildfolgen aus. Es ist klar, daß die Rekonstruktion (zeitlich ausgedehnter) dynamischer Szenen nur auf der Grundlage solcher Bildfolgen möglich ist.

[4] Vgl. zum folgenden Neumann 1993: 569–570.

nach Marr (1982) zu immer neuen Repräsentationen führen, bis am Ende eine Repräsentation der relevanten Umweltszene steht.[5]

Die *primäre Bildanalyse* hat die Segmentierung des Bildes zum Ziel. Dafür entscheidend ist die Herausarbeitung und Repräsentation der zentralen *Bildelemente* wie Kanten, homogene Bereiche, Textur, etc.

Bei der *niederen Bilddeutung* geht es schon darum, Bildelemente als Szenenelemente zu interpretieren, d. h. als Abbildungen von Teilen realer, dreidimensionaler Szenen. Beispielsweise können in diesem Schritt eine Bildkante als Schattengrenze, ein roter Bereich als Hauswand, ein grüner texturierter Bereich als Grasfläche gedeutet werden.

Der dritte Verarbeitungsschritt gilt der *Objekterkennung*. Objekte werden in den bisher extrahierten Bilddaten und auf der Basis der Szenenelemente erkannt. Dabei ist erhebliches Vorwissen über das Aussehen von Objekten bei Betrachtung aus unterschiedlichen Blickrichtungen erforderlich. Denn Objekterkennung ist wesentlich eine Umkehrung des Abbildungsprozesses.

Weitere Verarbeitungschritte werden unter dem Stichwort *Höhere Bilddeutung* zusammengefaßt. „Sie haben in der Regel das Ziel, objekt- und zeitübergreifende Zusammenhänge zu erkennen, z. B. interessante Objektkonfigurationen, spezielle Situationen, zusammenhängende Bewegungsabläufe u. a. Ähnlich wie bei der Objekterkennung spielt hier modellhaftes Vorwissen über das, was man erkennen will, eine wichtige Rolle." (Neumann 1993: 567)

Schematisch läßt sich der Ablauf visueller Informationsverarbeitung so darstellen:

Rohbild
↓
Primäre Bildanalyse
(Bildsegmentierung: Kanten, homogene Flächen, Textur, etc.)
↓
Niedere Bilddeutung
(Szenenelemente)
↓
Objekterkennung
↓
Höhere Bilddeutung
(Objektkonfigurationen, Situationen, Ereignisse)

[5] Die folgenden vier Stufen werden auf jeden Fall im Bereich der KI bei der Konstruktion künstlicher visueller Systeme zugrundegelegt. Ob sie in derselben Form auch bei natürlichen Systemen zu finden sind, ist nicht ganz klar. Vgl. zu diesem Punkt unten Abschnitt 3.

3. Aus Gründen möglichst ökonomischer Speichernutzung versucht man in der KI, die Abfolge der gerade angeführten Schritte so zu organisieren, daß alle späteren Schritte der Bildverarbeitung ihre Aufgabe möglichst auf der Basis von Bildelementen, also den Ergebnissen der primären Bildanalyse, ohne Rückgriff auf Rohbilder erledigen können. Eine Farbbildfolge mit der Auflösung eines Fernsehbildes und der Dauer von 10 Sekunden z. B. hat das beträchtliche Datenvolumen von 220 GByte. Es scheint deshalb sinnvoll, den verfügbaren Speicher nur so kurz wie möglich mit Rohbildern zu belasten.

Für die visuelle Informationsverarbeitung bei uns Menschen scheint jedoch charakteristisch zu sein, daß „Rohbilder" nach der primären Bildanalyse nicht einfach „weggeworfen" werden. Wenn wir unserer Introspektion trauen können, scheint es sogar so, daß bei der Verarbeitung von Netzhautbildern nicht sukzessiv immer neue, *voneinander unabhängige* Repräsentationen erzeugt werden. Vielmehr scheinen die Ergebnisse der einzelnen Verarbeitungsschritte sehr eng miteinander verwoben zu sein und jeweils auch zu einer Verbesserung oder Schärfung des Ausgangsbildes zu führen.

Nach Beendigung der primären Bildanalyse erscheinen die homogenen Bildflächen schärfer herausgearbeitet, die Konturen zwischen diesen Flächen sind klarer, das Bild sieht insgesamt „schärfer" aus. Mit der niederen Bilddeutung beginnt die Interpretation des Bildes; wir sehen nicht mehr weiße, graue und unterschiedlich farbige Flächen bzw. verschieden konturierte Kanten, sondern die Oberfläche eines Tisches, eine Schattenlinie oder einen grüngrau gemusterten Hintergrund. Es ist so, als würden den einzelnen Bildelementen Etiketten angeheftet, auf denen jeweils vermerkt ist, welche Szenenelemente für diese Bildelemente verantwortlich sind. Nach dem Schritt der Objekterkennung haben wir gar nicht mehr den Eindruck, ein Bild zu sehen. Stattdessen schauen wir gewissermaßen durch das Bild hindurch auf die Objekte, die es erzeugt haben.[6] Wir sehen jetzt einen

---

[6] Van Gulick spricht (vgl. z. B. 1989: 223 ff.) in diesem Zusammenhang von der hohen „semantischen Transparenz" phänomenaler Repräsentationen. Dies ist sicher ein sehr suggestiver Ausdruck; allerdings entspricht er nicht ganz den hier vorgetragenen Überlegungen. Denn nach Van Gulick gilt für den Begriff der semantischen Transparenz:
„*[The notion of semantic transparency concerns the] extent to which a system can be said to understand the content of the internal symbols or representations on which it operates.*" (a.a.O.: 223).
Damit ist aber vorausgesetzt, daß es sich bei Wahrnehmungsbildern um Repräsentationen handelt, und dies ist eine Annahme, die ich zumindest für diskussionswürdig halte (vgl. unten Abschnitt 4). Darüber hinaus kommt Van Gulick mit seiner Auffassung einer homunculus-Annahme zumindest gefährlich nahe. Denn wenn man von einem System sagt, es verstehe die Bedeutung

Tisch, ein Glas, eine gemusterte Tapete usw. Auch dieser Schritt führt in der Regel zu einer schärferen Konturierung des Ausgangsbildes. Denn das Vorwissen darüber, wie die wahrgenommenen Objekte unter den gegebenen Bedingungen *im allgemeinen* aussehen, wird offenbar dazu verwendet, das aktuelle Ausgangsbild entsprechend zu ergänzen und zu verbessern. Schließlich sehen wir sogar die Eigenschaften und räumlichen Beziehungen der wahrgenommenen Objekte – wir sehen, daß der Tisch weiß ist, daß das Glas auf dem Tisch steht etc.

Auch unsere eigene Introspektion belegt somit, daß es im Wahrnehmungsprozeß Stufen gibt, die den von an KI-Modellen orientierten Kognitionswissenschaftlern postulierten vier Stufen visueller Informationsverarbeitung entsprechen.[7] Aber diese Stufen führen nicht zu *voneinander unabhängigen* Repräsentationen. D.h. nicht, daß nicht jede dieser Stufen auch zu neuen Repräsentationen führt, sondern nur daß diese Repräsentationen nicht unabhängig voneinander sind; sie modifizieren immer auch die jeweils vorangehenden Repräsentationen, und alle führen zu einer Veränderung und Verschärfung des Ausgangsbildes. Am Ende haben wir sogar den Eindruck, nicht mehr ein Bild, sondern direkt Objekte und Szenen zu sehen. Und dies trifft in gewissem Sinne sicher zu. Aber das modifizierte Ausgangsbild ist offenbar ebenfalls immer noch verfügbar.

Es ist diese Tatsache, auf die uns Sinnesdatentheoretiker immer wieder aufmerksam zu machen versucht haben. G.E. Moore z.B. lädt uns dazu ein, einmal darauf zu achten, was genau passiert, wenn wir – ohne den Blick auf eine Szene zu verändern – mit einem Finger auf einen Augapfel drücken. An der wahrgenommenen Szene ändert sich in der Regel nichts. Wir sehen immer noch dieselben Gegenstände in derselben Konfiguration. Aber irgend etwas ändert sich eben doch; und das ist das Ausgangsbild der Szene. Genauer gesagt, die Bildelemente (homogene Flächen, Kanten, etc.), die von den einzelnen Objekten hervorgerufen werden, verändern ihre Form. Worauf Moore hinweist, ist ein Vorgang, der uns eigentlich allen vertraut ist. Wir sehen zwar normalerweise „durch die Bilder hindurch“ die

---

seiner internen Repräsentationen, muß es in diesem System dann nicht eine interpretierende Instanz – eine Art homunculus – geben? Die Auffassung, die hier von mir vertreten wird, kommt jedoch völlig ohne homunculus-Annahmen aus. Denn ihr zufolge besteht die „semantische Transparenz“ von Wahrnehmungsbildern nur darin, daß diese Bilder vom System auf's Engste mit expliziten Repräsentationen der wahrgenommenen Szene verknüpft werden.

[7] Daß auch bei uns der Prozeß der visuellen Informationsverarbeitung verschiedene Stufen durchläuft, zeigt sich in voller Deutlichkeit allerdings erst in psychologischen Experimenten, in denen Wahrnehmungsreize nur für den Bruchteil einer Sekunde dargeboten werden.

Objekte und Szenen, auf die diese Bilder zurückgehen, aber wir *können* uns auch auf die Bilder und ihre Elemente konzentrieren. Mit anderen Worten, die Ausgangsbilder sind, wenn auch in modifizierter Form, am Ende des Wahrnehmungsprozesses immer noch vorhanden.

4. Nach dem bisherigen Gang der Argumentation ist es wohl kaum erstaunlich, daß meine zentrale These lautet: Wenn Prozesse visueller Informationsverarbeitung auf die gerade geschilderte Weise strukturiert sind, d. h. wenn während dieser Prozesse Repräsentationen der Ursprungsszene aus Ausgangsbildern (re)konstruiert werden, und zwar auf eine Weise, die nicht zum Verlust, sondern nur zur Modifikation der Ausgangsbilder führt, dann besitzen solche Prozesse auch einen phänomenalen Aspekt. Zumindest gilt dies, wenn die modifizierten Ausgangsbilder dem entsprechenden System genauso zugänglich sind wie die expliziten Repräsentationen der wahrgenommenen Szene. Entscheidend für den phänomenalen Charakter entsprechender Wahrnehmungsprozesse ist meiner Ansicht nach also, daß neben expliziten Repräsentationen der wahrgenommenen Szene auch die Ausgangsbilder erhalten bleiben.

Angesichts dieser These kann allerdings sofort wieder die Frage gestellt werden, ob es nicht auch in diesem Fall zumindest vorstellbar ist, daß in einem System Prozesse visueller Informationsverarbeitung in der gerade geschilderten Weise ablaufen, ohne überhaupt einen phänomenalen Gehalt zu besitzen, bzw. daß diese Prozesse mit phänomenalen Gehalten verbunden sind, die sich von den üblichen erheblich unterscheiden. Bevor ich versuche, eine Antwort auf diese Frage zu geben, möchte ich noch drei Bemerkungen machen, die jedoch ebenfalls den Zweck haben, die Plausibilität meiner Hauptthese zu erhöhen. Die erste Bemerkung betrifft die Beziehungen, die zwischen dieser These und anderen in der Literatur vertretenen Positionen bestehen.

In der Literatur ist immer wieder die Ansicht vorgetragen worden, daß phänomenale Zustände im Rahmen einer repräsentationalistischen Theorie des Geistes durch eine bestimmte Art von Repräsentationen erklärt werden können – nämlich durch analoge bzw. piktorielle Repräsentationen.[8] Der Zusammenhang dieser Auffassung mit der hier vertretenen These ist offensichtlich. Allerdings bin ich – im Gegensatz zu vielen anderen Autoren – nicht der Meinung, daß Bilder ohne weiteres als Repräsentationen angesehen werden können.

Während von den meisten Autoren externe Bilder (Photographien, Zeichnungen, etc.) geradezu als Paradefälle von Repräsentationen betrachtet werden, sind Bilder in dem hier einschlägigen Sinn zunächst einmal

[8] Siehe z. B. Nelkin 1989, 1994.

nichts weiter als zweidimensionale Projektionen dreidimensionaler Szenen. Darüber hinaus sind Netzhautbilder und Rohbilder sicher auch kausale Spuren der Szenen, durch die sie hervorgerufen werden. Allerdings sind sie wegen der charakteristischen Vieldeutigkeit von Bildern in der Regel nicht einmal natürliche Zeichen dieser Szenen. Was spricht also dafür, Netzhautbilder und Rohbilder überhaupt als Repräsentationen anzusehen? Meiner Meinung nach nichts. Sie sind kausale Spuren, aber auch nicht mehr als das.

Auf der anderen Seite ist es allerdings die natürliche Aufgabe unseres Wahrnehmungssystems, Netzhautbilder zu interpretieren,[9] d. h. aus Netzhautbildern Repräsentationen der zugrunde liegenden Szenen zu rekonstruieren, wobei die Gesetze der inversen Optik sowie bestimmte Hintergrundannahmen ausgenutzt werden. Und genau dies ist die Grundlage dafür, daß für uns auch die Interpretation externer Bilder im allgemeinen kein Problem ist. Denn externe Bilder erzeugen – *cum grano salis* – dieselben Netzhautbilder wie die dazugehörigen Szenen. Also ist es alles andere als erstaunlich, daß unser Wahrnehmungssystem aus diesen Netzhautbildern mehr oder weniger automatisch die entsprechenden Außenweltszenen rekonstruiert.[10]

Aber der Unterschied zwischen Bildern und analogen Repräsentationen ist an dieser Stelle nicht entscheidend. Denn die meisten Autoren, die phänomenale Zustände im Rahmen einer repräsentationalistischen Theorie des Geistes durch eine bestimmte Art von Repräsentationen erklären wollen, denken dabei an piktorielle oder bildhafte Repräsentationen. Es scheint also die gemeinsame Intuition zu geben, daß Bilder oder bildhafte Repräsentationen gute Kandidaten für die Erklärung phänomenaler Zustände sind – und zwar besonders dann, wenn es um die Erklärung visueller Eindrücke geht.

Eine weitere Position, zu der ich eine große Nähe sehe, ist die, die N. Humphrey in seinem neuen Buch *A History of the Mind* entwickelt hat. Nach Humphrey haben Lebewesen bei der Entwicklung der Fähigkeit zur Wahrnehmung im Laufe der Evolution zwei deutlich unterschiedene Repräsentationssysteme entwickelt.

---

[9] Aus dieser Tatsache allein kann man sicher nicht schließen, daß Netzhautbilder und Rohbilder Repräsentationen sind. Denn auch der Kriminalist interpretiert die ihm zugänglichen Spuren, ohne daß diese deshalb Repräsentationen der Tat oder des Täters sein müßten.

[10] Aus dieser einfachen Beobachtung ergibt sich meiner Meinung nach auch eine ebenso einfache Antwort auf die viel diskutierte Frage, worin eigentlich die Ähnlichkeit von Bildern mit den von ihnen „dargestellten" Szenen bestehen soll. Bilder sind den von ihnen dargestellten Szenen in dem Maße ähnlich, in dem sie dieselben Netzhautbilder erzeugen.

Seit es Lebewesen gibt, stehen sie in ständigem direkten Kontakt mit ihrer Umwelt, d. h. diese Umwelt wirkt unmittelbar auf sie ein. Licht fällt auf sie, sie stoßen mit anderen Dingen zusammen, ihre Oberfläche wird durch Druckwellen erschüttert oder kommt in Kontakt mit fremden chemischen Stoffen. Einige dieser Ereignisse sind gut für das Lebewesen, andere schlecht. Also bietet es einen evolutionären Vorteil, wenn das Lebewesen lernt, gute von schlechten Ereignissen zu unterscheiden oder, um es am Anfang noch nicht zu kompliziert werden zu lassen, in den verschiedenen Fällen verschieden zu reagieren. „Natural selection was therefore likely to select for ‚sensitivity'" (Humphrey 1993: 18). Diese Reaktionen sind zu Beginn noch lokal: die Oberfläche zieht sich zusammen, verhärtet sich oder sondert bestimmte chemische Stoffe ab. Eine wichtige neue Stufe ist erreicht, wenn Lebewesen die Fähigkeit entwickeln, Signale von einem Teil ihrer Oberfläche zu anderen Teilen weiterzuleiten, so daß es erst dort zu entsprechenden Reaktionen kommt. Dies ermöglicht es ihnen nämlich, z. B. wegzuschwimmen statt einfach nur zurückzuzucken.

Auch jetzt gibt es aber noch einen direkten Zusammenhang zwischen Reizung („Wahrnehmung") auf der einen und Reaktion („Handlung") auf der anderen Seite. Die nächste Stufe ist erreicht, wenn dieser direkte Zusammenhang aufgehoben wird. Signale von Teilen der Oberfläche werden zwar weitergeleitet, aber sie führen jetzt nicht mehr automatisch zu bestimmten Reaktionen. Ob und in welcher Weise reagiert wird, hängt vielmehr von einer Reihe von anderen Faktoren ab, die, wenn man so will, in einem zentralen „Entscheidungsprozeß" zur Geltung kommen. An dieser Stelle ist es offenbar noch nicht nötig, die von der Oberfläche kommenden Informationen zu speichern oder dauerhaft zu repräsentieren. Dies wird aber zwingend erforderlich, wenn es sich als sinnvoll erweist, die Reaktion auch *zeitlich* vom Reiz abzukoppeln. Erste Repräsentationssysteme entwickeln sich in Lebewesen also, wenn nicht nur der direkte Zusammenhang zwischen Reiz und Reaktion aufgebrochen ist, sondern die Reaktionen auch zeitlich von den Reizen abgekoppelt sind.

Auf die Frage, *was* diese Repräsentationen repräsentieren, liegt zunächst die Antwort nahe: Die proximalen Reize, das Licht, das auf die Oberfläche fällt, die Objekte, an die das Lebewesen stößt, die chemischen Stoffe, mit denen es in Berührung kommt. Wenn man sich die Sache etwas genauer ansieht, wird aber klar, daß die Informationen, die von der Oberfläche weitergeleitet werden, weniger von den proximalen Reizen abhängen als von den Wirkungen, die diese Reize auf die Oberfläche haben. Wenn zwei verschiedene Objekte die Oberfläche in der gleichen Weise verändern, werden gleiche Informationen weitergeleitet; wenn dasselbe Objekt zu verschiedenen Zeitpunkten verschiedene Wirkungen hat, werden dagegen auch die weitergeleiteten Informationen verschieden sein. Mit anderen Worten: Die

ersten zentralen Repräsentationen, die sich im Laufe der Zeit in Lebewesen ausbilden, tragen Informationen weniger über die proximalen Reize als über den Zustand des Lebewesens selbst bzw. – um genauer zu sein – über den Zustand der verschiedenen Teile seiner Oberfläche, die von den proximalen Reizen verändert werden, über Drücke, über Wärme, über Berührungsveränderungen. Es ist daher kein Wunder, daß Humphrey diese Repräsentationen mit den „raw sensations" identifiziert.

> So the phenomenology of sensory experiences came first. Before there were any kinds of phenomena there were ‚raw sensations' – tastes, smells, tickles, pains, sensations of warmth, of light, of sound and so on. (a.a.O.: 21)

Die ersten Repräsentationen geben dem Lebewesen also nur Antworten auf die Frage ‚What is happening to me?'. Ohne Zweifel ist es jedoch für das Überleben der meisten Lebewesen ebenso wichtig zu wissen, was um sie herum vorgeht. D.h., für sie sind auch Antworten auf die Frage ‚What is happening out there?' von äußerstem Interesse. Wie lassen sich aber Repräsentationssysteme entwickeln, die Antworten auf diese Frage enthalten?

Kurz gesagt lautet Humphreys These, daß sich im Laufe der Evolution neben dem ersten Repräsentationssystem ein zweites Repräsentationssystem entwickelt hat, das ebenfalls von den von der Oberfläche kommenden Signalen ausgeht, diese Signale aber auf eine ganz andere Weise verarbeitet.

> By the end of the first stage of evolution sense organs existed with connections to a central processor, and most of the requisite information about potential signs was being received as ‚Output'. But the subsequent processing of this information, leading to subjective sensory states, had to do with quality rather than quantity, the transient present rather than permanent identity, meness rather than otherness. In order that the same information could now be used to represent the outside world, a whole new style of processing had to evolve, with an emphasis less on the subjective present and more on object permanence, less on immediate responsiveness and more on future possibilities, less on what it is like for me and more on how what ‚it' signifies fits into the larger picture of a stable external world.
>
> To cut a long story short, there developed in consequence two distinct kinds of mental representation, involving very different styles of information processing. While one path led to the qualia of subjective feelings and first-person knowledge of the self, the other led to the intentional objects of cognition and objective knowledge of the external physical world. (a.a.O.: 22)

Der zentrale Punkt der Theorie Humphreys läßt sich also so zusammenfassen:

1. Im Laufe der Evolution haben sich zwei grundsätzlich verschiedene Weisen entwickelt, Reize, die auf die Oberfläche von Lebewesen gelangen, zu verarbeiten (sensation and perception). Die erste Weise führt

zu Repräsentationen, in denen Antworten auf die Frage ‚What is happening to me?' kodiert sind; die zweite zu Repräsentationen, aus denen hervorgeht ‚what is happening out there'.
2. Repräsentationen der ersten Art machen den phänomenalen Aspekt von Wahrnehmungsprozessen aus.
3. Auch wenn beide Arten der Informationsverarbeitung in den verschiedenen Wahrnehmungsprozessen normalerweise eng miteinander verknüpft sind, handelt es sich doch um verschiedene Prozesse, die auch unabhängig voneinander auftreten können.

Ohne Zweifel gibt es gewisse Unterschiede zwischen dem hier entwickelten Modell und dieser Theorie Humphreys. Aber es gibt offenbar auch viele Parallelen. Z. B. können die Repräsentationen von Ausgangsbildern, die der hier vertretenen Theorie zufolge für den phänomenalen Aspekt visueller Wahrnehmung verantwortlich sind, ohne weiteres als Repräsentationen dessen aufgefaßt werden, was auf der Retina geschieht – also als Repräsentationen des ersten Typs der Humphreyschen Theorie. Aber ich will diese Details hier nicht weiter verfolgen. Denn die grundsätzliche Verwandtschaft der beiden Ansätze ergibt sich, denke ich, auch schon aus diesen kurzen Erläuterungen.

Die zweite Bemerkung bezieht sich auf die viel diskutierte Frage, welchen evolutionären Vorteil phänomenales Bewußtsein eigentlich mit sich bringt. Ich kann diese Frage hier natürlich nicht allgemein beantworten, möchte aber auf ein sehr wichtiges Einzelphänomen eingehen – die Rolle von Wahrnehmungsbildern bei der Steuerung von Verhalten, die sich schon an sehr einfachen Beispielen demonstrieren läßt. Jeder kennt die Situation, in der es darum geht, z. B. auf der Autobahn den Abstand zum vorausfahrenden Fahrzeug konstant zu halten. Dies kann etwa so geschehen, daß in den *höheren* Stufen der Bildverarbeitung regelmäßig der Abstand des eigenen zum vorausfahrenden Fahrzeug berechnet wird und daß dann – je nachdem, ob sich dieser Abstand vergrößert oder verringert hat oder ob er unverändert geblieben ist – die Geschwindigkeit erhöht, verringert oder konstant gehalten wird. Derselbe Effekt läßt sich aber ebensogut auch mit sehr viel geringerem Aufwand erreichen – nämlich einfach dadurch, daß das System darauf achtet, daß sich die *Größe* des *Bildes* des vorausfahrenden Fahrzeugs nicht verändert. D. h. dasselbe der Situation angepaßte Verhalten kann *ohne Inanspruchnahme der Ergebnisse höherer Verarbeitungsstufen* nur durch Rückgriff auf die Resultate des Prozesses der *primären Bildanalyse*, d. h. allein durch Bezugnahme auf Bildelemente erzeugt werden. Es ist sogar anzunehmen, daß der zweite Weg sehr viel schneller und im allgemeinen auch verläßlicher ist. Selbst das Finden von Wegen und die Vermeidung von Hindernissen läßt sich häufig schon durch relativ einfache Bildanalyse kontrollieren.

Nehmen wir als zweites Beispiel eine Greifhandlung. Auch diese Handlung kann man offenbar steuern, indem man auf Ergebnisse der *höheren* Stufen der Bildverarbeitung zurückgreift. Man berechnet zuerst die Koordinaten des Ortes, an dem sich das zu greifende Objekt befindet, und entwirft dann einen Plan, mit Hilfe welcher Muskelanspannungen man seine Hand an diesen Ort bringen kann. Dies ist jedoch ein langwieriger und – etwa bei dem Versuch, ein Objekt zu greifen, das sich bewegt – auch ein sehr komplexer Prozeß mit häufig unsicherem Ergebnis. Wie schwierig diese Art von Verhaltenssteuerung tatsächlich ist, kann man sich dadurch veranschaulichen, daß man sich vorstellt, man solle bei *geschlossenen* Augen einen Gegenstand greifen, wobei man über keine anderen Informationen verfügt als die, wo sich das zu greifende Objekt befindet, wo sich die eigene Hand befindet und welche Auswirkungen bestimmte Muskelbewegungen für die Position der eigenen Hand haben.

Offensichtlich spielt sich Handlungssteuerung bei uns Menschen auf ganz andere Weise ab. Und daß Wahrnehmungsbilder dabei eine zentrale Rolle spielen, zeigt sich eben genau daran, wie schlecht diese Steuerung funktioniert, wenn man die Augen schließt oder verbindet. Auch Greifbewegungen werden also zumindest zu großen Teilen durch Wahrnehmungsbilder und Bildelemente gesteuert. Die Einzelheiten sind sicher nicht völlig klar. Aber man kann sich z. B. vorstellen, daß es beim Greifen nicht darum geht, die Koordinaten meiner Hand mit den Koordinaten des zu greifenden Objekts zur Deckung zu bringen, sondern darum, die Bildelemente, die von meiner Hand und dem zu greifenden Objekt hervorgerufen werden, einander anzunähern. (Zusätzlich muß es natürlich auch noch einen Mechanismus geben, der der Tiefendimension Rechnung trägt.) Vielleicht wirkt das Bildelement des zu greifenden Objekts sogar als eine Art Attraktor, durch den das Bildelement meiner Hand gewissermaßen „angezogen" wird.[11] Auf jeden Fall ist es meiner Meinung nach sehr wahrscheinlich, daß wir bei Handlungsplanungen nicht nur Wissen darüber verwenden, welche Veränderungen bestimmte Muskelbewegungen in der Welt draußen bewirken, sondern insbesondere auch Wissen darüber, wie sich unsere Wahrnehmungsbilder aufgrund dieser Bewegungen verändern. Die Effizienz dieser Bewegungen können wir daher häufig allein schon aufgrund primärer Bildanalyse beurteilen – genau so wie in dem zuvor angeführten Beispiel, in dem es darum ging, den Abstand zu einem vorausfahrenden Fahrzeug konstant zu halten.

Die dritte Bemerkung schließlich bezieht sich auf die Tatsache, daß die hier vertretene Auffassung auch mit den bekannten neurobiologischen Befunden in guter Übereinstimmung steht. Bekanntlich werden die Axone der

[11] Diese interessante Idee habe ich zum ersten Mal in einer Diskussion mit dem Bremer Psychologen Michael Stadler gehört.

retinalen Ganglienzellen zum Sehnerven gebündelt, der etwa in Höhe der Fovea centralis das Auge verläßt. Die Sehnerven der beiden Augen laufen an der Schädelbasis aufeinander zu und tauschen in der Sehkreuzung die Hälfte ihrer Nervenfasern. Nach der Sehkreuzung verlaufen die Ganglienaxone zum Corpus geniculatum laterale, wo sie einmal umgeschaltet werden d.h. der „Ausgang" des Corpus geniculatum laterale projiziert direkt zur primären Sehrinde im Hinterhauptslappen der Großhirnrinde (Area 17). Die Area 17 projiziert ihrerseits in geordneter Weise – nämlich Punkt für Punkt – auf die Area 18, und diese wiederum auf mindestens drei andere Gebiete: auf ein MT genanntes Feld, auf die Area 19 und das vierte visuelle Feld V4. In ähnlicher Weise sind auch die weiteren Verschaltungen organisiert. Immer projiziert eine Area auf mehrere andere. Jedes dieser Felder sendet aber auch Signale zurück an die Areae, von denen es Eingänge bekommt. Außerdem projizieren die einzelnen Areae auch noch auf tiefer im Gehirn liegende Strukturen – beispielsweise die Colliculi superiores und verschiedene Teile des Thalamus. Schließlich erhalten sämtliche visuellen Felder Eingänge von Untereinheiten des Thalamus: so wie das Corpus geniculatum laterale auf den primären visuellen Cortex projiziert, sind andere Thalamusteile mit anderen Areae verbunden.

In diesem Zusammenhang sind zwei Punkte besonders relevant. Erstens die Tatsache, daß die Projektion der Axone der retinalen Ganglienzellen – auch nach der Umschaltung im Corpus geniculatum laterale – *retinotop* organisiert ist. Benachbarte Ganglienzellen projizieren auf benachbarte Bereiche der Area 17, so daß die topologische Struktur der Feuerungsmuster der Ganglienzellen unter dieser Projektion erhalten bleibt. Mit anderen Worten: *Auch das Feuerungsmuster der Neurone in der primären Sehrinde kann als eine Kodierung des Netzhautbildes angesehen werden*. Zweitens: Die Tatsache, daß nachgeschaltete visuelle Felder Signale zu den Bereichen zurückschicken, von denen sie Eingänge bekommen, kann leicht mit der Beobachtung in Zusammenhang gebracht werden, daß spätere Stufen visueller Informationsverarbeitung immer auch zu einer Veränderung und Verbesserung des Ausgangsbildes führen. *Die Vermutung, daß die primäre Sehrinde das physische Substrat des phänomenal relevanten Wahrnehmungsbildes* (oder zumindest einen wesentlichen Teil dieses physischen Substrats) *darstellt, scheint daher nicht unplausibel*.

Diese Vermutung wird auch durch pathophysiologische Befunde gestützt. Es ist seit langem bekannt, daß kleine Läsionen, umgrenzte Infarkte oder kleine Tumore im Bereich der primären Sehrinde zu Blindheit innerhalb eines umschriebenen Teils des Gesichtsfeldes oder, wenn die gesamte Area 17 betroffen ist, zur vollständigen Blindheit führen. Man spricht in diesen Fällen von kortikaler bzw. zentraler Blindheit, um sie von den Fällen zu entscheiden, in denen die Blindheit auf einen Defekt der Augen oder

Sehnerven zurückgeht. Offensichtlich ist kortikale Blindheit genau das, was man erwarten würde, falls die primäre Sehrinde tatsächlich das physische Substrat des phänomenalen Wahrnehmungsbildes darstellt. Jede Schädigung des Substrats müßte zu einem Ausfall innerhalb dieses Bildes führen und zu einem Ausfall aller höheren Repräsentationen, die vom fehlerfreien Funktionieren dieses Wahrnehmungsbildes abhängen.

In den letzten Jahren ist in diesem Zusammenhang das Phänomen der „blindsight" viel diskutiert worden. Dieses Phänomen manifestiert sich insbesondere darin, daß Patienten, die unter kortikaler Blindheit leiden und deshalb auf Befragen angeben, in einem bestimmten Bereich ihres Gesichtsfeldes buchstäblich nichts zu sehen, offenbar trotzdem über gewisse „Informationen" über das verfügen, was sie nicht zu sehen behaupten. Denn z. B. unter *forced choice* Bedingungen geben sie weit häufiger richtige Antworten, als dies bei völlig fehlender Information statistisch zu erwarten wäre. In dem oben erläuterten Modell ließe sich dieser Effekt unter anderem so erklären, daß spätere Verarbeitungsstufen, die ja aufgrund der neuronalen Verschaltungen nicht nur Eingänge von der primären Sehrinde (oder den nachgeschalteten Areae), sondern z. B. auch vom Thalamus erhalten, ihre Aufgaben auch bei einer Schädigung der primären Sehrinde zumindest noch teilweise erfüllen können, daß aber in diesen Fällen das (an die primäre Sehrinde geknüpfte) phänomenale Wahrnehmungsbild geschädigt bleibt, was dazu führt, daß die betroffenen Personen subjektiv völlig gerechtfertigt berichten, nichts zu sehen.

Auch bestimmte Phänomene, die bei *split-brain* Patienten, denen das corpus callosum durchtrennt wurde, beobachtet werden können, lassen sich im Rahmen des Modells erklären. Wenn man einen Gegenstand nur in der linken Hälfte des Gesichtsfeldes dieser Patienten präsentiert, sind sie nicht in der Lage, verbal die Frage zu beantworten, welcher Gegenstand ihnen gezeigt wurde. Wie allgemein angenommen wird, ist dies darauf zurückzuführen, daß Reizungen der rechten Hälfte der Retina nur zur rechten Hirnhälfte gelangen und daß bei *split-brain* Patienten daher die an der Sprachproduktion beteiligten Zentren der linken Hirnhälfte keine (direkten) Informationen über Gegenstände erhalten, die nur in der linken Hälfte des Gesichtsfeldes gezeigt werden. Im Rahmen des hier vorgeschlagenen Modells würde das bedeuten, daß split-brain Patienten zwar in gewisser Weise über ein intaktes phänomenales Wahrnehmungsbild der präsentierten Gegenstände verfügen, daß aber höhere (insbesondere Sprach-)Zentren keinen Zugang zu diesem Bild haben und die Patienten daher keine entsprechenden Angaben machen können.

5. Zurück zu der entscheidenden Frage nach der Plausibilität der Auffassung, daß Prozesse visueller Informationsverarbeitung, die auf die zuvor

geschilderte Weise strukturiert sind, immer auch einen phänomenalen Aspekt besitzen. Warum soll die Tatsache, daß Wahrnehmungsbilder in diesen Prozessen eine zentrale Rolle spielen, für diesen phänomenalen Aspekt entscheidend sein?

Wenn zwei Systeme *A* und *B* funktional äquivalent sind, d. h., wenn sie in ihrer funktionalen Struktur übereinstimmen und sich zum Zeitpunkt *t* in denselben funktionalen Zuständen befinden, dann impliziert dies auch, daß sich die beiden Systeme zum Zeitpunkt *t* in derselben Weise verhalten. Denn funktionale Zustände sind durch ihre kausalen Beziehungen zu Inputs, Outputs und anderen funktionalen Zuständen bestimmt. Wenn sich zwei funktional äquivalente Systeme in denselben funktionalen Zuständen befinden, bewirken gleiche Inputs daher identische interne Prozesse, die am Ende auch zu demselben Verhalten führen müssen. Vertreter von „absent-“ oder „inverted-qualia“-Argumenten müssen daher die Annahme in Kauf nehmen, daß es möglich ist, daß sich zwei Systeme, die in ihrem Verhalten ununterscheidbar sind, trotzdem im Hinblick auf den phänomenalen Gehalt ihrer Zustände deutlich voneinander unterscheiden, ja daß es sogar möglich ist, daß von zwei Systemen, die sich völlig gleich verhalten, nur das eine überhaupt über phänomenale Zustände verfügt, während das andere diese Zustände gänzlich entbehrt.

Trotz aller damit verbundenen Schwierigkeiten wird diese Konsequenz von vielen Philosophen akzeptiert, da es intuitiv durchaus plausibel scheint, daß das Verhalten eines Organismus allein nicht ausreicht, um zu entscheiden, in welchem phänomenalen Zustand er sich befindet. S. Shoemaker hat jedoch schon vor 20 Jahren darauf hingewiesen, daß die Dinge erheblich anders aussehen, wenn zwei Systeme nicht nur in ihrem Verhalten übereinstimmen, sondern auch in allen Überzeugungen, die ihre eigenen und insbesondere ihre eigenen phänomenalen Zustände betreffen. Denn diejenigen, die die Auffassung vertreten, daß das Verhalten eines Organismus nicht ausreicht, um zu entscheiden, welchen phänomenalen Gehalt die Zustände dieses Organismus haben bzw. ob diese Zustände überhaupt einen phänomenalen Gehalt besitzen, meinen in der Regel zugleich, daß es einen anderen, direkten Weg zur Entscheidung dieser Frage gibt – den Weg der *Introspektion*. Introspektion, so Shoemaker, beruht aber auf dem, was ein Wesen über seine eigenen Zustände glaubt bzw. weiß. Wenn sich zwei Systeme also nicht nur beide genauso verhalten, wie es für Systeme typisch ist, die Schmerzen empfinden, wenn sie vielmehr auch beide in der gleichen Weise davon überzeugt sind, daß sie Schmerzen empfinden, dann spricht nicht nur ihr Verhalten, sondern auch ihre Introspektion dafür, daß sie tatsächlich Schmerzen empfinden. Und was könnte dann noch dafür sprechen, daß dies nicht so ist? Was könnte es in diesem Fall überhaupt heißen, daß dies nicht so ist?

> So one way of putting our question is to ask whether anything could be evidence (for anyone) that someone was not in pain, given that it follows from the states he is in … that the totality of possible behavioral evidence *plus* the totality of possible introspective evidence points unambiguously to the conclusion that he is in pain? I do not see how anything could be. (Shoemaker 1975: 189–190)

Wenn zwei Systeme nicht nur in ihrem Verhalten übereinstimmen, sondern auch in allen Überzeugungen, die ihre eigenen und insbesondere ihre eigenen phänomenalen Zustände betreffen, gibt es also – zumindest epistemisch gesehen – keine Möglichkeit mehr, zwischen ihnen im Hinblick auf den phänomenalen Gehalt ihrer mentalen Zustände zu unterscheiden.

Ein System, in dem visuelle Informationsverarbeitung auf die zuvor geschilderte Weise organisiert ist und das außerdem über Metarepräsentationen verfügt, die nicht nur die Ergebnisse des Informationsverarbeitungsprozesses, sondern auch die früheren Stufen dieses Prozesses und insbesondere die Repräsentationen des modifizierten Ausgangsbildes betreffen, erfüllt aber – zumindest im Hinblick auf die phänomenalen Aspekte der visuellen Wahrnehmung – genau diese Bedingung. Auf jeden Fall wird es auf Befragen Antworten nicht nur über die Dinge und Situationen geben, die ihm aufgrund seines visuellen Systems zugänglich sind, sondern auch über die visuellen Eindrücke, die es in diesem Zusammenhang erworben hat. Es wird nicht nur sagen, daß der Tisch vor ihm quadratisch ist und das Glas auf dem Tisch steht; sondern z.B. auch, daß die obere Öffnung des Glases, obwohl sie kreisrund ist, ellipsenförmig aussieht, daß die Tischoberfläche leicht rötlich aussieht oder daß die Tapete im Hintergrund so verwaschen aussieht, daß es ihr Muster nicht genau erkennen kann.

Wenn man diesem System Überzeugungen zuschreibt (und ich sehe keinen Grund, dies nicht zu tun), beziehen sich diese Überzeugungen also nicht nur auf seine Umwelt, sondern auch auf die eigenen Zustände und insbesondere auf etwas, was *wir* visuelle Eindrücke nennen würden und was das System selbst *genau so* beschreibt, wie wir unsere visuellen Eindrücke beschreiben. Was könnte also dafür sprechen, daß das System im Gegensatz zu uns doch keine visuellen Eindrücke hat? Welche Anhaltspunkte könnte es hier geben, die uns nicht zugleich auch dazu veranlassen müßten, daran zu zweifeln, daß unsere Mitmenschen bzw. sogar wir selbst tatsächlich visuelle Eindrücke haben?

Mit diesem Punkt hängt ein zweiter eng zusammen. Systeme wie das gerade beschriebene können nämlich genau wie wir unterscheiden zwischen der Art, wie die Dinge wirklich sind, und der Art, wie sie zu sein scheinen. Einige vom späten Wittgenstein inspirierte Philosophen haben die Auffassung vertreten, daß Sätze wie „Die Wand sieht rot aus“ ebenso wie „Die Wand scheint rot zu sein“ nichts mit phänomenalen Qualitäten zu tun ha-

ben; daß sie vielmehr in einem performativen Sinne zu verstehen sind. Wenn ich diese Sätze äußere, so meinen diese Philosophen, mache ich damit nur deutlich, daß ich mir meiner Sache nicht sicher bin und daß ich nicht bereit bin, irgendwelche Verpflichtungen im Hinblick auf den Wahrheitsgehalt meiner Aussagen zu übernehmen. Wer aufgrund meiner Äußerungen glaubt, was ich sage, tut dies auf eigenes Risiko. Wenn ich dagegen sage „Die Wand ist rot" oder sogar „Ich weiß, daß die Wand rot ist", dann kann man mich auf den Inhalt dieser Aussagen festnageln, und dann gehe ich daher das Risiko ein, zur Rechenschaft gezogen zu werden, wenn sich herausstellt, daß es nicht so ist, wie ich sage.

R. Chisholm hat jedoch überzeugend nachgewiesen, daß der performative Gebrauch von „scheint" bzw. „sieht aus" bestenfalls *einen* möglichen Gebrauch darstellt (vgl. etwa Chisholm 1989: 20–22). Es gibt auch andere Verwendungsweisen, die offenbar doch etwas mit phänomenalen Qualitäten zu tun haben. So gibt es z. B. eine Verwendungsweise, in der es sinnvoll ist zu sagen:

(1) Die Wand scheint mir in diesem Licht grau; ich weiß aber, daß sie rot ist.

Oder sogar:

(2) Die Wand scheint mir in diesem Licht rot; und ich weiß auch, daß sie tatsächlich rot ist.

Besonders das letzte Beispiel macht klar, daß hier das „scheint" nicht den angesprochenen performativen Sinn haben kann. Denn wenn das so wäre, würde der zweite Teilsatz die Pointe des ersten aufheben, was offenbar nicht der Fall ist. Wenn jedoch „aussehen" oder „scheinen" in diesen Beispielen nicht im performativen Sinn gebraucht werden, in welchem dann?

Für die Beantwortung dieser Frage ist es wichtig zu sehen, in welchen Situationen wir „aussehen" oder „scheinen" in diesem Sinn verwenden. Paradigmatisch sind wohl Fälle wie der, daß ich in einem psychophysischen Experiment vom Versuchsleiter dazu aufgefordert werde, nur darauf zu achten, welche Farbe die Wand vor mir unter verschiedenen Bedingungen zu haben *scheint*, eine Wand, von der ich mich vorher habe überzeugen können, daß sie weiß ist. Was will der Versuchsleiter hier von mir wissen?

Er will sicher nicht wissen, was ich über die wirkliche Farbe der Wand glaube. Denn ich weiß, daß die Wand weiß ist, und an dieser Überzeugung ändert sich während des gesamten Experiments nichts. Also möchte er offenbar etwas über die subjektiven *Eindrücke* erfahren, die sich in mir unter den verschiedenen Versuchsbedingungen einstellen. Wenn er fragt, welche Farbe die Wand zu haben scheint, geht er davon aus, daß sich nicht meine objektiven Überzeugungen, wohl aber meine visuellen Eindrücke während des Experiments verändern. Und über diese Veränderungen möchte er et-

was erfahren. Wir sind hier wieder an demselben Punkt, den ich am Ende von Abschnitt 3 schon einmal angesprochen hatte. Aufgrund der Art und Weise, in der unser Wahrnehmungssystem organisiert ist, sind wir nicht nur dazu in der Lage, etwas über die wahrgenommenen Objekte und Szenen zu sagen, wir können uns auch auf das (modifizierte) Ausgangsbild konzentrieren und berichten, wie sich dieses Ausgangsbild verändert, ohne damit zu implizieren, daß diese Veränderungen Veränderungen in der wahrgenommenen Szene voraussetzen. Aber das bedeutet auch, daß jedes System, in dem der Prozeß der visuellen Informationsverarbeitung in derselben Weise strukturiert ist, ebenfalls zwischen solchen Veränderungen unterscheiden kann, die die wahrgenommene Szene betreffen, und solchen, die sich nur auf das *Bild* dieser Szene beziehen, d. h. darauf, wie diese Szene dem System *erscheint*. In diesem Sinn ist also für ein solches System auch die Unterscheidung zwischen dem, wie die Dinge wirklich sind und wie sie nur zu sein scheinen, ganz natürlich.

Wenn in Systemen die visuelle Informationsverarbeitung so organisiert ist, daß aus Ausgangsbildern Repräsentationen der Ursprungsszenen rekonstruiert werden (und zwar auf eine Weise, die nicht zum Verlust, sondern nur zur Modifikation dieser Ausgangsbilder führt), dann verhalten sich diese Systeme also nicht nur genauso wie wir (z. B. greifen sie in derselben schnellen und eleganten Weise nach Gegenständen), dann haben sie vielmehr im Hinblick auf sich und ihre Wahrnehmungen auch dieselben Überzeugungen wie wir; zumindest reden sie dann über sich und ihre Wahrnehmungen in derselben Weise wie wir. Ebenso wie wir können solche Systeme zwischen wahrgenommenen Szenen und visuellen Eindrücken unterscheiden; ebenso wie wir können sie einen Unterschied machen zwischen dem, wie die Dinge wirklich sind und wie sie nur zu sein scheinen. Und darin liegt meiner Meinung nach ein kaum widerlegbares Argument für die Auffassung, daß die Wahrnehmungszustände dieser Systeme ebenfalls einen phänomenalen Gehalt besitzen. Denn welchen Grund könnte es für die Annahme geben, daß sich diese Systeme – im Gegensatz zu uns – beständig irren, wenn sie sich selbst Zustände mit phänomenalem Gehalt zuschreiben?

## Literatur

Chisholm, R. (1989) *Theory of Knowledge*. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Humphrey, N. (1993) *A History of the Mind*. London: Random House.

Marr, D. (1982) *Vision*. San Francisco: Freemann & Co.

Nelkin, N. (1989) „Unconscious Sensations“. *Philosophical Psychology* 2, 129–141.

Nelkin, N. (1994) „Phenomena and Representation“. *British Journal for the Philosophy of Science* 45, 527–547.

Neumann, B. (1993) „Bildverstehen – ein Überblick“. In: G. Görz (Hg.) *Einführung in die künstliche Intelligenz*. Addison-Wesley: Bonn/Paris/Reading MA, 559–588.

Shoemaker, S. (1975) „Functionalism and Qualia“. *Philosophical Studies* 27, 291–315. Wiederabgedruckt in S. Shoemaker, *Identity, Cause, and Mind*. Cambridge: Cambridge University Press 1984, 184–205.

Van Gulick, R. (1989) „What Difference Does Consciousness Make?“ *Philosophical Topics* 17, 211–230.

# 12

## Könnte es sein, dass ich ein Zombie bin?*

Gibt es Aspekte des Bewusstseins, die der naturwissenschaftlichen Erklärung grundsätzlich entzogen sind?

Die meisten Menschen kennen Zombies aus Büchern und Filmen:

> Als **Zombie** wird die fiktive Figur eines zum Leben erweckten Toten (Untoter) oder eines seiner Seele beraubten, willenlosen Wesens bezeichnet. Der Begriff leitet sich von dem Wort *nzùmbe* aus der zentralafrikanischen Sprache Kimbundu ab und bezeichnet dort ursprünglich einen Totengeist. (http://de.wikipedia.org/wiki/Zombie – Abruf 16.02.2011)

Diese Zombies unterscheiden sich deutlich von „normalen" Menschen. Oft sind sie blutüberströmt oder halbverwest, immer haben sie einen starren, seelenlosen Blick. Philosophische Zombies sind anders. Als „philosophischen Zombie" bezeichnet man den *physischen Doppelgänger* eines realen Menschen; ein Wesen, das aus den gleichen Molekülen aufgebaut ist wie ein realer Mensch, wobei diese Moleküle auch genau so angeordnet sind wie bei diesem Menschen. Der Zombie sieht also genau so aus wie dieser Mensch. Aber in gewisser Weise fehlt auch ihm eine Seele; denn ein philosophischer Zombie fühlt nichts, ihm fehlt jede Art bewusster Erlebnisse. Kann es ein solches Wesen wirklich geben? Ist es z. B. möglich, dass es einen physischen Doppelgänger von mir gibt, der überhaupt nichts fühlt? Die Auseinandersetzungen um diese Frage, die in der gegenwärtigen Diskussion des Leib-Seele-Problems eine zentrale Rolle spielen, lassen sich nur nachvollziehen, wenn man ihre argumentative Rolle versteht. Ich möchte deshalb zu Beginn einen kurzen Überblick über die Geschichte und die aktuelle Diskussion des Leib-Seele-Problems geben.[1]

Menschen sind auf der einen Seite physische Wesen: Sie haben einen Ort in Raum und Zeit, sie haben eine Gestalt und ein Gewicht. Sie sind auch biologische Wesen: Sie atmen und nehmen Nahrung zu sich, sie paaren sich und pflanzen sich fort, sie wachsen, altern und sterben. Und schließlich haben sie ein mentales Leben: Sie nehmen wahr und erinnern sich, sie denken nach und fällen Entscheidungen, sie freuen und ärgern sich, sie fühlen Schmerz und Freude. Die entscheidende Frage der Philosophie des Geistes lautet: Wie verhält sich das mentale Leben eines Menschen zu sei-

---

* Originalbeitrag. Druckfassung eines Vortrags, den ich am 31.01.2011 in Freiburg gehalten habe.

1 Eine ausführlichere Version dieses doch sehr skizzenhaften Überblicks findet sich in Beckermann (2008a, Kap.1).

nen biologischen und zu seinen physischen Eigenschaften? Dabei muss man jedoch zwei Teilfragen unterscheiden:

1. Hat der Mensch eine von seinem Körper unabhängige (immaterielle) Seele, die den Tod des Körpers überleben kann? (Problem mentaler Substanzen)
2. Wie verhält sich das mentale Leben eines Menschen – sein Wahrnehmen, Erinnern, Fühlen, Denken und Entscheiden – zu dem, was in seinem ZNS vorgeht? (Problem mentaler Eigenschaften und Prozesse)

In der Antike wurde die erste Frage mit einem klaren „Ja" beantwortet. Nicht nur jeder Mensch, sondern jedes Lebewesen hat eine Seele; denn ohne Seele, so die antike Vorstellung, gibt es kein Leben. Die Seele ist das Prinzip des Lebens – das, was Leben verleiht. In diesem Punkt waren sich alle antiken Autoren einig. Große Unterschiede gab es jedoch bei der Beantwortung der Frage nach der Natur der Seele. Für Demokrit, Epikur und Lukrez – die Hauptvertreter des antiken Atomismus – war die Seele selbst etwas Materielles; ihrer Ansicht nach besteht sie aus sehr kleinen, sehr schnellen Atomen, die einerseits in der Mitte des Körpers ein Zentrum haben, andererseits über den ganzen Körper verteilt sind. Für Platon ist die Seele dagegen ein immaterielles Wesen, das schon vor der Geburt des Körpers existiert und wohl auch nach seinem Tod. Für Aristoteles schließlich ist die Seele die Form des mit den richtigen Organen ausgestatteten – und daher im Prinzip lebensfähigen – Körpers.

Mit Descartes beginnt wie in so vielen Bereichen der Philosophie auch in der Philosophie des Geistes eine neue Epoche. Descartes bricht radikal mit dem antiken Seelenverständnis. In seinen Augen hat die Seele mit dem Phänomen des Lebens nichts zu tun. Tiere haben keine Seele; ihre Lebensvorgänge lassen sich alle rein mechanisch – durch Bezug auf ihre physischen Teile, deren Anordnung und deren Interaktionen – erklären, so wie „die Bewegungen einer Uhr oder eines anderen Automaten [allein auf die] Anordnung ihrer Gewichte und ihrer Räder" zurückgeführt werden können (Descartes *Über den Menschen*, 136). Trotzdem hat auch Descartes zufolge Platon in einem Punkt Recht: Zumindest jeder Mensch hat eine Seele. Denn Menschen können *sprechen* und *denken*; und diese beiden Fähigkeiten lassen sich *nicht* mechanisch erklären. Deshalb muss jeder Mensch außer einem Körper auch eine Seele besitzen. Diese Seele ist eine immaterielle Substanz, die vom Körper prinzipiell unabhängig ist, die das eigentliche Selbst jedes Menschen ausmacht und die nach dem Tod des Körpers weiter existieren kann.

Dieser Platonisch-Cartesische Dualismus war zwar lange Zeit kaum umstritten, ist aber im letzten Jahrhundert stark in Frage gestellt worden. Deshalb steht heute meist die zweite Teilfrage des Leib-Seele-Problems im Vordergrund: Wie verhält sich das mentale Leben eines Menschen – sein

Wahrnehmen, Fühlen, Denken und Entscheiden – zu dem, was in seinem ZNS und insbesondere in seinem Gehirn vorgeht?

Diese Frage wird häufig als Frage nach dem Status *mentaler Eigenschaften* formuliert. *Eigenschaftsdualisten* vertreten die These, dass mentale Eigenschaften ontologisch eigenständig sind. *Eigenschaftsphysikalisten* dagegen sehen eine ontologische Abhängigkeit der mentalen von den physischen Eigenschaften. Der Eigenschaftsphysikalismus wird heute in drei Versionen vertreten:

1. Mentale Eigenschaften sind mit physischen Eigenschaften *identisch* – so wie die Temperatur eines Gases identisch ist mit der mittleren kinetischen Energie seiner Moleküle, Wasser identisch ist mit $H_2O$ und Blitze identisch sind mit elektrischen Entladungen.
2. Mentale Eigenschaften sind auf neuronale Eigenschaften *reduzierbar* – so wie die Wasserlöslichkeit von Salz auf seine molekulare Struktur (und die molekulare Struktur von Wasser) oder die Fähigkeit von Seife, Schmutz zu lösen, auf die Struktur der Seifenmoleküle zurückgeführt werden kann.[2]
3. Mentale Eigenschaften *supervenieren* über physischen Eigenschaften in folgendem Sinn: Es kann keinen Unterschied in den mentalen Eigenschaften einer Person geben, wenn es nicht auch einen Unterschied in ihren physischen Eigenschaften gibt. Bzw.: Wenn zwei Dinge (Welten) in allen physischen Eigenschaften übereinstimmen, stimmen sie notwendigerweise auch in ihren mentalen Eigenschaften überein.

Zwischen den drei Versionen des Eigenschaftsphysikalismus bestehen folgende Beziehungen: (i) Sowohl Identität als auch Reduzierbarkeit implizieren Supervenienz. Wenn jede mentale mit einer physischen Eigenschaft identisch ist, kann es ganz offensichtlich keinen Unterschied in den mentalen Eigenschaften einer Person geben, ohne dass es auch einen Unterschied in ihren physischen Eigenschaften gibt. Und wenn eine Eigenschaft *F* eines Gegenstandes reduktiv – d. h. allein durch Bezugnahme auf seine physischen Teile und deren Anordnung – erklärt werden kann, heißt das, dass alle Gegenstände, die aus den gleichen Teilen bestehen, die auf die gleiche Weise angeordnet sind, notwendigerweise ebenfalls die Eigenschaft *F* besitzen. (ii) Die These, dass mentale Eigenschaften auf physischen Eigenschaften supervenieren, obwohl sie *weder* mit ihnen identisch *noch* auf sie reduzierbar sind, wird von niemandem vertreten. Letzten Endes sind deshalb die Identitätstheorie und die Reduzierbarkeitsthese die einzigen beiden Versionen des Eigenschaftsphysikalismus, die gegenwärtig eine Rolle spielen. Die Beziehung zwischen diesen Positionen wird im Augenblick ebenso heftig diskutiert wie die Frage, welche der beiden die plausiblere Position

[2] Ich werde auf diese beiden Beispiele gleich noch ausführlicher eingehen.

ist. In diesem Zusammenhang spielt dies aber keine Rolle, da es im Folgenden um Argumente geht, die sowohl gegen die Identitätstheorie als auch gegen die Reduzierbarkeitsthese vorgebracht wurden. Festgehalten werden muss an dieser Stelle aber, dass sowohl die Identitätstheorie als auch die Reduzierbarkeitsthese mit der Möglichkeit von Zombies *unvereinbar* sind. Denn beide Theorien implizieren Supervenienz, und Supervenienz impliziert, dass jeder physische Doppelgänger auch ein mentaler Doppelgänger ist. Die *Möglichkeit* philosophischer Zombies ist daher für den Eigenschaftsphysikalismus von entscheidender Bedeutung. Wenn solche Zombies möglich sind, ist der Eigenschaftsphysikalismus falsch (zumindest sehen das viele Philosophen so). Und wenn der Eigenschaftsphysikalismus wahr ist, kann es keine philosophischen Zombies geben. Schauen wir uns also die Hauptargumente an, die *gegen* den Eigenschaftsphysikalismus vorgetragen worden sind. Zuvor muss aber noch die Frage beantwortet werden, welche Arten mentaler Eigenschaften oder Zustände es gibt und durch welche Merkmale diese Zustände jeweils charakterisiert sind.[3]

Heute ist es üblich, zwei große Gruppen von mentalen Zuständen zu unterscheiden: *Empfindungen* auf der einen und *intentionale Zustände* (propositionale Einstellungen) auf der anderen Seite. Zu den Empfindungen gehören körperliche Empfindungen wie Schmerzen, Kitzel und Übelkeit ebenso wie Wahrnehmungseindrücke – wie der Eindruck einer bestimmten Farbe, das Klangerlebnis des Brummens einer Hummel oder das Geschmackserlebnis beim Essen einer süßen Birne. Zwischen diesen beiden Gruppen gibt es zwar eine Reihe von Unterschieden; trotzdem ist es sinnvoll, sie zusammenzufassen. Denn alle Empfindungen scheinen im Wesentlichen durch ihre phänomenalen Eigenschaften definiert zu sein. Für Empfindungen ist ihr qualitativer Gehalt charakteristisch – das, *was* man fühlt, wenn man eine solche Empfindung hat; die Art, *wie* es ist, eine solche Empfindung zu haben. (In der neueren Literatur spricht man hier häufig von den mit Empfindungen verbundenen *Qualia*).

Die zweite große Gruppe mentaler Zustände ist die Gruppe der Zustände, die einen intentionalen Inhalt haben und bei deren Zuschreibung wir deshalb „dass"-Sätze verwenden – etwa, wenn wir Hans eine bestimmte Überzeugung zuschreiben, indem wir sagen „Hans glaubt, *dass* die Erde rund ist". Auch innerhalb der Gruppe der intentionalen Zustände gibt es erhebliche Unterschiede, z. B. zwischen kognitiven Einstellungen wie Überzeugungen auf der einen und Einstellungen wie Wünschen, Absichten und Befürchtungen auf der anderen Seite, die auch eine konative oder affektive Komponente haben. Allen intentionalen Zuständen ist aber gemeinsam, dass sie durch jeweils zwei Aspekte gekennzeichnet sind: durch die Art des

[3] Zu den nächsten beiden Absätzen vgl. Beckermann 2003, 212.

Zustandes – glauben, wünschen, hoffen, etc. – und durch ihren intentionalen Gehalt – das, was geglaubt, was gewünscht oder was gehofft wird.

Die meisten Argumente gegen den Eigenschaftsphysikalismus setzen am phänomenalen Charakter von Empfindungen an. Die wichtigsten dieser Argumente sind Thomas Nagels *Argument der Subjektivität von Empfindungen*, Frank Jacksons *Argument des unvollständigen Wissens* und Joseph Levines *Argument der Erklärungslücke*. Ich werde im Folgenden nur auf Levines Argument der Erklärungslücke eingehen, weil dieses Argument für das Zombie-Problem besonders relevant ist.[4]

Levine beginnt mit einem Vergleich der folgenden beiden Aussagen:

(1) Schmerz ist identisch mit dem Feuern von C-Fasern.[5]
(2) Temperatur[6] ist identisch mit der mittleren kinetischen Energie der Moleküle eines Gases.

Zwischen diesen beiden Aussagen besteht, so Levine, ein bemerkenswerter Unterschied. Auf der einen Seite ist es nämlich in einem bestimmten Sinn *undenkbar*, dass in einem Gas die mittlere kinetische Energie der Moleküle einen bestimmten Wert (sagen wir, $6.21 \times 10^{-21}$ Joule) hat, dass dieses Gas aber nicht die entsprechende Temperatur von 300 K besitzt, während es auf der anderen Seite sehr wohl denkbar zu sein scheint, dass in meinem Körper die C-Fasern feuern, ich aber keinen Schmerz empfinde. Nach Levine liegt dies daran, dass die Aussage (2) *vollständig explanatorisch* ist, die Aussage (1) dagegen nicht. Was ist damit gemeint?

Wenn man uns fragen würde, was wir mit dem Ausdruck „Temperatur" meinen, würden wir, wieder Levine zufolge, wahrscheinlich antworten:

(2′) Temperatur ist die Eigenschaft von Körpern, die in uns bestimmte Wärme- bzw. Kälteempfindungen hervorruft, die dazu führt, dass die Quecksilbersäule in Thermometern, die mit diesen Körpern in Berührung kommen, steigt oder fällt, die bestimmte chemische Reaktionen auslöst, und so weiter.

---

[4] Nagels *Argument der Subjektivität von Empfindungen* wird in Beckermann 2008a, 100ff. und in Beckermann 2008b, 410–413 diskutiert, Frank Jacksons *Argument des unvollständigen Wissens* in Beckermann 2008a, 102–105 und in Beckermann 2008b, 413–427.

[5] In der frühen Diskussion der Identitätstheorie wurde tatsächlich die Frage diskutiert, ob Schmerz eventuell mit dem Feuern von C-Fasern identisch sein könne. Heute ist klar, dass C-Fasern zwar Signale von Nozizeptoren weiterleiten, aber sicher nicht das zentrale neuronale Korrelat von Schmerz ausmachen. Ich verwende daher im Folgenden neutral den Ausdruck „Feuern der S-Neuronen" für das, was auch immer tatsächlich das neuronale Korrelat von Schmerzen ist.

[6] Levine spricht nicht von „Temperatur", sondern von „Hitze" („heat"). Der Sache nach geht es aber tatsächlich um Temperatur.

Mit anderen Worten: Wir würden Temperatur durch ihre *kausale Rolle* charakterisieren. Das ist der entscheidende Grund für den explanatorischen Charakter von (2):

> Eine solche Aussage ist insofern explanatorisch, als unser physikalisches und chemisches Wissen verständlich macht, wie so etwas wie die Bewegung von Molekülen diejenige kausale Rolle spielen kann, die wir mit Hitze assoziieren. Darüber hinaus *erschöpft sich* bereits vor unserer Entdeckung der wesentlichen Eigenschaften von Hitze unser Begriff von Hitze in unserem Wissen um deren kausale Rolle, wie wir sie mit Aussagen wie [(2′)] ausdrücken. Sobald wir jedoch wissen, wie diese kausale Rolle ausgefüllt wird, gibt es nichts mehr, was wir noch verstehen müßten. (Levine 1983, 95 f.; meine Hervorhebung)

Genau genommen hat der explanatorische Charakter von (2) also zwei Gründe:

1. Unser Begriff von Temperatur erschöpft sich vollständig in ihrer kausalen Rolle.
2. Physik und Chemie können verständlich machen, dass die mittlere kinetische Energie der Moleküle eines Gases genau diese kausale Rolle spielt.

Aber könnten diese beiden Punkte – *mutatis mutandis* – nicht auch auf Schmerzen zutreffen? Könnte nicht auch die Aussage (1) vollständig explanatorisch sein? Mit dem Ausdruck „Schmerzen" assoziieren wir doch ebenfalls eine kausale Rolle. Schmerzen werden durch die Verletzung von Gewebe verursacht, sie führen dazu, dass wir schreien oder wimmern, und sie bewirken in uns den Wunsch, den Schmerz so schnell wie möglich loszuwerden. Dies bestreitet auch Levine nicht. Und er bestreitet auch nicht, dass das Feuern von C-Fasern den Mechanismus erklären könnte, auf dem die kausale Rolle von Schmerzen beruht. Dennoch gibt es seiner Meinung nach einen entscheidenden Unterschied.

> [U]nser Begriff von Schmerz umfaßt […] mehr als nur die kausale Rolle, er umfaßt den qualitativen Charakter von Schmerz, der bestimmt, wie Schmerz sich anfühlt. Und was durch die Entdeckung von C-Faserreizungen unerklärt bleibt, ist, *warum sich Schmerzen so anfühlen, wie sie sich anfühlen*! Denn nichts an C-Faserreizungen scheint auf natürliche Weise zu den phänomenalen Eigenschaften von Schmerzen zu „passen", jedenfalls nicht besser, als es zu irgendeiner anderen Art von phänomenalen Eigenschaften passen würde. Anders als bei der funktionalen Rolle bleibt bei der Identifikation der qualitativen Eigenschaften von Schmerzen mit C-Faserreizungen (oder einer Eigenschaft von C-Faserreizungen) die Verbindung zwischen diesen Eigenschaften und dem, womit wir sie identifizieren, völlig rätselhaft. Man könnte sagen, diese Identifikation macht die Art und Weise, wie sich Schmerzen anfühlen, zu einem *factum brutum*. (ebd., 96)

Ein erster Grund dafür, dass die Aussage (1) in Levines Augen nicht vollständig explanatorisch ist, ist also:

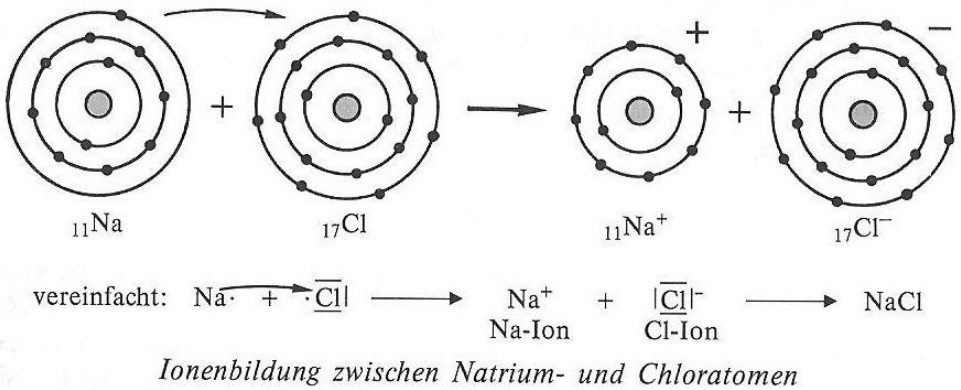
3. Unser Begriff von Schmerzen erschöpft sich nicht in einer kausalen Rolle; er umfasst auch einen qualitativen Aspekt – die Art, wie es sich anfühlt, Schmerzen zu haben.

Aber dies allein ist noch nicht entscheidend. Denn (1) könnte trotzdem vollständig explanatorisch sein, *wenn* die Neurobiologie verständlich machen könnte, dass sich das Feuern von C-Fasern genau so anfühlt, wie dies für Schmerzen charakteristisch ist. Für den nicht-explanatorischen Charakter von (1) ist deshalb der folgende Punkt noch wichtiger:

4. Die Neurobiologie kann *nicht* verständlich machen, dass sich das Feuern von C-Fasern genau so anfühlt, wie dies für Schmerzen charakteristisch ist.

Wenn Levine davon spricht, dass unser „physikalisches und chemisches Wissen *verständlich* macht, wie so etwas wie die Bewegung von Molekülen diejenige kausale Rolle spielen kann", durch die der Begriff der Temperatur charakterisiert ist, bzw. dass die Neurobiologie *nicht verständlich* machen kann, dass sich das Feuern von C-Fasern genau so anfühlt, wie dies für Schmerzen charakteristisch ist, verbirgt sich dahinter ein Begriff der *reduktiven Erklärung*, der für seine ganze Argumentation entscheidend ist. Ich möchte diesen Begriff zunächst an zwei Beispielen erläutern.

Beispiel 1: Die *Festigkeit* und *Wasserlöslichkeit* von Kochsalz. Kochsalz ist bekanntlich Natriumchlorid; es besteht aus Natrium- und Chloratomen, d. h. genauer aus Natrium- und Chlorionen. Natriumatome besitzen in ihrer äußersten Schale nur ein einziges Elektron, das leicht abgespalten werden kann. Chloratome dagegen besitzen in ihrer äußersten Schale sieben Elektronen; diese Atome sind daher „bestrebt", ihre äußerste Schale mit einem weiteren Elektron aufzufüllen, um so auf die Idealzahl von acht Elektronen

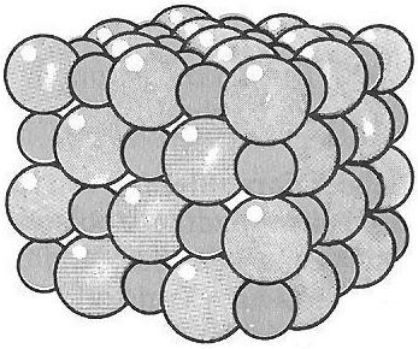



*Ionenbildung zwischen Natrium- und Chloratomen*

Abbildung 1 (Lossow/Wernet 1998, 73)

zu kommen. Wenn Natrium- und Chloratome miteinander reagieren, geschieht deshalb folgendes. Das Natriumatom gibt sein äußerstes Elektron ab, und dieses Elektron wird vom Chloratom aufgenommen.

So entstehen positiv geladene Natrium- und negativ geladene Chlorionen, die sich aufgrund der zwischen ihnen bestehenden elektromagnetischen Anziehungskräfte in einer Gitterstruktur anordnen.



*Kochsalzgitter*

Abbildung 2 (ebd.)

Entscheidend sind hier zunächst diese Anziehungskräfte. Sie erklären, warum Kochsalz unter normalen Bedingungen fest ist. Diese Kräfte sind nämlich so groß, dass die einzelnen Ionen an ihren relativen Positionen „festgezurrt" sind. Wenn sich ein Stück Kochsalz bewegt, bewegt sich daher immer das ganze Stück. Seine Teile verändern ihre relativen Positionen nicht, und deshalb behält das Stück Kochsalz seine Form. Die starken Anziehungskräfte sind auch dafür verantwortlich, dass es immer eines gewissen Kraftaufwands bedarf, um ein Stück Kochsalz zu zerteilen. Mit anderen Worten: Aus den allgemeinen Naturgesetzen ergibt sich, welche Kräfte zwischen den Natrium- und Chlorionen wirken, aus denen Kochsalz besteht. Und aus diesen Kräften ergibt sich, dass sich Ionengitter aus Natrium- und Chlorionen genau so verhalten, wie dies für feste Körper charakteristisch ist.

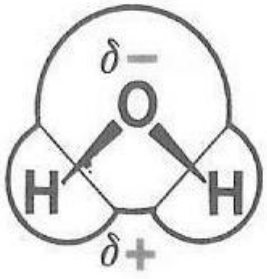



Abbildung 3 (Lossow/Wernet 1998, 69)

Und warum ist Kochsalz wasserlöslich? Das liegt zum einen wieder daran, dass Kochsalz aus einem Gitter von positiv und negativ geladenen Ionen besteht. Zum anderen liegt es an der Dipolstruktur der $H_2O$-Moleküle. Denn aufgrund dieser Struktur können $H_2O$-Moleküle die einzelnen Ionen aus ihrer Position im Gitter herauslösen, so dass sich diese zwischen den Wassermolekülen verteilen.

Beispiel 2: Die *Fähigkeit* von Seife, *Schmutz zu lösen*. „Seifen sind eine Mischung verschiedener, längerkettigen Alkalisalze der Fettsäuren und zählen zu den Tensiden, genauer zu den anionischen Tensiden. Die Seifenmoleküle verdanken ihre Eigenschaften der Tatsache, dass sie aus einer langen, wasserabweisenden (hydrophoben) Kohlenwasserstoffkette und einem wasseranziehenden (hydrophilen) Teil, der sogenannten Carboxylatgruppe ($–COO^-$) bestehen. […].“ (http://de.wikipedia.org/wiki/Seife – Abruf 22.04.2011)

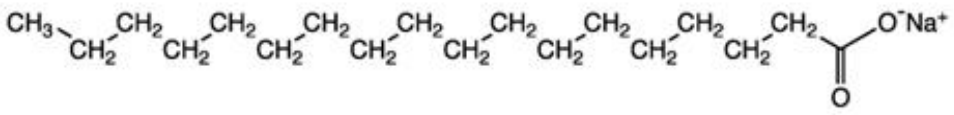



Abbildung 4

Das „Lösen von Fett“ (Öl, Staub, Schmutz) von der zu reinigenden Fläche und die Abführung dieser über das Waschwasser ist die eigentliche reinigende Wirkung der Seifen. Die langen Kohlenwasserstoffketten der Seifenmoleküle lösen sich leicht in kleinen Fetttropfen […]. Die polaren Enden ragen jedoch in das umgebende Wasser hinaus. Der Fetttropfen wird von den Seifenmolekülen schließlich vollständig umhüllt und von der zu reinigenden Fläche abgelöst. (ebd.)

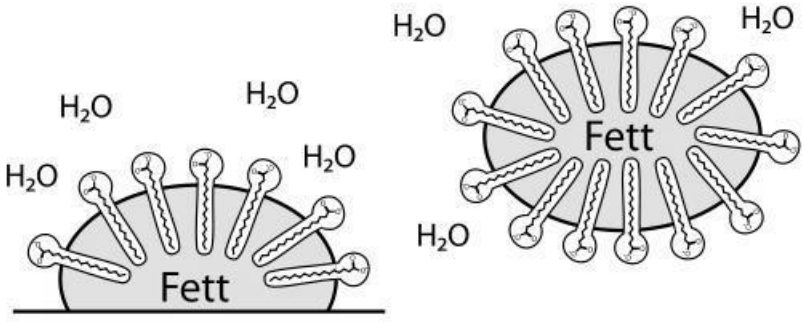



Abbildung 5

Die Vielzahl der so mit Seifenmolekülen ummantelten Fett- und Öltropfen bildet im Wasser eine sogenannte Emulsion, die am Ende des Waschvorganges durch Abspülen mit frischem Wasser abgeführt werden kann. (ebd.)

In beiden Beispielen zeigt sich offensichtlich dieselbe Struktur: Die Festigkeit und Wasserlöslichkeit von Kochsalz wird ebenso wie die Fähigkeit von Seife, Schmutz zu lösen, auf die Eigenschaften der Atome bzw. Moleküle, aus denen diese Stoffe bestehen, zurückgeführt sowie auf deren räumliche Anordnung. Der Begriff der reduktiven Erklärung, von dem Levine ausgeht, lässt sich also so fassen:

(RE) Die Eigenschaft F eines Stoffes oder Gegenstandes, der aus den Teilen $C_1$, …, $C_n$ besteht, die in der Weise R angeordnet sind, ist genau dann reduktiv erklärbar, wenn sich aus den allgemeinen grundlegenden Naturgesetzen ergibt, dass ein Stoff oder Gegenstand, der aus solchen Teilen besteht, die in dieser Weise angeordnet sind, alle Merkmale aufweist, die für F charakteristisch sind.

In diesem Sinne folgt aus den *allgemeinen grundlegenden* Naturgesetzen, dass ein Gitter aus Natrium- und Chlorionen sich genau so verhält, wie es für einen festen und wasserlöslichen Stoff charakteristisch ist. Und in diesem Sinne folgt aus den allgemeinen grundlegenden Naturgesetzen auch, dass ein Stoff, dessen Moleküle aus einer langen, wasserabweisenden Kohlenwasserstoffkette und einer wasseranziehenden Carboxylatgruppe bestehen, Schmutz z. B. von Kleidungsstücken ablöst. Allerdings, so Levine, folgt aus den Gesetzen der Neurobiologie eben *nicht*, dass sich ein Hirnzustand, bei dem bestimmte Neuronen (nennen wir sie „S-Neuronen“) feuern, für den Betroffenen schmerzhaft anfühlt. Das ist der Grund dafür, dass es – vor dem Hintergrund der grundlegenden Naturgesetze – *undenkbar* ist, dass ein Gitter aus Natrium- und Chlorionen nicht fest oder nicht wasserlöslich ist bzw. dass Seifen die Fähigkeit, Schmutz zu lösen, nicht besitzen, während es auf der anderen Seite sehr wohl *denkbar* ist, dass in meinem Hirn die S-Neuronen feuern, ich aber keinen Schmerz empfinde. Dabei kann es allerdings durchaus sein, dass aus den Gesetzen der Neurobiologie sehr wohl folgt, dass das Feuern der S-Neuronen genau die kausale Rolle spielt, die normalerweise Schmerzen spielen – es wird durch Gewebeverletzungen verursacht und verursacht seinerseits bestimmte Verhaltensweisen (ich stöhne; versuche die schmerzende Stelle zu kühlen; gehe ins Bad, um mir eine Schmerztablette zu holen). Diese kausale Rolle, so Levine, ist von dem qualitativen Charakter des Sich-schmerzhaft-Anfühlens strikt zu trennen. Und etwas ist eben nur dann ein Schmerz, wenn es sich wie ein Schmerz anfühlt. Natürlich kann es sein, dass es *zusätzlich* zu den allgemeinen grundlegenden Naturgesetzen ein *spezielles* Gesetz gibt, dem zufolge das Feuern von S-Neuronen immer zu einem Erlebnis führt, das sich schmerzhaft anfühlt. Doch dieses Gesetz könnte auch anders sein, ohne dass sich an den allgemeinen grundlegenden Gesetzen etwas ändern müsste.

Dieser eigenartig „lose“ Zusammenhang zwischen Gehirnzuständen und Empfindungen wird in der Philosophie schon sehr lange diskutiert – am häufigsten am Beispiel von Farbempfindungen. Viele stellen sich den Wahrnehmungsprozess in etwa so vor: Wenn ich eine rote Blume sehe, fällt das von der Blume reflektierte Licht auf meine Retina und erzeugt dort bestimmte chemisch-elektrische Impulse, die über den Sehnerv in die visuellen Areale des Gehirns gelangen und dort bestimmte neuronale Prozesse auslösen (nennen wir sie „R-Prozesse“). Damit ist der Sehvorgang, so diese Auffassung, aber noch nicht abgeschlossen; vielmehr verursachen am Ende die R-Prozesse bestimmte Rotempfindungen in meinem Bewusstsein.

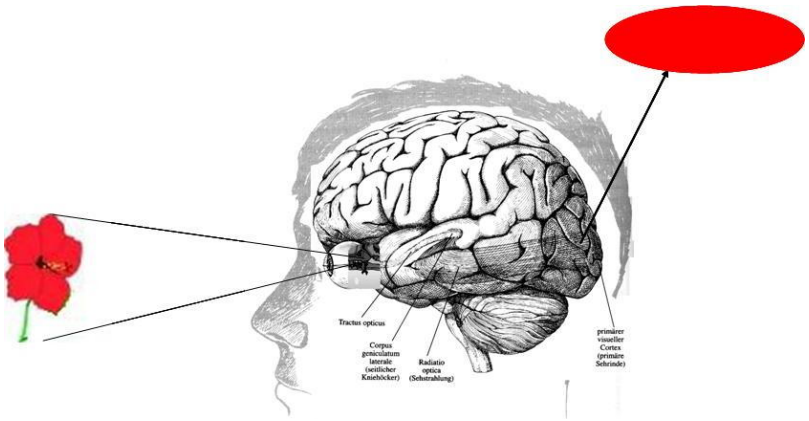



Abbildung 6

Aber können wir eigentlich ausschließen, dass es Menschen gibt, bei denen die Farbempfindungen gegenüber den „normalen“ vertauscht sind? Die beim Sehen roter Dinge (Rosen, Feuerwehrautos, Sonnenuntergänge) grüne

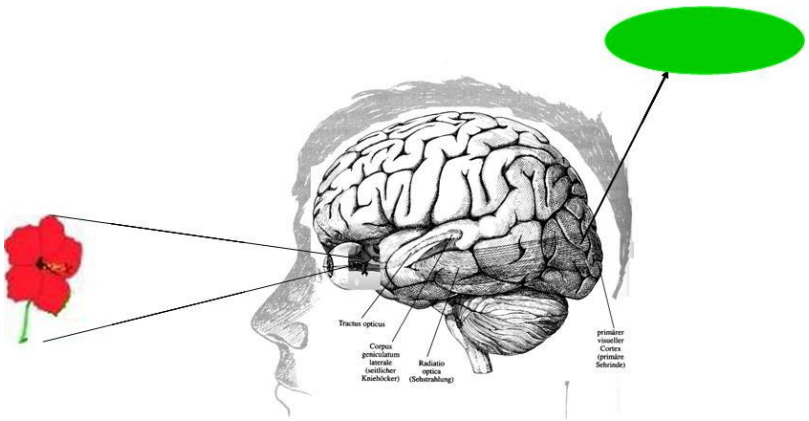



Abbildung 7

Farbempfindungen haben und beim Sehen grüner Dinge (Gurken, Wiesen, Laubfrösche) rote Farbempfindungen? Mit anderen Worten: Sind *vertauschte Qualia* möglich? Und wenn vertauschte Qualia möglich sind, könnte es dann nicht vielleicht sogar sein, dass bei manchen Menschen R-Prozesse überhaupt nicht mit einem Quale verbunden sind? Sind vielleicht sogar *fehlende Qualia* möglich?

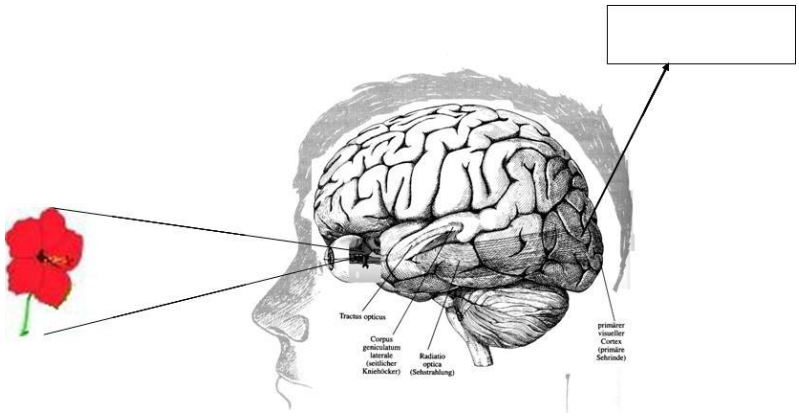



Abbildung 8

Eins scheint klar: Es ist nicht zu sehen, wie wir *vertauschte* Qualia feststellen können. Wenn jemand auf rote Dinge mit Grünempfindungen reagiert, dann wird er trotzdem sagen, diese Dinge seien rot. Denn er hat ja gelernt, Dinge, die in ihm Grünempfindungen auslösen, „rot" zu nennen. Und wenn jemand, wenn die Ampel auf Rot springt, ein Grünerlebnis hat, wird er trotzdem anhalten; denn er hat ja gelernt, genau dies zu tun, wenn die Ampel Rot zeigt. Aber wie ist es bei *fehlenden* Qualia? Müsste man die nicht feststellen können?

Das hängt offenbar davon ab, ob schon die neuronalen Korrelate die kausalen Rollen innehaben, die wir normalerweise den Empfindungen zuschreiben. Nehmen wir noch einmal den Fall des Schmerzes. Schmerzen haben typische Ursachen (z. B. Gewebeverletzungen) und sie haben typische Wirkungen (Stöhnen oder Schreien, Versuche zur Linderung des Schmerzes, Unaufmerksamkeit). So wie wir die Dinge bisher analysiert haben ist klar, dass die typischen Ursachen zunächst zum Feuern der S-Neuronen führen (dem neuronalen Korrelat von Schmerzempfindungen), das dann seinerseits Erlebnisse hervorruft, die sich schmerzhaft anfühlen. Und wie ist es mit den typischen Wirkungen? Hier sind zwei Möglichkeiten denkbar: a) Die typischen Wirkungen werden schon vom Feuern der S-Neuronen hervorgerufen; b) diese Wirkungen gehen erst auf die Schmerzempfindungen selbst zurück. Grafisch kann man das so darstellen:

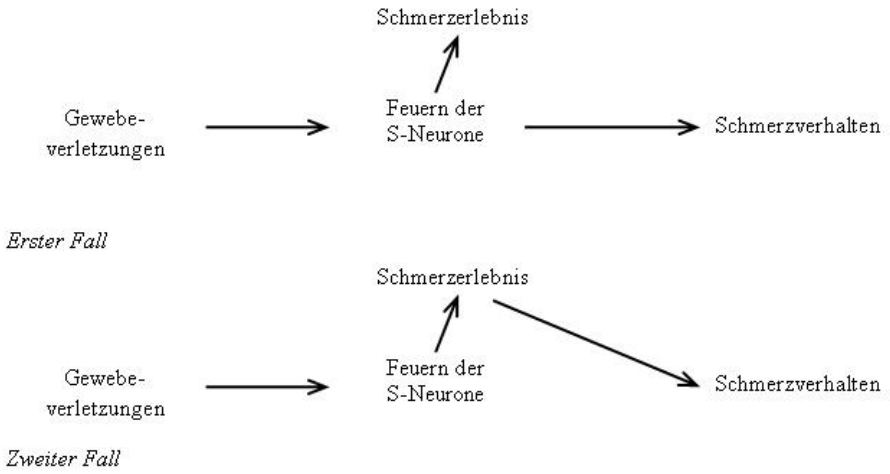



Abbildung 9

Was würde passieren, wenn wir einem Wesen mit *fehlenden* Qualia begegnen, bei dem das Feuern der S-Neuronen überhaupt nicht mit einem Erlebnis verbunden ist – weder mit einem, das sich schmerzhaft anfühlt, noch mit einem, das sich irgendwie anders anfühlt (vielleicht wie ein Kitzel)? Im ersten Fall würde sich dieses Wesen offenbar genau so verhalten wie jedes „normale Wesen"; es wäre von einem normalen Wesen ununterscheidbar. Im zweiten Fall dagegen würde es ein gänzlich anderes Verhalten zeigen, so dass wir schnell erkennen könnten, dass diesem Wesen die Qualia fehlen.

Damit sind wir zurück bei der Titelfrage. Denn philosophische Zombies sind offenbar genau die Wesen, denen *alle Qualia fehlen* – Wesen, die physisch völlig einem normalen Menschen gleichen, bei denen aber kein Gehirnzustand mit irgendwelchen qualitativen Erlebnissen verbunden ist. Wirklich interessante Zombies sind allerdings erst die Wesen, bei denen darüber hinaus die Gehirnzustände, die normalerweise zu qualitativen Erlebnissen führen, schon genau die kausale Rolle innehaben, die für diese Erlebnisse charakteristisch ist – ganz im Sinne des gerade geschilderten ersten Falls. Wenn wir diese Wesen „Typ 1-Zombies" nennen, ist die Frage also, ob es wirklich möglich ist, dass ich ein Typ 1-Zombie bin.

Ich habe da meine Zweifel – hauptsächlich aus epistemischen Gründen. Angenommen, ich wäre ein Typ 1-Zombie, könnte jemand anderes herausfinden, dass das so ist? Könnte jemand anderes z. B. herausfinden, dass ich keine wirklichen Schmerzempfindungen habe? Ich sehe nicht wie. Wenn ich ein Typ 1-Zombie wäre, dann würden bei mir die S-Neuronen in genau denselben Situationen feuern, in denen sie bei normalen Menschen feuern.

Und wenn meine S-Neuronen feuern, verhalte ich mich exakt genau so wie ein normaler Mensch, wenn er Schmerzen hat: Ich schreie „Aua“; ich versuche, die schmerzende Stelle zu behandeln; vielleicht gehe ich ins Bad und nehme eine Schmerztablette. Selbst wenn es möglich wäre, in mein Gehirn hineinzuschauen, würde niemand etwas finden, das mich von einem normalen Menschen unterscheidet. Und schließlich: Dass ich mich, wenn meine S-Neuronen feuern, genau so verhalte wie ein normaler Mensch, der Schmerzen empfindet, heißt auch, dass ich genau so *rede* wie dieser normale Mensch. Wenn ich gefragt werde, ob es weh tut, antworte ich z. B. „Ja sehr, ich bin in eine Scherbe getreten“. Wenn man mich dann fragt, ob sich mein Schmerz schmerzhaft anfühlt, werde ich selbstverständlich mit „Ja“ antworten. Und ich bin auch durchaus in der Lage, meinen Schmerz als „bohrend“, „stechend“, „dumpf“ oder „pochend“ zu beschreiben. Offenbar gibt es wirklich nichts, das jemand anderen dazu berechtigen könnte, mich für einen Typ 1-Zombie zu halten.

Und wie steht es mit mir selbst? Könnte wenigstens ich feststellen, dass ich ein Typ 1-Zombie bin? Auf den ersten Blick scheint das völlig unproblematisch. Ich selbst weiß doch auf Grund von Introspektion, ob ich etwas fühle und was ich fühle, d. h., welchen qualitativen Charakter meine Erlebnisse haben. Aber ganz so einfach sind die Dinge nicht. Schließlich: Wenn ich ein Typ 1-Zombie bin, stelle ich per Introspektion fest, dass ich gar nichts fühle, höre mich aber zugleich sagen: „Ja, ich habe starke Schmerzen“, „Sicher fühlen sich diese Schmerzen schmerzhaft an“ und „Es ist ein eher stechender Schmerz“. Ein Fall von Schizophrenie?

Normalerweise stellen wir uns die Dinge so vor: Per Introspektion stelle ich fest, was ich fühle – ob ich Schmerzen habe, ob mir übel ist oder ob ich mich freue, d. h., per Introspektion kenne ich auch den qualitativen Charakter meiner Erlebnisse. Was ich sage, hängt dann davon ab, was ich per Introspektion über mich herausfinde. Wenn ich herausfinde, dass ich Schmerzen fühle, sage ich, dass ich Schmerzen habe, usw. Doch dieses Bild kann nicht zutreffen, wenn ich ein Typ 1-Zombie bin. Denn bei einem solchen Zombie hängt *per definitionem* alles, was er sagt, allein von dem ab, was in seinem Gehirn vorgeht. Introspektion kann hier also nicht so funktionieren, wie wir uns das normalerweise denken. Generell ist es eher so: Menschen haben in der Regel nicht nur Empfindungen; sie wissen auch, welche Empfindungen sie haben, und das impliziert z. B., dass sie, wenn sie Schmerzen haben, auch davon überzeugt sind, dass sie Schmerzen haben, dass sie, wenn ihnen übel ist, auch davon überzeugt sind, dass ihnen übel ist, usw. Dabei hängt, was sie über sich sagen, von den Überzeugungen ab, die sie im Bezug auf ihre Empfindungen haben. Weiter ist es nicht unplausibel, anzunehmen, dass der Zusammenhang zwischen Empfindungen und Überzeugungen kausal ist; d. h., mein Schmerz verursacht meine Überzeugung,

dass ich Schmerzen habe, und meine Übelkeit verursacht meine Überzeugung, dass mir übel ist. Wenn das so ist, wird die Überzeugung, dass er Schmerzen hat, bei einem Typ 1-Zombie aber voraussetzungsgemäß durch das Feuern seiner S-Neuronen verursacht. Mit anderen Worten: Wenn bei einem Typ 1-Zombie die S-Neuronen feuern, dann glaubt er, dass er Schmerzen hat, und dann sagt er deshalb auch, dass er Schmerzen hat. Für einen Typ 1-Zombie stellt sich deshalb sein Innenleben genau so dar wie für einen normalen Menschen. Wenn ein normaler Mensch Schmerzen empfindet, weil seine S-Neuronen feuern, glaubt er, dass er Schmerzen hat; aber auch der Typ 1-Zombie glaubt, dass er Schmerzen hat, wenn seine S-Neuronen feuern. Auch er hat daher keine Möglichkeit, herauszufinden, dass er in Wirklichkeit nur ein Zombie ist.

Wenn jedoch andere nicht herausfinden können und wenn nicht einmal der Zombie selbst herausfinden kann, dass er ein Zombie ist, scheint mir die Idee, er sei tatsächlich ein Zombie, absurd. Wahrscheinlich haben wir etwas für möglich gehalten, was tatsächlich nicht möglich ist – nämlich dass die neuronalen Korrelate qualitativer Erlebnisse auch dann mit keinerlei qualitativen Erlebnissen verbunden sind, wenn diese qualitativen Erlebnisse selbst keinerlei kausale Effekte in der Welt haben, wenn alle kausalen Wirkungen vielmehr allein auf die neuronalen Korrelate zurückgehen.

## Literatur

Beckermann, A. (2003): „Mentale Eigenschaften und mentale Substanzen – Antworten der Analytischen Philosophie auf das ‚Leib-Seele-Problem'". In: U. Lorenz (Hg.) *Philosophische Psychologie*. Freiburg/München: Karl Alber, 203–221.

– (2008a): *Das Leib-Seele-Problem. Reihe: Kurs Philosophie*. (utb 2983) Paderborn: Wilhelm Fink Verlag.

– (2008b): *Analytische Einführung in die Philosophie des Geistes*. 3. Aufl., Berlin/New York: de Gruyter.

Descartes, R., *Über den Menschen (1632) sowie Beschreibung des menschlichen Körpers (1648)*. Nach der ersten französischen Ausgabe von 1664 übersetzt und mit einer historischen Einleitung versehen von K.E. Rothschuh. Heidelberg: Schneider 1969.

Levine, J. (1983): „Materialism and Qualia: The Explanatory Gap". *Pacific Philosophical Quarterly* 64, 354–361. (Dt. in: Michael Pauen & Achim Stephan (Hg.) *Phänomenales Bewusstsein – Rückkehr zur Identitätstheorie?* Paderborn: mentis 2002, 91–102)

Lossow, C. & H. Wernet (1998): *Telekolleg II, Chemie, Bd. 1*. 5. Aufl. München: TR Verlagunion.

# Ich, Selbst, Selbstbewusstsein

# 13

## Selbstbewusstsein in kognitiven Systemen[*1]

1.

Etwas provokant könnte man die Ausgangsfrage meiner Überlegungen so formulieren: „Welche kognitiven Systeme haben ein Selbst?“ Oder: „Können auch künstliche kognitive Systeme ein Selbst haben?“ Doch die Rede von *einem* oder *dem* Selbst ist höchst fragwürdig. Das Wort ‚selbst‘ ist in erster Linie ein Pronomen wie in den Sätzen „Er kam selbst“ oder „Ich weiß nicht, ob der Bundespräsident selbst kommen wird“. Es kann auch ein Adve-rb sein wie in „Selbst er konnte die Niederlage nicht verhindern“. Das Substantiv ‚Selbst‘ kommt dagegen nur in bestimmten Redewendungen vor wie in „Da zeigte sich sein wahres Selbst“. Die Überhöhung dieses Substantivs, die sich in Sätzen wie „Das Selbst jedes Menschen ist unzerstörbar“ zeigt, ist dagegen eine philosophische Erfindung, die wohl im Wesentlichen auf John Locke zurückgeht. Im zweiten Buch des *Essay Concerning Human Understanding* jedenfalls schreibt Locke:

> Self is that conscious thinking thing, (whatever Substance made up of whether Spiritual or Material, Simple or Compounded, it matters not), which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern’d for it *self*, as far as that consciousness extends. (ECHU II xxvii 17)

Der linguistische Hintergrund dieser ungewöhnlichen Verwendung von ‚Selbst‘ bzw. ‚self‘ ist offenbar die Tatsache, dass man früher im Englischen Wörter wie ‚my self‘ oder ‚it self‘ auseinander schreiben konnte, was natürlich die Vermutung zumindest begünstigt, es gäbe da so etwas wie mein Selbst. Dass diese Lesart jedoch zumindest eigenartig ist, zeigt sich sehr deutlich in einer neueren englischen Ausgabe von Descartes’ *Meditationen*, wo man als Übersetzung des eigentlichen harmlosen Satzes „Nunquid *me ipsum* non tantum multo verius, multo certius, sed etiam multo distinctius evidentiusque, cognosco?“ (Descartes *Meditationes* 33 – Hervorh. vom Verf.) Folgendes finden kann: „Surely my awareness of *my*

---

* Erstveröffentlichung in: Markus F. Peschl (Hg.) *Die Rolle der Seele in der Kognitionswissenschaft und der Neurowissenschaft. Auf der Suche nach dem Substrat der Seele*. Würzburg: Königshausen & Neumann 2005, 171–187.

1 Bei diesem Aufsatz handelt es sich um die deutsche Fassung von Beckermann 2003. Für äußerst hilfreiche Kommentare zu einer früheren Fassung möchte ich Christian Nimtz herzlich danken.

*own self* is not merely much truer and more certain but also much more distinct and evident.“[2]

Auf der anderen Seite ist die ungewöhnliche neue Verwendung des Wortes ‚Selbst‘ bzw. ‚self‘ bei Locke zwar ärgerlich, aber nicht unbedingt schädlich; denn Locke sagt ja ausdrücklich, was er unter einem Selbst verstehen will: Ein Selbst ist das Ding, das Lust und Schmerz (bewusst) erfahren kann, das zu Glück und Unglück fähig ist und das insofern um sich selbst besorgt ist. Das Selbst scheint also gar nichts anderes zu sein als der Mensch oder die Person.[3] Und Locke lässt bewusst offen, ob dieses Selbst etwas Materielles oder etwas (immateriell) Geistiges ist. Er will also bzgl. des Cartesischen Dualismus keine Stellung beziehen, lässt aber ausdrücklich doch die Möglichkeit offen, dass das eigentliche Selbst (!) des Menschen eine Cartesische *res cogitans* ist. Es ist daher nicht verwunderlich, dass das Reden von einem ‚Selbst‘ oft mit impliziten oder expliziten Cartesischen Konnotationen verbunden ist.

Trotzdem: Ein Selbst ist nach Locke einfach das, was zu angenehmen und unangenehmen Empfindungen und damit zu Glück und Unglück fähig ist. In der neueren Literatur steht neben diesem jedoch häufig noch ein anderer Aspekt im Vordergrund. So schreibt z. B. E. J. Lowe:

> [Selves] are subjects of experience which have the capacity to recognise themselves as being individual subjects of experience. Selves possess reflexive self-knowledge. By ‚reflexive self-knowledge‘ I mean, roughly speaking, knowledge of one’s own identity and conscious mental states – knowledge of who one is and of what one is thinking and feeling. … [R]oughly speaking – having the kind of reflexive self-knowledge which makes one a person goes hand-in-hand with possessing a ‚first-person‘ concept of oneself, the linguistic reflection of which resides in an ability to use the word ‚I‘ comprehendingly to refer to oneself. (Lowe 2000, 264 f.)

Außer durch die Fähigkeit, Erfahrungen machen zu können, sind Selbste (diesen Plural akzeptiert selbst der Duden nicht!) nach Lowe also dadurch charakterisiert, dass sie reflexive Selbstkenntnis besitzen – und d. h. offenbar im besonderen Wissen, dessen Inhalt der Wissende selbst adäquat nur unter Verwendung des Wortes ‚Ich‘ formulieren kann. So verstanden sollte man sich durch das Wort ‚Selbst‘ nicht zu Fragen verführen lassen wie „Welche kognitiven Systeme haben ein Selbst?“. Denn erstens müsste die

---

[2] Descartes, René *Selected Philosophical Writings* (86 – Hervorhebung vom Verf.). In der Haldane/Ross Ausgabe von 1911 hieß es noch: „[D]o I not know *myself*, not only with much more truth and certainty, but also with much more distinctness and clearness?“ (156 – Hervorhebung vom Verf.)

[3] Vgl. z. B. die folgende entlarvende Bemerkung Disraelis: „Self is the only *person* whom we know anything about.“ *The Oxford English Dictionary*, 2nd ed., s.v. ‚self‘, C.I.1.e. (Hervorhebung vom Verf.)

Frage eigentlich lauten: „Welche kognitiven Systeme sind Selbste?“. Und zweitens lässt sich der Inhalt dieser Frage auch in normalem Deutsch formulieren: „Welche kognitiven Systeme verfügen über reflexives Wissen, dessen Inhalt adäquat nur unter Verwendung des Wortes ‚Ich‘ formuliert werden kann?“[4] Dieser durchaus vernünftigen Frage, die nicht gleich insinuiert, dass Selbste eine besondere, mystische Art von Entitäten darstellen, will ich im Folgenden nachgehen.

## 2.

Kognitive Systeme sind Systeme, die versuchen, sich ein Bild von der Welt zu machen, in der sie leben. Sie repräsentieren ihre Umwelt, um in dieser Umwelt besser zurechtzukommen. Grundsätzlich können diese Repräsentationen sehr unterschiedliche Formen annehmen. Im Folgenden werde ich allerdings nur solche kognitive Systeme betrachten, die ihre Umwelt explizit in Form von Listen, also in einer bestimmten Spielart einer *lingua mentis*, repräsentieren. Meine These ist, dass solche kognitiven Systeme genau dann über die für ein Selbst erforderliche reflexive Selbstkenntnis verfügen, wenn sie über eine ganze bestimmte Art von Repräsentationen verfügen – Repräsentationen, die von ihnen selbst handeln und die darüber hinaus auf eine ganz bestimmte Weise von ihnen selbst handeln. Über kognitive Systeme anderer Art werde ich hier nichts sagen. Allerdings bin ich davon überzeugt, dass auch diese anderen kognitiven Systeme genau dann über die für ein Selbst erforderliche reflexive Selbstkenntnis verfügen, wenn es in ihnen *analoge* Strukturen oder Fähigkeiten gibt. Ich bin also davon überzeugt, dass sich die folgenden Überlegungen – mit mehr oder weniger großem Aufwand – auf andere kognitive Systeme übertragen lassen. Mehr will ich darüber hier aber nicht sagen. Konzentrieren wir uns also auf die Frage: Welche Repräsentationen müssen die von mir behandelten kognitiven Systeme aufbauen, um über die für ein Selbst erforderliche reflexive Selbstkenntnis zu verfügen?[5]

John Perry hat in „Myself and I“ eine für diese Frage äußerst wichtige Unterscheidung zwischen drei Arten selbstbezogenen Wissens getroffen – die Unterscheidung zwischen akteurzentriertem Wissen (*agent-relative knowledge*), eigentlichem Selbstwissen (*self-attached knowledge*) und Wissen, das sich auf eine Person bezieht, die (zufällig) man selbst ist (*know-*

---

[4] Genauer muss es natürlich heißen: „Welche kognitiven Systeme verfügen über reflexives Wissen, dessen Inhalt adäquat nur unter Verwendung von Quasi-Indikatoren im Sinne Castañedas formuliert werden kann?“

[5] Die Geschichte, mit der ich diese Frage beantworten werde, ist natürlich nicht neu. Vgl. z. B. Rosenberg 1986, bes. Kap. VI und VII.

*ledge of the person one happens to be*). Akteurzentriertes Wissen liegt dann vor, wenn ein System die Umwelt von seiner eigenen Perspektive aus repräsentiert. Es setzt nicht voraus, dass dieses System, der Akteur, über Repräsentationen verfügt, die sich explizit auf es selbst beziehen. Das System braucht keinen Begriff von sich selbst zu haben. In unserem Kontext bedeutet das: Die Listen, in denen akteurzentriertes Wissen repräsentiert wird, müssen keine Symbole enthalten, die sich explizit auf das System beziehen. Letzten Endes besteht akteurzentriertes Wissen also in Repräsentationen, die in der Kognitionswissenschaft schon lange bekannt sind – Repräsentationen, in denen die Umwelt nicht in Welt-, sondern in beobachterzentrierten Koordinaten repräsentiert wird. Was das heißt, lässt sich an folgendem Beispiel gut veranschaulichen.

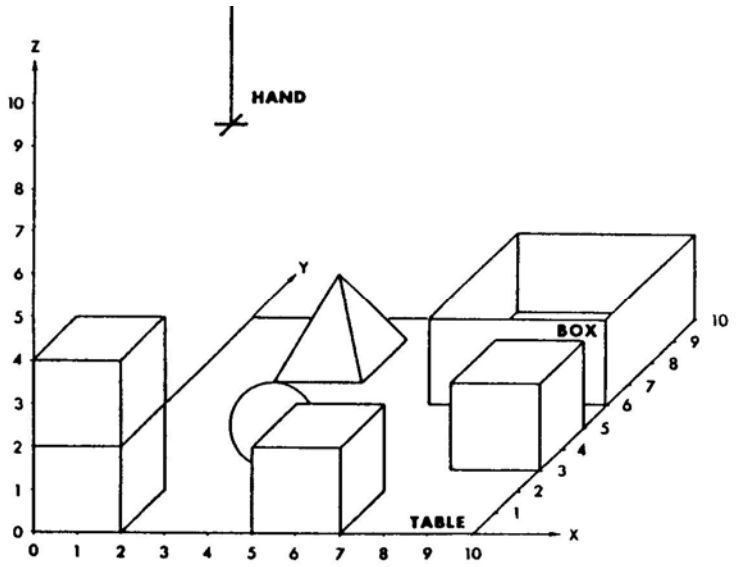



Abbildung 1

Eines der frühen Erfolgsprogramme der KI war das Programm SHRDLU – ein Programm, in dem Terry Winograd versuchte, Sprachverstehen und Handlungsplanung in ein System zu integrieren. SHRDLU ‚lebt' in einer Art Mikrowelt, in der es eine Reihe unterschiedlicher Gegenstände gibt – Blöcke, Kugeln, Pyramiden und eine Schachtel. Die Aufgaben des Programms bestehen in der Regel darin, diese Gegenstände auf bestimmte Weise neu anzuordnen – also z. B. die grüne Pyramide auf den roten Block zu stellen oder die Kugel in die Schachtel zu legen. Für uns ist hier nur wichtig, wie SHRDLU seine ‚Umwelt' repräsentiert. Die Repräsentation einer Situation wie der, die in Abbildung 1 dargestellt ist, hat z. B. die folgende Form:

```
(IST-EIN      OBJEKT-1    BLOCK)
(FARBE        OBJEKT-1    GRÜN)
(ORT          OBJEKT-1    (1  1  2))
(GRÖßE        OBJEKT-1    (2  2  2))
…
(IST-EIN      OBJEKT-5    KUGEL)
(FARBE        OBJEKT-5    GRÜN)
(ORT          OBJEKT-5    (4  3  0))
(GRÖßE        OBJEKT-5    (2  2  2))
…
(HÄLT         HAND        NICHTS)
(ORT          HAND        (2  5  7))
```

Entscheidend ist hier, dass alle Informationen über den Ort der beteiligten Objekte in so genannten *Weltkoordinaten* repräsentiert sind – sogar der Ort, an dem sich die Greifhand des Systems befindet. Wenn SHRDLU einen Gegenstand *a* greifen will, muss es also überlegen: Wo befindet sich *a*? Wo befindet sich die Hand? Wie kommt die Hand von dort zum Ort von *a*? Das erscheint nicht nur extrem unnatürlich. Es stellt sich auch die Frage, wie kognitive Systeme eigentlich Wissen darüber erwerben sollen, an welchen *objektiven* Koordinaten sich die Gegenstände in ihrer Umgebung und sie selbst befinden. Wir Menschen jedenfalls nehmen die Welt ganz offensichtlich nicht in Weltkoordinaten wahr. Wenn ich von der Tür aus in mein Arbeitszimmer schaue, sehe ich vielmehr ungefähr folgendes:

> Ca. 5 Schritte geradeaus befindet sich ein Schreibtisch.
> Direkt hinter dem Schreibtisch ist ein Fenster.
> Links auf dem Schreibtisch steht ein Bildschirm.
> Davor eine Tastatur.
> Mitten auf dem Schreibtisch steht eine Tasse.
> Links neben dem Schreibtisch ist ein Regal.
> In dem Regal steht halbhoch ein Drucker.

Ich repräsentiere den Ort der Dinge in meiner Umgebung also durch die räumlichen Relationen, in denen sie *zu mir* und *zueinander* stehen. Und dafür gibt es zwei sehr gute Gründe. Zum ersten dieser Gründe schreibt Perry:

> Everything we learn about other objects we learn by employing methods that are appropriate because those objects stand in certain relations to us. … [The objects in our vicinities have] *agent-relative roles*: roles that other individuals play in the lives of agents. These are agent-relative roles, because an object plays or doesn't play such a role relative to a given agent, at a given time. For example, my computer is playing the role of *object in front* right now, relative to me, but not relative to you. … This is the first of two very general facts I

> want to emphasize: any object we learn about plays some agent-relative role, basic or derived, in our life. We learn about the object by using an epistemic method connected to the role, a way of finding out about the object or person playing that role. The way to find out about the object in front of you is to look at it, or perhaps to walk up to it and touch it. (Perry 1998, 84 f.)

Mit anderen Worten: Wir können nur etwas über die Dinge unserer Umwelt erfahren, weil sie zu uns in bestimmten Beziehungen stehen, weil sie uns gegenüber bestimmte Rollen einnehmen. Es ist daher kein Wunder, dass wir diese Dinge zunächst als Träger dieser Rollen repräsentieren – als den Tisch 5 Schritte vor uns, die Person rechts neben uns, der Boden unter unseren Füßen, usw. Noch wichtiger ist aber ein anderer Punkt. Nur wenn ich die Dinge meiner Umwelt als Träger der Rollen repräsentiere, die sie mir gegenüber spielen, weiß ich unmittelbar, was von dem, was um mich herum vorgeht, für mich relevant ist. Nehmen wir einen Ball, der auf mich zufliegt. Wenn ich die Bewegung des Balles als ein Auf-mich-Zufliegen repräsentiere, folgt unmittelbar, dass ich etwas tun muss – den Ball auffangen oder mich bücken oder was auch immer. Wenn ich die Bewegung des Balles dagegen in Weltkoordinaten repräsentiere, folgt daraus nur dann etwas, wenn ich weiß, an welchen Weltkoordinaten ich mich selbst befinde und ob dieser Ort auf der Bahn des Balles liegt. Offenbar erleichtern akteurzentrierte Repräsentationen es uns also erheblich, das, was um uns herum passiert, in seiner Handlungsrelevanz richtig einzuschätzen. Dies führt auch sofort zu dem zweiten Grund, über den Perry schreibt:

> … [E]verything we do comes down to performing operations on the objects around us – objects in front of us, behind us, above us; objects we are holding; objects we can see. By doing these things, we do things to objects in less basic relations to us. … I know how to move my body so as to effect objects around me, and I know how effecting those objects will effect other objects related to them in certain ways. (Perry 1998, 85)

Wenn wir handeln, führen wir immer Körperbewegungen mit Bezug auf Dinge aus, die uns gegenüber eine bestimmte Rolle einnehmen. Es gibt einen bestimmten Typ von Körperbewegung, mit dem wir eine Tasse greifen, die vor uns auf dem Tisch steht; es gibt einen bestimmten Typ von Körperbewegung, mit dem wir eine Fliege von unserer Wange verscheuchen; und es gibt – Perrys Lieblingsbeispiel – einen bestimmten Typ von Körperbewegung, mit dem wir die Person links neben uns erstechen. Aber es gibt *keinen* bestimmten Typ von Körperbewegung, mit dem wir die Pyramide am Ort (a b c) auf den Block am Ort (d e f) setzen. D. h., wenn wir überlegen, mit welchem Typ von Körperbewegung wir eine bestimmte Handlung vollziehen können, ist es überaus hilfreich, wenn die Orte der Dinge in unserer Umwelt in akteurzentrierten Koordinaten repräsentiert sind. Ich weiß, mit welcher Körperbewegung ich eine Tasse von mir auf dem Schreibtisch

greifen kann; aber ich weiß nicht ohne weiteres, mit welcher Körperbewegung ich eine Tasse greifen kann, die sich an der Stelle (7 10 3) befindet.

> There are then two kinds of methods connected with agent-relative roles, epistemic methods and pragmatic methods. These two kinds of methods are the key to all human intelligence and purposive activity. We know how to find out what kinds of objects occupy these roles, and we know how to perform various operations on them. … Our practical knowledge then, the knowledge that enables us to do things, forms a structure at whose base is information about the objects that play relatively basic agent-relative roles in our lives. (Perry 1998, 85)

Warum ist das für uns wichtig? Nun, weil akteurzentrierte Repräsentationen immer einen Bezug auf den Akteur selbst beinhalten. Nicht ohne Grund sprechen Kognitionswissenschaftler in diesem Zusammenhang von einem ‚*ego*-centric reference system'. Wenn ich einen Apfel auf dem Tisch liegen sehe, in die Hand nehme und aufesse, beruhen die komplexen Körperbewegungen, mit deren Hilfe ich den Apfel in meinen Mund befördere, darauf, dass ich durch Wahrnehmung gelernt habe, dass sich der Apfel in einer bestimmten räumlichen Beziehung zu *mir* befindet. Wenn ich die Tastatur meines Computers nutze, muss ich meine Finger in einer bestimmten Entfernung und Richtung von *mir* bewegen. „It isn't enough to know where the buttons were relative to one another, or where the [keyboard] was in the building or the room. I had to know where these things were relative to *me*. It seems then, that these basic methods already require me to have some notion of myself." (Perry 1998, 86)

Schon akteurzentrierte Repräsentationen scheinen also zu den Repräsentationen zu zählen, deren Inhalt nur unter Verwendung des Wortes ‚ich' sprachlich angemessen wiedergegeben werden kann. Doch dieser Schein trügt. Wie auch Perry betont, ist das Wort ‚ich' bei der sprachlichen Formulierung des Inhalts dieser Repräsentationen immer verzichtbar. Wenn ich die Situation repräsentieren möchte, dass sich 50 cm vor mir leicht nach rechts versetzt ein Apfel befindet, dann muss ich nicht repräsentieren, dass sich der Apfel in einer bestimmten räumlichen *Relation* zu *mir* befindet, in der er sich auch zu anderen Objekten befinden könnte. Vielmehr geht es mir in diesem Fall nur darum, dass der Apfel eine bestimmte Eigenschaft hat – nämlich die Eigenschaft, sich in einer bestimmten räumlichen *Relation-zu-mir* zu befinden. Da das zweite Relationsglied – ich – in all diesen Fällen gleich bleibt, kann ich es sozusagen festhalten und mich auf die Frage konzentrieren, welche Dinge in meiner Umwelt haben die Eigenschaft, in dieser Relation-zu-mir zu stehen. Es ist so, als würde ich die Situation, dass sich 50 cm vor mir leicht nach rechts versetzt ein Apfel befindet, in Polarkoordinaten $r$ und $\alpha$ repräsentieren:

(IST-EIN OBJEKT-A APFEL)
(ORT OBJEKT-A (0.5 8°))

Auch in einer solchen Repräsentation wird ganz offensichtlich nicht auf den Ursprung des Koordinatensystems – auf mich – Bezug genommen. Diese Bezugnahme ist, wie gesagt, verzichtbar. Perry selbst drückt das so aus:

> The general point is this. Sometimes all of the facts we deal with involving a certain *n*-ary relation involve the same object occupying one of the argument roles. In that case, we don't need to worry about that argument role; we don't need to keep track of its occupant, because it never changes. (Perry 1998, 87)

Akteurzentrierte Repräsentationen stellen also eine Form von Selbstkenntnis dar; sie beinhalten Wissen darüber, wie sich die Dinge um den Akteur herum zu ihm selbst verhalten. Insofern beinhalten sie immer auch Wissen über den Akteur selbst. Allerdings setzen diese Repräsentationen doch noch nicht voraus, dass der Akteur einen *Begriff von sich* hat. Denn sie müssen keinen Ausdruck enthalten, der sich explizit auf den Akteur bezieht. Dies gilt im Übrigen auch für alle Repräsentationen, die sich auf Eigenzustände des Systems beziehen. Empfindungen etwa können auf die folgende Weise repräsentiert werden:

Schmerzen im linken Knie.
Leichtes Kribbeln in der Magengegend.
Jucken auf dem Kopf.

Auch hier ist der explizite Bezug auf das System verzichtbar. Denn in dieser Form wird ein System Empfindungen niemals anderen Wesen, sondern immer nur sich selbst zuschreiben. Also muss es selbst nicht eigens erwähnt werden.

Erst Repräsentationen, die Selbstkenntnis im strengen Sinne (*self-attached knowledge* im Sinne Perrys) darstellen, erfordern, dass das System einen *Begriff von sich* entwickelt. Was sind das für Repräsentationen und wie können sie entstehen?

## 3.

Kommen wir noch einmal auf den Grundsatz zurück: Kognitive Systeme sind Systeme, die versuchen, sich ein Bild von der Welt zu machen, in der sie leben, d. h. Systeme, die versuchen, ihre Umwelt zu repräsentieren. Wie geht das vor sich? Was ist dafür erforderlich, dass es einem kognitiven System – nennen wir es ‚AL' – gelingt, seine Umwelt zu repräsentieren? Welche Prozesse sind hier beteiligt?

Offenbar ist es zunächst so, dass die Umwelt in AL kausale Spuren hinterlässt und dass AL das Problem lösen muss, aus diesen Spuren seine Umwelt zu rekonstruieren. Genauer heißt das, dass es AL durch eine Auswertung der kausalen Spuren gelingen muss, die folgenden Fragen zu beantworten:

1. Wie viele *Objekte* gibt es in der aktuellen Szene?
2. Zu welcher *Art* von Dingen gehören diese Objekte?
3. *Wo* befinden sich diese Objekte?
4. Welche *Eigenschaften* haben sie und in welchen *Relationen* stehen sie zueinander?

Wenn AL diese Fragen beantwortet hat, kann er auf die folgende Weise Repräsentationen seiner Umwelt aufbauen.

- Nach der Beantwortung der ersten Frage gibt AL jedem der an der Szene beteiligten Objekte einen internen Namen (etwa ‚Objekt-36', ‚Objekt-37', usw.).
- Nach Beantwortung der zweiten Frage fügt AL der Liste seiner Repräsentationen für jedes Objekt eine Repräsentation der Form

  (IST-EIN OBJEKT-*X* *TYP*)

  hinzu.
- Nach Beantwortung der dritten Frage fügt AL der Liste seiner Repräsentationen für jedes Objekt eine Repräsentation der Form

  (ORT OBJEKT-*X* *KOORDINATEN*)

  hinzu.
- Nach der Beantwortung der vierten Frage schließlich fügt AL der Liste seiner Repräsentationen für jedes Objekt geeignete Repräsentationen der Form

  (FARBE OBJEKT-*X* *FARBE*)
  (GRÖßE OBJEKT-*X* *GRÖßE*)
  (*RELATION* OBJEKT-*X* OBJEKT-*Y*)
  usw.

  hinzu.

Soweit die Grundidee. Aber das alleine reicht häufig noch nicht aus. Viele kognitive Systeme müssen darüber hinaus das Problem lösen, Objekte *wiederzuerkennen*, denen sie früher schon einmal begegnet sind. Wenn sich ein anderer Storch seinem Nest nähert, muss der Storch im Nest nicht nur entscheiden: Ist das ein weiblicher oder ein männlicher Storch? Er muss auch eine Antwort auf die Frage finden: Ist das meine Partnerin oder ein fremder Storch? Wenn AL bei der Szenenanalyse jedem Objekt, das er wahrnimmt, einen *neuen* internen Namen gäbe, wäre das aber so, als wäre ihm dieses Objekt noch nie begegnet, als wäre dieses Objekt völlig neu für ihn. Doch das würde im Zweifelsfall zu ganz und gar unangemessenen Re-

aktionen führen. Also muss AL auch die Aufgabe lösen, – etwa durch Vergleich von typischen Merkmalen – herauszufinden, welche der in der analysierten Szene vorgefundenen Objekte mit Dingen identisch sind, denen er schon einmal begegnet ist und denen er daher schon früher einen internen Namen gegeben hat. Für diese Objekte muss AL in allen neuen Repräsentationen den schon vorhandenen Namen verwenden.[6]

Für uns ist aber folgendes von besonderer Bedeutung: Bis jetzt gab es für AL noch keinerlei Grund, in den gespeicherten Repräsentationen einen Namen für sich selbst zu verwenden. Es kann durchaus sein, dass AL den Ort, an dem sich die von ihm wahrgenommenen Dinge befinden, in akteurzentrierten Koordinaten repräsentiert und dass AL repräsentiert, in welchen Relationen-zu-ihm sich diese Dinge befinden. Doch all dies erfordert, wie wir schon gesehen haben, keineswegs, dass die entsprechenden Repräsentationen einen Ausdruck enthalten, der sich explizit auf AL bezieht. Daher ist die Frage jetzt: Was muss passieren, damit es für AL sozusagen unvermeidlich wird, einen solchen Ausdruck zu verwenden? Diese Frage lässt sich, wie mir scheint, am besten beantworten, wenn wir ALs kognitive Geschichte noch ein wenig weiter erzählen.

Der nächste wichtige Schritt beruht auf der Beobachtung, dass in der Umgebung kognitiver Wesen häufig andere kognitive Wesen vorkommen. Und auch von AL wollen wir annehmen, dass es in seiner Umgebung andere kognitive Systeme gibt, denen er von Zeit zu Zeit begegnet. Das Problem mit diesen kognitiven Wesen ist, dass ihr Verhalten nicht nur von ihren ‚natürlichen' Eigenschaften abhängt, sondern ganz entscheidend auch davon, welches Bild *sie* sich von der Umwelt machen, wie *sie* die Welt repräsentieren. Wenn AL das Verhalten seiner Mitwesen richtig voraussagen möchte, bleibt ihm daher gar nicht anderes übrig als Repräsentationen zu entwickeln, die die Repräsentationen seiner Kumpane zum Gegenstand haben – Repräsentationen, die man gemeinhin ‚Metarepräsentationen' nennt. Diese Repräsentationen haben generell die Form:

(GLAUBT OBJEKT-*X* *REPRÄSENTATION*)

oder

(WÜNSCHT OBJEKT-*X* *REPRÄSENTATION*)

usw.

---

[6] Perry geht in entsprechenden Überlegungen nicht davon aus, dass Objekte interne Namen erhalten, sondern dass für jeden Gegenstand ein Ordner angelegt wird, in dem alle Informationen über diesen Gegenstand gesammelt werden. Für ihn stellt sich das Problem daher so dar: Muss AL für einen Gegenstand, den er bei der Analyse einer Szene entdeckt, einen neuen Ordner anlegen oder kann er die neu gewonnenen Informationen über diesen Gegenstand in einem schon vorhandenen Ordner ablegen? (Cf. Perry 1998, 89 ff.)

Es ist eine interessante Frage, wie die Repräsentationen, die für die Variable „*Repräsentation*“ eingesetzt werden können, im Einzelnen aussehen. Manchmal wird AL zu der Überzeugung kommen, dass eines seiner Mitwesen – etwa das Wesen mit dem internen Namen ‚Objekt-111‘ – eine Überzeugung über einen Gegenstand hat, der AL selbst bekannt ist und für den er intern den Namen ‚Objekt-7‘ verwendet. Dann würde die entsprechende Metarepräsentation so aussehen:

(GLAUBT OBJEKT-111 (FARBE OBJEKT-7 GRÜN)).

Mit dieser Metarepräsentation würde AL seinem Mitwesen sozusagen eine *de re* Überzeugung zuschreiben; denn diese Repräsentation hätte den Inhalt: Objekt-111 glaubt *von dem Gegenstand*, den AL unter dem Namen ‚Objekt-7‘ kennt, dass er grün ist.

Es kann aber auch vorkommen, dass ALs Mitwesen auf etwas starrt, das AL nicht sehen kann, und dass das Mitwesen sich dabei so verhält, wie es sich sonst nur verhält, wenn es eine Spinne sieht. Wie wird AL diese Situation repräsentieren? Nun, erstens sieht AL das Ding, das sein Mitwesen offenbar vor sich hat, nicht selbst. Also muss AL einen *neuen* internen Namen verwenden – sagen wir ‚Objekt-57‘. Zweitens weiß AL von dem durch diesen neuen Namen bezeichneten Gegenstand nichts – außer, dass er sich direkt vor seinem Mitwesen befindet und dass sein Mitwesen offenbar glaubt, dass dieser Gegenstand eine Spinne ist. Also wird AL seinem Repräsentationssystem die folgenden beiden Listen hinzufügen:

(VOR OBJEKT-57 OBJEKT-111)
(GLAUBT OBJEKT-111 (IST-EIN OBJEKT-57 SPINNE)).

Es kann natürlich auch noch sein, dass sich AL die Überzeugung seines Mitwesens zu eigen macht und daher zusätzlich die Repräsentation ausbildet

(IST-EIN OBJEKT-57 SPINNE).

Aber das steht hier nicht zur Debatte. Wichtig ist nur, dass AL, wenn er Überzeugungen seiner Kumpane repräsentieren will, die sich auf Dinge beziehen, die er nicht kennt, *neue* interne Namen verwenden muss.

Und wichtig ist darüber hinaus, dass AL gut daran tut, auch die Empfindungen seiner Mitwesen zu repräsentieren – etwa so:

(SCHMERZEN-IM-KNIE OBJEKT-111).

Denn auch Schmerzen, Freude und andere Empfindungen sind ja verhaltensrelevant. Allerdings: Der wichtigste Schritt kommt erst noch.

Irgendwann nämlich wird AL darauf kommen, dass er für die anderen kognitiven Wesen nichts anderes ist als sie für ihn. Diese anderen Wesen repräsentieren ihre Umwelt nämlich so, dass in ihr ein Wesen vorkommt, das seinerseits die Welt repräsentiert und das *de facto* niemand anderes ist

als – AL. Irgendwann wird AL etwa merken, dass eines seiner Mitwesen IHN anstarrt oder sich IHM nähert oder etwas von IHM will, z. B. Essen. Und: Irgendwann wird es nicht mehr ausreichen, solche Situationen auf akteurzentrierte Weise zu repräsentieren. Mit anderen Worten: Um repräsentieren zu können, welche Repräsentationen seine Mitwesen über ihn, AL, haben, muss AL einen internen Namen – sagen wir ‚Objekt-100' – für sich selbst verwenden. Erst mit Hilfe dieses Namens kann er nämlich den Wunsch seines Mitwesens angemessen repräsentieren:

(WÜNSCHT OBJEKT-111 (GIBT-ESSEN OBJEKT-100 OBJEKT-111)).

Ein ähnlicher Fall würde eintreten, wenn AL merkt, dass das Wesen mit dem Namen ‚Objekt-111' offenbar von ihm selbst, AL, glaubt, er habe Schmerzen im Knie; dies müsste AL nämlich so repräsentieren:

(GLAUBT OBJEKT-111 (SCHMERZEN-IM-KNIE OBJEKT-100)).

Offenbar gibt es keine Möglichkeit, die Überzeugung seines Mitwesens auf akteurzentrierte Weise zu repräsentieren. AL benötigt also einen internen Namen für sich selbst – einen Namen, der zunächst allerdings nur ein Name ist wie jeder andere. Irgendwann wird es AL aber wohl dämmern, dass der Ausdruck ‚Objekt-100' ein Name für ihn selbst, AL, ist. Sicher, das ist nur eine metaphorische Ausdruckweise. Denn wie kann AL merken, dass ‚Objekt-100' ein Name für ihn selbst ist, wenn er noch gar keinen Begriff von sich selbst hat? Wir müssen also herausfinden, was mit dieser Metapher gemeint sein kann, was es heißen kann, dass AL merkt, dass ‚Objekt-100' ein Name für ihn selbst ist – allgemeiner: was es heißen kann, dass AL einen Begriff von sich ausbildet. Drei Schritte sind hier von besonderer Bedeutung:

1. AL beginnt, in seinen Repräsentationen einen neuen Namen zu verwenden, wenn es darum geht, Repräsentationen anderer kognitiver Wesen zu repräsentieren, die AL selbst zum Gegenstand haben.
2. AL beginnt damit, diesen Namen auch zu verwenden, wenn er sich z. B. im Spiegel sieht.
3. In ALs kognitiver Architektur entwickelt sich eine systematische Verbindung zwischen Repräsentationen, die diesen neuen Namen enthalten, und den schon früher erzeugten akteurzentrierten Repräsentationen, in denen nur implizit auf AL Bezug genommen wird.

Dieser letzte Punkt ist entscheidend. Denn es ist wirklich von allergrößter Bedeutung, wenn in ALs kognitiver Architektur die Repräsentation

(SITZE-AUF OBJEKT-3)

anfängt, dieselbe Rolle zu spielen wie die Repräsentation

(SITZT-AUF OBJEKT-100 OBJEKT-3),

und wenn die Repräsentation

(SCHMERZEN-IM-KNIE)

beginnt, dieselbe Rolle zu spielen wie die Repräsentation

(SCHMERZEN-IM-KNIE OBJEKT-100).

Das Ergebnis dieses Prozesses ist, dass sich in ALs kognitiver Architektur eine Äquivalenz entwickelt zwischen älteren akteurzentrierten Repräsentationen und Repräsentationen, die sich mit Hilfe des neuen Namens ‚Objekt-100' explizit auf AL beziehen. Und das bedeutet auch, dass von diesem Zeitpunkt an der gesamte Input aus ALs Körper – d. h., alles was ihm über Propriozeption zugänglich ist – nicht nur zu akteurzentrierten Repräsentationen, sondern auch zu Repräsentationen führt, die sich explizit auf AL beziehen. Mit anderen Worten: AL entwickelt ein Körperschema.

Früher repräsentierte AL seine Umgebung anhand der Frage: Ich welcher Relation-zu-mir steht das Objekt *x*? Und der Grund dafür war, dass das zweite Relationsglied – AL selbst – in allen Fällen gleich blieb. Jetzt macht er diese Fixierung rückgängig, weil sich herausgestellt hat, dass er nur eines von vielen Dingen ist, die in derselben Relation zu *x* stehen können. Metaphorisch: AL beginnt sich mit den Augen der anderen zu sehen.

Dies hat noch eine weitere Wirkung: AL beginnt, Metarepräsentationen mit Bezug auf sich selbst zu entwickeln. Früher war es für ihn einfach nicht nötig zu wissen, was er selbst glaubt und wünscht. Jetzt jedoch, wo er beginnt, sich mit den Augen der anderen zu sehen, wird alles anders. Denn das Verhalten der anderen hängt auch davon, was sie darüber denken, wie er, AL, die Welt repräsentiert. Also muss er anfangen, sich um seine eigenen intentionalen Zustände zu kümmern. Und das führt schließlich dazu, dass AL beginnt, die für ein Selbst notwendige Selbstkenntnis zu entwickeln. Denn die hatte Lowe ja so charakterisiert:

> [By] ‚reflexive self-knowledge' I mean, roughly speaking, knowledge of one's own identity and conscious mental states – knowledge of who one is and of what one is thinking and feeling. (Lowe 2000, 264 f.)

## 4.

Aber, so könnte man einwenden, der Name ‚Objekt-100' ist doch ein interner Name wie jeder andere. Wie kann es sein, dass Repräsentationen, in denen dieser Name vorkommt, einen so speziellen Status haben, dass sich gerade aus diesen Repräsentationen die Art von Selbstkenntnis ergibt, die dafür sorgt, dass AL über Selbstbewusstsein verfügt oder – in Lowes Redeweise – dass AL ein Selbst ist?

Nun, zunächst ist sicher vorstellbar, dass AL zusammen mit seinen Kumpanen beginnt, eine Sprache zu entwickeln, in der auch indexikalische Ausdrücke wie ‚ich', ‚du', ‚dort' usw. vorkommen. (Vielleicht ist das sogar notwendig dafür, dass AL die entsprechenden Repräsentationen und Metarepräsentationen entwickelt.) Weiter ist auch vorstellbar, dass AL lernt, nur solche Repräsentationen, in denen der Name ‚Objekt-100' vorkommt, mit ‚ich'-Sätzen auszudrücken. Aber wäre das nicht ein bloßer Zufall? Ist nicht ebenso gut vorstellbar, dass AL sich angewöhnt, nur Repräsentationen, in denen der Name ‚Objekt-13' vorkommt, mit ‚ich'-Sätzen auszudrücken, wobei ‚Objekt-13' ein Name für einen beliebigen anderen Gegenstand ist? Die Antwort lautet: Nein. Den Gebrauch des Wortes ‚ich' zu lernen heißt unter anderem zu lernen, dass sich jedes Mitglied der Sprachgemeinschaft mit diesem Wort nur auf sich selbst beziehen kann. AL hat die Bedeutung von ‚ich' also nur dann gelernt, wenn er gelernt hat, mit Hilfe dieses Wortes nur Repräsentationen auszudrücken, die sich auf ihn selbst beziehen. Doch diese Antwort reicht noch nicht aus. Denn an dieser Stelle spielt Perrys Unterscheidung zwischen *self-attached knowledge* und *knowledge of the person one happens to be* eine entscheidende Rolle. Erinnern wir uns kurz an eines der einschlägigen Beispiele.

Gleich am Beginn seines Aufsatzes „The Problem of the Essential Indexical" schreibt Perry:

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. … Finally it dawned on me. I was the shopper I wanted to catch. (Perry 1979, 33)

Die Pointe dieser Geschichte ist klar. Zunächst glaubte Perry, dass jemand anderes – eine Person, über die er sonst nichts wusste – mit einem Einkaufswagen durch den Supermarkt fuhr, in dem sich eine beschädigte Packung Zucker befand. Doch nach einer gewissen Zeit bemerkte er, dass diese Person niemand anderes war als er selbst. Und dies brachte ihn zu der Überzeugung, dass sich die beschädigte Packung Zucker in seinem eigenen Einkaufswagen befinden musste. Wenn man diese beiden Überzeugungen im Sinne Russells analysiert, scheinen sie denselben Inhalt zu besitzen – die singuläre Proposition ‚hat eine beschädigte Packung Zucker im Einkaufswagen, Perry'. Doch die beiden Überzeugungen müssen verschieden sein; denn ihre Verhaltenskonsequenzen unterscheiden sich ganz erheblich. Die erste Überzeugung veranlasste Perry, die Person mit der beschädigten Packung Zucker im Einkaufswagen zu suchen, um sie darauf hinzuweisen, welche Unordnung sie verursache. Die zweite Überzeugung führte dagegen zu einem ganz anderen Verhalten. „I stopped following the trail around the counter and rearranged the torn sack in my cart." (Perry 1972, 33) Da beide

Überzeugungen denselben Russellschen Inhalt haben, können sie sich aber nur in ihren Fregeschen Inhalten – in verschiedenen Arten des Gegebenseins – unterscheiden. Und: Offenbar haben wir es bei der zweiten Überzeugung mit einer ganz besonderen Art zu tun, in der Perry sich selbst ‚gegeben' ist – einer Art, die man als ‚EGO-Art des Gegebenseins' bezeichnen könnte. Die Frage, die wir uns stellen müssen, lautet daher: Was hat der interne Name ‚Objekt-100' mit dieser EGO-Art des Gegebenseins zu tun?

Zunächst gilt es festzuhalten, dass in ALs Repräsentationen verschiedenen Arten des Gegebenseins normalerweise durch verschiedene interne Namen Rechung getragen wird. Angenommen, AL sieht in einem Spiegel, dass hinter einem kognitiven Wesen ein Bär auftaucht, wobei er jedoch nicht erkennt, dass es sich bei diesem kognitiven Wesen *de facto* um ihn selbst handelt. Bei der Repräsentation dieser Situation muss AL für das beobachtete kognitive Wesen daher einen neuen internen Namen, sagen wir den Namen ‚Objekt-213', einführen. (Dasselbe gilt natürlich für den Bären, es sei denn, AL wäre genau diesem Bären früher schon einmal begegnet.) Auf diese Weise wird AL z. B. Repräsentationen wie diese entwickeln:

(IST-EIN OBJEKT-511 BÄR)
(IST-EIN OBJEKT-213 KOGNITIVES SYSTEM)
(HINTER OBJEKT-511 OBJEKT-213)

Den Namen ‚Objekt-100' würde AL nur verwenden, wenn er bemerken würde, dass sich der Bär tatsächlich hinter seinem *eigenen* Rücken befindet.

Zweitens, und das ist noch wichtiger: Verschiedenen Arten des Gegebenseins entsprechen verschiedene Arten der kognitiven Verarbeitung, d. h. verschiedene funktional/computationale Rollen. Wenn zwei interne Ausdrücke für AL kognitiv äquivalent sind, entsprechen sie also derselben Art des Gegebenseins. Und umgekehrt: Zwei Ausdrücke mit unterschiedlichen computationalen Rollen entsprechen verschiedenen Arten des Gegebenseins. Die Frage ist also: Welche spezifischen Merkmale der computationalen Rolle des Namens ‚Objekt-100' sind dafür verantwortlich, dass dieser Name einer EGO-Art des Gegebenseins entspricht?

Nun, die Art und Weise, in der der Name ‚Objekt-100' in ALs kognitiver Architektur verarbeitet wird, unterscheidet sich in der Tat ganz wesentlich von der Verarbeitung aller anderen Namen. Schließlich haben wir angenommen, dass Repräsentationen, in denen der Name ‚Objekt-100' vorkommt, die einzigen Repräsentationen sind, die zu akteurzentrierten Repräsentationen äquivalent sind – wobei zwei Repräsentationen genau dann ‚äquivalent' heißen sollen, wenn sie dieselbe funktionale/computationale Rolle besitzen. Das hat zwei sehr wichtige Konsequenzen. Erstens: Der gesamte propriozeptive Input führt in ALs kognitiver Architektur nur zu akteurzentrierten Repräsentationen und – wegen der angenommenen Äquiva-

lenz – zu Repräsentationen mit dem Namen ‚Objekt-100'. Selbst wenn ‚Objekt-213' ein Name ist, der sich – ohne dass AL das weiß – tatsächlich ebenfalls auf AL bezieht, würde dieser Name in ALs kognitiver Architektur niemals dazu verwendet, Informationen über ALs Kopfschmerzen oder die Position seiner Glieder zu speichern. Jedenfalls dann nicht, wenn diese Informationen nicht von den äußeren Sinnen, sondern aus dem propriozeptiven System ALs stammen. Die besondere Art und Weise, in der AL viele seiner eigenen Körperzustände gegeben sind, ist also nur mit dem Namen ‚Objekt-100' und mit keinem anderen Namen verknüpft.

Zweitens, und mindestens ebenso wichtig: Akteurzentrierte Repräsentationen haben eine charakteristische unmittelbare Handlungswirksamkeit. Denken wir noch einmal an die Repräsentation mit dem Inhalt: Da fliegt ein Ball direkt auf mich zu. Diese Repräsentation wird mich sofort zum Handeln veranlassen. Ich werde mich ducken oder versuchen, den Ball zu fangen, oder was auch immer. Ganz anders bei einer Repräsentation mit dem Inhalt: Da bewegt sich ein Ball mit der Geschwindigkeit *v* von Ort *a* zu Ort *b*. Diese Repräsentation bewirkt alleine wahrscheinlich gar nichts. Erst wenn außerdem klar ist, dass ich selbst mich zwischen *a* und *b* befinde, wird mich das wahrscheinlich zum Handeln veranlassen. Und: Wie wir gesehen haben, besitzen akteurzentrierte Repräsentationen darüber hinaus ein weiteres charakteristisches Merkmal. Wenn ein kognitives System die Gegenstände in seiner Umwelt mit Hilfe der Relationen repräsentiert, in denen diese Gegenstände zu ihm *selbst* stehen, gibt es für eine Vielzahl von Handlungen klar definierte Typen von Körperbewegungen, mit deren Hilfe diese Handlungen ausgeführt werden können. Wenn hinter meinem Rücken ein Bär auftaucht, muss ich mich umdrehen und – sofern ich ein geeignetes Messer habe – versuchen zuzustechen. Wegen der Äquivalenz zwischen akteurzentrierten und Repräsentationen mit dem Namen ‚Objekt-100' erben viele Repräsentationen mit diesem Namen diese beide Merkmale akteurzentrierter Repräsentationen. Mit anderen Worten: Repräsentationen mit dem Namen ‚Objekt-100' haben – zumindest häufig – dieselbe unmittelbare Handlungswirksamkeit wie die entsprechenden akteurzentrierten Repräsentationen. Und sie sind in derselben unmittelbaren Weise mit bestimmten Typen von Körperbewegungen verbunden, durch die AL bestimmte Handlungen ausführen kann.

Dies ist entscheidend, da für *de se* Einstellungen ebenfalls charakteristisch ist, dass sie eine spezifische kognitive Rolle und eine spezifische kausale Rolle im Hinblick auf die Handlungen einer Person spielen. Zumindest in vielen Fällen führen sie zu typischen egozentrischen Reaktionen dieser Person. Nehmen wir noch einmal Perrys Zuckerbeispiel. Bevor er bemerkte, dass er selbst den ganzen Schlammassel verursacht hatte, brachte ihn seine Überzeugung dazu, die Person mit der beschädigten Packung Zucker

im Einkaufswagen zu suchen, um sie darauf hinzuweisen, welche Unordnung sie verursache. Danach änderte sich sein Verhalten auf typische Weise. Perry hörte auf, nach der anderen Person zu suchen und versuchte stattdessen, die beschädigten Packung Zucker in *seinem eigenen* Einkaufswagen in Ordnung zu bringen. Oder nehmen wir den Fall Ernst Machs, über den dieser selbst schreibt:

> Ich stieg einmal nach einer anstrengenden nächtlichen Eisenbahnfahrt sehr ermüdet in einen Omnibus, eben als von der anderen Seite auch ein Mann hereinkam. „Was steigt doch da für ein herabgekommener Schulmeister ein", dachte ich. (Mach 1911, 3 Fn. 1)

Dieser Gedanke wird bei Mach zu kaum mehr geführt haben als vielleicht zu einer gewissen Geringschätzung oder im besten Fall zu Mitleid. Doch plötzlich stellte Mach fest, dass der Mann am anderen Ende des Busses niemand anderes war als er selbst.

> Ich war es selbst, denn mir gegenüber befand sich ein großer Spiegel. Der Klassenhabitus war mir also viel geläufiger als mein Specialhabitus. (ebd.)

Diese Erkenntnis änderte die Dinge erheblich. Denn als er bemerkte, dass der Mann am anderen Ende des Busses niemand anderes als er selbst war, wird Mach wohl gedacht haben: *Ich* sehe aus wie herabgekommener Schulmeister. Und dieser Gedanke hatte ganz andere Konsequenzen als der erste – vielleicht begann Mach sich zu schämen oder zu versuchen, seine Kleider ein bisschen zu säubern.

Die Antwort auf die Frage, welche spezifischen Merkmale der computationalen Rolle des Namens ‚Objekt-100' dafür verantwortlich sind, dass dieser Name eine EGO-Art des Gegebenseins verkörpert, lautet also: Repräsentationen, die diesen Namen enthalten, spielen genau die kausale Rolle, die für *de se* Einstellungen charakteristisch ist – für Einstellungen, deren Inhalt angemessen nur unter Verwendung des Wortes ‚ich' wiedergegeben werden kann.

## 5.

Kognitive Systeme sind Wesen, die versuchen, Wissen über ihre Umwelt zu erwerben, um in dieser Umwelt besser zurechtzukommen. Zum effektiven Handeln benötigen sie vielfach jedoch nicht nur Wissen über ihre Umgebung, sondern auch Wissen über sich selbst. Jedes kognitive System muss wissen, ob *es* bedroht ist, wo *es* sich befindet, welche *seiner* Glieder einsatzfähig sind, was *es* braucht (Wasser, Energie, Ruhe), usw. In den meisten Fällen reicht es allerdings aus, diese Art von Selbstkenntnis in akteurzentrierten Repräsentationen zu speichern – Repräsentationen, in denen es zwar auch um das jeweilige kognitive System geht, in denen jedoch

nicht explizit auf dieses System Bezug genommen wird. Repräsentationen mit expliziter Selbstbezugnahme werden für ein kognitives System erst unabdingbar, wenn es beginnt, Objekte in seiner Umwelt als Wesen zu repräsentieren, die ihrerseits kognitive Systeme sind, d. h. die ihrerseits über Repräsentationen ihrer Umwelt verfügen. Denn in den Repräsentationen seiner Mitwesen kommt das kognitive System selbst als Objekt vor – als etwas, von dem die Repräsentationen der Mitwesen handeln. Explizite Selbstkenntnis erwächst also aus der Erkenntnis eines kognitiven Systems, dass die anderen es genau so repräsentieren, wie es selbst die anderen repräsentiert. Spätestens wenn es repräsentieren will, welche Repräsentationen andere Wesen über es selbst haben, braucht ein kognitives System einen internen Namen für sich selbst; und damit wird es sich selbst zum möglichen Objekt der Repräsentation. Der letzte Schritt zu genuiner Selbstkenntnis besteht schließlich darin, dass das System beginnt, einen Zusammenhang zwischen den akteurzentrierten Repräsentationen, in denen es nur implizit um es selbst geht, und den neuen expliziten Selbstrepräsentationen herzustellen. Auf diese Weise erhalten diese neuen Repräsentationen genau die Rolle, die genuine *de se* Einstellungen charakterisiert.

## Literatur

Beckermann, A. (2003) „Self-Consciousness in Cognitive Systems“. In: Ch. Kanzian, J. Quitterer & E. Runggaldier (Hg.) *Persons. An Interdisciplinary Approach.* Wien: öbv&hpt, 174–188

Descartes, R. *Meditationes de prima philosophia*. *OEuvres des Descartes VII.* Publiées par C. Adam et P. Tannery. Nouvelle Présentation. Paris, J. Vrin, 1964–1976.

Descartes, R. (1988) *Selected Philosophical Writings*. Transl. by J. Cottingham, R. Stoothoff & D. Murdoch. Cambridge: Cambridge University Press.

Descartes, R. (1911) *The Philosophical Works of Descartes*. *Vol. 1.*. Transl. by E. S. Haldane & G. R. T. Ross. Cambridge: Cambridge University Press.

Locke, J. *An Essay Concerning Human Understanding*. Ed. by P. H. Nidditch. Oxford: Clarendon Press 1975.

Lowe, E. J. (2000) *An Introduction to the Philosophy of Mind*. Cambridge: Cambridge University Press.

Mach, E. (1911) *Die Analyse der Empfindungen und das Verhältnis des Physischen zum Psychischen*, 6. Aufl. (1. Aufl. 1885), Jena.

Perry, J. (1979) „The Problem of the Essential Indexical“. In: J. Perry, *The Problem of the Essential Indexical and Other Essays*. Oxford: Oxford University Press 1993, 33–52.

Perry, J. (1998) „Myself and I“. In: M. Stamm (Hg.) *Philosophie in synthetischer Absicht (Festschrift für Dieter Henrich)*. Stuttgart: Klett-Cotta, 83–103.

Rosenberg, J. F. (1986) *The Thinking Self*. Philadephia: Temple University Press.

# 14

# Es gibt kein Ich, doch es gibt mich[*1]

## *1. Gehirn und Ich*

Siegmund Freud hat von drei „Kränkungen ihrer naiven Eigenliebe“ gesprochen, die „die Menschheit im Laufe der Zeiten von der Wissenschaft [hat] erdulden müssen.“

> Die erste, als sie erfuhr, daß unsere Erde nicht der Mittelpunkt des Weltalles ist, sondern ein winziges Teilchen eines in seiner Größe kaum vorstellbaren Weltsystems. Sie knüpft sich für uns an den Namen Kopernikus, obwohl schon die alexandrinische Wissenschaft ähnliches verkündet hatte. Die zweite dann, als die biologische Forschung das angebliche Schöpfungsvorrecht des Menschen zunichte machte, ihn auf die Abstammung aus dem Tierreich und die Unvertilgbarkeit seiner animalischen Natur verwies. Diese Umwertung hat sich in unseren Tagen unter dem Einfluss von Ch. Darwin, Wallace und ihren Vorgängern nicht ohne das heftigste Sträuben der Zeitgenossen vollzogen. Die dritte und empfindlichste Kränkung aber soll die menschliche Größensucht durch die heutige psychologische Forschung erfahren, welche dem Ich nachweisen will, daß es nicht einmal Herr ist im eigenen Hause, sondern auf kärgliche Nachrichten angewiesen bleibt von dem, was unbewusst in seinem Seelenleben vorgeht. (Freud 1917, 294 f.)

Heute ist oft von einer weiteren grundlegenden, vermutlich letzten Kränkung des Selbstwertgefühls des Menschen die Rede. Denn Neurobiologie und Neurophilosophie hätten, so etwa Siefer und Weber (2006, 252), gezeigt, dass das Ich nicht nur nicht Herr im eigenen Haus ist, dass es vielmehr gar kein Ich gibt. Das Ich sei nichts als eine Illusion. Auf welche Befunde gründet sich diese These? Halten wir uns an einen Gewährsmann von Siefer und Weber – Gerhard Roth. Roth schreibt in *Fühlen, Denken, Handeln*:

> [Die] erlebte Welt wird von unserem Hirn in mühevoller Arbeit über viele Jahre hindurch konstruiert und besteht aus den Wahrnehmungen, Gedanken, Vorstellungen, Erinnerungen, Gefühlen, Wünschen und Plänen, die unser Gehirn hat. Innerhalb dieser Welt bildet sich […] langsam ein Ich aus, das sich zunehmend als vermeintliches Zentrum der Wirklichkeit erfährt, indem es den Eindruck entwickelt, es „habe“ Wahrnehmungen (d. h. dass Wahrnehmungen

---

[*] Erstveröffentlichung in: M. Fürst, W. Gombocz & C. Hiebaum (Hg.) *Gehirne und Personen*. Frankfurt/M.: ontos Verlag 2009, 1–17.

[1] Eine ausführlichere Version dieses Aufsatzes findet sich in Beckermann (2008, Kap. 2).

> auf es bezogen sind), es sei Autor der eigenen Gedanken und Vorstellungen, es rufe aktiv die Erinnerungen auf, es bewege den Arm, die Lippen, es besitze diesen bestimmten Körper, und so fort. Selbstverständlich ist dies eine Illusion, denn Wahrnehmungen, Gefühle, Intentionen und motorische Akte entstehen innerhalb der Individualentwicklung lange bevor das Ich entsteht. (Roth 2003, 395 f.)

Da ist sie also – die These, das Ich sei nichts als eine Illusion. Oder? Wenn man genau hinschaut, sieht man, dass Roth gar nicht behauptet, dass das Ich eine Illusion ist. Vielmehr vertritt er zwei verwandte Thesen: a) Das Ich selbst bildet sich erst langsam in der erlebten Welt aus, die von unserem Gehirn konstruiert wird. b) Es ist eine Illusion anzunehmen, dieses Ich „‚habe' Wahrnehmungen […], es sei Autor der eigenen Gedanken und Vorstellungen, es rufe aktiv die Erinnerungen auf, es bewege den Arm, die Lippen, es besitze diesen bestimmten Körper, und so fort".

Aber was besagt die zweite These eigentlich? Auch wenn es kein *Ich* gibt, scheint es doch *mich* zu geben. Also können Sätze wie

- Ich sehe den blauen Himmel
- Ich erinnere mich an meine erste Liebe
- Ich bewege meine Hand
- Ich schreibe jetzt diesen Text

durchaus wahr sein. Aber Roth sagt, es sei falsch anzunehmen, das Ich habe Wahrnehmungen, sei Autor der eigenen Gedanken und Vorstellungen, rufe aktiv die Erinnerungen auf, bewege den Arm, die Lippen, besitze diesen bestimmten Körper etc. Und er scheint das ernst zu meinen. Seiner Meinung nach ist es eigentlich das Gehirn, das Wahrnehmungen, Gedanken, Vorstellungen, Erinnerungen, Gefühle, Wünsche und Pläne hervorbringt und zugleich hat. D. h., er scheint sagen zu wollen, dass die gerade angeführten Sätze alle falsch sind und dass es richtig heißen müsste:

- Mein Gehirn sieht den blauen Himmel
- Mein Gehirn erinnert sich an meine erste Liebe
- Mein Gehirn bewegt meine Hand
- Mein Gehirn schreibt jetzt diesen Text

Dass dies nicht nur äußerst befremdlich klingt, sondern so auch nicht stimmen kann, zeigt sich aber schon an dem Ausdruck „mein Gehirn". Denn wie kann, wenn man Roth folgt, ein Gehirn *mein* Gehirn sein? Und was heißt es, dass es *meine* Hand bewegt?

Wenn Roth tatsächlich sagen will, dass der Satz „Ich bewege meine Hand" immer falsch und dass statt dessen höchstens der Satz „Mein Gehirn bewegt meine Hand" wahr ist, dann ist das nur zu verstehen, wenn man davon ausgeht, dass Roth einerseits den Rahmen des Cartesischen Dualismus

akzeptiert, andererseits aber Descartes' These ablehnt, dass die menschliche Seele auf den Körper kausal einwirkt.

Descartes zufolge besteht jeder Mensch aus einem biologischen Körper und einer immateriellen Seele, die das eigentliche Selbst (Ich) des Menschen ausmacht. Wie stellte sich Descartes das Zusammenwirken von Körper und Seele vor? Der Mensch muss sich in seiner Umwelt orientieren; also muss er seine Umwelt wahrnehmen. Dabei spielt der Körper mit seinen Sinnesorganen und dem Gehirn eine wichtige Rolle; doch das eigentliche Wahrnehmen geschieht in der Seele.

> Wenn wir zum Beispiel ein Tier auf uns zukommen sehen, malt das Licht, das von seinem Körper reflektiert wird, zwei Bilder von ihm, eines in jedem unserer Augen. Diese beiden Bilder bilden davon zwei weitere mittels der optischen Nerven auf der Innenwand des Gehirns ab. Von da aus strahlen diese Bilder durch Vermittlung der Lebensgeister, von denen diese Kammern erfüllt sind, derart gegen die kleine Drüse, welche von Lebensgeistern umgeben ist, daß die Bewegung, die jedem Punkt von einem jeden dieser Bilder darstellt, auf denselben Punkt der Drüse zielt, den die Bewegung, die den Punkt des anderen Bildes wiedergibt, anzielt, und so denselben Teil des Tieres darstellt. Dadurch bilden die beiden Bilder im Hirn nur ein einziges auf der Drüse ab, das unmittelbar auf die Seele einwirkt und sie die Gestalt des Tieres sehen läßt. (Descartes 1984, 59 ff.)

Beim Sehen werden also die durch das vom wahrgenommenen Tier reflektierte Licht hervorgerufenen beiden Netzhautbilder mittels des *nervus opticus* ins Gehirn weitergeleitet; dort werden sie in einem weiteren neuronalen Prozess zu einem einzigen Bild auf der Zirbeldrüse vereint. Dieses Bild wirkt auf die *Seele* und lässt *dort* den Wahrnehmungseindruck eines auf uns zu kommenden Tieres entstehen. Dies ist das eigentliche Sehen. Sehen setzt voraus, dass in der Seele ein Wahrnehmungseindruck entsteht.

Und wie geht es weiter? Wie können wir uns das Verhalten eines Menschen erklären? Bei reflexhaftem Handeln kann man wohl davon ausgehen, dass das Bild auf der Zirbeldrüse im Gehirn selbst unmittelbar bewirkt, dass Lebensgeister über die efferenten Nerven zu bestimmten Muskeln geleitet werden, was seinerseits bewirkt, dass sich unsere Glieder auf eine bestimmte Weise bewegen – dass wir uns z. B. herumdrehen und vor dem Tier weglaufen. Bei überlegtem Handeln ist das nach Descartes anders; das Gehirn kann nicht überlegen, das kann nur die Seele. Die Seele betrachtet also den Wahrnehmungseindruck, versucht die Szene einzuschätzen (Ist das herannahende Tier bedrohlich?), überlegt, was zu tun ist, und kommt schließlich zu einer Entscheidung. Diese Entscheidung mündet in einen immer noch seelischen Willensakt, der nun seinerseits in der Lage ist, die Zirbeldrüse im Gehirn ein bisschen zu drehen. Aufgrund dieser Bewegung der Zirbeldrüse werden wieder Lebensgeister zu bestimmten Muskeln ge-

leitet, was dazu führt, dass sich unsere Glieder auf eine bestimmte Weise bewegen.

Für Descartes gibt es also eine klare Unterscheidung, ja sogar eine Konkurrenz zwischen Gehirn und Seele. Wenn eine Bewegung allein durch neuronale Prozesse hervorgerufen wird, dann hat die Seele mit dieser Bewegung nichts zu tun. Erst wenn Bewegungen auf neuronale Prozesse zurückgehen, die ihrerseits durch seelische Willensakte verursacht sind, kann man sagen, dass die Seele selbst etwas bewirkt hat.

Roths Überlegungen beruhen auf einer völlig analogen Annahme der Konkurrenz zwischen Gehirn und Ich. Alle meine Wahrnehmungen, Gedanken, Vorstellungen, Erinnerungen, Gefühle, Wünsche und Pläne werden durch mein Gehirn hervorgerufen; also können sie nicht auf mein Ich zurückgehen, also kann mein Ich nicht der „Autor" meiner Gedanken und Vorstellungen sein, Erinnerungen aktiv aufrufen, den Arm oder die Lippen bewegen. Im Zusammenhang mit seiner Diskussion des Willensfreiheitsproblems verwendet Roth genau dieselbe Argumentationsfigur. Frei sind in seinen Augen nur Handlungen, die durch immaterielle Willensakte hervorgerufen werden. Besonders die Libet-Experimente zeigen aber, dass alle Handlungen durch Hirnprozesse und nicht durch Willensakte verursacht werden; denn die entsprechenden Willensakte treten immer erst auf, *nachdem* das Hirn schon angefangen hat, die Handlung zu initiieren. Sie kommen also zu spät. Sie verursachen keine Hirnprozesse (und keine Handlungen), sondern sind selbst Wirkungen dieser Hirnprozesse. Dies ist eine epiphänomenalistische Position. Der Epiphänomenalist bestreitet nicht, dass der Mensch eine Cartesische Seele hat; er bestreitet nur, dass das, was in der Seele stattfindet, irgend einen kausalen Einfluss hat auf das, was in der physischen Welt vorgeht.

Doch der Epiphänomenalismus ist genau so verfehlt wie der Cartesische Interaktionismus. Es gibt eine ganze Reihe gewichtiger Gründe, die gegen die Annahme sprechen, dass wir eine immaterielle Seele besitzen, die kausal mit unserem Körper interagieren kann:[2]

- Es gibt auch nicht den kleinsten *empirischen* Hinweis auf das Eingreifen einer immateriellen Seele in die neuronalen Prozesse in unseren Hirnen.
- Es gibt keine befriedigende Antwort auf die Frage, wie materielle Körper und immaterielle Seelen überhaupt kausal aufeinander einwirken können. Was bestimmt den Ort der Einwirkung der Seele auf den Körper? Wie verträgt sich diese Einwirkung mit den Erhaltungssätzen der Physik? Wie kommt eine immaterielle Seele zu der Energie, die sie benötigt, um physische Wirkungen erzielen zu können?

[2] Vgl. zum Folgenden Beckermann (2008, Kap. 1).

- Descartes' Vorstellung von Wahrnehmen, Denken und Handeln sieht im Kern so aus: 1. Der Körper versorgt über die Sinnesorgane die Seele mit Wahrnehmungseindrücken, also mit Informationen über die Außenwelt. 2. Die Seele ordnet diese Eindrücke, macht sich ein Bild von der Umwelt, überlegt und fällt dann eine Entscheidung. 3. Diese Entscheidung wird über die Zirbeldrüse und die Nerven an die Muskeln weitergegeben, die am Ende die entsprechenden Bewegungen ausführen. Alles was zwischen der Aufnahme von Sinneseindrücken und der Ausführung von Bewegungen liegt, ist also Aufgabe der Seele. Warum haben wir dann aber ein so großes Gehirn, das im Wesentlichen ebenfalls damit beschäftigt zu sein scheint, zwischen sensorischem Input und motorischem Output zu vermitteln?
- Dem Cartesianismus zufolge können immaterielle Seelen auch ohne einen Körper existieren. Aber wie hat man sich das Leben solcher reiner Geister vorzustellen, wenn sie sich vom Körper getrennt haben? Was können sie wahrnehmen? Wie kommunizieren sie miteinander und mit uns? Wie kann man verschiedene reine Geister voneinander unterscheiden? Kann sich ein reiner Geist in seiner Identität irren? Kann er unter Amnesie leiden? Alle diese Fragen zeigen, dass die Idee immaterieller Geister inkohärent ist.

Diese Punkte sprechen dafür, dass die Annahme, es gäbe Cartesische Seelen, zu einer Unzahl unlösbarer Rätsel, wahrscheinlich sogar zu Widersprüchen führt. Mir scheint daher, dass wir Darwin und den ihm folgenden Naturwissenschaften in einem zentralen Punkt Recht geben sollten: Auch Menschen sind durch und durch natürliche Wesen. Es wäre schon sehr merkwürdig, sich die Evolution als einen Prozess vorzustellen, bei dem sich nach und nach aus komplizierten Makromolekülen immer komplexere Lebewesen entwickeln, dass aber Menschen erst entstehen, wenn den am höchsten entwickelten Lebewesen zusätzlich eine immaterielle Seele eingehaucht wird. Nichts spricht für diese Annahme. Menschen sind ebenfalls Produkte der Evolution; alles, was sie zu Menschen macht, hat eine rein biologische Grundlage.

Aber was ist, wenn der cartesische Dualismus falsch ist? Folgt dann nicht erst recht, dass unsere Seele, unser Ich niemals etwas bewirken kann und dass Sätze wie „Ich bewege meine Hand" daher immer falsch sind? Sicher würde das folgen, wenn das Ich nichts weiter wäre als eine Cartesianische Seele in anderer Verkleidung – ein immaterieller Personenkern, der, wenn es gut geht, unser Wollen und Tun bestimmt. Aber das Ich ist nichts dergleichen; dieses Ich gibt es in der Tat nicht. Salopp gesagt, es gibt kein Ich; es gibt nur mich. Und wenn es mich gibt, kann ich auch meinen Arm heben, nachdenken und mich erinnern.

Um zu verstehen, wie das möglich ist, müssen wir uns endgültig aus dem Griff der Cartesianischen Bilder befreien. Für einen Cartesianer ist klar: Der Satz „*Ich* bewege meine Hand" ist genau dann wahr, wenn meine Seele, mein Ich – vermittels eines Willensaktes – *kausal bewirkt*, dass sich meine Hand bewegt. Für den Anticartesianer kann das nicht die richtige Antwort sein. Er muss deshalb versuchen, alternative Wahrheitsbedingungen für diesen Satz zu finden. Bevor wir uns dieser Aufgabe zuwenden, sind aber zunächst einige Bemerkungen zur Grammatik und Semantik des Ausdrucks „ich" nötig, die besonders in der philosophischen Diskussion immer wieder missverstanden werden.

## *2. Die Semantik von „ich" und „selbst"*

In der Philosophie der letzten vier Jahrhunderte ist es bedauerlicherweise üblich geworden, die Ausdrücke „Ich" und „Selbst" als Gattungsnamen für bestimmte Arten von Entitäten zu gebrauchen wie etwa „Tier" oder „Zahl". Jeder Mensch hat einen Körper; und er hat, so sagt man, auch ein „Ich" oder ein „Selbst". Das so verstandene Ich oder Selbst ist eine bestimmte Art von „Gegenstand". Und von dieser Art von Gegenstand kann man dann natürlich fragen, ob es ihn gibt oder nicht. Diese Entwicklung beginnt wohl mit René Descartes; aber richtig voran kommt sie erst mit John Locke. Descartes versucht in den *Meditationen* zu beweisen, dass zumindest er selbst existiert, d. h., dass der Satz „Ich bin" notwendigerweise wahr ist, solange er ihn denkt. Nachdem dies erreicht ist, stellt Descartes aber sofort die nächste Frage: Was ist das für ein Ding, dessen Existenz ich da gerade bewiesen habe? Und diese Frage drückt er an mehreren Stellen so aus:

> Ich bin mir aber noch nicht hinreichend klar darüber, wer denn Ich bin – *jener Ich*, der notwendigerweise ist. (Nondum vero satis intelligo, quisnam sim *ego ille*, qui jam necessario sum […].) (Descartes 1986, 78 f. – meine Hervorh.)

> Ich weiß, daß ich bin, und ich frage mich, was *dieser Ich* sei, den ich kenne. (Novi me existere; quaero quis sim *ego ille* quem novi.) (ebd., 84 f. – meine Hervorh.)

„Ego ille" – eine sprachliche Entgleisung mit schrecklichen Folgen.[*] „Jener Ich"? Oder vielleicht sogar „jenes Ich"? Descartes ist meines Wissens der erste, der das Wort „ich" mit einem Demonstrativpronomen verbindet, und das muss in den Ohren seiner Zeitgenossen genau so schief geklungen haben, wie für uns die Verbindung „jener Ich" noch heute klingt. Von da an war es aber nur noch ein kleiner Schritt bis zur Verbindung von „ich" mit einem Artikel – „das Ich" oder „ein Ich". Dass dies in höchstem Maße

---

[*] Diese – leider unzutreffende – Einschätzung habe ich im nächsten Aufsatz (siehe unten S. 292 f.) korrigiert.

sprachwidrig ist, wird geflissentlich übersehen; aber es ist so – man kann ein Personalpronomen weder mit einem Demonstrativpronomen noch mit einem Artikel verbinden.

Was Descartes mit dem Wort „ich" anstellt, tut Locke dem Wort „selbst" an. Im zweiten Buch des *Versuchs über den menschlichen Verstand* schreibt er:

> Self is that conscious thinking thing, (whatever Substance made up of whether Spiritual or Material, Simple or Compounded, it matters not), which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern'd for it *self*, as far as that consciousness extends. (Locke 1975, II, xxvii, 17)[3]

Der linguistische Hintergrund dieser ungewöhnlichen Verwendung von „self" ist offenbar die Tatsache, dass man früher im Englischen Wörter wie „my self" oder „it self" auseinander schreiben konnte, was natürlich die Vermutung zumindest begünstigt, es gäbe da so etwas wie mein Selbst. Dass „self" kein Substantiv ist, hätte Locke aber durchaus bemerken können. Schließlich schreibt er selbst „concern'd for it *self*" und nicht „concern'd for its *self*". Trotzdem, das Englische begünstigt hier ein sprachliches Missverständnis, das z. B. im Lateinischen unmöglich gewesen wäre. Descartes wäre nie auf die Idee gekommen „ipse" mit „ille" zu verbinden. Das wäre nun wirklich zu ungrammatisch gewesen. Und wenn er schreibt „Nunquid *me ipsum* non tantum multo verius, multo certius, sed etiam multo distinctius evidentiusque, cognosco?", dann meint er natürlich „[S]ollte ich nicht *mich selbst* nicht nur viel wahrer und gewisser, sondern auch viel deutlicher und evidenter erkennen?" (Descartes *Meditationen* 94 f. – meine Hervorh.). Und nicht etwa, wie man in einer neueren englischen Übersetzung lesen kann: „Surely my awareness of *my own self* is not merely much truer and more certain but also much more distinct and evident."[4]

Dass „ich" und „selbst" keine Ausdrücke sind, die Dinge einer bestimmten Art bezeichnen, wird sofort deutlich, wenn man sich klar macht, was man alles sagen könnte, wenn sie es wären. „Ich komme heute Abend zur Party; aber ob mein Ich mitkommt, weiß ich nicht." „Natürlich wird die

---

[3] Bemerkenswert die deutsche Übersetzung: „Das *Ich* ist das bewußt denkende Wesen, gleichviel aus welcher Substanz es besteht (ob aus geistiger oder materieller, einfacher oder zusammengesetzter), das für Freude und Schmerz empfindlich und sich seiner bewußt ist, das für Glück und Unglück empfänglich ist und sich deshalb soweit um sich selber kümmert, wie jenes Bewußtsein sich erstreckt." (Hervorh. im Original)

[4] Descartes, René *Selected Philosophical Writings* (86 – meine Hervorh.). In der Ausgabe von Haldane und Ross aus dem Jahre 1911 hieß es noch: „[D]o I not know *myself*, not only with much more truth and certainty, but also with much more distinctness and clearness?" (156 – meine Hervorh.)

Bundeskanzlerin selbst kommen; aber ihr Selbst lässt sie zu Hause.“ Wenn „ich“ und „selbst“ keine Ausdrücke sind, die Dinge einer bestimmten Art bezeichnen, welche Bedeutung haben diese Ausrücke dann? Das Wort „selbst“ hat überhaupt keine eigenständige Bedeutung. Es bezeichnet nichts; es ist, technisch gesprochen, ein synkategorematischer Ausdruck. Im Zusammenhang mit anderen Wörtern hat es aber eine Vielzahl sehr verschiedener Funktionen. Als Fokuspartikel kann „selbst“ dazu dienen, bestimmte Teile eines Satzes ins Zentrum der Aufmerksamkeit zu bringen, wobei diese Teile gegenüber anderen Möglichkeiten hervorgehoben oder eingeschränkt werden. („Alle amüsierten sich. Selbst seine sonst so mürrische Tochter hat gelacht.“ „Selbst ein Wunder hätte ihm nicht mehr helfen können.“) Als Demonstrativpronomen kann „selbst“ eingesetzt werden, um anzugeben, dass nur das Wort gemeint ist, auf das sich „selbst“ bezieht; andere oder anderes sind ausdrücklich ausgeschlossen. („Der Fahrer selbst blieb unverletzt.“, „Importe aus dem Land selbst“, „Das hat er sich selbst zuzuschreiben.“) Schließlich können mit „selbst“ Reflexivpronomina verstärkt werden. („Er rasiert sich.“ – „Er rasiert sich selbst.“ „Sie adressieren den Brief an sich.“ – „Sie adressieren den Brief an sich selbst.“)

Anders als „selbst“ gehört „ich“ zu den Ausdrücken, die etwas bezeichnen, aber es steht – im Gegensatz zu den Wörtern „Tier“ und „Zahl“ – nicht für eine bestimmte Art von Dingen. Um die Semantik von „ich“ zu verstehen, muss man zunächst sehen, dass dieses Wort ein indexikalischer Ausdruck ist. Indexikalische Ausdrücke sind Ausdrücke, die keinen feststehenden Bezug haben, deren Bezug sich vielmehr in Abhängigkeit vom Äußerungskontext ändert. „Sokrates“ bezeichnet (wenn man einmal davon absieht, dass es viele Menschen gibt, die „Sokrates“ heißen) den Philosophen, der mit Xanthippe verheiratet war und der 399 v. Chr. den Schierlingsbecher trinken musste. Der Satz „Sokrates wurde 399 v. Chr. zum Tode verurteilt und hingerichtet“ ist daher immer wahr, unabhängig davon, in welchen Umständen er geäußert wird. Und der Satz „Sokrates war der Lehrer von Alexander dem Großen“ ist immer falsch. „Sokrates“ ist kein indexikalischer Ausdruck; er bezeichnet immer dieselbe Person.

Wenn ich dagegen sage „Dieser Tisch ist rund“, dann hängt die Wahrheit einer Äußerung dieses Satzes davon ab, welchen Tisch „dieser Tisch“ bezeichnet; und das hängt von der Situation ab, in der der Ausdruck geäußert wird. Wenn ich den Satz äußere, während ich auf den runden Tisch vor mir zeige, wenn „dieser Tisch“ also den vor mir stehenden runden Tisch bezeichnet, ist die Äußerung wahr. Wenn der Tisch, auf den ich zeige, aber quadratisch ist, ist sie falsch. „Dieser Tisch“ ist ein indexikalischer Ausdruck, der seinen Bezug ändert, wenn er in unterschiedlichen Situationen geäußert wird. Ebenso sind die Wörter „ich“ und „du“ indexikalische Ausdrücke. Wenn ich sage „Ich bin 1,83 m groß“, ist dieser Satz wahr; wenn

jedoch Angela Merkel diesen Satz äußert, ist er falsch. Wenn jemand, an mich gerichtet, sagt „Du hast nur noch wenige Haare", hat er Recht. Wenn er dies zu Angela Merkel sagen würde, hätte er Unrecht. Auch „ich" und „du" ändern also, je nach Äußerungskontext, ihren Bezug. Und dabei gilt offenbar: „ich" bezieht sich immer auf den, der diesen Ausdruck äußert, während „du" die Person bezeichnet, an die eine Äußerung adressiert ist. Wir können also festhalten: „ich" und „selbst" sind keine Gattungsnamen, die irgendwelche mysteriösen Entitäten bezeichnen. „ich" ist ein indexikalischer Ausdruck, der jeweils die Person bezeichnet, die diesen Ausdruck äußert, und „selbst" ist eine Partikel, die gar nichts bezeichnet, sondern eine Vielzahl sehr unterschiedlicher semantischer und pragmatischer Funktionen besitzt. Auch der Ausdruck „Selbstbewusstsein" steht nicht für ein Wissen, das ich von meinem Selbst habe, sondern einfach für ein Wissen *von mir selbst*.[5]

### *3. Die Wahrheitsbedingungen von „Ich hebe meinen Arm"*

Damit haben wir eine erste Antwort auf die Frage, was es denn heißen kann, dass ich etwas denke oder tue. Wenn Almut sagt „Ich habe meinen Arm gehoben", dann bezieht sie sich mit dem Personalpronomen „ich" nicht auf ihre immaterielle Seele, auf ihren Personenkern, auf ihr Ich, sondern schlicht auf sich selbst – die Person, das Lebewesen, die bzw. das sie ist. Der von Almut geäußerte Satz ist also wahr, wenn Almut selbst ihren Arm gehoben hat. Doch was heißt das, wenn es nicht heißt, dass Almuts Selbst das Heben des Arms verursacht hat? Kann es auch dann wahr sein, dass Almut ihren Arm gehoben hat, wenn die Bewegung ihres Arms letzten Endes auf neuronale Prozesse in ihrem Gehirn zurückgeht? Wann kann man überhaupt sagen, dass eine Person etwas tut und nicht eines ihrer Organe?

Ohne Frage hat das Gehirn sehr viel mit dem zu tun, was wir wahrnehmen, denken und fühlen. Aber ist es wirklich das *Gehirn*, das wahrnimmt, denkt und fühlt? Autoren wie Gerhard Roth ist immer wieder vorgeworfen worden, so zu reden sei ein Kategorienfehler. Und dieser Vorwurf ist berechtigt. Das Gehirn ist, das wird wohl niemand bestreiten, ein Organ eines Lebewesens sowie das Herz, die Leber oder der Magen. Manchmal sagen wir, dass *das ganze Lebewesen* etwas tut: „Der Hund jagt die Katze", „Hans hat Frieda etwas zugeflüstert". Manchmal sagen wir, dass *ein Organ* etwas tut: „Sein Herz schlägt unregelmäßig", „Seine Hände zittern". Und manchmal sagen wir, dass *in einem Organ* etwas geschieht: „In der Niere

[5] „Selbstbewusstsein" ist hier natürlich im philosophischen Sinne gemeint; es geht nicht darum, dass wir Personen, die ein besonderes Selbstvertrauen oder eine besondere Selbstsicherheit an den Tag legen, „selbstbewusst" nennen.

wird das Blut von Giftstoffen gereinigt“, „In der Lunge nimmt das Blut Sauerstoff auf“. Sätze der zweiten Art haben immer etwas Merkwürdiges an sich. Sie sind nicht sprachwidrig; aber der Sache nach stellt sich jedes Mal die Frage, ob Organe tatsächlich zu der Kategorie von Dingen gehören, die selbstständig handeln können. Ist es wirklich das Herz, das schlägt? Sind es wirklich die Hände, die zittern? Oder ist es nicht vielmehr auch in diesen Fällen so, dass mit dem Herzen bzw. mit den Händen etwas passiert?

Doch lassen wir diese Frage beiseite und fragen: Wie ist es mit dem Wahrnehmen, Erinnern, Denken und dem Sich-Bewegen? Ist es das Gehirn, das wahrnimmt, sich erinnert, denkt und Bewegungen ausführt? Oder ist es nicht doch das ganze Lebewesen, dem wir diese Tätigkeiten zuschreiben müssen. Einige Dinge sind klar: Es sind nicht die Beine, die laufen, sondern das Lebewesen, das mit Hilfe seiner Beine läuft; es ist nicht das Auge, das sieht, sondern das Lebewesen, das mit Hilfe seiner Augen sieht. Und genauso ist es auch mit dem Gehirn. Es ist nicht das Gehirn, das sich erinnert, sondern das Lebewesen, das sich mit Hilfe seines Gehirns erinnert; nicht das Gehirn, das überlegt, sondern das Lebewesen, das mit Hilfe seines Gehirns überlegt. Im Gehirn laufen neuronale Prozesse ab, ohne die wir nicht wahrnehmen, uns erinnern, denken oder unsere Hand bewegen können. Aber das bedeutet nicht, dass es das Gehirn selbst ist, das wahrnimmt, sich erinnert, denkt oder meine Hand bewegt. Wahrnehmen, Erinnern, Denken und sich Bewegen sind Tätigkeiten des ganzen Lebewesens und nicht Tätigkeiten eines seiner Organe.

Doch beantwortet dies schon die Frage, ob Sätze wie „Ich denke nach“, „Ich erinnere mich an meine erste Liebe“, „Ich bewege meine Hand“ und „Ich schreibe jetzt diesen Text“ jemals wahr sein können? Bisher haben wir nur gesehen, dass es Tätigkeiten gibt, die nur ganzen Lebewesen und nicht ihren Organen zugeschrieben werden. Noch haben wir aber keine Antwort auf die Frage, was einen Satz wie „Hans hat Simon geschlagen“ wahr macht. Wann können wir sagen, dass es wirklich die Person (das Lebewesen) selbst ist, die (das) etwas tut? Dies ist letzten Endes die Frage nach der Unterscheidung zwischen Aktiv und Passiv, zwischen dem, was ein Wesen tut, und dem, was ihm widerfährt. Diese Unterscheidung ist so fundamental für unser Weltverständnis, dass sie zum grundlegenden Bestandteil der Grammatik unserer Sprache geworden ist. Aber was liegt ihr zugrunde?

Schon bei Tieren unterscheiden wir zwischen dem, was das Tier tut, und dem, was ihm zustößt. Wir unterscheiden den Fall, dass ein Hund ein Kaninchen jagt, von dem, dass er an der Leine von seinem Lieblingsbaum weggezogen wird. Manchmal bewegt sich der Hund selbst, manchmal wird er von etwas oder jemand anderem bewegt. Genauso bei Menschen. Wenn jemand meine rechte Hand fasst und nach oben zieht, dann bewegt sich

mein rechter Arm nach oben; nicht ich bewege in diesem Fall meinen Arm, er wird bewegt – von jemand anderem. Auf der anderen Seite kann ich ihn aber auch selbst bewegen. Ich kann meinen rechten Arm heben, und zwar *direkt*, ohne dass ich etwa mit der linken Hand meine rechte Hand fasse und nach oben ziehe. Was ist der Unterschied zwischen diesen Fällen?

Wenn ich von etwas anderem bewegt werde, wird meine Bewegung von diesem anderem verursacht. Deshalb liegt es nahe zu sagen, dass, wenn *ich mich selbst* bewege, *ich selbst* es bin, der diese Bewegung *kausal hervorruft*. Doch dieses – wieder Cartesianische – Bild ist unangemessen. Dies wird sofort klar, wenn wir uns zunächst auf den Fall von Tieren konzentrieren. Nehmen wir an, mein Hund läuft zu seinem Lieblingsbaum, und zwar von sich aus, ohne dass ihn jemand schubst oder zerrt. Ist es vernünftig anzunehmen, dass dies genau dann der Fall ist, wenn die Bewegungen des Hundes durch *ihn selbst* und durch niemand anderen verursacht werden? Was sollte es überhaupt heißen, dass *der Hund selbst* etwas verursacht? Man wird kaum bezweifeln können, dass, auch wenn der Hund von sich aus zu einem Baum läuft, die Bewegungen der Beine des Hundes durch neuronale Prozesse in seinem ZNS verursacht werden und dass diese neuronalen Prozesse selbst ganz natürliche Ursachen haben. In diesem Verursachungsprozess kommt an keiner Stelle *der Hund selbst* (oder das *Selbst* des Hundes) vor, der (das) in der Lage wäre, von sich aus bestimmte neuronale Prozesse in Gang zu setzen. Bei Tieren kommt uns diese Vorstellung ganz absurd vor. Aber das ändert nichts daran, dass wir auch bei Tieren *berechtigterweise* zwischen Fällen unterscheiden, in denen das Tier selbst etwas tut und in denen es – wir sagen sogar: gegen seinen Willen – bewegt wird.

Dafür dass wir bestimmte Bewegungen mancher Wesen als etwas klassifizieren, was sie selbst tun, ist zunächst zentral, dass diese Bewegungen nicht auf äußere Kräfte zurückgehen. Wenn ich meinen Hund an der Leine ziehe oder ihm einen Schubs gebe, dann wirken äußere Kräfte auf ihn, und seine Bewegungen sind nichts, was ihm zugerechnet werden kann. Genau so, wenn ein Eisenstück von einem Magneten angezogen wird. Auch hier wird die Bewegung durch eine äußere Kraft hervorgerufen; also kann man eigentlich nicht sagen, dass das Eigenstück *sich* bewegt, vielmehr wird es bewegt. Sehr viele Bewegungen von Tieren gehen aber nicht in diesem Sinne auf äußere Kräfte zurück. Tiere verfügen über eigene Energiequellen und setzen die so gewonnene Energien ein, um sich zu bewegen. Wesen mit der Fähigkeit zur Selbstbewegung müssen also über eigene Energieressourcen verfügen.

Hinzu kommt ein zweiter Punkt: Wesen, die selbst etwas tun können, handeln in der Regel nicht reflexhaft; vielmehr verfügen sie über mehrere Handlungsoptionen, zwischen denen eine Wahl getroffen werden muss. Ein

Hund, der von einem anderen Hund angegriffen wird, kann sich dem Kampf stellen, er kann aber auch weglaufen. Also muss eine Entscheidung getroffen werden; und da gibt es zwei Möglichkeiten. Entweder wird das Wesen, das eine Entscheidung zu treffen hat, fremdgesteuert; oder es verfügt über einen *internen Entscheidungsmechanismus*. Paradigmatische Beispiele für fremdgesteuerte Wesen sind Marionetten, aber auch ferngesteuerte Kleinflugzeuge oder Schiffe. Tiere sind nicht in diesem Sinne fremd- oder außengesteuert. Niemand gibt ihnen durch direkte Manipulation oder Fernsteuerung ein, was sie tun sollen. Sie verfügen über einen internen Steuerungsmechanismus, der die zu treffenden Entscheidungen fällt. Gerade weil Tiere nicht reflexhaft handeln, ist vor jeder Handlung eine Entscheidung nötig. Auf irgendeine Weise muss ja bestimmt werden, welche der möglichen Handlungen unter den gegebenen Umständen ausgeführt wird. Zu sagen, dass interne Steuerungsmechanismen Entscheidungen fällen, heißt also nichts anderes, als dass sie dafür sorgen, dass diese und keine andere Handlung initiiert wird. Wenn die Tatsache, dass ein Tier eine bestimmte Handlung ausführt, auf dem dafür zuständigen inneren Steuerungsmechanismus beruht, sagt man: Das Tier selbst hat diese Entscheidung gefällt. Wenn jedoch jemand von außen – zum Beispiel durch Funksignale oder andere Manipulationen – eingreift und so eine Entscheidung herbeiführt, dann handelt es sich um eine fremdbestimmte Entscheidung, die das Wesen nicht selbst getroffen hat. Wenn die Tatsache, dass ein angegriffener Hund sich nicht dem Kampf stellt, sondern wegläuft, auf neuronale Vorgänge in seinem Gehirn zurückgeführt werden kann, bedeutet das also *nicht*, dass es *nicht* der Hund war, der diese Entscheidung getroffen hat. Denn die Entscheidung ist nicht fremdgesteuert; vielmehr ist im Gehirn des Hundes genau die Art von internem Entscheidungsmechanismus realisiert, der für eine Eigensteuerung sorgt. Dass, wie man bei manchen Neurobiologen lesen kann, die Entscheidung wegzulaufen, vom „Gehirn des Hundes getroffen wurde“, heißt also nicht, dass sie nicht vom Hund getroffen wurde. Ganz im Gegenteil: Da diese Entscheidung weder auf direkter Manipulation noch auf Fernsteuerung beruht, da diese Entscheidung also nicht fremdbestimmt ist, handelt es sich gerade deshalb um eine Entscheidung des Hundes selbst.

In der Konsequenz heißt das: *Es gibt keine Konkurrenz zwischen meinem Gehirn und mir*. Mein Gehirn ist in der Regel weder von außen manipuliert noch fremdgesteuert. In meinem Gehirn ist der interne Entscheidungsmechanismus realisiert, der dafür sorgt, dass das, was ich tue, von mir ausgeht; wenn mein Handeln in geeigneter Weise auf diesen Entscheidungsmechanismus zurückgeht, werde ich weder durch äußere Kräfte bewegt noch werden mir meine Entscheidungen von außen eingeflößt. Wenn mein Handeln auf diesem Entscheidungsmechanismus beruht, bin ich es, der

handelt. In diesem Fall können meine Handlungen mir zugerechnet werden. Mit anderen Worten: Wenn sich der Arm von Hans hebt und wenn diese Bewegung durch den internen Entscheidungsmechanismus hervorgerufen wird, der in seinem Gehirn realisiert ist, dann ist es Hans, der seinen Arm hebt.

Ein letzter Punkt: Kognitive Wesen sind Wesen, die in der Lage sind, sich ein Bild von ihrer Umwelt zu machen, d. h., ihre Umwelt intern zu repräsentieren. Wenn solche Wesen nicht nur ihre Umwelt repräsentieren (in der sich in der Regel auch andere kognitive Wesen aufhalten), sondern auf eine bestimmte Art und Weise auch *sich selbst in dieser Umwelt*, verfügen sie über die Art von Selbstbewusstsein, die Voraussetzung dafür ist, dass sie über sich reden können, indem sie den indexikalischen Ausdruck „ich" verwenden.[*] Wenn wir dies mit den vorhergehenden Überlegungen verbinden, ergibt sich Folgendes: Wenn ein mit Selbstbewusstsein ausgestattetes kognitives Wesen von sich sagt „Ich habe meinen Arm gehoben", dann ist das wahr, wenn es selbst seinen Arm gehoben hat. Und dies wiederum ist wahr, wenn die Bewegung seines Arms nicht auf äußere Kräfte zurückgeht und auch nicht auf Manipulation und Fremdsteuerung beruht, sondern auf einem unabhängigen internen Entscheidungsmechanismus. Also: Wenn sich der Arm von Hans hebt und wenn diese Bewegung durch den internen Entscheidungsmechanismus hervorgerufen wird, der in seinem Gehirn realisiert ist, dann ist es wahr, wenn Hans sagt „Ich habe meinen Arm gehoben". Es gibt also eine Analyse der Wahrheitsbedingungen von Sätzen wie „Ich hebe meinen Arm", der zufolge diese Sätze auch wahr sein können, wenn es nicht mein Ich ist, das meine Armbewegung kausal hervorruft.

### *4. Being No One*

Mit diesem Ergebnis erledigt sich auch die recht mystische Spekulation, wir seien eigentlich niemand; denn das Ich sei nichts als ein Produkt unseres (!) Gehirns. Diese These ist in letzter Zeit besonders von Thomas Metzinger (1993, 2003) vertreten worden, hat aber auch unter Wissenschaftsjournalisten einige Anhänger gefunden (siehe etwa Siefer/Weber 2006).[6] Bei Siefer/Weber wird der Punkt, um den es geht, besonders dramatisch formuliert. Gleich auf der ersten Seite im Vorwort „Warnung vor Nebenwirkungen" kann man lesen:

> Wer bin ich, warum bin ich so und nicht anders? Auf diese uralten Fragen gibt unser Buch Antworten. Es ist eine Reise zum Mittelpunkt des Menschen, zu

---

[*] Vgl. zu dieser Überlegung den Beitrag 13 in diesem Band.

[6] Eine sehr hilfreiche Analyse und Kritik dieser These findet sich in Lenzen 2006.

> unserem Selbst. Dorthin, wo ein jeder nicht mehr ist als nur noch ein Ich. Doch Vorsicht! Dieser Ort heißt Nirgendwo. Und diesmal ist das keine besonders kitschige Phrase aus einem deutschen Schlager. Denn: Sie sind Niemand! Kein Ich, nirgends. Sie erfinden sich, jetzt, in diesem Augenblick, da Sie diesen Text lesen. Hinter Ihren Augen ist ein Nichts. (Siefer/Weber 2006, 7)

Schon auf den ersten Blick ist die Absurdität dieser Formulierung mit Händen zu greifen. Ich bin eine bloße Erfindung; als Fiktion gibt es mich nicht mehr als Sherlock Holmes oder Adrian Leverkühn. Doch wer ist der Erfinder? Nach Auskunft von Siefer/Weber: ich. Wenn ich etwas erfinde, muss es mich aber geben (s. Lenzen 2006, 163). Wenn ich niemand wäre, könnte ich auch nichts erfinden, nicht einmal mich.

Besonders bemerkenswert ist der Kontrast zu Descartes' Überlegungen am Anfang der zweiten Meditation – dem berühmten cogito-Argument, in dem Descartes ganz wesentlich von den semantischen Eigenschaften des Wortes „ich" Gebrauch macht.[7]

> Aber ich habe in mir die Annahme gefestigt, es gebe gar nichts in der Welt, keinen Himmel, keine Erde, keine Geister, keine Körper: also bin doch auch ich nicht da? Nein, ganz gewiß war Ich da, wenn ich mich von etwas überzeugt habe. Aber es gibt irgendeinen sehr mächtigen, sehr schlauen Betrüger, der mit Absicht mich immer täuscht. Zweifellos bin also auch Ich, wenn er mich täuscht; mag er mich nun täuschen, soviel er kann, er wird doch nie bewirken können, daß ich nicht sei, solange ich denke, ich sei etwas. Nachdem ich so alles genug und übergenug erwogen habe, muß ich schließlich festhalten, daß der Satz, *Ich bin, ich existiere*, sooft ich ihn ausspreche oder im Geiste auffasse, notwendig wahr sei. (Descartes 1986, 79)

Die Pointe dieser Passage ist nicht, wie oft geglaubt wird, dass die Überzeugung, dass ich existiere, aus der Überzeugung, dass ich denke, *erschlossen* oder *abgeleitet* wird (cogito, *ergo* sum). Die Pointe ist eine ganz andere, wie sich besonders an dem Satz zeigt: „[M]ag er mich nun täuschen, soviel er kann, er wird doch nie bewirken können, daß ich nicht sei, *solange ich denke, ich sei etwas*". Worauf Descartes offenbar hinaus will, ist der *selbstverifizierende Charakter* der Überzeugung, dass ich existiere. Selbstverifizierend sind Überzeugungen, die wahr sein müssen, wenn (und solange) man sie hat. Ist die Überzeugung, dass ich existiere, wirklich selbstverifizierend? Gibt es überhaupt selbstverifizierende Überzeugungen und Gedanken? Offensichtlich ja; denn der Gedanke *Ich denke* ist ganz offenkundig selbstverifizierend. Ich kann diesen Gedanken einfach dadurch wahr machen, dass ich ihn denke. *Wenn* und *solange* ich denke, dass ich denke, ist es notwendig wahr, dass ich denke. Für den Gedanken *Ich existiere* lie-

[7] Vgl. zu den nächsten Abschnitten Beckermann 2004, 216 ff.

gen die Dinge etwas komplizierter; doch auch dieser Gedanke ist tatsächlich selbstverifizierend.

Wir haben schon gesehen, dass sich das Wort „ich“ immer auf die Person bezieht, die einen Satz äußert, der dieses Wort enthält, oder einen entsprechenden Gedanken denkt. Wenn jemand denkt *Ich bin reich*, dann kann das, was er denkt, falsch sein; denn es kann sein, dass er nicht reich ist. Es ist aber unmöglich, dass sich das Wort „ich“ in diesem Gedanken auf nichts bezieht. Denn es bezieht sich automatisch auf den, der diesen Gedanken hat. Wenn jemand denkt *Ich existiere*, kann er sich somit überhaupt nicht irren. Denn das Wort „ich“ in diesem Gedanken bezieht sich, wie gesagt, automatisch auf den, der diesen Gedanken hat – also auf etwas, das existiert. Auch der Gedanke *Ich existiere* garantiert somit seine eigene Wahrheit. Mit anderen Worten: Schon aus semantischen Gründen muss der Satz „Ich bin niemand“ falsch sein, wenn ihn jemand äußert. Und genau deshalb ist die These, wir seien niemand, von vornherein absurd.

Allerdings hat Wolfgang Lenzen (2006, 162) darauf hingewiesen, dass die Metzingersche These vielleicht gar nicht wörtlich gemeint ist. Vielleicht will Metzinger nur sagen, dass wir im ‚Inneren‘ kognitiver Wesen, so tief wir auch vordringen mögen, niemals eine Seele oder ein Selbst finden werden. Dabei wäre ihm natürlich aus vollem Herzen zuzustimmen. Ich habe ja selbst dafür argumentiert, dass es keine Seelen, ‚Iche‘ oder ‚Selbste‘[8] gibt, die sozusagen den eigentlichen Wesenskern denkender und handelnder Wesen ausmachen.

Allerdings: Ein System, das in der Lage ist, den Gedanken zu fassen, dass es selbst existiert, kann sich in diesem Gedanken nicht irren. Wenn es denkt „Ich existiere“, dann existiert es auch, dann ist es also etwas. Ich z. B. bin nicht nichts und auch nicht niemand; *ich* bin Ansgar Beckermann, 1945 in Hamburg geboren, 1,83 m groß, *ich* bin ein Mensch und Professor für Philosophie an der Universität Bielefeld. Wer wollte bestreiten, dass all das wahr ist?

## Literatur

Beckermann, A. (2004) „René Descartes: Die Suche nach den Grundlagen sicherer Erkenntnis“, in: A. Beckermann & D. Perler (Hg) *Klassiker der Philosophie heute*. Stuttgart: Philipp Reclam jun., 208–229. (Auch in A. Beckermann, *Aufsätze 2*, Universitätsbibliothek Bielefeld 2012, Beitrag 7)

Beckermann, A. (2008) *Gehirn, Ich, Freiheit*, Paderborn: mentis.

[8] Dass man im Deutschen diese Plurale nicht bilden kann, ist ein deutliches Indiz, dass mit dem zuvor kritisierten Verständnis der Worte „ich“ und „selbst“ etwas nicht stimmt.

Descartes, R. (1986) *Meditationes de prima philosophia*. *Meditationen über die erste Philosophie*. Lateinisch-Deutsch. Übers. und hrsg. von Gerhart Schmidt. Stuttgart: Reclam.

Descartes, R. (1984) *Les passions de l'âme*. *Die Leidenschaften der Seele*. Französisch-Deutsch. Herausgegeben und übersetzt von Klaus Hammacher. Hamburg: Felix Meiner.

Descartes, R. (1977) *The Philosophical Works of Descartes. Vol. I*. Transl. by E.S. Haldane & G.R.T. Ross. Cambridge: Cambridge University Press.

Descartes, R. (1988) *Selected Philosophical Writings.* Transl. by J. Cottingham, R. Stoothoff & D. Murdoch. Cambridge: Cambridge University Press.

Freud, S. (1917) *Vorlesungen zur Einführung in die Psychoanalyse*. In: S. Freud *Gesammelte Werke, Band 11*, Frankfurt/M.: Fischer Verlag 1969.

Lenzen, W. (2006) „Auf der Suche nach dem verlorenen „Selbst" – Thomas Metzinger und die ‚letzte Kränkung' der Menschheit". *Facta Philosophica* 8, 161–192.

Locke, J. (1975) *An Essay Concerning Human Understanding*. Ed. by P.H. Nidditch, Oxford: Clarendon Press. (Dt.: *Versuch über den menschlichen Verstand*. 4., durchgesehene Auflage in 2 Bänden. Hamburg: Felix Meiner 1981)

Metzinger, T. (1993) *Subjekt und Selbstmodell – Die Perspektivität phänomenalen Bewußtseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation*. Paderborn: Schöningh.

Metzinger, T. (2003) *Being No One – The Self-Model Theory of Subjectivity*. Cambridge MA: MIT Press.

Roth, G. (2003) *Fühlen, Denken, Handeln*. Neue, vollständig überarbeitete Ausgabe. Frankfurt/M.: Suhrkamp.

Siefer, W. & C. Weber (2006) *Ich – Wie wir uns selbst erfinden*. Frankfurt/M.: Campus.

# 15

## Die Rede von *dem* Ich und *dem* Selbst
## Sprachwidrig und philosophisch höchst problematisch*

> [Dieses Buch] ist eine Reise zum Mittelpunkt des Menschen, zu unserem Selbst. Dorthin, wo ein jeder nicht mehr ist als nur noch ein Ich. Doch Vorsicht! Dieser Ort heißt Nirgendwo. Und diesmal ist das keine besonders kitschige Phrase aus einem deutschen Schlager. Denn: Sie sind Niemand! Kein Ich, nirgends. Sie erfinden sich, jetzt, in diesem Augenblick, da Sie diesen Text lesen. Hinter Ihren Augen ist ein Nichts. (Siefer/Weber 2006, 7)

Mit diesen reißerischen Worten beginnen die Autoren Werner Siefer und Christian Weber ihr Buch *Ich – Wie wir uns selbst erfinden*. Was für ein unglaublicher Unsinn, sollte man denken. Doch die meisten Leser sehen das offenbar anders – sie halten diese Sätze nicht für ausgemachten Blödsinn, sondern für tiefgründige philosophische Einsichten. Wie konnte es dazu kommen? Warum kann heute jedermann ungestraft davon reden, dass es einen Mittelpunkt des Menschen gibt – sein Selbst? Und dass es einen Ort gibt, an dem jeder Mensch nicht mehr ist als ein Ich?

Offenbar setzen diese Formulierungen voraus, dass die Wörter „Ich“ und „Selbst“ als Gattungsbegriffe verwendet werden können, die man problemlos mit Demonstrativ- und Possessivpronomina sowie mit bestimmten und unbestimmten Artikeln verbinden kann: „Mein Ich“, „jenes Ich“, „das Ich“, „ein Ich“, „mein Selbst“, „das Selbst“, „ein Selbst“ – so wie „mein Buch“, „jenes Buch“, „das Buch“, „ein Buch“. Aber natürlich sind „ich“ und „selbst“ – zumindest ursprünglich – keine Gattungsbegriffe. „ich“ ist, wie jeder weiß, das Personalpronomen der ersten Person Singular. Es bezieht sich als singulärer Term immer auf den, der das Wort äußert. Das Wort „selbst“ hat dagegen keine eigenständige Bedeutung. Es bezeichnet nichts; es ist, technisch gesprochen, ein synkategorematischer Ausdruck. Im Zusammenhang mit anderen Wörtern hat es aber eine Vielzahl unterschiedlicher Funktionen. Als Fokuspartikel kann „selbst“ dazu dienen, bestimmte Teile eines Satzes ins Zentrum der Aufmerksamkeit zu bringen, wobei diese Teile gegenüber anderen Möglichkeiten hervorgehoben oder eingeschränkt werden. („Alle amüsierten sich. Selbst seine sonst so mürrische Tochter hat gelacht“, „Selbst ein Wunder konnte ihm nicht mehr helfen“.) Als Demonstrativpronomen kann „selbst“ eingesetzt werden, um anzugeben, dass nur der, die oder das gemeint ist, auf das sich „selbst“ bezieht; andere oder anderes sind ausdrücklich ausgeschlossen. („Der Fahrer selbst blieb unverletzt“, „Importe aus dem Land selbst“, „Das hat er sich

---

* Überarbeitete Version des Aufsatzes Beckermann 2010.

selbst zuzuschreiben“.) Schließlich können mit „selbst“ Reflexivpronomina verstärkt werden. („Er rasiert sich“ – „Er rasiert sich selbst“, „Sie adressieren den Brief an sich“ – „Sie adressieren den Brief an sich selbst“.)

„Philosophische Probleme entstehen, wenn die Sprache feiert“,[1] d. h., wenn harmlose sprachliche Ausdrücke, die im normalen Sprachgebrauch eine sinnvolle Rolle spielen, aus dem Zusammenhang gerissen werden und wenn man ihnen Eigenschaften andichtet, die sie im normalen Sprachgebrauch nicht haben. Lässt sich sagen, wer damit angefangen hat, mit den Wörtern „ich“ und „selbst“ feiern zu gehen? Ich glaube ja: Mit einigen zu Missverständnissen einladenden Formulierungen Descartes’ beginnt eine Umdeutung des Wortes „ich“, und was im Anschluss an Descartes dem Wort „ich“ angetan wurde, das hat Locke dem Wort „selbst“ zugemutet. Dabei fängt alles scheinbar harmlos an.

Descartes versucht zu Beginn der zweiten Meditation zu beweisen, dass zumindest er selbst existiert, d. h., dass der Satz „Ich bin“ notwendigerweise wahr ist, solange er ihn denkt.[2] Nachdem er diesen Beweis geführt hat, fragt Descartes aber sofort weiter: Was für eine Art Ding ist das eigentlich, dessen Existenz ich da gerade bewiesen habe? Und diese Frage drückt er so aus:

> Nondum vero satis intelligo, quisnam sim *ego ille*, qui jam necessario sum […]. (Descartes, *Meditationes*, AT VII, 25 – meine Hervorh.)

Zwei Seiten später schreibt er:

> Novi me existere; quaero quis sim *ego ille* quem novi. (ibid., 27 – meine Hervorh.)

„Ego ille“ – das mag auf den ersten Blick schief klingen, ist aber schon im antiken Latein durchaus üblich.[3] Eine schöne Stelle findet sich in einem Brief von Plinius an Tacitus:

> Ridebis, et licet rideas. ego ille, quem nosti, apros tres et quidem pulcherrimos cepi. ‚ipse?‘ inquis. ipse, non tamen ut omnino ab inertia mea et quiete discederem. (Plinius *Briefe*, 78)
>
> Du wirst lachen, und Du kannst auch lachen. [Ausgerechnet ich], den Du kennst, habe drei und zwar ganz prächtige Eber gefangen. „Selbst?“ fragst Du. Ja, selbst, doch ohne dabei auf meine Bequemlichkeit und Ruhe ganz zu verzichten. (ebd., 79)

---

1 Wittgenstein, *Philosophische Untersuchungen*, Teil I, § 38.

2 Beckermann 2004, 216 ff.

3 In der ersten Version dieses Aufsatzes hatte ich noch behauptet, „ego ille“ sei im Lateinischen ungrammatisch. Auf diesen Fehler hat mich in seiner charmanten Art Andreas Kemmerling hingewiesen.

Hier kommen beide Ausdrücke „ego ille“ und „ipse“ vor; aber diese Vorkommnisse sind alle völlig harmlos. „Ego ille“ heißt offenbar „ausgerechnet ich“ oder „gerade ich“ (wie in „das stößt ausgerechnet mir zu“), und „ipse“ ist elliptisch – einmal steht es für „Du selbst?“, das zweite Mal für „[ja,] ich selbst“.

Aber was bedeutet „ego ille“ bei Descartes? „Jener Ich“? Oder vielleicht sogar „jenes Ich“? Dann wäre es nur noch ein kleiner Schritt zu „das Ich“ oder „mein Ich“. Dabei hätte Descartes auf das „ille“ ohne Weiteres verzichten können. Die Sätze „Nondum vero satis intelligo, quisnam sim *ego*, qui jam necessario sum […]“ und „Novi me existere; quaero quis sim *ego* quem novi“ hätten ziemlich dasselbe geleistet. Aber Descartes wollte „ego“ ganz offensichtlich besonders betonen und dazu benutzt er eben jenes „ille“. Dass diese Ausdrucksweise zumindest problematisch ist, zeigt sich deutlich, wenn man sieht, welche Schwierigkeiten sich ergeben, wenn man versucht, die angeführten Sätze in andere Sprachen zu übersetzen.

In der neuen englischen Standardübersetzung von Cottingham, Stoothoff und Murdoch (CSM) heißt es:

> But I do not yet have a sufficient understanding of what this ‚I‘ is, that now necessarily exists. (CSM II, 17)
>
> I know that I exist; the question is, what is this ‚I‘ that I know? (ibid., 18)

Interessant sind hier die auf den ersten Blick völlig unmotivierten Anführungszeichen. Offenbar wollte Descartes keine Aussage über den Buchstaben „I“ machen; das glauben auch die Übersetzer nicht. Aber warum dann die Anführungszeichen? In meinen Augen sind sie ein Zeichen der Unsicherheit und der Distanzierung; die Übersetzer fühlen sich nicht wohl mit der naheliegenden Übersetzung „this I“. Sie spüren, dass „this I“ kein korrektes Englisch ist.[4] Die Anführungszeichen sollen wohl sagen: Wir, die Übersetzer, wissen, dass „this I“ kein korrektes Englisch ist, aber wir können es nicht ändern; so steht es im Original! Eine andere Möglichkeit wäre, das CSM an eine metasprachliche Lesart denken, dass sie meinen, Descartes habe in etwa sagen wollen: „Noch sehe ich aber nicht hinreichend ein, was denn derjenige ist, auf den ich mich bisher mit dem Wort ‚ich‘ bezogen habe, und von dem ich jetzt sicher bin, dass er existiert.“[5] Das wäre nicht unvernünftig; aber ist „that ‚I‘ “ (ganz zu schweigen von „ego ille“)

---

[4] Vgl. auch: „The elusive ‚I‘ that shows an alarming tendency to disappear when we try to introspect it. […]“ (Blackburn, *Oxford Dictionary*, 344); siehe unten, S. 7.

[5] Diese Überlegung verdanke ich Rüdiger Bittner, der mir auch mit einer Reihe anderer Hinweise sehr geholfen hat.

ein angemessener Ausdruck für „derjenige, auf den ich mich bisher mit dem Wort ‚ich' bezogen habe"?[6]

Es ist aus vielen Gründen hilfreich und nützlich, die CSM-Übersetzung mit der älteren Standardübersetzung von Haldane und Ross (HR) zu vergleichen. Haldane und Ross übersetzen die angeführten Sätze so:

> But I do not yet know clearly enough what I am, I who am certain that I am […]. (HR 1, 150)
>
> I know that I exist, and I inquire what I am, I whom I know to exist. (ibid., 152)[7]

Natürlich kann man über die Übersetzung von „qui jam necessario sum" durch „who am certain that I am" und die Übersetzung von „quem novi" durch „whom I know to exist" streiten. Darauf komme ich gleich zurück. Aber ansonsten wählen Haldane und Ross einen sauberen Weg. Angesichts der Sprachwidrigkeit von „that I" lassen sie das „ille" einfach unter den Tisch fallen (was ja, wie schon gesagt, gar nichts ausmacht) und kommen so zu einer ebenso einfachen wie sachlich angemessenen Übertragung ins Englische.

Dabei hatten sie allerdings ein prominentes Vorbild, an dem sie sich orientieren konnten – die französische Übersetzung von Louis Charles d'Albert, Duc de Luynes, die besonders interessant ist, weil sie 1647 noch zu Descartes' Lebzeiten erschien und von ihm autorisiert wurde. Auch d'Albert stand als Übersetzer vor einer schwierigen Situation. Wie sollte er „ego ille" ins Französische übertragen? „ce je" oder „ce moi" – offenbar hat ihm das nicht gefallen.[8] So entscheidet sich schon d'Albert dazu, das „ille" einfach zu ignorieren – mit der Zustimmung Descartes'!

---

[6] Wenn man in diese Richtung weiter denkt, zeigt sich meiner Meinung nach, dass „ille" in den zitierten Passagen eine rückverweisende Funktion hat und dass es deshalb wohl am besten wäre, „ego ille" mit „ich, der" zu übersetzen. „Noch sehe ich aber nicht hinreichend ein, wer ich denn nun bin – ich, der nunmehr notwendig ist" und „Ich weiß, dass ich bin, und ich frage mich, was ich sei – ich, der, den ich kenne (bzw. ich, der, von dem ich weiß, dass er existiert)". Diesen Hinweis verdanke ich Heike Wiese.

[7] Ähnlich Andreas Schmidt: „Aber ich verstehe noch nicht genug, wer ich denn bin, der ich nunmehr notwendigerweise bin". (Descartes, *Meditationen*, A. Schmidt, 73) Aber: „Ich habe erkannt, dass ich existiere; ich frage, wer ich denn bin, jener Ich, den ich erkannt habe." (ibid., 79)

[8] Allerdings scheut sich Descartes selbst nicht, im *Discours* zu schreiben: „En sorte que *ce moi* […] est entierement distincte du corps […]." (*Discours* IV 2, AT VI 33 – meine Hervorh.) Und in der sechsten Meditation heißt es auch in der französischen Übersetzung des Duc de Luynes: „[…] il est certain que *ce moy* […] est entierement & veritablement distincte de mon corps […]." (*Meditations*, AT IX, 62 – meine Hervorh.)

> Mais ie ne connois pas encore assez clairement ce que ie suis, moy qui suis certain que ie suis […]. (Descartes, *Meditations*, AT IX, 19)
>
> […] i'ay reconnu que i'etois, & ie cherche quel ie suis, moy que i'ay reconnu estre. (ibid., 21)

D'Albert übersetzt „ego ille" schlicht mit „moy". Haldane und Ross folgen ihm darin, so wie sie ihm auch bei der Übersetzung von „qui jam necessario sum" und „quem novi" folgen, was in meinen Augen durchaus vertretbar ist.

Wie kann man „ego ille" ins Deutsche übertragen? Christian Wohlers schreibt in seiner neuen Übersetzung der *Meditationen*:

> Noch sehe ich aber nicht hinreichend ein, wer ich denn nun bin, jenes Ich, der ich nunmehr notwendig bin. (Descartes, *Meditationen*, Wohlers, 49)
>
> Mir ist bekannt, daß ich existiere; ich frage, was ich bin, jenes Ich, das mir bekannt ist. (ibid., 55)

Auch hier ist der Vergleich mit einer älteren Übersetzung hilfreich. Ich wähle die Übersetzung von Gerhard Schmidt bei Reclam:

> Ich bin mir aber noch nicht hinreichend klar darüber, wer denn Ich bin – jener Ich, der notwendigerweise ist. (Descartes, *Meditationen*, G. Schmidt, 79)
>
> Ich weiß, daß ich bin, und ich frage mich, was dieser Ich sei, den ich kenne. (ibid., 85)

Auffällig ist sofort, dass beide bei der Übersetzung von „ego ille" das groß geschriebene „Ich" wählen, wofür der lateinische Text eigentlich keine Grundlage bietet. Noch interessanter ist aber die Übersetzung von „ille". Wohlers wählt – gegen den lateinischen Text (schließlich heißt es nicht „ego illud") – das Neutrum „jenes", während Schmidt die dem Originaltext entsprechende männliche Form „jener" verwendet. Was ist hier passiert? In meinen Augen lesen beide Übersetzer (Wohlers noch mehr als Schmidt) Descartes' Text durch eine moderne Brille, sie lesen ihn als Autoren, für die die Rede vom *dem Ich* oder *dem Selbst* völlig selbstverständlich geworden ist; sie zeigen ja auch gar keine Scheu, „ego ille" relativ wörtlich zu übersetzen. Außerdem scheinen sie zu glauben, auch Descartes rede über *das Ich*. Aber dafür gibt es keinen Anhaltspunkt. Wenn man dem Wortlaut folgt, redet Descartes ganz eindeutig über *sich*, nicht über sein Ich. Es gibt keinerlei Anzeichen dafür, dass Descartes – wie die heutigen – „ego" auch mit einem bestimmten oder unbestimmten Artikel verbunden hätte (wobei klar ist, dass es solche Artikel im Lateinischen nicht gibt). Es gibt also keinen Hinweis darauf, dass Descartes „ego" – von „ipse" ganz zu schweigen, auch darauf komme ich noch zurück – als Gattungsbegriff verwendet. Aber

ganz offensichtlich hat er all denen den Weg geebnet, die dies inzwischen völlig unbefangen tun.[9]

Wahrscheinlich wäre der Philosophie viel erspart geblieben, wenn Descartes schon in der lateinischen Version der *Meditationen* auf „ille" verzichtet hätte. Die Missverständnisse kommen aber erst richtig in Gang mit der Behandlung, die John Locke dem Wort „self" angedeihen lässt. Im 27. Kapitel des zweiten Buches des *Essay Concerning Human Understanding* schreibt er:

> *Self* is that conscious thinking thing, (whatever Substance made up of whether Spiritual or Material, Simple or Compounded, it matters not), which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern'd for it *self*, as far as that consciousness extends. (Locke, *Essay,* II, xxvii, § 17 – Hervorh. im Original)

Was will Locke hier sagen? Offenbar: 1. Es gibt so etwas wie ein Selbst; es gibt Dinge, die ein Selbst sind. 2. Und das sind folgende: die bewussten denkenden Dinge, die Lust und Schmerz fühlen bzw. sich dieser Zustände bewusst sein können, die glücklich oder unglücklich sein können und die sich deshalb um sich selbst sorgen, soweit dieses Bewusstsein reicht. Locke behandelt „Selbst" also tatsächlich als Gattungsbegriff – es gibt Dinge, die ein Selbst sind, und ich sage euch auch, welche das sind. Dabei ist besonders pikant, dass „selbst" in der zitierten Passage zweimal vorkommt – sprachwidrig als Subjekt des ganzen Satzes und völlig sprachkonform in der Klausel „concern'd for it self". (Merke: Locke schreibt nicht „*its* self"!)

Die Sprachwidrigkeit des ersten Vorkommnisses von „selbst" kommt wieder besonders klar zum Ausdruck, wenn man mögliche Übersetzungen betrachtet. So heißt es in der Meiner'schen Übersetzung von C. Winckler:

> Das *Ich* ist das bewußt denkende Wesen, gleichviel aus welcher Substanz es besteht (ob aus geistiger oder materieller, einfacher oder zusammengesetzter), das für Freude und Schmerz empfindlich und sich seiner bewußt ist, das für Glück und Unglück empfänglich ist und sich deshalb soweit um sich selber kümmert, wie jenes Bewußtsein sich erstreckt. (Locke, *Versuch*, 428 – Hervorh. im Original)

Auf den ersten Blick eine ziemliche Dreistigkeit. Wie kann Winckler „self" einfach mit „das Ich" übersetzen? Aber welche Alternativen hätte er gehabt? Hätte er die Formulierung wählen können: „*Selbst* ist das bewußt denkende Wesen, …"? Das ist kein korrektes Deutsch. Und auch „Ein (Das) *Selbst* ist das bewußt denkende Wesen, …" scheint nicht viel besser.

---

[9] Siehe etwa Pascal, der in den *Pensées* ganz unbefangen von dem Ich redet: „[…] car le moi consiste dans ma pensée" (120f.), „Le moi est haïssable" (344), „Qu'est-ce que le moi?" (377). Diesen Hinweis verdanke ich Frau Prof. Dr. Gisela Schlüter aus Erlangen.

Winckler wählt „das Ich“, weil es Anfang des 20. Jahrhunderts schon überhaupt kein Problem mehr war, von dem Ich zu reden. Überhaupt kann man ja feststellen, dass man im Deutschen eher vom Ich und im Englischen eher vom Selbst redet.

Aufschlussreich ist auch die Frage, wie man Lockes Passage ins Lateinische übersetzen könnte. Vielleicht: „*Ipse* est res cogitans illa quae …“. Da läuft einem wirklich ein Schauder über den Rücken. Im Lateinischen lässt sich mit „ipse“ einfach nicht so umgehen, wie es im Englischen mit „self“ zumindest versucht worden ist. Dies zu ignorieren, ist mehr als frech. Gegen Ende der zweiten Meditation schreibt Descartes:

> Nunquid *me ipsum* non tantum multo verius, multo certius, sed etiam multo distinctius evidentiusque, cognosco? (Descartes, *Meditationes*, AT VII, 33 – meine Hervorh.)

Haldane und Ross übersetzen korrekt:

> [D]o I not know *myself*, not only with much more truth and certainty, but also with much more distinctness and clearness? (HR 1, 156 – meine Hervorh.)

CSM allerdings trauen sich was:

> Surely my awareness of *my own self* is not merely much truer and more certain […], but also much more distinct and evident. (CSM II, 22 – meine Hervorh.)

In meinen Augen ist das nicht nur eine Frechheit, sondern eine veritable Fehlübersetzung.

Auch bei den deutschen Übersetzungen fällt z. B. bei Wohlers eine gewisse Unentschiedenheit auf. Den zitierten lateinischen Satz aus der zweiten Meditation übersetzt er korrekt:

> Sollte ich *mich selbst* nicht nur viel wahrer, viel sicherer, sondern auch viel deutlicher und evidenter erkennen? (Descartes, *Meditationen*, Wohlers, 65 – meine Hervorh.)

Anders sieht es allerdings bei der folgenden Passage aus den *Principia* aus:

> Et quamvis sibi certius esse putarint, *se ipsos* existere, quam quidquam aliud, non tamen adverterunt, per *se ipsos*, mentes solas hoc in loco fuisse intelligendas […] (Descartes, *Principia*, AT VIII-1, 9 – meine Hervorh.)

Hier wählt Wohlers die folgende Übersetzung:

> Und sosehr sie auch vermeinten, sich *ihrer eigenen* Existenz sicherer zu sein als irgend etwas anderem, so haben sie dennoch nicht bemerkt, daß sie unter *ihrem Selbst* an dieser Stelle allein ihren Geist hätten verstehen müssen. (Descartes, *Prinzipien*, Wohlers, 21 – meine Hervorh.)

„*se ipsos* existere“ übersetzt er korrekt mit „*ihrer eigenen* Existenz“; aber bei der Übersetzung von „per *se ipsos*“ schleicht sich wieder das Selbst ein – „unter *ihrem Selbst*“. Wahrscheinlich gefiel Wohlers die schlichte Über-

setzung „daß sie unter *sich selbst* an dieser Stelle allein ihren Geist hätten verstehen müssen“ nicht. Und zugegebenermaßen klingt das ein bisschen eigenartig; aber nur so wäre es richtig.

Im Lateinischen kann man „ipse“ nicht als Substantiv und damit auch nicht als Gattungsbegriff verwenden. „meus ipse“ ist von vornherein völliger Unsinn; „ille ipse“ ist zwar korrektes Latein, heißt aber nicht „jenes Selbst“, sondern einfach „er selbst“. Ähnlich ist es im Französischen; auch „mon même“ ist offensichtlich unkorrekt. Im Französischen ist es deshalb üblich, das Lockesche „self“ mit „soi“ zu übersetzen.[10] Auch das ist ein Akt der Verzweiflung. Denn wie steht es z. B. mit „mon soi“, „le soi“ oder „ce soi“? Sicher kein korrektes Französisch, auch wenn etwa Sartre mit „le soi“ keine Probleme zu haben scheint. Allerdings gibt es im Französischen die Ausdrücke „ce même“ und „le même“. Das stimmt, aber die bedeuten eben nicht „das Selbst“, sondern „derselbe“ bzw. „dasselbe“.

Zurück zu Locke. Wie konnte es dazu kommen, dass er offenbar keine großen Probleme damit hatte, „self“ als Gattungsbegriff zu verwenden? Ich denke, dass es mindestens zwei linguistische Gründe gibt, die diesen Sprachgebrauch befördert haben. Zunächst ist bemerkenswert, dass es im Englischen lange Zeit üblich war, Wörter wie „myself“ auseinander zu schreiben – „my self“ –, was natürlich die Vermutung zumindest begünstigt, es gäbe da so etwas wie mein Selbst. Wichtiger noch scheint mir aber, dass die Verwendung von „self“ im Englischen bei genauerer Betrachtung ‚unlogisch‘ wirkt. Wenn man sagt „Ich komme selbst“, dann gebraucht man im Deutschen den Ausdruck „ich selbst“ so wie man im Lateinischen „ego ipse“ verwendet, was bei Descartes oft genug vorkommt. Zur Verstärkung von „ich“ bzw. „ego“ verbindet man den Nominativ des Personalpronomens (erste Person Singular) mit dem Ausdruck „selbst“ bzw. „ipse“. Im Englischen ist das anders; dort sagt man „I am coming *my*self“. Dort verbindet man den Nominativ des Personalpronomens (erste Person Singular) also nicht einfach mit „self“, sondern mit „myself“. Noch interessanter sind die anderen Kasus. „Ich erkenne mich selbst“, im Lateinischen „Cognosco me ipsum“. Hier bildet „mich selbst“ bzw. „me ipsum“ das grammatische Objekt. Entsprechend verbindet man zur Verstärkung von „mich“ den Akkusativ des Personalpronomens mit dem Ausdruck „selbst“ bzw. mit dem Akkusativ von „ipse“ – „ipsum“. Wieder ist das Englische anders: „I know myself“ bzw. „I know my self“. Das ist wirklich merkwürdig; denn logischerweise würde man doch auch hier den Akkusativ des

[10] Jean-Michel Vienne etwa übersetzt „And by this every one is to himself, that which he calls *self* […]“ (Locke, *Essay,* II, xxvii, § 9) und „[…] that makes every one to be, what he calls *self* […]“ (ibid.) durch „[…] et par là chacun est pour soi-même ceci, qu’il appelle *soi* […]“ und „[…] fait de chacun ce, qu’il appelle soi […]“.

Personalpronomens erwarten: „I know me self“. Und so verfährt man im Englischen ja auch bei der dritten Person Singular (und Plural): „He knows himself“. „Myself“ ist also eine äußerst eigenartige Verbindung des *Possessivums* „my“ mit „self“ – ein Sprachgebrauch, für den es in meinen Augen keine logische Erklärung gibt. Ähnlich übrigens bei der zweiten Person Singular – „You know yourself“ anstatt „You know you self“. Nur in der dritten Person Singular und Plural ist es so, wie man es erwarten würde. Ein irritierender Befund. Denn natürlich verführt die Tatsache, dass es nicht „I know me self“, sondern „I know myself“ heißt und dass in „myself“ das Possessivum „my“ mit „self“ verbunden ist, dazu anzunehmen, es gäbe da etwas, was meins ist – nämlich ein Selbst. Man kann also durchaus verstehen, was Locke dazu gebracht hat, „self“ als Gattungsbegriff zu verwenden.

Dabei ist Lockes Sprachgebrauch philosophisch gesehen noch relativ harmlos. Locke sagt ja nur, dass es Wesen gibt, die denken, die sich ihrer bewusst sind, die Lust und Schmerz fühlen können, die glücklich oder unglücklich sein können und die sich daher um sich selbst sorgen. Natürlich gibt es solche Wesen; nur sollte man sie weder „Selbste“ noch „selves“ nennen. Philosophisch bedenklich wird die Rede von dem Ich oder dem Selbst erst, wenn man unter einem Ich oder einem Selbst etwas anderes versteht als Locke.

Sehr ähnlich wie Locke charakterisiert allerdings auch E. Jonathan Lowe den Begriff des Selbst in *The Oxford Companion to Philosophy*:

> **self.** The term ‚self‘ is often used interchangeably with ‚person‘, though usually with more emphasis on the ‚inner‘, or psychological, dimension of personality than on outward bodily form. Thus a self is conceived to be a subject of consciousness, a being capable of thought and experience and able to engage in deliberative action. More crucially, a self must have a capacity for self-consciousness, which partly explains the aptness of the term ‚self‘. Thus a self is a being that is able to entertain first-person thoughts. (Lowe 1995, 816 f.)

Dieser Gebrauch von „self“ oder „Selbst“ ist, wiewohl immer noch sprachwidrig, relativ harmlos. Dies zeigt sich schon daran, dass bei diesem Verständnis Thesen wie „Es gibt kein Selbst“ uninteressant, weil trivialerweise falsch sind. Natürlich gibt es Wesen, die über Bewusstsein verfügen, die denken, Erfahrungen machen sowie überlegt handeln können und die *de se*-Überzeugungen haben – z. B. Menschen wie du und ich. Es scheint mir ziemlich ausgeschlossen, dass Siefer und Weber dies bestreiten wollen. Doch es gibt noch ein anderes Verständnis von Selbst, auf das Blackburn in dem folgenden Artikel seines *Oxford Dictionary of Philosophy* anspielt.

> **self** The elusive ‚I‘ that shows an alarming tendency to disappear when we try to introspect it […]. (Blackburn 1994, 344)

Hier deutet Blackburn zumindest an, dass sich „self“ und „I“ nicht auf ganze Personen, sondern auf etwas *in* diesen Personen beziehen sollen, das bei näherem Hinsehen die Tendenz hat zu verschwinden. Deutlicher wird dieses Verständnis in dem Lexikonartikel von Roland Henke:

> **Selbst** Bezeichnung für den innersten Wesenskern der Persönlichkeit (→Ich), der auf der Möglichkeit, sich seiner selbst bewusst zu werden […], beruht. Insofern kennzeichnet der Begriff in religiöser Hinsicht auch den wahren alle wechselnden Lebenserscheinungen überdauernden Kern des Menschen – als *atman* oder →Seele. (Henke 2003, 609 f.)

Entsprechend schreibt Thomas Blume in demselben Handwörterbuch zum Begriff des Ich:

> **Ich** […] An der Auffassung einer allen Bewusstseinszuständen zugrunde liegenden *Seelensubstanz*, welche mit dem Ich identifiziert wird, entzündet sich die Kritik →Humes. (Blume 2003, 394 – meine Hervorh.)

Hier wird der Zusammenhang zur Seelentheorie Descartes’ offenkundig. Ein paar Zeilen früher wird Blume sogar noch deutlicher: „Entsprechend der aristotelischen Substanzontologie fasst Descartes *das Ich* als eine →Substanz auf, und zwar als eine denkende Substanz […], die allen Denkakten als Träger zugrunde liegt.“ (ibid. – meine Hervorh.) Wie wir inzwischen wissen, ist das völliger Unsinn. Descartes vertritt nicht die These, *sein Ich* sei eine denkende Substanz; vielmehr behauptet er, *er selbst* (Descartes) sei eine denkende Substanz.[11]

---

[11] „Sed quid igitur sum? Res cogitans“. (Descartes, *Meditationes*, AT VII, 28) Es kann gut sein, dass sich viele Übersetzer mit der Übertragung von „ego ille“ deshalb so schwer tun, weil sie nicht glauben wollen, dass Descartes tatsächlich diese These vertritt. Ist Descartes nicht ein Mensch? Und hat ein Mensch nicht eine Seele *und* einen Körper? Wenn Descartes sagt „(Ego) sum res cogitans“, kann er daher doch offenbar nur meinen, dass zwar sein „Wesenskern“, nicht aber dass er selbst eine *res cogitans* ist. Mir scheint jedoch, dass Descartes das tatsächlich anders sieht. Für ihn ist ein Mensch kein aus Seele und Körper „zusammengesetztes“ Wesen, ein Mensch ist in seinen Augen vielmehr eine Seele (*res cogitans*), die während ihres Erdenlebens sehr eng mit einem Körper verbunden ist. Vgl. z. B. die folgende Passage aus der sechsten Meditation, in der wieder von *seiner* (*Descartes’*) Natur die Rede ist: „Et quamvis fortasse (vel potius, ut postmodum dicam, pro certo) habeam corpus, quod *mihi* valde arcte conjunctum est, quia tamen ex una parte claram & distinctam habeo ideam *mei ipsius*, quatenus sum tantum res cogitans, non extensa, & ex alia parte distinctam ideam corporis, quatenus est tantum res extensa, non cogitans, certum est *me* a corpore meo revera esse distinctum, & absque illo posse existere.“ (ibid., 78 – meine Hervorh.) All dies hängt offensichtlich eng zusammen mit Descartes’ keineswegs immer klarer Position zur *unio substantialis*. In einem sehr schönen Artikel hat Stephen Voss aber ge-

Es zeigt sich, dass sehr viele – vielleicht die meisten – Philosophen nicht der Meinung sind, „Ich“ und „Selbst“ seien gleichbedeutend mit „Person“. Sie verwenden diese Wörter vielmehr, um über den *inneren Wesenskern* von Personen zu reden, wobei allgemein angenommen wird, dieser Wesenkern sei etwas Immaterielles – eine Art Cartesische *res cogitans*. Außerdem scheinen viele zu glauben, dass es nicht die ganze Person, sondern nur dieser innere Wesenkern ist, der über Bewusstsein verfügt, der Erfahrungen macht, nachdenkt und überlegt handelt. So heißt es in dem von G. Schischkoff herausgegebenen *Philosophischen Wörterbuch*:

> **Ich** (lat. *ego*) Ausdruck für den Bewusstseinskern, für den Träger des Selbstbewusstseins der leiblich-seelischen-geistigen Ganzheit des Menschen […]. (Schischkoff, *Philosophisches Wörterbuch*, 319)

Und sogar in der *Microsoft Encarta* kann man lesen:

> **Ich** (lateinisch *ego*), Ausdruck für das Bewusstsein von der eigenen Person in Abgrenzung von der Umwelt, auch Persönlichkeitskern genannt. Das Ich als Träger allen Fühlens, Denkens und Handelns besitzt sowohl Mechanismen zur Kontaktaufnahme wie auch zur Abwehr der Außenwelt. (© 1993–2003 Microsoft Corporation. Alle Rechte vorbehalten.)

Es zeigt sich also, dass die Rede von dem Ich und dem Selbst eine enge Verbindung mit dem Cartesianischen Dualismus eingegangen ist. Zumindest machen die meisten Philosophen, die so reden, einen Unterschied zwischen dem Menschen als ganzem und seinem inneren Wesenskern, wobei nur dieser Wesenskern als Träger von Bewusstsein, Fühlen, Denken und überlegtem Handeln angesehen wird. Nun ist es jedem unbenommen, an die Existenz eines solchen Wesenskerns oder einer Cartesischen Seele zu glauben. Aber man sollte zur Bezeichnung dieses Wesenskerns eben nicht die Wörter „Ich“ und „Selbst“ verwenden (es gibt ja auch genügend andere geeignete Wörter); denn damit öffnet man schwerwiegenden Missverständnissen Tür und Tor.

Die Probleme, die sich ergeben, wenn man versucht, Descartes’ „ego ille“ oder Lockes „self“ in andere Sprachen zu übersetzen, zeigen zunächst, dass die Entwicklung, die hier in Gang gekommen ist, grammatisch gesehen alles andere als unproblematisch ist. Ein Personalpronomen ist nun einmal kein Gattungsbegriff, den man problemlos mit Demonstrativ- und Possessivpronomina sowie mit bestimmten und unbestimmten Artikeln verbinden kann. Dasselbe gilt für die Partikel „selbst“. Wenn man in der

---

zeigt, dass Descartes spätestens seit 1641 nicht mehr die These vertritt, dass der Mensch ein aus Seele und Körper zusammengesetztes Wesen ist. Diese These wäre ja auch mit seiner metaphysischen Grundannahme nicht vereinbar, dass es genau zwei Arten von geschaffenen Substanzen gibt. Und sie würde die Unsterblichkeit des Menschen zum Problem werden lassen.

sprachlichen Entwicklung, die mit Descartes und Locke einsetzt, überhaupt einen Sinn finden will, muss man daher annehmen, dass im Anschluss an diese beiden Autoren die Wörter „ich“ und „selbst“ eine weitere *neue* Bedeutung und Grammatik bekommen haben. „ich“ und „selbst“ werden *auch* zu Gattungsbegriffen, die für eine bestimmte Art von Dingen stehen. (Ihre alte Bedeutung und Grammatik behalten sie trotzdem bei.) Nun könnte man fragen: Was ist eigentlich so schlimm daran? Veränderungen und Weiterentwicklungen der Sprache sind ja nicht verboten. Aber so einfach ist die Sache nicht. Erstens: Es geht hier nicht um eine der üblichen Weiterentwicklungen der Umgangssprache, wie wir sie auch heute tagtäglich beobachten können. Diese werden in aller Regel von den normalen Sprechern vorangetrieben. (Man denke an „simsen“ oder „abhängen“.) Bei der Rede von *dem* Ich und *dem* Selbst handelt es sich dagegen um die Einführung einer neuen *philosophischen Fachterminologie*, die allerdings zugegebenermaßen sehr schnell in die Umgangssprache übernommen wurde. Bei einer solchen fachterminologischen Neuerung kann man aber sehr wohl nach ihrer Berechtigung fragen. Welche Funktion haben die neuen Ausdrücke? Gibt es Probleme oder Sachverhalte, die sich einzig – oder wenigstens deutlich besser – mit Hilfe der neuen Termini darstellen oder analysieren lassen?

Hier kann und muss man, denke ich, erhebliche Zweifel haben. Warum reichen die Begriffe „Person“ und auch „Seele“ nicht aus? Warum redet man nicht einfach wie bisher von Wesen mit der Fähigkeit zur Selbsterkenntnis oder mit der Fähigkeit, Überzeugungen über sich selbst, die eigenen Gedanken und Gefühle zu haben? Das Fehlen einer befriedigenden Antwort auf diese Fragen wiegt umso schwerer, als, zweitens, allgemein gilt: Es ist immer äußerst riskant, *schon vorhandene* Wörter mit einer eingespielten Grammatik und Semantik *zusätzlich* mit einer *neuen* Grammatik und Semantik zu versehen. Unklarheiten und Missverständnisse sind geradezu vorprogrammiert. Wenn Descartes schreibt „Ego sum res cogitans“, wie ist dieser Satz zu verstehen? Ist das „ego“ als Personalpronomen aufzufassen, so dass man den Satz so lesen muss „Ich (Descartes) bin ein denkendes Ding“? Oder steht „ego“ hier für Descartes’ Wesenskern, so dass der Satz bedeutet „Mein Ich (meine Seele) ist ein denkendes Ding“?

Noch deutlicher werden die Probleme, wenn wir noch einmal auf die Kernaussagen von Weber und Siefer zurückkommen – „Es gibt kein Ich“ und „Ich bin niemand“. Was ist mit der ersten Aussage gemeint? Nach dem, was wir bisher über „Ich“ und „Selbst“ gehört haben, gibt es zunächst zwei Lesarten: (a) „Es gibt keine Wesen, die zu Gedanken über sich selbst fähig sind“ und (b) „Menschen haben keinen immateriellen Wesenskern“. In der Lesart (a) ist der Satz „Es gibt kein Ich“ ganz sicher falsch. Denn wir alle wissen von uns selbst ebenso wie von unseren Mitmenschen, dass wir

uns Gedanken über uns selbst machen und uns um uns selbst sorgen können. Leugnen wird das nur, wer fälschlicherweise glaubt, Selbstbewusstsein sei das Bewusstsein, das wir von *unserem Selbst* haben, (und wer zugleich glaubt, dass es kein Selbst gibt). Aber das ist nicht so. Selbstbewusstsein ist das Bewusstsein, das wir von *uns selbst* haben – und das schließt körperliche Eigenschaften genauso ein wie mentale. In der Lesart (b) „Menschen haben keinen immateriellen Wesenskern" kann der Satz „Es gibt kein Ich" aber durchaus wahr sein. Ich selbst würde ihn in dieser Lesart akzeptieren, da ich davon überzeugt bin, dass es keine cartesischen Seelen gibt.[12] Doch das ändert nichts daran, dass es wirklich keine gute Idee ist, die Leugnung des Cartesischen Dualismus in die Formel zu kleiden: *Es gibt kein Ich, es gibt kein Selbst*. Denn diese Sätze können eben auch in der Lesart (a) verstanden werden, in der sie sicher falsch sind.

Schließlich gibt es auch noch eine dritte Lesart, die unter anderem durch den zweiten Satz „Ich bin niemand" nahegelegt wird: (c) „Es gibt mich nicht". Diese Lesart wird insbesondere durch die irrige semantische Annahme gestützt, dass sich das Wort „ich" immer auf ein Ich – auf einen immateriellen Wesenskern – bezieht. Denn wenn das so wäre und wenn es keine solchen immateriellen Wesenskerne gäbe, dann wäre „ich" genauso bezugslos wie die Wörter „Einhorn" und „Pegasus". Doch das ist offenkundiger Unsinn.

Man kann diesen Punkt auch so fassen: Können die folgenden Sätze auch dann wahr sein, wenn es keine immateriellen Wesenskerne oder cartesischen Seelen gibt? Oder sind sie in diesem Fall alle falsch oder sinnlos?

- Ich bin 63 Jahre alt.
- Ich bin Professor für Philosophie an der Universität Bielefeld.
- Ich fühle einen bohrenden Zahnschmerz.
- Ich erinnere mich an den letzten Urlaub.
- Ich überlege, in welches Restaurant wir am Wochenende gehen könnten.
- Ich hebe jetzt meinen Arm.

Jeder, der einigermaßen bei Verstand ist, wird der Auffassung zustimmen, dass diese Sätze zumindest wahr sein können. Aber können sie das wirklich, wenn Siefer und Weber recht haben? Die Schlussfolgerung, dass diese Sätze *nicht* wahr sein können, ergibt sich sofort, wenn man die folgenden Prämissen akzeptiert:

1. Semantische Prämisse: Der Satz „Ich hebe meinen Arm" ist genau dann wahr, wenn mein Ich oder mein Selbst die Aufwärtsbewegung meines Arms (kausal) hervorruft.
2. Metaphysische Prämisse: Es gibt weder ein Ich noch ein Selbst.

[12] Zur Begründung dieser Auffassung vgl. Beckermann 2008, Abs. 1.5.

Doch die semantische Prämisse ist falsch! Dies zeigt sich sofort, wenn man sich die wesentlichen grammatischen und semantischen Tatsachen noch einmal vor Augen führt:

- „ich“ ist als *Personalpronomen* der ersten Person Singular ein *singulärer Term*;
- „ich“ ist ein *indexikalischer* Ausdruck, dessen Bezug sich in Abhängigkeit vom Äußerungskontext ändert;
- „ich“ *bezeichnet immer den Sprecher*, der diesen Ausdruck äußert.

Das Wort „ich“ bezieht sich also nicht auf einen inneren Wesenskern, sondern auf den Sprecher, der dieses Wort äußert. (Und dieser Sprecher existiert natürlich; sonst könnte er das Wort ja nicht äußern. „ich“ kann also gar nicht bezugslos sein.[13])

Wenn Anna sagt: „Ich hebe meinen Arm“, ist dieser Satz also *nicht* dann wahr, wenn die Aufwärtsbewegung ihres Arms durch ihr Ich oder ihr Selbst (kausal) hervorgerufen wird, sondern dann, wenn Anna selbst ihren Arm hebt. Und das ist auch möglich, wenn sie keine cartesische Seele besitzt.[14] In Sätzen wie „Ich hebe meinen Arm“ oder „Ich denke nach“ ist nicht von einem ominösen inneren Wesenskern die Rede, sondern von den Wesen, die diese Sätze äußern – mögen sie nun rein biologische Lebewesen oder Wesen mit einer cartesischen Seele sein.

Trotzdem glauben manche Autoren – allein weil das Wort „ich“ inzwischen sowohl als Personalpronomen als auch als Gattungsbegriff verwendet wird –, „ich“ bezöge sich doch auf einen ominösen Wesenskern – nicht auf mich, sondern auf mein Ich. In der neueren Diskussion des Willensfreiheitsproblems etwa finden sich immer wieder entsprechende Aussagen. So schreibt Gerhard Roth in seinem Buch *Aus Sicht des Gehirns*:

> Heißt dies, dass wir für das, was wir tun, nicht verantwortlich sind? Etwa in dem Sinne: Nicht ich bin es, sondern unbewusst arbeitende Mechanismen in meinem Gehirn sind es gewesen! Die Antwort auf diese Frage ist eindeutig: Das bewusste, denkende und wollende Ich ist nicht im *moralischen* Sinne verantwortlich für dasjenige, was das Gehirn tut, auch wenn dieses Gehirn „perfiderweise“ dem Ich die entsprechende Illusion verleiht. […] Wenn also Verantwortung an *persönliche moralische Schuld* gebunden ist, wie es im deutschen Strafrecht der Fall ist, dann können wir nicht subjektiv verantwortlich sein, weil niemand Schuld an etwas sein kann, das er gar nicht begangen hat und auch gar nicht begangen haben *konnte*. Das Gefühl der persönlichen Schuld, das wir häufig empfinden, wenn wir etwas Unrechtes getan haben, re-

[13] Das ist die Kernidee des Descartesschen *Cogito*; vgl. Beckermann 2004 und Beckermann 2008, Abs. 2.5.

[14] Zu den Bedingungen, die erfüllt sein müssen, damit das der Fall ist, vgl. Beckermann 2008, Abschnitt 2.3, und den Beitrag 16 in diesem Band.

> sultiert aus der irrtümlichen Annahme, wir als bewusstes Ich hätten das Unrecht verursacht. (Roth 2003, 180)

Ich kann nicht für mein Handeln verantwortlich sein, weil dieses Handeln nicht auf mein Ich, sondern auf mein Gehirn zurückgeht (und mir dieses Gehirn darüber hinaus den Streich spielt, mich – mein Ich – glauben zu machen, ich – mein Ich – sei doch der Urheber meiner Handlungen). Das Hinundherspringen zwischen „ich" und „mein Ich" ist unübersehbar. Und es macht deutlich, dass Roth offenbar meint, mit dem Wort „ich" würde sich jeder auf sein Ich beziehen. Demgegenüber muss man noch einmal auf die Semantik von „ich" verweisen. Wenn Hans sagt: „Ich habe mich entschieden, nach Hause zu gehen", dann ist dieser Satz nicht dann wahr, wenn sein innerster Wesenskern einen immateriellen Willensakt vollzogen hat, sondern dann, wenn Hans selbst (nicht: Hans' Selbst) diese Entscheidung getroffen hat – Hans, der ganze Mensch oder die ganze Person. Dabei mag Hans' Gehirn durchaus eine zentrale Rolle spielen; aber es ist eben nicht das Gehirn, das entscheidet, sondern Hans mit Hilfe seines Gehirns. Sowie ja auch nicht die Beine von Hans gehen, sondern Hans mit Hilfe seiner Beine.

Mir scheint, dass das Hauptproblem der neuen Verwendungsweisen von „ich" und „selbst" darin liegt, dass diese Verwendungsweisen zu einer Verdopplung führen – neben dem Wort „ich" gibt es jetzt auch den Ausdruck „mein Ich", neben „ich selbst" auch „mein Selbst". Gefährlich ist diese Verdopplung, weil sie mit unbeantwortbaren Fragen und unlösbaren Problemen verbunden ist. Erstens: Beziehen sich „ich" und „mein Ich" auf dasselbe oder (zumindest potentiell) auf Verschiedenes? Beziehen sich „ich selbst" und „mein Selbst" auf dasselbe oder (zumindest potentiell) auf Verschiedenes? Für die erste Antwort scheint *prima facie* die Verwendung derselben Wörter zu sprechen. Aber auf der anderen Seite: Wie kann etwas *mein* Ich sein, wenn es nicht von mir verschieden ist? Wie kann etwas *mein* Selbst sein, wenn ich nicht etwas anderes bin als mein Selbst? Neben einer Verdopplung der Ausdrücke scheint es also auch eine Verdopplung der Entitäten zu geben. Neben mir (dem, der diese Zeilen schreibt) scheint es auch noch mein Ich zu geben, neben mir selbst auch noch mein Selbst. Doch das ist eine substantielle These, deren Wahrheit nicht einfach durch die Einführung eines neuen Sprachgebrauchs garantiert werden kann.

Zweitens: Wie steht es mit den Sätzen

- Ich bin mein Ich. (Ich = mein Ich)
- Ich bin ein Ich.
- Ich habe ein Ich.
- Ich bin mein Selbst. (Ich = mein Selbst)
- Ich bin ein Selbst.
- Ich habe ein Selbst.

Welche dieser Sätze sind wahr? Welche sind überhaupt sinnvoll? Offenbar ist dies alles andere als klar; und solange das so ist, stiften Ausdrücke wie „mein Ich“ und „mein Selbst“ nur Verwirrung. Natürlich kann man mit Locke und Lowe sagen: „Ich bin ein Wesen, das sich seiner selbst bewusst ist und das sich deshalb um sich selbst sorgt“. Aber ist es wirklich sinnvoll, dies so auszudrücken: „Ich bin ein Selbst“ oder „Ich habe ein Selbst“? Und natürlich kann man mit Descartes sagen: „Ich habe zwar einen Körper, aber eigentlich bin ich nur eine denkende Substanz“. Doch macht es wirklich Sinn, dies so zu formulieren: „Ich bin ein Ich“ oder „Ich habe ein Ich“?

Kurz: Es ist unübersehbar, dass die Tatsache, dass die Wörter „ich“ und „selbst“ inzwischen nicht nur in ihrer ursprünglichen Bedeutung, sondern auch als Gattungsbegriffe verwendet werden, zu einer großen Zahl von Unklarheiten, Missverständnissen und Scheinproblemen führt. Ich denke, wir sollten deshalb in diesem Fall tatsächlich versuchen, das Rad der Geschichte zurückzudrehen. Hören wir auf, „Ich“ und „Selbst“ als Gattungsbegriffe zu verwenden und von *dem* Ich oder *dem* Selbst zu reden. Geben wir den Wörtern „ich“ und „selbst“ – ohne Demonstrativpronomina und Artikel – ihre natürliche Rolle in der Sprache zurück. Schicken wir die Sprache wieder zurück an die Arbeit. Die Rede von *dem* Ich oder *dem* Selbst hat keinen Nutzen, sie erzeugt nur (Schein)-Probleme, wo es in Wirklichkeit gar keine Probleme gibt.

## Literatur

Beckermann, A. (2004) „René Descartes: Die Suche nach den Grundlagen sicherer Erkenntnis“, in: Ansgar Beckermann, Dominik Perler (Hg.), *Klassiker der Philosophie heute*. Stuttgart: Reclam, 208–229. (Auch in A. Beckermann, *Aufsätze 2*, Universitätsbibliothek Bielefeld 2012, Beitrag 7)

– (2008) *Gehirn, Ich, Freiheit. Neurowissenschaften und Menschenbild*. Paderborn: mentis.

– (2010) „Die Rede von *dem* Ich und *dem* Selbst. Sprachwidrig und philosophisch höchst problematisch“, in: K. Crone, R. Schnepf & J. Stolzenberg (Hg.) *Über die Seele*. Frankfurt/M.: Suhrkamp, 458–473.

Blackburn, S. (1994) *The Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

Blume, T. (2003) ‚Ich‘, in: Wulff D. Rehfus (Hg.) *Handwörterbuch Philosophie*. Göttingen: Vandenhoeck & Ruprecht.

Descartes, R. *Discours de la méthode*, in: *Œuvres de Descartes*, herausgegeben von Charles Adam und Paul Tannery, Bd. VI, Paris: Vrin 1983.

– *Meditationes de prima philosophia*, in: *Œuvres de Descartes*, herausgegeben von Charles Adam und Paul Tannery, Bd. VII, Paris: Vrin 1983.

– *Méditations Métaphysique*. Traduction du Duc de Luynes. in: *Œuvres de Descartes*, herausgegeben von Charles Adam und Paul Tannery, Bd. IX, Paris: Vrin 1982.
– *Meditationes de prima philosophia*. *Meditationen über die erste Philosophie*. Lat.-Deutsch, übers. u. hrsg. v. Gerhart Schmidt, Stuttgart: Reclam 1986.
– *Meditationen*. Latein-Französisch-Deutsch, hrsg. v. Andreas Schmidt, Göttingen: Vandenhoeck & Ruprecht 2004.
– *Meditationes de prima philosophia. Meditationen über die Grundlagen der Philosophie*. Lateinisch-Deutsch, übertr. u. hrsg. v. Christian Wohlers, Hamburg: Meiner 2008.
– *Principia Philosophiae*, in: *Œuvres de Descartes*, herausgegeben von Charles Adam und Paul Tannery, Bd. VIII-1, Paris: Vrin 1996.
– *Die Prinzipien der Philosophie*. Lateinisch-Deutsch, übers. u. hrsg. v. Christian Wohlers. Hamburg: Meiner 2005.
– *The Philosophical Works of Descartes*. Vol. I. Transl. by Elizabeth S. Haldane, G. R. T. Ross, Cambridge: Cambridge University Press [11]1977.
– *The Philosophical Writings of Descartes*. Vol. II. Transl. by John Cottingham, Robert Stoothoff & Dugald Murdoch, Cambridge: Cambridge University Press 1984.

Henke, R.W (2003) ‚Selbst', in: Wulff D. Rehfus (Hg.) *Handwörterbuch Philosophie*. Göttingen: Vandenhoeck & Ruprecht.

Locke, J., *An Essay Concerning Human Understanding*. Ed. by Peter H. Nidditch, Oxford: Clarendon Press 1975.
– *Versuch über den menschlichen Verstand*. Übers. von C. Winckler, 4., durchgesehene Auflage in 2 Bänden, Hamburg: Meiner 1981.
– *Essai sur l'entendent humain*, livres I et II, traduction par Jean-Michel Vienne, Paris 2001.

Lowe, E. J., (1995) ‚self', in: Ted Honderich (Hg.) *The Oxford Companion to Philosophy*. Oxford: Oxford University Press.

Pascal, B., *Pensées*, presentation et notes par Gérard Ferreyrolles, texte établi par Philippe Sellier. Paris: Librairie Général Francaise 2000.

Gaius Plinius Caecilius Secundus, *Briefe. Epistularum libri*. Lateinisch-Deutsch. Ausgewählt u. auf d. Grundlage d. Ausg. von Helmut Kasten neu übers. u. hrsg. von Rainer Nickel. Düsseldorf/Zürich: Artemis & Winkler 2000.

Roth, G. (2003) *Aus Sicht des Gehirns*. Frankfurt/M.: Suhrkamp.

Schischkoff G. (Hg.), *Philosophisches Wörterbuch*. Stuttgart: Kröner [22]1991.

Siefer, W. & C. Weber (2006) *Ich – Wie wir uns selbst erfinden*. Frankfurt/M.: Campus.

Voss, S. (1994) „Descartes: The End of Anthropology", in: J. Cottingham (Hg.) *Reason, Will, and Sensation.* Oxford: Clarendon Press, 273–315.

Wittgenstein, L., *Philosophische Untersuchungen*, in: Ludwig Wittgenstein, *Werkausgabe, Band 1*. Frankfurt/M.: Suhrkamp 1984.

# 16

# *Ich* sehe den blauen Himmel, *ich* hebe meinen Arm[*1]

## 1. Steile Thesen und besonnene Reaktionen

In letzter Zeit hat es auch in der Philosophie Kollegen gegeben, die versucht haben, mit gewagten und zum Teil absurden Thesen die Aufmerksamkeit der Medien zu erlangen, die an nüchterner philosophischer Analyse kaum interessiert sind. Thomas Metzinger etwa ist es mit spektakulären Formulierungen wie „Sie sind niemand" oder „Jeder Mensch verwechselt sich mit seinem Selbstmodell" gelungen, auch in einer breiteren Öffentlichkeit Resonanz zu finden.[2] Solch steile Thesen findet man ansonsten bei einigen Naturwissenschaftlern, denen es ebenfalls um öffentliche Aufmerksamkeit geht,[3] oder bei Wissenschaftsjournalisten, die sich offenbar hohe Auflagen versprechen.[4]

Wolfgang Lenzen gebührt das Verdienst, diese zum Teil wirklich schrägen Thesen nicht einfach zu ignorieren, sondern sorgfältig zu analysieren, Richtiges von Falschem zu unterscheiden und damit einen wichtigen Beitrag zur Aufklärung darüber zu leisten, wie man die Dinge sinnvoll und nüchtern betrachten sollte.[5] In seinem Aufsatz „Auf der Suche nach dem verlorenen ‚Selbst'" setzt sich Lenzen unter anderem mit den Thesen auseinander, das Ich sei eine bloße Illusion; das Ich sei nichts als ein Konstrukt unseres Gehirns; die Annahme, wir hätten ein Ich, beruhe allein darauf, dass wir uns ständig mit unserem Selbstmodell verwechseln.

Lenzen zeigt, dass diese Thesen häufig auf einigen leicht zu durchschauenden Missverständnissen beruhen. Am Beginn der Überlegungen der Autoren, die er kritisiert, stehen häufig Annahmen, die man – so oder in leicht variierter Form – durchaus akzeptieren kann.

---

* Erstveröffentlichung in: Ch. Lumer & U. Meyer (Hg.) *Geist und Moral*. Paderborn: mentis 2011, 19–34.

1 In diesem Aufsatz nehme ich Überlegungen aus Beckermann 2008, 2009, 2010 auf, um sie ein Stück voranzubringen. Allerdings befürchte ich, dass es auch am Schluss dieses Aufsatzes Fragen geben wird, auf die noch keine vollständig befriedigende Antwort gefunden wurde.

2 Metzinger 2003.

3 Z. B. Roth 2003.

4 Etwa Siefer/Weber 2006.

5 Lenzen 2002, 2005, 2006.

(1) Die äußere Welt wird von einem Subjekt S in Gestalt eines Weltmodells mental repräsentiert.

(2) Das Subjekt selbst (und insbesondere der Körper von S) wird durch ein entsprechendes Selbstmodell repräsentiert. (Lenzen 2006, 168)

In einem zweiten Schritt stellen diese Autoren dann fest, dass es sich bei diesen Modellen oder Repräsentationen um ‚Konstrukte' handelt. Dieser Konstruktionsgedanke ist in gewissem Sinn trivial. Natürlich muss sich ein Subjekt die Repräsentationen seiner Umwelt und seines eigenen Körpers selbst erarbeiten. Sie fliegen ihm nicht auf magische Weise zu. Sie beruhen auf internen Informationsverarbeitungsprozessen, die dazu dienen, aus den kausalen Spuren, die z. B. das einfallende Licht auf der Netzhaut der Augen hinterlässt, die Umwelt, von der das reflektierte Licht ausgeht, zu *re*konstruieren, wie ich lieber sagen würde. Wahrnehmung, auch Selbstwahrnehmung, ist also ohne Zweifel ein aktiver Prozess. Aber berechtigt dies die kritisierten Autoren, von den Annahmen (1) und (2) zu den folgenden Thesen überzugehen?

(1′) Die äußere Welt ist eine Simulation (des menschlichen Gehirns).

(2′) (Auch) Menschen sind Simulationen (ihrer eigenen Gehirne). (Ebd., 163)

Der Ausdruck ‚Simulation' suggeriert zumindest, dass die hier in Rede stehenden Modelle und Repräsentationen sämtlich falsch oder illusionär sind. Die Annahmen (1) und (2) wären also letztlich so zu lesen:

(1″) Das Subjekt erzeugt ein illusionäres Bild seiner Außenwelt.

(2″) Das Subjekt erzeugt ein illusionäres Bild seines Körpers und seiner selbst.

Oder sogar:

(1‴) Die Außenwelt ist eine vom Subjekt erzeugte Illusion.

(2‴) Der Körper des Subjekts oder sogar das Subjekt selbst ist eine vom Subjekt erzeugte Illusion.

Die Tatsache, dass Modelle und Repräsentationen ‚konstruiert', d. h. aktiv erarbeitet werden, berechtigt jedoch ganz sicher nicht zu dem Schluss, dass die Ergebnisse dieses Konstruktionsprozesses allesamt illusionär sind. Auch die Tatsache, dass wir bei der Konstruktion von Modellen und Repräsentationen manchmal Fehler machen, rechtfertigt diesen Schluss nicht. Erstens, so Lenzen völlig einleuchtend, sollten wir die Außenwelt nicht mit unserem Bild der Außenwelt verwechseln; dieses Bild kann fehlerhaft oder illusionär sein, doch dadurch wird die Außenwelt selbst nicht zu einer Illusion. Und zweitens: Die Absurdität besonders der These (2‴) ist geradezu

mit Händen zu greifen: Wie kann eine Illusion eine Illusion von sich selbst erzeugen? Wenn sich irgendein X ein Modell von seiner Umwelt und von sich selbst macht, ist eines sicher – X existiert und ist nicht niemand![6] Man könnte es mit dieser schlichten Feststellung bewenden lassen. Ich möchte im Folgenden aber der Frage nach der Existenz des Ich und der These, dieses Ich sei nichts als eine Illusion, noch etwas weiter nachgehen.

## 2. ‚Ich' und ‚Selbst'

„In der klassischen antiken und mittelalterlichen Philosophie ist der philosophische Begriff des Ich kaum vorhanden […]." (Herring/Schönpflug 1976, 1) Mit dieser ebenso lapidaren wie zutreffenden Bemerkung beginnt der Artikel „Ich, Abschn. I" von H. Herring im *Historischen Wörterbuch der Philosophie*. Erst in der Neuzeit beginnt die Rede von *dem* Ich und *dem* Selbst erst zögernd, dann aber rasant um sich zu greifen. Dabei scheint es keine Rolle zu spielen, dass auf diese Weise dem eher unschuldigen Personalpronomen ‚ich' und der ebenso unschuldigen Partikel ‚selbst' erhebliche Gewalt angetan wird.[7] Was ist der Grund für diese Entwicklung? Herring deutet an, dass die für Menschen charakteristische Fähigkeit der reflexiven Selbsterkenntnis, des Selbstbewusstseins eine entscheidende Rolle gespielt hat. Besonders deutlich wird das bei Locke, der den Ausdruck ‚Selbst' so einführt:

> *Self* is that conscious thinking thing, (whatever Substance made up of whether Spiritual or Material, Simple or Compounded, it matters not), which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern'd for it *self*, as far as that consciousness extends. (Locke, *Essay*, II, xxvii, §17)

Für Locke ist ein Selbst also einfach ein bewusstes denkendes Wesen, das Lust und Schmerz fühlen sowie glücklich und unglücklich sein kann und das sich deshalb um sich selbst sorgt. Ein Selbst oder Ich scheint somit ein Wesen zu sein, dass zu reflexiver Erkenntnis und Selbstbewusstsein fähig ist und dass sich deshalb über sich selbst Gedanken machen kann – Gedanken, die es sprachlich unter Verwendung des Wortes ‚ich' ausdrückt. So ähnlich sieht es auch E. J. Lowe, der in einem Lexikonartikel schreibt:

> **self** The term ‚self' is often used interchangeably with ‚person', though usually with more emphasis on the ‚inner', or psychological, dimension of personality than on outward bodily form. Thus a self is conceived to be a subject of consciousness, a being capable of thought and experience and able to engage in deliberative action. More crucially, a self must have a capacity

---

[6] Lenzen 2006, 163.

[7] Vgl. Beckermann 2010; in diesem Band Beitrag 15.

> for self-consciousness, which partly explains the aptness of the term ‚self'. Thus a self is a being that is able to entertain first-person thoughts (Lowe 1995, 816 f.)

Locke lässt bewusst offen, ob es sich bei einem Selbst um ein geistiges oder ein körperliches Wesen handelt. Und auch Lowe äußert sich in dem zitierten Artikel nicht zu dieser Frage. Von sehr vielen wird die Fähigkeit zur Selbsterkenntnis aber allein der Seele zugesprochen. Leibniz spricht sie sogar nur den Seelen vernünftiger Lebewesen zu, den von ihm so genannten ‚Geistern':

> [M]ais ceux [animaux], qui connoissent ces verités necessaires, sont proprement ceux qu'on appelle *Animaux Raisonnables*, et leurs ames sont appellées *Esprits*. Ces Ames sont capables de faire des Actes reflexifs, et de considerer ce qu'on appelle Moy, Substance, Ame, Esprit […]. (Leibniz 1714, 158)

So ist es kein Wunder, dass es neben der Auffassung von Locke und Lowe auch noch einen engeren Begriff von Ich und Selbst gibt, der inzwischen vielleicht sogar der Mehrheitsmeinung entspricht. R.W. Henke etwa charakterisiert den Begriff des Selbst in einem neueren Lexikon so:

> **Selbst** Bezeichnung für den innersten Wesenskern der Persönlichkeit (→Ich), der auf der Möglichkeit, sich seiner selbst bewusst zu werden […], beruht. Insofern kennzeichnet der Begriff in religiöser Hinsicht auch den wahren, alle wechselnden Lebenserscheinungen überdauernden Kern des Menschen – als *atman* oder →Seele. (Henke 2003, 609 f.)

Entsprechend schreibt Thomas Blume in demselben Lexikon zum Begriff des Ich:

> **Ich** […] An der Auffassung einer allen Bewusstseinszuständen zugrunde liegenden Seelensubstanz, welche mit dem Ich identifiziert wird, entzündet sich die Kritik →Humes. (Blume 2003, 394)

Blume und Henke zufolge bezeichnen ‚Ich' und ‚Selbst' also nicht einfach zur Selbsterkenntnis fähige Wesen, sondern den seelischen Wesenskern solcher Wesen, wobei offenbar wie bei Leibniz vorausgesetzt wird, dass es dieser Wesenskern ist, der ein reflexives sich auf sich selbst Beziehen überhaupt erst möglich macht.

## 3. ‚Ich' und ‚ich'

Dass sich in der Neuzeit die Rede von *dem* Ich und *dem* Selbst erstaunlich schnell verbreitet, bedeutete zunächst einmal einen erheblichen Eingriff in die bestehenden Umgangssprachen. Das Personalpronomen ‚ich' hatte und hat eine klare und eingespielte Rolle im Deutschen (ebenso wie die entsprechenden Ausdrücke ‚ego', ‚I', ‚je' bzw. ‚moi' im Lateinischen, Engli-

schen und Französischen); dasselbe gilt für die Partikel ‚selbst', die zwar nur ein synkategorematischer Ausdruck ist, aber als Fokuspartikel, Pronomen und Adverb ebenfalls klar umrissene semantische Funktionen besitzt.[8] Mit der Rede von *dem* Ich und *dem* Selbst bekommen die Ausdrücke ‚ich' und ‚selbst' jedoch eine neue Bedeutung – und das heißt auch: Sie werden mehrdeutig. Wenn man beginnt, ‚ich' und ‚selbst' mit bestimmten und unbestimmten Artikeln sowie mit Demonstrativ- und Possessivpronomina zu verbinden, heißt das, dass man sie als Artbegriffe verwendet – wie ‚Hund', ‚Haus' oder ‚Buch'.[9] Erst dadurch werden Sätze möglich wie „Welche Wesen haben ein Selbst?", „Jeder Mensch hat ein Ich", „Können Computer ein Ich ausbilden?" usw. Und genau so hatte Locke ja ‚self' auch eingeführt als einen Ausdruck, der eine bestimmte Art von Wesen bezeichnet.[10]

Nun könnte man – gegen orthodoxe Vertreter der *ordinary language* Philosophie – ins Feld führen, dass Veränderungen und Weiterentwicklungen der Sprache natürlich nicht verboten sind. Aber so einfach ist die Sache nicht. Erstens: Es geht hier nicht um eine der üblichen Weiterentwicklungen der Umgangssprache, wie wir sie auch heute tagtäglich beobachten können. Diese werden in aller Regel von den normalen Sprechern vorangetrieben. (Man denke an ‚simsen' oder ‚abhängen'.) Bei der Rede von *dem* Ich und *dem* Selbst handelt es sich dagegen um die Einführung einer neuen *philosophischen Fachterminologie*, die allerdings zugegebenermaßen sehr schnell in die Umgangssprache übernommen wurde. Bei einer solchen fachterminologischen Neuerung kann man aber sehr wohl nach ihrer Berechtigung fragen. Welche Funktion haben die neuen Ausdrücke? Gibt es Probleme oder Sachverhalte, die sich einzig – oder wenigstens deutlich besser – mit Hilfe der neuen Termini darstellen oder analysieren lassen? Hier kann man, denke ich, erhebliche Zweifel haben. Warum reichen die Begriffe ‚Person' und eventuell auch ‚Seele' nicht aus? Warum redet man nicht einfach wie bisher von Wesen mit der Fähigkeit zur Selbsterkenntnis oder mit der Fähigkeit, Überzeugungen über sich selbst, die eigenen Gedanken und Gefühle zu haben? Das Fehlen einer befriedigenden Antwort

---

[8] Das Wort ‚ich' ist als Personalpronomen ein *singulärer* Ausdruck, der sich immer auf eine Sprecherin/einen Sprecher bzw. eine Denkerin/einen Denker bezieht; als indexikalischer Ausdruck bezieht sich ‚ich' immer auf die/den, die/der diesen Ausdruck äußert oder einen entsprechenden Gedanken fasst. Zur Grammatik und Semantik von ‚selbst' vgl. Beckermann 2010, 458 f., in diesem Band S. 291 f.

[9] Im Deutschen geht dies damit einher, dass beide Wörter in der neuen Bedeutung groß geschrieben werden.

[10] Bemerkenswerterweise ist es aber im Deutschen nicht möglich, die entsprechenden Pluralformen zu bilden (‚Iche', ‚Selbste'), was bei ‚self' im Englischen durchaus möglich ist (‚selves').

auf diese Fragen wiegt umso schwerer, als, zweitens, allgemein gilt: Es ist immer äußerst riskant, *schon vorhandene* Wörter mit einer eingespielten Grammatik und Semantik zusätzlich mit einer *neuen* Grammatik und Semantik zu versehen. Unklarheiten und Missverständnisse sind geradezu vorprogrammiert. Wenn Descartes schreibt „Sum res cogitans", was bedeutet dieser Satz? Ich (Descartes) bin ein denkendes Ding; oder: Mein Ich (meine Seele) ist ein denkendes Ding? Wenn George Berkeley schreibt, „What I am myself – that which I denote by the term *I* – is the same with what is meant by *soul* or *spiritual substance*" (Principles, sec. 139), bedeutet das: Wenn George Berkeley sagt „Ich bin George Berkeley", bezeichnet das Wort ‚ich' in dieser Äußerung eine Seele bzw. eine geistige Substanz? Oder: Das, was ich mit dem Wort ‚ich' bezeichne, nämlich meine Seele, ist eine geistige Substanz? Wenn jemand sagt „Es gibt kein Ich", was will er damit ausdrücken? Es gibt keine Wesen, die zu Gedanken über sich selbst fähig sind; oder: Menschen haben keinen immateriellen Wesenskern; oder: Es gibt mich nicht?

Besonders kritisch wird die Situation, wenn die Wörter ‚ich' und ‚selbst' in einem Satz sowohl in der alten als auch in der neuen Bedeutung vorkommen. Lockes Definition ist ein schönes Beispiel dafür. Denn dort heißt es gegen Ende ja „is concern'd for it *self*" und nicht „is concern'd for its *self*" (was auch ganz unsinnig wäre). Auch „Das Ich ist eine Illusion, die ich jederzeit selbst hervorrufe" gehört in diese Reihe. Und wie ist es schließlich mit dem Satz „Ich bin niemand"? Soll das wirklich heißen, dass ich nicht existiere, oder dass ich mich mit dem Wort ‚ich' nicht auf mich beziehen kann, oder dass ich keine auf mich selbst bezogenen Gedanken haben kann, oder nur, dass ich keinen immateriellen Wesenskern besitze? Mit einem Wort: Wenn wir zusätzlich zu der herkömmlichen Verwendung der Wörter ‚ich' und ‚selbst' die Rede von *dem* Ich und *dem* Selbst einführen, handeln wir uns eine Unmenge selbst gemachter Probleme ein, die allesamt durch einen Verzicht auf das groß geschriebene ‚Ich' und das groß geschriebene ‚Selbst' vermieden werden könnten. Die Rede von *dem* Ich und *dem* Selbst hat keinen erkennbaren Nutzen; sie schafft nur Probleme, die wir ohne sie nicht hätten.

## 4. Das Gehirn, die Welt, das Ich

In Gerhard Roths Buch *Fühlen, Denken, Handeln* findet sich die folgende bemerkenswerte Passage:

> [Die] erlebte Welt wird von unserem Hirn in mühevoller Arbeit über viele Jahre hindurch konstruiert und besteht aus den Wahrnehmungen, Gedanken, Vorstellungen, Erinnerungen, Gefühlen, Wünschen und Plänen, die unser Gehirn hat. Innerhalb dieser Welt bildet sich […] langsam ein Ich aus, das sich zu-

> nehmend als vermeintliches Zentrum der Wirklichkeit erfährt, indem es den Eindruck entwickelt, es „habe" Wahrnehmungen (d. h. dass Wahrnehmungen auf es bezogen sind), es sei Autor der eigenen Gedanken und Vorstellungen, es rufe aktiv die Erinnerungen auf, es bewege den Arm, die Lippen, es besitze diesen bestimmten Körper, und so fort. Selbstverständlich ist dies eine Illusion, denn Wahrnehmungen, Gefühle, Intentionen und motorische Akte entstehen innerhalb der Individualentwicklung, lange bevor das Ich entsteht. (Roth 2003, 395 f.)

Roth beschreibt hier einen realen Vorgang; aber er beschreibt ihn so, dass man keine Chance hat zu verstehen, was wirklich vorgeht. Zunächst finden wir wieder die von Wolfgang Lenzen diagnostizierte Verwechslung unserer Repräsentationen der Welt mit der Welt selbst. Die Welt selbst besteht *nicht* aus Wahrnehmungen, Gedanken, Vorstellungen, Erinnerungen, Gefühlen, Wünschen und Plänen – auch die erlebte Welt nicht. Wahrnehmungen, Gedanken, Vorstellungen und Erinnerungen haben eher etwas mit unseren Repräsentationen der Welt zu tun. Das Gehirn ‚konstruiert' also keine Welt; es versucht vielmehr, wenn man überhaupt so reden will,[11] sich ein Bild von dieser Welt zu machen. Innerhalb dieses Bildes, so Roth weiter, taucht irgendwann ein Ich auf, das sich selbst für das Zentrum der Welt hält und das irrtümlicherweise der Meinung ist, es selbst nehme wahr, habe Gedanken und Vorstellungen, spreche und bewege die Gliedmaßen seines Körpers. Wieder die Verwechslung von Repräsentation und Repräsentiertem. Es ist ja nicht das Ich, das auftaucht. Vielmehr erweitert das Gehirn sein *Bild* der Welt. Es entwickelt die Idee, dass in der Welt ein Wesen vorkommt, das sich im Zentrum der Welt sieht und das von sich glaubt, die Wahrnehmungen zu haben und die Bewegungen zu initiieren, die nach Roth letztlich auf das Gehirn zurückgehen. Spätestens jetzt kommt man nicht mehr weiter, ohne das Verhältnis von Mensch und Gehirn zu klären. Zuvor aber noch eine Bemerkung zu Roths Hauptargument. Roth sagt, es könne gar nicht sein, dass es das Ich ist, das Wahrnehmungen hat, das der Autor der eigenen Gedanken und Vorstellungen ist, das aktiv Erinnerungen aufruft und das den Arm und die Lippen bewegt, weil „Wahrnehmungen, Gefühle, Intentionen und motorische Akte […] innerhalb der Individualentwicklung [entstehen], lange *bevor* das Ich entsteht" (meine Hervorh.). Hier fällt Roth offenbar in die selbst gegrabene Grube. Denn, wie gesagt, nicht das Ich entsteht relativ spät, sondern die Repräsentation des Ich. Und damit ist natürlich nicht ausgeschlossen, dass das Ich selbst schon früher Gedanken hatte, Erinnerungen hervorrief und Bewegungen initiierte. Offenbar ist es höchste Zeit, dieses ganze Gewirr von irreführenden Formulierungen und nur scheinbar plausiblen Argumenten aufzulösen.

---

[11] Ich übernehme hier zunächst Roths Redeweise von dem Ich und vom Gehirn als Akteur, um sie später zurechtzurücken.

Was genau passiert, wenn sich, so die Rothsche Formulierung, innerhalb der vom Gehirn konstruierten Welt langsam ein Ich ausbildet? Ich denke, dass Roth nichts anderes meinen kann als das Folgende: Zunächst repräsentiert das Gehirn seine Umwelt; es entwickelt Gedanken wie „Da drüben steht ein Baum", „Hinter dem Baum fließt ein Fluss" und „An dem Fluss stehen einige Hütten" und es kann Erinnerungen abrufen wie „Gestern stand ein Reh am Fluss".[12] Im Laufe der Zeit kommt eine neue Art von Gedanken hinzu wie „Zwischen *mir* und dem Fluss steht eine Mauer" oder „*Ich* war gestern am Fluss". Worauf bezieht sich der Ausdruck ‚ich' in diesen Gedanken? Von wem oder was handeln diese Gedanken? Roth scheint davon auszugehen, dass sich das ‚ich' hier – im Sinne der ‚neuen' Lesart – auf einen immateriellen Wesenskern bezieht. Und da es seiner Meinung nach einen solchen Wesenskern zwar gibt (schließlich wird das Ich vom Gehirn ‚konstruiert'), dieser aber weder Gedanken hat noch Erinnerungen abruft noch die Lippen bewegt (denn das alles tut allein das Gehirn), sind alle genannten Gedanken illusionär. Aber warum sollen wir glauben, dass sich das ‚ich' in dem Gedanken „Ich war gestern am Fluss" auf einen immateriellen Wesenskern bezieht? Viel näher liegt doch, das ‚ich' in diesem Gedanken ganz traditionell als Personalpronomen zu verstehen, so dass man den Inhalt dieses Gedankens auch so ausdrücken kann „Ich – der Denker, dieses Gedankens – war gestern am Fluss". Roth wird ja wohl nicht bestreiten, dass es dort, wo es Wahrnehmungen, Gedanken und Erinnerungen gibt, auch jemanden (etwas?) gibt, der wahrnimmt, denkt, sich erinnert. Sein Default-Wert für dieses Etwas scheint das Gehirn zu sein. Kommen wir also endlich zum Verhältnis von Mensch und Gehirn.

Ohne Frage haben Menschen ein Gehirn, so wie sie ein Herz, eine Leber und einen Magen haben; das Gehirn ist eines der Organe des Menschen.[13] Und ebenfalls ohne Frage hat das Gehirn sehr viel mit dem zu tun, was wir wahrnehmen, denken und fühlen. Aber ist es wirklich das *Gehirn*, das wahrnimmt, denkt und fühlt? Wenn wir uns an unserem alltäglichen Sprachgebrauch orientieren, können wir feststellen, das wir manchmal sa-

---

[12] Wenn ich von dem Gedanken „…" spreche, steht „…" für den Inhalt dieses Gedankens.

[13] Da nach Roth die Wirklichkeit ein ‚Konstrukt' des Gehirns ist, gesteht er zu, dass zumindest das Gehirn nicht Teil der Wirklichkeit, sondern Teil der Realität ist. Das Gehirn gibt es also, unabgängig von der subjektiven Wirklichkeit. Lenzen hat jedoch klar gemacht, dass das Gehirn allein wohl kaum existieren kann. „Aufgrund elementarer biologischer Annahmen können Gehirne kaum ohne entsprechenden *Körper* existieren: […] In der Realität wird das *reale* Gehirn von einem *realen* Organismus mit *realem* Blut versorgt." (Lenzen 2006, 175 f.) Wir können also getrost davon ausgehen, dass nicht nur das Gehirn, dass vielmehr der ganze Mensch mit all seinen Organen *real* ist.

gen, dass *der ganze Mensch* etwas tut: „Hans sitzt auf einem Hocker", „Anna hat Frieda etwas zugeflüstert". Manchmal sagen wir auch, dass *ein Organ* etwas tut: „Sein Herz schlägt unregelmäßig", „Seine Hände zittern". Und manchmal sagen wir, dass *in einem Organ* etwas geschieht: „In der Niere wird das Blut von Giftstoffen gereinigt", „In der Lunge nimmt das Blut Sauerstoff auf".

Wie ist es nun mit dem Wahrnehmen, Erinnern, Denken und dem Sich-Bewegen? Ist es das Gehirn, das wahrnimmt, sich erinnert, denkt und Bewegungen ausführt? Oder ist es nicht doch der ganze Mensch, dem wir diese Tätigkeiten zuschreiben müssen? Einige Dinge sind klar: Offenbar sind es nicht die Beine, die laufen, sondern der Mensch, der mit Hilfe seiner Beine läuft; es ist nicht das Auge, das sieht, sondern der Mensch, der mit Hilfe seiner Augen sieht. Und natürlich ist es nicht die Lunge, die atmet, sondern der ganze Mensch, der mit Hilfe seiner Lunge atmet. In all diesen Fällen handelt es sich um Tätigkeiten, die den ganzen Menschen betreffen. Wenn ein Mensch läuft, bewegen sich nicht nur seine Beine, sondern der ganze Mensch; wenn ein Mensch sieht, ist auch das ein Vorgang, der den ganzen Menschen betrifft; und ähnlich ist es beim Atmen. Atmen heißt Luft in die Lunge ziehen, damit das Blut im gesamten Organismus mit Sauerstoff angereichert werden kann, und diese Luft wieder auszustoßen, um das bei Verbrennungsvorgängen entstandene $CO_2$ zu entsorgen. Mit dem Wahrnehmen, Erinnern und Denken ist es so wie mit dem Laufen, Sehen und Atmen. Es ist nicht das Gehirn, das wahrnimmt, sondern der ganze Mensch, der mit Hilfe seiner Sinnesorgane und seines Gehirns wahrnimmt; es ist nicht das Gehirn, das sich erinnert, sondern der ganze Mensch, der sich mit Hilfe seines Gehirns erinnert; und es ist nicht das Gehirn, das überlegt, sondern der ganze Mensch, der mit Hilfe seines Gehirns überlegt. Im Gehirn laufen neuronale Prozesse ab, ohne die wir nicht wahrnehmen, uns erinnern, denken oder unsere Hand bewegen können. Aber das bedeutet nicht, dass das Gehirn selbst wahrnimmt, sich erinnert, denkt oder meine Hand bewegt. Wahrnehmen, Erinnern, Denken und sich Bewegen sind Tätigkeiten des ganzen Menschen und nicht Tätigkeiten eines seiner Organe. Allerdings soll damit nicht geleugnet werden, dass für das, was wir wahrnehmen, erinnern, denken und tun, die Prozesse, die in unserem Gehirn oder dem ganzen ZNS und den Sinnesorganen ablaufen, von entscheidender Bedeutung sind.

Nach diesen Klarstellungen können wir den Vorgang, auf den sich Roth in der am Anfang des Abschnitts zitierten Passage bezieht, auf andere und, wie ich hoffe, zutreffendere Weise darstellen:

1. Menschen sind kognitive Wesen – Wesen, die versuchen, sich ein Bild ihrer Umwelt zu machen, ihre Umwelt zu repräsentieren. Bei diesem Repräsentationsprozess, der ganz ohne Zweifel auf neuronalen Prozessen in ihren Gehir-

> nen beruht, geht es primär um die Fragen: Welche Dinge befinden sich in der Umgebung? Was sind das für Dinge (Bäume, Flüsse, Tiere)? Welche Eigenschaften haben diese Dinge? In welchen Beziehungen stehen diese Dinge zueinander und zu mir? 2. Menschen repräsentieren ihre Umwelt, um in ihr handeln zu können. Sie haben Wünsche und Ziele und sind in der Lage zu überlegen, wie sich diese Wünsche und Ziele am besten erreichen lassen. Auch die Wünsche und Ziele und die Überlegensprozesse haben eine neuronale Grundlage. 3. Irgendwann realisieren Menschen, dass *sie selbst Teil der Welt* sind, die sie zu repräsentieren versuchen;[14] sie erkennen, dass sie selbst in der Umwelt situiert sind, dass sie selbst sich in dieser Umwelt bewegen und dass sie selbst es sind, die die Umwelt repräsentieren. Technisch gesprochen: Sie entwickeln ein Selbstmodell, das in ihr Weltmodell eingebettet ist.

Auch wenn man zugesteht, dass die Repräsentationen der Welt, die Menschen entwickeln, ‚konstruiert', d. h. aktiv erarbeitet werden, und auch wenn man zugesteht, dass es bei diesem Prozess zu Fehlern kommen kann, spricht nichts dafür, dass unsere Repräsentationen *insgesamt* fehlerhaft oder illusionär sind. Schließlich handeln wir auf der Grundlage unserer Repräsentationen, d. h., wir würden ständig scheitern, wenn sie alle falsch wären. (Nur Mister Magoo gelingt es, trotz ständig fehlerhafter Repräsentationen in der Welt zurecht zu kommen.[15]) Und es spricht auch nichts dafür, dass unser Selbstmodell grundsätzlich illusionär ist. Wir müssen nur verstehen, dass dieses Selbstmodell nicht von einem in der Tat illusionären immateriellen Wesenskern handelt, sondern von uns selbst – Menschen aus Fleisch und Blut, die allerdings in der Lage sind, sich ein Bild von ihrer Welt und von sich selbst zu machen. Noch einmal: Es ist nicht richtig zu sagen, dass sich in der „Welt […] langsam ein Ich aus[bildet], das sich zunehmend als vermeintliches Zentrum der Wirklichkeit erfährt, indem es den Eindruck entwickelt, es ‚habe' Wahrnehmungen […], es sei Autor der eigenen Gedanken und Vorstellungen, es rufe aktiv die Erinnerungen auf, es bewege den Arm, die Lippen, es besitze diesen bestimmten Körper, und so fort." Richtig ist vielmehr, dass wir bei dem Versuch, uns ein Bild von der Umwelt zu machen, erst allmählich ein Selbstmodell als Teil unseres Weltmodells entwickeln – ein Selbstmodell, das sich nicht auf irgendein

---

[14] In Beckermann 2005 habe ich diesen Prozess genauer analysiert; in diesem Band Beitrag 13.

[15] Quincy Magoo ist eine Zeichentrickfigur, die 1949 vom Animationsstudio United Productions of America geschaffen wurde. Magoo ist ein älterer, extrem kurzsichtiger Mann, der aufgrund seiner Sehbehinderung in die unglaublichsten Situationen gerät; so hält er in der ersten Episode einen Bären für seinen Neffen Waldo. Trotz seiner Fehlwahrnehmungen gelingt es ihm aber immer wieder, alle misslichen Situationen zu meistern. (http://de.wikipedia.org/wiki/Mister_Magoo – 5.8.2010, 17.15 Uhr)

mysteriöses Ich, sondern schlicht auf uns selbst bezieht. Genau so ist es nicht richtig, zu behaupten, wir glaubten nur deshalb, wir hätten ein Ich, weil wir uns ständig mit unserem Selbstmodell verwechseln. Mag sein, dass manche glauben, sie hätten ein Ich, einen immateriellen Wesenskern; doch das liegt nicht daran, dass sie sich mit ihrem Selbstmodell verwechseln, sondern daran, dass manche Philosophen ihnen das eingeredet haben. Auch wenn es keinen solchen immateriellen Wesenskern gibt, ist es aber völlig legitim, wenn ich jetzt denke „Ich sitze hier am Schreibtisch", „Ich sehe, dass die Sonne scheint" und „Ich denke darüber nach, warum manche meinen, sie hätten einen immateriellen Wesenskern". Auch wenn es kein Ich gibt, sind solche *de se*-Gedanken, Gedanken, die sich auf den Denker selbst beziehen, absolut in Ordnung. Denn sie beziehen sich nicht auf mein Ich, sondern auf mich – ein Wesen aus Fleisch und Blut, das sich Gedanken über die Welt und sich selbst macht. Und dieses Wesen existiert!

## 5. Was man selbst tut

Nach dem bisher Gesagten ist eine Äußerung des Satzes „Ich hebe meinen Arm" genau dann wahr, wenn der Mensch, der diesen Satz äußert, selbst seinen Arm hebt. Aber wann genau ist das der Fall? Wenn man die Bewegungen eines Menschen betrachtet, gibt es solche, von denen man zu Recht sagen kann, dass der Mensch selbst sie ausführt – er hebt seinen Arm; er singt ein Lied; er kratzt sich am Kopf. Aber es gibt auch Bewegungen, bei denen das offensichtlich nicht der Fall ist – jemand nimmt meinen Arm und zieht ihn nach oben; jemand stößt mich zu Boden. In beiden Fällen können die Bewegungen meines Arms oder meines Körpers *nicht* mir selbst zugeschrieben werden. Was ist der Unterschied zwischen diesen beiden Arten der Bewegung? Wer an die Existenz immaterieller Wesenskerne glaubt, scheint hier eine auf den ersten Blick plausible Antwort parat zu haben: Ein Mensch führt genau die Bewegungen selbst aus, die durch sein immaterielles Ich oder Selbst verursacht werden.[16] Aber diese Antwort leidet nicht nur daran, dass es viele Gründe gibt, nicht an die Existenz eines solchen immateriellen Selbst zu glauben; die Antwort passt auch nicht zu unseren alltäglichen Unterscheidungen zwischen dem, was jemand selbst tut, und dem, was mit ihm geschieht. Im Alltag sagen wir nämlich nicht nur „Hans (selbst) hebt seinen Arm", „Hans geht über die Straße" und „Hans hält einen Vortrag", sondern auch „Hans stolpert", „Hans atmet" und „Hans niest". Sollen wir wirklich annehmen, dass auch das Stolpern, das Atmen und das Niesen durch Hans' immaterielles Selbst verursacht werden? Oder

[16] Diese Antwort kann man z. B. bei Descartes finden (*Leidenschaften der Seele*, Teil I, Artikel 41); allerdings spricht Descartes weder vom Ich noch vom Selbst, sondern schlicht von der Seele.

dass die zuletzt genannten Sätze alle falsch sind? Beides klingt nicht besonders plausibel.

Wenn wir Beispiele dafür suchen, was Menschen selbst tun, denken wir oft an Fälle eindeutig absichtlicher Handlungen, die jemand um eines bestimmten Zweckes willen ausführt – er geht in die Küche, um sich ein Glas Wasser zu holen; er hebt den Arm, um auf sich aufmerksam zu machen; er redet auf einen anderen ein, um ihn von etwas abzubringen. Dies legt eine Antwort auf die gestellte Frage nahe, wie man sie vor einigen Jahren in der Handlungstheorie gegeben hat: Eine Körperbewegung ist genau dann eine Handlung (etwas, was der Handelnde selbst tut), wenn sie in der richtigen Weise durch die Wünsche und Überzeugungen des Handelnden verursacht wurde.[17] Die gerade schon genannten Fälle machen aber deutlich, dass auch diese Annahme nicht unserem alltäglichen Reden entspricht. Wir sagen selbst in einigen Fällen, in denen ganz offensichtlich kein absichtliches Handeln vorliegt, dass jemand selbst etwas getan hat – er hustet; er niest; er stolpert. Ein solches Tun unterliegt oft nicht einmal seiner willkürlichen Kontrolle. Wir sagen auch in Fällen, in denen wir unser Tun zwar kontrollieren können, aber trotzdem mit diesem Tun keinen Zweck verfolgen, dass wir selbst etwas tun – wir pfeifen eine Melodie; wir gestikulieren beim Reden mit den Armen; wir wippen nervös mit den Beinen. Außerdem tun wir manchmal Dinge, die zwar einem Zweck dienen, aber trotzdem nicht auf einer Absicht beruhen – wir kratzen uns am Kopf, weil es juckt. Schließlich darf man nicht übersehen, dass wir schon bei Tieren zwischen dem unterscheiden, was sie selbst tun, und dem, was mit ihnen geschieht – Fido läuft selbst zu seinem Lieblingsbaum, wird aber von seinem Herrchen von diesem Baum weggezerrt.

Mir scheint, dass der oder zumindest ein wesentlicher Unterschied zwischen aktiver Bewegung und passivem Bewegtwerden darin liegt, dass das Bewegtwerden auf *äußere Kräfte* zurückzuführen ist. Ein Windstoß wirft mich um, Fido wird von seinem Lieblingsbaum weggezogen – in beiden Fällen wirken Kräfte von außen auf mich bzw. Fido ein, die die entsprechenden Bewegungen kausal hervorrufen. Ganz anders, wenn Fido von seinem Kissen aufsteht und zur Tür läuft. Natürlich kann auch dies (unter anderem) eine äußere Ursache haben – z. B. das Geräusch, das beim Aufschließen der Tür entsteht. Aber diese Ursache wirkt ganz anders als der Windstoß oder das Herrchen, das an der Leine zieht. Fidos Laufen beruht auf der Bewegung seiner Beine; aber das Geräusch an der Tür übt keine Kräfte auf diese Beine aus. Vielmehr beruht die Bewegung von Fidos Beinen in diesem Fall allein auf der Kontraktion und Relaxation bestimmter Muskeln. Äußere Kräfte spielen hier überhaupt keine Rolle. Dass Tiere

[17] Siehe Beckermann 1985.

sich selbst bewegen, heißt auch, dass die Energie, die zur Ausführung dieser Bewegungen nötig ist, aus ihnen selbst stammt.

Allerdings reicht dieser Aspekt noch nicht aus. Denken wir etwa an den Kniesehnenreflex. Der Arzt schlägt dem Patienten leicht auf die Patellarsehne unterhalb der Kniescheibe. Wenn alles in Ordnung ist, kommt es als Reflexantwort durch Kontraktion des Quadricepsmuskels zu einer Streckung des Kniegelenks – der Unterschenkel schwingt leicht nach vorn. Auch in diesem Fall wirken auf den Unterschenkel keine äußeren Kräfte; die zu seiner Bewegung nötige Energie stammt ganz aus dem Patienten selbst. Dennoch sagen wir in diesem Fall nicht: Der Patient hat seinen Unterschenkel gehoben. Allgemein gilt Folgendes: Intern induzierte Bewegungen der Glieder eines Menschen beruhen auf Muskelkontraktionen und -relaxationen. Diese Kontraktionen und Relaxationen werden ihrerseits durch das Feuern von Motoneuronen hervorgerufen, deren Zellkörper sich im Vorderhorn im Rückenmark befinden und deren Axone bis zu motorischen Endplatten direkt an den Muskelzellen reichen. Diese unteren Motoneurone können ihrerseits durch die oberen Motoneurone aktiviert werden, deren Zellkörper in der motorische Rinde im Gehirn liegen und deren Axone bis ins Vorderhorn zu den Zellkörpern der unteren Motoneurone reichen. Beim Kniesehnenreflex spielen die oberen Motoneurone aber keine Rolle. Vielmehr passiert Folgendes:

> Dehnungsrezeptoren (sogenannte Muskelspindeln) im Quadrizeps registrieren die Dehnung und melden sie an das Rückenmark. […] Beim Menschen ziehen die sensiblen Neurone (Afferenzen) zu den Lendensegmenten L2-L4 […]. Dort wird die Erregung über jeweils eine Synapse auf die motorischen Neurone (Efferenzen) umgeschaltet. Diese Neurone durchlaufen den Plexus lumbalis und im Nervus femoralis zurück zum Muskel, wo eine Kontraktion des Quadriceps femoris ausgelöst wird.[18]

Beim Kniesehnenreflex gibt es also eine Schleife von den Dehnungsrezeptoren im Quadriceps zum Rückenmark und von dort *direkt* zurück zum Quadriceps; höhere Regionen des ZNS sind nicht beteiligt. Das ist insofern interessant, als man in der einschlägigen Literatur immer wieder lesen kann, dass es die *oberen* Motoneurone sind, die für Willkürbewegungen zuständig sind. Mit anderen Worten: Bei Willkürbewegungen muss der neuronale Impuls der zu den entsprechenden Muskelkontraktionen führt, aus dem motorischen Kortex stammen. Das legt es nahe zu sagen: Ein Mensch bewegt eines seiner Glieder selbst, wenn die entsprechenden Mus-

[18] http://de.wikipedia.org/wiki/Patellarsehnenreflex (4.6.2010, 8.30 Uhr).

kelkontraktionen auf Nervenimpulse aus seinem motorischen Kortex zurückgehen.[19]

Bei Menschen gibt es eine Möglichkeit, die wir bei Tieren nicht haben – wir können sie fragen, ob sie etwas selbst getan haben, d. h., ob sie sich eine Bewegung selbst zuschreiben. Dies wurde von Roger Penfield ausgenutzt, der Mitte des vorigen Jahrhunderts bei Operationen am offenen Gehirn durch Reizung des prämotorischen und supplementärmotorischen Kortex komplette Bewegungen von Gliedmaßen induzieren konnte. Allerdings erlebten die Patienten diese Bewegungen als aufgezwungen, sie sagten von diesen Bewegungen nicht: „Ich habe das getan" (vgl. Roth 2003, 515 f.). Auf der anderen Seite bezieht sich Roth auf José Delgado, der „berichtete, dass unter ähnlichen Bedingungen wie bei Penfield die Stimulation des rostralen Anteils der so genannten internen Kapsel (d. h. der Faserbahnen, die vom Thalamus durch die Basalganglien hindurch zum Cortex ziehen) zu Bewegungen des Patienten führte, die er sich selbst zuschrieb" (Roth 2003, 516). Auch die Tatsache, dass der Nervenimpuls, der zu einer Bewegung der Gliedmaßen führt, aus dem motorischen Kortex eines Menschen stammt, reicht also nicht aus, damit er sich diese Bewegung selbst zuschreibt. Vielmehr scheint es so, dass dies nur dann der Fall ist, wenn der Impuls auf eine bestimmte Weise zustande gekommen ist. In der ersten Auflage von *Fühlen Denken, Handeln* äußert Roth dazu die folgende Vermutung:

> Das Gefühl der *Selbstveranlassung unserer Bewegungen im Willensakt* haben wir aus einem ganz anderen Grund. Dieses Gefühl ist für das Gehirn ein Zeichen, dass vor dem Starten der Bewegung die dorsale und ventrale cortical-limbische Schleife durchlaufen wurde und die exekutiven Zentren der Großhirnrinde zusammen mit dem limbischen System sich damit „ausreichend befasst" haben. In diesem Falle baut sich das symmetrische und dann das lateralisierte Bereitschaftspotential auf, und letzteres gibt den „Startschuss" für die Ausführung der intendierten Bewegung. Das Gefühl des *fiat*!, des *ich will das jetzt* ist demnach die bewusste Meldung dieses neurophysiologischen Vorgangs. (Roth 2001, 446)

---

[19] Christoph Lumer hat mich allerdings auf Folgendes hingewiesen: Wenn ich die Hand wegziehe, weil ich einen heißen Gegenstand berühre, ist diese Bewegung in der Regel auf Rückenmarksebene gesteuert; sie kann aber zentral unterdrückt werden. Außerdem rechne ich mir diese Bewegung selbst zu; ich sage also „Ich ziehe die Hand weg". Vielleicht ist dafür, dass jemand seine Glieder selbst bewegt, also nicht nötig, dass die entsprechenden Muskelkontraktionen auf Nervenimpulse aus seinem motorischen Kortex zurückgehen, vielleicht reicht es aus, dass diese Nervenimpulse zentral reguliert werden *können*.

Diese Vermutung findet sich in der zweiten Auflage in dieser Form nicht mehr. Allerdings kommt es auf die Details vielleicht auch gar nicht an. Vielleicht reicht es, festzustellen, dass Menschen sich offenbar genau die Bewegungen selbst zuschreiben, die auf neuronalen Impulsen aus dem motorischen Kortex beruhen, die ihrerseits auf eine bestimmte Weise neuronal erzeugt wurden. Diese Befunde passen gut zur folgenden Überlegung: Menschen (und manche Tiere) sind zwar im Wortsinn Automaten – Wesen, die sich selbst bewegen; aber sie sind keine Automaten in dem Sinne, dass sie immer mechanisch, reflexhaft oder unüberlegt – eben automatisch – handeln. Vielmehr sind sie *autonome Systeme*. Das bedeutet erstens, dass sie über ein Repertoire sehr unterschiedlicher Verhaltensweisen verfügen, das es ihnen gestattet, in Situationen derselben Art ganz unterschiedliche Dinge zu tun. Und es bedeutet zweitens, dass sie die – sicher neuronal realisierte – Fähigkeit besitzen, eine situationsangemessene Wahl zwischen diesen unterschiedlichen Verhaltensweisen zu treffen. Diese Fähigkeit beinhaltet zwei Teilfähigkeiten – die Fähigkeit, die Situation, in der sie sich befinden, angemessen zu analysieren (Welche Gegenstände befinden sich wo in Relation zu mir? Sind diese Gegenstände gefährlich oder nützlich? Etc.), und zweitens die Fähigkeit, eine Handlungsoption zu finden, die in der gegebenen Situation der Erreichung der eigenen Ziele dient. Mit anderen Worten: Menschen (und auch viele Tiere) verfügen über ein Steuerungssystem, das es ihnen gestattet, sich in sehr unterschiedlichen Situationen zurechtzufinden und jeweils die Handlungsoptionen zu wählen, die in diesen Situationen am nützlichsten erscheinen. Dieses Steuerungssystem ist nach allem, was wir wissen, neuronal realisiert. Deshalb ist meine Hypothese, dass die neuronalen Subsysteme des ZNS, die in Roths Überlegungen eine zentrale Rolle spielen, genau die Hirnbereiche sind, in denen das Steuerungssystem realisiert ist, das uns zu autonomen Systemen macht. Mit anderen Worten: Menschen führen genau die Bewegungen selbst aus, die auf neuronalen Impulsen aus dem motorischen Kortex beruhen, die ihrerseits von ihrem zentralen neuronalen Steuerungssystem kontrolliert werden.[20]

Allerdings muss man noch eine Einschränkung anfügen: Delgados Befunde machen deutlich, dass auch das zentrale neuronale Steuerungssystem von außen – z. B. über eingepflanzte Elektroden – manipuliert werden kann. Wenn es ihm gelungen ist, durch Stimulation bestimmter Hirnregionen bei seinen Patienten Bewegungen zu indizieren, die sie sich selbst zuschreiben, muss man, denke ich, zugestehen, dass sich diese Patienten geirrt haben. Tatsächlich haben sie diese Bewegungen nicht selbst durchge-

[20] Auch diese Annahme hat offenbar Schwierigkeiten mit Verhaltensweisen wie Stolpern, Husten oder Niesen; aber darauf kann ich an dieser Stelle nicht weiter eingehen.

führt. Deshalb sollte man letzten Endes wohl so formulieren: Menschen führen genau die Bewegungen selbst aus, die auf neuronalen Impulsen aus dem motorischen Kortex beruhen, die ihrerseits von ihrem zentralen neuronalen Steuerungssystem kontrolliert werden, sofern dieses Steuerungssystem nicht von außen manipuliert, die Menschen selbst also gewissermaßen ferngesteuert werden.

## Literatur

Beckermann, Ansgar (1985): „Handeln und Handlungserklärungen". In: Ansgar Beckermann (Hg.) *Analytische Handlungstheorie. Band 2. Handlungserklärungen*. Frankfurt/M.: Suhrkamp, 7–84.

Beckermann, Ansgar (2005): „Selbstbewusstsein in kognitiven Systemen". In: Markus Peschl (Hg.) *Die Rolle der Seele in der Kognitionswissenschaft und Neurowissenschaft*. Würzburg: Königshausen & Neumann, 171–187; in diesem Band Beitrag 13.

Beckermann, Ansgar (2008): *Gehirn, Ich, Freiheit. Neurowissenschaften und Menschenbild*. Paderborn: mentis.

Beckermann, Ansgar (2009): „Es gibt kein Ich, doch es gibt mich". In: Martina Fürst, Wolfgang Gombocz & Christian Hiebaum (Hg.) *Gehirne und Personen*. Frankfurt/M.: ontos Verlag, 1–17; in diesem Band Beitrag 14.

Beckermann, Ansgar (2010): „Die Rede von *dem* Ich und *dem* Selbst. Sprachwidrig und philosophisch höchst problematisch". In: Katja Crone, Robert Schnepf & Jürgen Stolzenberg (Hg.) *Über die Seele*. Frankfurt/M.: Suhrkamp 2010, 458–473; überarbeitete Fassung in diesem Band Beitrag 15.

Blume, Thomas (2003): ‚Ich'. In: Wulff D. Rehfus (Hg.) *Handwörterbuch Philosophie*. Göttingen: Vandenhoeck & Ruprech, 394–395.

Descartes, René: *Les passions de l'âme*. *Die Leidenschaften der Seele*. Französisch-Deutsch. Herausgegeben und übersetzt von Klaus Hammacher. Hamburg: Felix Meiner 1984.

Henke, Roland W. (2003): ‚Selbst'. In: Wulff D. Rehfus (Hg.) *Handwörterbuch Philosophie*. Göttingen: Vandenhoeck & Ruprecht, 609–610.

Herring, H. & U. Schönpflug (1976): ‚Ich'. In: Joachim Ritter & Karlfried Gründer (Hg.) *Historisches Wörterbuch der Philosophie. Band 4*. Basel: Schwabe. Sp. 1–18.

Leibniz, Gottfried Wilhelm (1714): „Principes de la nature et de la grâce fondés en raison". In: G.W. Leibniz, *Monadologie und andere metaphysische Schriften*. Französisch-Deutsch. Hg. und übersetzt von Ulrich J. Schneider. Hamburg: Meiner 2002. S. 152–173.

Lenzen, Wolfgang (2002): „Realität und ‚Wirklichkeit'. Kritische Bemerkungen zu Gerhard Roths ‚neurobiologischem Konstruktivismus'". In: Carlos U. Moulines & Karl-Georg Niebergall (Hg.) *Argument und Analyse*. Paderborn: mentis, 33–54.

Lenzen, Wolfgang (2005): „Alles nur Illusionen? Philosophische (In-)Konsequenzen der Neurobiologie". *Facta Philosophica* 7, 189–229.

Lenzen, Wolfgang (2006): „Auf der Suche nach dem verlorenen ‚Selbst'. Thomas Metzinger und die ‚letzte Kränkung' der Menschheit". *Facta Philosophica* 8, 161–192.

Locke, John: *An Essay Concerning Human Understanding*. Hg. von Peter H. Nidditch. Oxford: Clarendon Press 1975.

Lowe, E. Jonathan (1995): ‚self'. In: Ted Honderich (Hg.) *The Oxford Companion to Philosophy*. Oxford: Oxford University Press.

Metzinger, Thomas (2003): *Being No One. The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Penfield, Roger (1958): *The Excitable Cortex in Conscious Man*. Springfield, Ill.: Liverpool University Press.

Roth, Gerhard (2001): *Fühlen, Denken, Handeln*. 1. Auflage. Frankfurt/M.: Suhrkamp.

Roth, Gerhard (2003): *Fühlen, Denken, Handeln*. Neue, vollständig überarbeitete Ausgabe. Frankfurt/M.: Suhrkamp.

Siefer, Werner & Christian Weber (2006): *Ich. Wie wir uns selbst erfinden*. Frankfurt/M.: Campus.

# Miscellanea

# 17

# Wittgenstein, Wittgensteinianism and the Contemporary Philosophy of Mind – Continuities and Changes*

For a short period of time in the middle of the last century, at least in Europe, Wittgenstein was the measure of all things in philosophy and especially in the philosophy of mind. The private language argument had shown the conception of the mind going back to Descartes and Locke to be principally flawed – or so the consensus was. Mental phenomena are not essentially private, and there simply cannot be mental states without any observable criteria at all. Anyone who disagreed was in for a difficult time. Yet, only one or two decades later the discussion had moved on considerably. First, the identity theory overcame behaviourism. Second, functionalism superseded the identity theory, thereby paving the way for more specialist approaches such as Fodor's representational theory of mind. Finally, a new, post-Wittgensteinian orthodoxy developed with amazing swiftness. With hindsight, this seems a remarkable phenomenon in the sociology of philosophy. As interesting as it would be, I shall disregard the sociological side here. Instead, I am interested in the questions: What has changed? And, are there good reasons for these changes?

These questions are difficult to answer, not least, because it is far from clear which position Wittgenstein himself held with regard to the mind-body problem. Even today, articles and books are being published in an attempt to come closer to answering this question; but 50 years after Wittgenstein's death we are still far from a generally accepted consensus. Of course, this picture is painted a little bleakly: some points are clear – for instance Wittgenstein's rejection of the picture of an inner world of the mind and an external world of material things, which pervades all areas of philosophy since early modern times. According to Descartes, the mind is an immaterial substance of its own – a *res cogitans* – and thinking, feeling and remembering are occurrences in this substance accessible only to this substance itself. The consequence of this picture is that the mind is private.

---

* Erstveröffentlichung in: A. Coliva & E. Picardi (Hg.) *Wittgenstein Today*. Padua: Il Poligrafo 2004, 275–296.
I would like to thank Antonia Barke for translating this paper into English. I am very grateful to the audience of my lecture in Bologna and to my colleagues in Bielefeld, especially to Eike von Savigny, Joachim Schulte, Hanjo Glock and Christian Nimtz, for very helpful comments on earlier drafts of this paper.

Only the mind itself can know what happens within; others, at best, have indirect access through a kind of inductive inference. They have to infer from the person's behaviour what goes on in a person's mind. This, Wittgenstein says in a famous passage in the *Philosophical Investigations*, is complete nonsense:

> In what sense are my sensations *private*? – Well, only I can know whether I am really in pain; another person can only surmise it. – In one way this is wrong, and in another nonsense. If we are using the word „to know" as it is normally used (and how else are we to use it?), then other people very often know when I am in pain. – Yes, but all the same not with the certainty with which I know it myself! – It can't be said of me at all (except perhaps as a joke) that I *know* I am in pain. What is it supposed to mean – except perhaps that I *am* in pain? (PI § 246)

If the epistemic consequences of the Cartesian picture of the mind are nonsensical, the picture itself must be wrong: the mind is not private, but public. But what does this mean? The predominant view in the 50s and 60s was a view that one could call the „criteriological account".[1] According to the proponents of this view Wittgenstein has shown by means of considerations on the meaning of linguistic expressions in general that there can be no mental states without behavioural criteria. *Pain behaviour* is not just a *symptom* of the mental state pain, but a *criterion*. That is to say, pain behaviour is corrigible evidence that somebody is in pain, but for semantic reasons it is, in a certain way, also sufficient evidence. *For semantic reasons* it is true that if a person shows this behaviour and there is no evidence to the contrary, then this person is in pain. There is ample evidence that Wittgenstein held this view. Consider for example this passage from the *Blue Book*:

> When we learnt the use of the phrase „so-and-so has toothache" we were pointed out certain kinds of behaviour of those who were said to have toothache. As an instance of these kinds of behaviour let us take holding your cheek. […] Now one may […] ask: „How do you know that he has got toothache when he holds his cheek?" The answer to this might be, „I say, *he* has toothache when he holds his cheek because I hold my cheek when I have toothache." But what if we went on asking: – „And why do you suppose that toothache corresponds to his holding his cheek just because your toothache corresponds to your holding your cheek?" You will be at a loss to answer this question, and find that here we strike rock bottom, that is we have come down to conventions. (*Blue and Brown Books*, 24)

In other words: not all signs of the presence of pain can be mere symptoms; some must be criteria in the semantic sense, because otherwise we would

[1] See ter Hark 1995.

have no basis for the application of the concept ‚pain'. Obviously, this is exactly one of the points of the private language argument. However, the criteriological interpretation has been criticised in recent years, among other reasons, because it places Wittgenstein in great proximity to behaviourism, a theory which he explicitly rejected in many places.

What may have been his reasons for this rejection? Perhaps, as Hanjo Glock suggests, that the behaviourist is still sticking too closely to the Cartesian picture by construing the mental after the image of the physical.

> Wittgenstein's attack on the inner/outer dichotomy is often accused of reducing the inner to the outer, and thereby ignoring the most important aspects of human existence. Ironically, Wittgenstein in turn accuses the inner/outer conception of mistakenly assimilating the mental to the physical. It construes the relationship between mental phenomena and mental terms ‚on the model of' material ‚object and designation', and thereby turns the mind into a *realm* of mental entities, states, processes and events, which are just like their physical counterparts, only hidden and more ethereal […]. […] [T]his tendency is fuelled by the *Augustinian picture of language*, which suggests that all words stand for objects, and all sentences describe something – if not physical entities, then entities of a different kind. (Glock 1996, 175)

Indeed, there are a number of passages that suggest that it was Wittgenstein's opinion that it is a fundamental error to construe the use of mental terms after the model of the use of physical language. He writes, for example, in the *Philosophical Investigations*:

> How does the philosophical problem about mental processes and states and about behaviourism arise? – The first step is the one that altogether escapes notice. We talk of processes and states and leave their nature undecided. Sometime perhaps we shall know more about them – we think. But that is just what commits us to a particular way of looking at the matter. For we have a definite concept of what it means to learn to know a process better. (The decisive movement in the conjuring trick has been made, and it was the very one that we thought quite innocent.) – And now the analogy which was to make us understand our thoughts falls to pieces. So we have to deny the yet uncomprehended process in the yet unexplored medium. And now it looks as if we had denied mental processes. And naturally we don't want to deny them. (PI § 308)

In the second part of the *Philosophical Investigations*, Wittgenstein touches again on his views on behaviourism:

> Then psychology treats of behaviour, not of the mind?
> What do psychologists record? – What do they observe? Isn't it the behaviour of human beings, in particular their utterances? But *these* are not about behaviour. „I noticed that he was out of humour." Is this a report about his behaviour or his state of mind? („The sky looks threatening": is this about the

> present or the future?) Both; not side-by-side, however, but about the one *via* the other. […] It is like the relation: physical object – sense-impressions. Here we have two different language-games and a complicated relation between them. – If you try to reduce their relations to a *simple* formula you go wrong. (*PI*, 179 f.)

In my eyes, this remark is rather enigmatic. One thing is clear, however: for Wittgenstein the difference between speaking-of-behaviour and speaking-of-mental-states is similar to the difference of speaking-of-physical-objects and speaking-of-sense-impressions. Here we have two different language games, even if these language games – as he explicitly points out – are closely linked. But, and this is very regrettable, Wittgenstein says very little about how the language game of talking about the mental really works and how it differs from speaking about behaviour and from other more physical language games. The situation gets even more confusing because in the *Remarks on the Philosophy of Psychology I* and in *Zettel* we find a rather peculiar passage, according to which Wittgenstein sees the level of psychology and that of physiology as completely separate.

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should […] this order not proceed, so to speak, out of chaos? The case would be like the following – certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced – but *nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that it comes out of. […] [T]here is no reason why this should not really hold for our thoughts, and hence for our talking and writing. (*RPPI* 903; see *Z* 608)

> I saw this man years ago: now I have seen him again, I recognize him, I remember his name. And why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored-up there *in any form*? Why must a trace have been left behind? Why should there not be a psychological regularity to which *no* physiological regularity corresponds? If this upsets our concepts of causality then it is high time they were upset. (*RPPI* 905; see *Z* 610)

If we read this passage from today's perspective, Wittgenstein seems to say no more and no less than that the thesis of the *emergent nature* of mental phenomena appears extremely plausible to him.[2] To that end he is prepared

[2] „Emergence" here is to be understood as C. D. Broad developed the concept in 1925: A property *F* of a complex system is emergent if it *cannot be deduced*

to allow a causality between mental phenomena that is not mediated by physiological processes, even if this brings the concurrent danger that it may appear to count in favour of classical mind-body dualism.

> The prejudice in favour of psycho-physical parallelism is also a fruit of the primitive conception of grammar. For when one admits a causality between psychological phenomena, which is not mediated physiologically, one fancies that in doing so one is making an admission of the existence of a soul *alongside* the body, a ghostly mental nature. (*RPPI* 906; see *Z* 611)

Of course, all these quotations come from texts written at very different times and occasions. But even if this is so, it can not be disputed that we do not find a coherent account of the relations and the differences between the mental and the physical, between the use of mental terms and the use of physical terms in Wittgenstein's writings. We do find a straightforward rejection of the idea that the mind is a private inner theatre, and we also find hints in direction of the idea that there is a significant difference between the use of the mental and the physical vocabulary. At least, many philosophers understood Wittgenstein this way. And, what is more, in my view a large part of post-Wittgensteinian philosophy can be conceived of as an attempt to spell out the idea of two levels or two language games.

One reason may have been that this idea seems to allow the dissolution of the mind-body problem. This problem is widely held to be the problem „of accounting for the place of mind in a world that is essentially physical".[3] At least at first sight there seem to be mental items in the world: pains and thoughts, colour sensations and wishes, consciousness and perhaps even souls. How is this realm of the mental related to the realm of the physical? Are mental items in fact not so different, but only a special kind of physical items? Or does the mental constitute a special realm of non-physical entities that nonetheless causally interact with the physical? Wittgenstein's views seem to allow the answer: All these questions are ill conceived. One only has to notice that the mental language does work in a way very different from that of the physical language. Mental terms do not denote special states or processes in the way physical terms denote physical states or processes. Indeed, they do not denote at all. And this means that there simply are no mental items about which we can reasonably ask how they fit into an essentially physical world.

This attractive feature of Wittgenstein's views seems to have inspired for example Gilbert Ryle whose work may also be understood as an attempt to

---

from the properties of the parts of the system together with their spatial relations. See Beckermann 2000.

[3] Kim 1996, 9.

elaborate Wittgenstein's two languages account.[4] To be sure, *The Concept of Mind* was published four years before the *Philosophical Investigations*. However, Ryle's thoughts are so similar to the thoughts of the late Wittgenstein, and he was so much influenced by Wittgenstein through conversations and lecture-notes that I do not hesitate to treat Ryle as a Wittgensteinian here. Ryle, however, seems to have been concerned with a particular aspect of the mind-body problem – the question of how the mind can causally interact with the body. He also tries to dissolve this problem by a semantic argument. According to Ryle, mental terms refer to dispositions and not to events, and, that is, not to possible causes. The problem of how the mind causally interacts with the body simply disappears when we only acknowledge that mental explanations are dispositional explanations. I shall come back to this issue soon. But first I would like to say a few words on the general outline of Ryle's argument.

To begin with, Ryle, too, emphatically rejects the concept of mind that has become pervasive since early modern times. For Ryle, too, it is absurd to assume a Cartesian theatre, in which mental objects abound and mental occurrences take place, that only the mind knows about – occurrences, which moreover interact causally with each other and with occurrences in the physical world. According to Ryle, this assumption rests on one big misunderstanding, or, to put it more precisely, it rests on a category mistake. This mistake can briefly be characterised like this: Cartesians, but not only Cartesians, assume that mental expressions such as ‚to remember', ‚to think', ‚to perceive', ‚to believe' and ‚to want' refer to (hidden) *internal occurrences within a person's mind*, which *cause* the person's outward behaviour. In reality however, according to Ryle, we do not employ these expressions to refer to some ‚shadow actions', which are hidden antecedents of the overt behaviour. Instead we use the mental expressions to characterise the publicly observable actions in a different way. The mentalist thinks mental phenomena consist in enigmatic occurrences behind the observable actions, while in reality mental phenomena are nothing but a manner of organisation of these actions.

Ryle supports his view that mental expressions do not refer to hidden inner occurrences – among other arguments – with an analysis of intelligent and voluntary actions. His opponent, the mentalist, analyses intelligent, or voluntary behaviour like this:

– an action is intelligent if and only if it has been caused by a corresponding thought;
– an action is voluntary if and only if it has been caused by a corresponding act of the will.

[4] For the following see Beckermann 2001, sec. 4.1.3.

However, when we consider carefully under which circumstances we really call an action ‚intelligent', a completely different picture emerges. For, normally, we would say that someone acts intelligently if

– he normally does what he does correctly, well and successfully and if
– he is able to discover a mistake in his way of proceeding and to eliminate it, if he is able to repeat successes and improve on them, if he is able to learn from the example of others etc.

That is to say, a closer consideration of our actual use of language shows that we do not call an action intelligent if we can trace it back to a hidden inner process, but rather if the action does not stand alone, but is part of a pattern of actions and abilities.

‚Willing', too, according to Ryle, is certainly not a verb that we use to refer to occurrences (acts of the will); for if this were the case, these occurrences would have to be dateable and countable. However:

> No one ever says such things as that at 10 a.m. he was occupied in willing this or that, or that he performed five quick and easy volitions and two slow and difficult volitions between midday and lunch-time. An accused person may admit or deny that he did something, or that he did it on purpose, but he never admits or denies having willed. Nor do the judge and jury require to be satisfied by evidence, which in the nature of the case could never be adduced, that a volition preceded the pulling of the trigger. (Ryle 1963, 63)

Moreover: What properties could acts of the will have? Can we perform them fast or slowly? Can we do more than one at the same time? Can we interrupt an act of will and later pick it up where we left it? The impossibility of finding answers to these questions shows very clearly that in our everyday understanding ‚willing' does not stand for occurrences or actions. However, what *does* it stand for?

Ryle advocates carefully considering our everyday use of language. It shows that we normally use the adjectives ‚voluntary' or ‚involuntary' when we are interested in the question of whether a mistake deserves reproach. A sailor is asked to tie a reef-knot, but he ties a granny-knot. A pupil arrives late for school. These are typical cases in which we ask whether the respective actions were voluntary or not. It depends upon the answer to this question whether we reprimand the sailor or the pupil or even punish them. The decisive question is whether the person who committed the mistake could have avoided it. This, in turn, depends upon whether he or she had the knowledge and the ability to carry out the action properly, and whether external circumstances have prevented him or her from correctly carrying it out. However, both can be detected without recourse to some mysterious acts of will.

Therefore, voluntary actions, contrary to the official characterisation, are better analysed thus:

– a mistaken (wrong) action is voluntary if and only if the agent possesses the knowledge and the ability to perform the action correctly and if he or she is not prevented by external circumstances from the correct performance of the action.

The analysis of intelligent actions as well as the analysis of voluntary actions therefore shows that the official doctrine is mistaken. Both types of actions are *not* characterised by the ‚fact' that they are caused by hidden internal events in the agent's mind.

However, if this is so, the question arises how this mistaken impression could have come about. Why are we so prone to make category mistakes when thinking about the mind? Why do we assume that mental expressions refer to events that take place inside people's heads? According to Ryle, a central reason for the mistaken views of the official doctrine lies in the fact that the official doctrine construes mental explanations as causal explanations, and therefore views mental phenomena as (hidden) causes. In Ryle's view, however, mental states are really dispositions and hence mental explanations *dispositional explanations*.

Ryle provides a famous analysis of the sentence „He boasted from vanity."

> The statement „he boasted from vanity" ought, on one view, to be construed as saying that „he boasted and the cause of his boasting was the occurrence in him of a particular feeling or impulse of vanity". On the other view, it is to be construed as saying „he boasted on meeting the stranger and his doing so satisfies the law-like proposition that whenever he finds a chance of securing the admiration and envy of others, he does whatever he thinks will produce this admiration and envy." (Ryle 1963, 87)

According to Ryle, it is perfectly obvious that the first analysis is totally absurd. When we ask to which type of explanation the statement „he passed his neighbour the salt from politeness" belongs, it is immediately obvious that the causal analysis is not tenable. This is so because a polite person is one who has the disposition not to jump the queue, to let others pass before him, to help without being asked, to avoid making tactless remarks, not to make his hosts uncomfortable through inappropriate dress or inappropriate behaviour, and so forth. Furthermore, the dispositional character of this explanation also shows itself from the fact that it requires supplementing by a causal explanation.

> But the general fact that a person is disposed to act in such and such ways in such and such circumstances does not by itself account for his doing a particular thing at a particular moment; any more than the fact that the glass was brittle accounts for its fracture at 10 p.m. As the impact of the stone at 10 p.m.

> caused the glass to break, so some antecedent of an action causes or occasions the agent to perform it when and where he does so. For example, a man passes his neighbour the salt from politeness; but his politeness is merely his inclination to pass the salt when it is wanted, as well as to perform a thousand other courtesies of the same general kind. So besides the question „for what reason did he pass the salt?" there is the quite different question „what made him pass the salt at that moment to that neighbour?" This question is probably answered by „he heard his neighbour ask for it", or „he noticed his neighbour's eye wandering over the table", or something of the sort. (Ryle 1963, 109)

The emerging picture is this: According to Ryle, dispositional statements like „this plane is brittle" or „John is polite" are encapsulated laws or law-like statements. Within the context of explanation dispositional statements, therefore, never express antecedent conditions and, that is, causes. Their role is the role of statements expressing laws. This, in turn, implies that dispositional explanations are, in a sense, incomplete. In addition to the relevant dispositions, i.e. laws, we also need to know the relevant antecedent conditions and, that is, the causes of the action to be explained.

In Ryle's eyes there can be no doubt that the explanation „he passed his neighbour the salt from politeness" – just as the explanation „he boasted from vanity" – is, in this sense, a dispositional explanation. And this, in his view, implies that the mental phenomena, to which the overt behaviour is traced back in these explanations, are not mysterious inner processes in the agent's mind, but dispositional properties, which are just as publicly accessible as are the dispositions of brittleness or of being soluble in water.

With this dispositional analysis Ryle aims at the same thing as with his alternative analyses of intelligent and voluntary behaviour. Firstly, he wishes to show that mental expressions do not refer to hidden inner occurrences in a person's mind, but to circumstances whose public observability is beyond doubt. Secondly, he wants to show that mental concepts *as dispositional concepts* do not refer to the causes of actions and that mental explanations, therefore, never compete for instance with physiological explanations.

A closer look reveals, however, that in Ryle's considerations there are at least two accounts of what the use of mental vocabulary amounts to: the pattern account and the dispositional account. According to the pattern account using mental terms to describe and explain behaviour is nothing but regarding the behaviour as an integral part of a wider pattern of behaviour and capacities. The dispositional account is more straightforward. According to the dispositional account, in using mental concepts we do nothing but ascribe certain behavioural dispositions and capacities to the person in question.

In the history of the reception of Ryle's views the dispositional account was obviously the more prominent one. However, it has been precisely this aspect of Ryle's thought that proved to be particularly vulnerable to objections. This is so for three reasons:

(1) Dispositional predicates are not just used when we characterise a person's mental life; they are just as much at home in the vocabulary of the natural sciences. Mass or weight, for instance, are classical examples of dispositions. That an object has the weight of 10 kp means, among other things, that if placed on scales it will generate a certain movement of the pointer. That an object has the mass of 10 kg means, among other things, that it experiences a certain amount of acceleration if a certain force acts upon it. Furthermore, natural dispositions are multi-track, too – i.e., they are characterised not just by one occasion-reaction pair but by quite a number of these pairs. Therefore, the dispositional account is not very helpful when we are trying to draw the line between the mental language game and the physical language game. At most, we can distinguish between the language games of events and the language games of dispositions in this way; this distinction, however, stands in an orthogonal relation to that between the mental and the physical language game.

(2) Ryle's thesis that a categorical difference exists between dispositional explanations and causal explanations, was highly controversial from the outset. Let us just listen to one of the voices of the diverse chorus of critical contributions to this question.

> [A]ctually, it is a mistake to suppose that only events may be properly spoken of as causes, for we frequently refer to states, dispositional properties and even the failure of events to occur as causes. For example, given appropriate circumstances, we might speak of a bent rail, an icy track or the failure of the brakeman to signal as the cause of a given train accident. (Gean 1965/1966, 677)

It is possible that Ryle might have felt vindicated in his view through Davidson's thoughts on the concept of causality according to which the relata of causal relations are events and nothing else. However, many doubts remain.

(3) The strongest argument against Ryle's dispositional account arose from developments in the theory of scientific concept formation around that time. As early as 1936/1937 in his paper „Testability and Meaning" Carnap broke with the old thesis of Logical Empiricism that all scientific concepts have to be definable exclusively in observation terms and logical vocabulary. It was the dispositional terms that had been the downfall of this thesis. Carnap therefore suggested analysing

> the recalcitrant dispositional concepts by means of so-called reduction sentences. In the final consequence, however, this suggestion proved to be only the first step to a complete dissolution of the empiricist criterion of meaning. Once the step was taken, the view gained hold that most concepts of scientific interest could not be defined in purely observational terms or even be adequately analysed by means of reduction sentences. Rather, according to the new insight, central concepts such as length, mass, temperature and charge are *theoretical concepts*, which receive their meaning on the one hand through their relations to other theoretical concepts and on the other hand through a number of correspondence rules, which – rather loosely – connect certain theoretical concepts with observational terms. An early canonical formulation of this new view can be found in Hempel's *Fundamentals of Concept Formation in Empirical Science*, which was published in 1952.

After this theory, which Carnap endorsed too, had gained fast acceptance, it was natural to conceive of mental terms no longer as dispositional predicates, but also as theoretical terms – especially so, since Ryle had already spoken of multi-track dispositions in this context. This view was defended by, e.g., Fodor and Chihara as well as by Brandt and Kim, who immediately drew a conclusion that is very uncomfortable for the dispositional account: if mental concepts really are theoretical concepts, then they are indeed not different from the concepts of natural science and everything suggests that mental explanations are completely normal causal explanations much like „the iron glowed red because it was heated to 750° C" or „this piece of iron attracts iron filings because it is magnetic". In other words, if mental concepts really are theoretical concepts, then statements like „John is angry" are *not* encapsulated laws or law-like statements, but statements that express antecedent conditions – and that is, causes – of actions.

This conclusion was strongly supported by the resurrection of realism going back to Carnap's and Tarski's work on pure semantics. Actually, this major shift in the fundamental ideas of Logical Empiricism also took place about 1950. Physical objects were no longer regarded as logical constructs out of sense data and unobservable states and properties no longer as logical constructs out of observables. Physical objects and unobservable states and properties were again considered as perfectly *real* though not *given*. In his article „The Mind-Body Problem in the Development of Logical Empiricism" Herbert Feigl writes:[5]

[5] See also Feigl 1950b, and Sellars 1948.

> The slogan of Vienna Logical Positivism: „The meaning of a statement is the method of its verification"; and the slogan of Bridgman's operationism: „A concept is synonymous with the set of operations [which determine its applications]" were excellent preventives of the transcendent type of metaphysical speculations. […] Logical empiricism in its later development, however, had to replace these radical principles by more conservative ones. […] [T]he meaning of scientific statements cannot in general be identified with their confirming evidence. […] For a […] very simple example we may refer to the concept of the temperature of a body. As ordinary and scientific commonsense […] would put it, thermometer (or pyrometer) readings, spectroscopic findings, and other types of measurement merely indicate something about the body in question, namely the intensity of heat which is a state of that body. No matter whether this heat intensity is construed in terms of classical (macro-) thermodynamics or in terms of statistical (micro- or molecular) thermodynamics, it is in any case only *evidenced by but not identical with* those indications. Similarly for psychology: The overt symptoms and behavior that indicate an emotion, like e.g., anxiety, are confirmable and measurable in terms of skin-temperature, endocrine secretions, psychogalvanic reflexes, verbal responses, etc. but must not be confused with the emotion itself. Generally, the „theoretical constructs," i.e., the hypothetically assumed entities of the sciences cannot be identified with (i.e., explicitly defined in terms of) concepts which apply to the directly perceptible facts as they are manifest in the contexts of ordinary observation or of experimental operations. (Feigl 1950a, 617 f.)

One page later Feigl continues:

> The realistic correction of positivism consists in the identification of meaning with factual reference. This conforms well with customary usage according to which a statement *means* a state of affairs; and is *true* if that state of affairs is fulfilled („is real," „exists"). This is the obvious grammar of „meaning," „truth," and „reality." (Feigl 1950a, 619)

Feigl could have added: According to customary usage, even the meaning of predicates as ‚temperature' and ‚pain' cannot be identified with the methods we use in trying to find out what temperature a body has or whether a person is in pain. Even these predicates *refer* – to the internal state of a body or the mental state of a person which are „evidenced by but not identical with" thermometer readings or certain symptoms and ways to behave.[6]

---

[6] In recent times this line of reasoning has been pushed even further by the work of Kripke and Putnam on the semantics of natural kind terms. Nonetheless, I would like to stress that even this first change in the semantics of theoretical terms was an indispensable step on the way towards the identity theory. Only if we assume that these terms, one way or another, *refer* to states or properties,

Perhaps these considerations contributed to the fact that other authors preferred the pattern account, which, as already mentioned, can also be found in Ryle. As early as the 50s and 60s of the last century Melden formulated thoughts in this direction. But, as far as I know, this account was only thoroughly spelled out by Eike von Savigny in his interpretation of the philosophy of the *Philosophical Investigations*. Central to this interpretation is the thesis that with regard to mental states it is not only the behaviour of the person who has the mental state in question that counts, but also the behaviour of the members of the society in which this person lives. Von Savigny believes that Wittgenstein held the following view:

> The fact that someone imagines something, expects something, wishes something, feels something, thinks of something or intends to do something etc. does not concern that person in isolation. Rather, this fact consists in that the *patterns* of this person's individual behaviour are in a certain way embedded in the *pattern* of the social behaviour of the community to which he belongs. (von Savigny 1994, 10 – italics mine)

Let us suppose someone has a headache and he displays the corresponding behaviour: he tries to hold his head still, he presses his hands against his temples, cools the forehead with a wet cloth and retires to a dark room. But not only that, his fellow humans, too, react in a way that we are used to: they offer him aspirin, they slink around noiselessly, they show sympathy towards him etc. This person is in pain. Let us now consider someone who behaves in exactly the same way as this person, but his environment reacts in a completely different manner:

> Moaning surprises the people; one does not receive any aspirin, and nobody calls the doctor. What we regard as pain behaviour, would there be treated as if the people were affected by some passing peculiarity. Their behaviour irritates the others, but they put up with it. The person with the headache has not changed in himself. (von Savigny 1994, 11 f.)

According to von Savigny, Wittgenstein would say, relative to this second environment, that the person no longer has a headache. That someone has a headache means that his own behaviour shows a certain pattern, and that the reactions of his fellow humans have a certain pattern. The extreme – and in my eyes very implausible – externalism of this view does not have to concern us here since the basic ideas of the pattern account may well be separated from it. For our purposes it is only important that according to the pattern account the following is true:

---

we can ask whether two of these terms refer to the *same* state or the *same* property.

– If one says that a person is in pain, one does not say that she is in any particular inner state, but that this person's behaviour (and the behaviour of his fellow humans) has a certain pattern.
– If one says a person is holding her cheek because she has a toothache, one does not causally explain this behaviour by tracing it back to a cause – the state of pain – but rather one explains this behaviour by pointing out that it forms part of a certain pattern of behaviour.
– If one says that a person simulates pain, one says that the same behaviour – holding one's cheek – for this person is embedded in a different pattern of behaviour. (An actor who acts being in pain, under certain circumstances behaves differently from someone who really is in pain.) If the same behaviour is embedded in exactly the same pattern of behaviour, it does not make sense to say in one case that the person is in pain and in the other she is not.

Let us recapitulate here. Wittgensteinians and the proponents of the new orthodoxy agree in one point – their rejection of Cartesianism. The mind is not a non-material substance and the mental is not an inner world of occurrences to which only the mind itself has access. Mental terms such as ‚to remember', ‚to think', ‚to perceive', ‚to believe' and ‚to want' do not refer to hidden occurrences inside a person or inside a person's mind. The mental is just as public as the physical. However, here the agreement ends. For the proponents of the new orthodoxy spell out the rejection of Cartesianism differently from the Wittgensteinians. For the new orthodoxy the mental is public because mental concepts are theoretical concepts – concepts with which we ascribe theoretical properties or states to persons and which – as concepts – have the same status as the concepts ‚is magnetic' or ‚has a mass of 10 kg'. Even if in mental explanations the behaviour of persons is not traced back to hidden mental occurrences, these explanations are every bit as causal as the corresponding physical explanations.

The characteristic view of most Wittgensteinians on the other hand is that this picture is fundamentally wrong. Mental expressions do not refer to any states – not even to theoretical ones. By means of mental expressions we characterise persons whose behaviour shows a certain pattern; but with these expressions we do not ascribe states to these persons which give rise to these patterns. For this reason, mental explanations are certainly not causal explanations. Given these views it is no wonder that the battleground for the fight between Wittgensteinians and the proponents of the new orthodoxy became the question of whether mental explanations are causal explanations or explanations of a completely different kind. What were the reasons which the proponents of the new orthodoxy put forward for their case?

To start with: Wittgensteinians who are committed to the pattern account claim that people who are in pain show a very specific behaviour. But they do not say anything about the causes of this behaviour; sometimes it even seems as if they regard even asking the question of the cause as illegitimate or nonsense. *Prima facie*, however, nothing counts against the view that pain behaviour has a cause, too. And don't we say „He is holding his cheek *because* he is in pain"? So what seems more natural than to assume that with the expression ‚pain' we do not refer to a pattern of behaviour, but to the *cause* of this behaviour?

Let us consider the parallel physical case of being magnetic. The behaviour of magnetic objects also forms a characteristic pattern: they attract iron filings in their proximity and induce an electrical current in coils which they pass through, the needle of a compass near them tends to point in their direction etc. However, when we say of an object *a* that it is magnetic then we thereby do not say that *a* shows the behaviour that is characteristic of magnetic objects, but that it has the property that is causally responsible for this pattern of behaviour. Why should this case differ so fundamentally from the case of pain?

Secondly, even proponents of the new orthodoxy do not deny that there is a relatively close relationship between mental states and typical behaviours. But they point out that it is simply not the case that all persons who are in pain show the same behavioural pattern. Compare, e.g., a person who cuts her finger during the usual morning toilet with a person who cuts her finger while hiding in a cave to escape her persecutors. The respective behaviour will be as different as you can imagine. The simple explanation for this is straightforward: Pain causes behaviour, but it does not do so in isolation, but only in the context of the respective (relevant) beliefs, wishes, ideological attitudes etc. Thus, in the context of different mental states pain may give rise to different patterns of behaviour.

Thirdly, pain itself can obviously be causally influenced in a multitude of ways. On the one hand pains are caused – through stabbing or beating, too much alcohol or muscle cramps and many other things. On the other hand, one can causally fight pain – through painkillers, through a cold compress, sometimes through warming the affected part, through acupuncture or through relaxation techniques etc. How can a proponent of the pattern account integrate this? Only by claiming that stabbings, beatings or too much alcohol causally lead to the person showing the pattern of behaviour characteristic of pain, and that painkillers, a cold compress on the forehead or acupuncture causally lead to the person ceasing to display this behaviour. This seems at least implausible. Causing pain is something different from causing pain behaviour, and fighting pain causally is something different from preventing pain behaviour. If pain was nothing

but a certain pattern of behaviour people should get rid of their pain by being paralysed.[7] Remember that in anaesthesia usually three different kinds of drugs are used: one that induces unconsciousness, another by which the muscles get paralysed, and, finally, a third against pain. On the pattern account it is completely incomprehensible what additional role the third kind of drug is supposed to play.

The strongest argument for the new orthodoxy, however, flows from the direct analysis of mental explanations. At least if one follows Mackie's or Lewis' considerations, these have exactly the features that are characteristic of causal explanations. For Mackie a cause of an effect *e* is, at least in principle, an INUS condition for *e* – an insufficient but necessary part of a condition which is itself unnecessary but sufficient for *e*.[8] And indeed, if someone says „John is holding his cheek because he has a strong toothache", we accept this explanation only if we are convinced that the toothache together with other conditions is sufficient for John's holding his cheek and if we furthermore believe that John would not hold his cheek if he had no toothache. If we learn that John would hold his cheek even if he had no ache whatsoever simply because someone told him to do so, we would immediately reject the explanation given. This also is in complete accordance with David Lewis' counterfactual account of causation.[9]

Moreover, causal explanations provide an answer to the question of why a certain event took place at all. Pattern explanations, however, are not suited to this job. If saying „John is holding his cheek because he has a strong toothache" would just mean „John is holding his cheek, and this behaviour is part of a certain behavioural pattern characteristic of pain" then, after this explanation, we would still not know why John is holding his cheek, because the explanation remains silent on the question what leads to John showing this behavioural pattern. Mental explanations, however, *do* tell us why a person behaves the way he or she does. If I learn that John crossed the road in order to buy something at the grocer's, then I also learn why John did exactly this and not something else, that is to say, I learn why this behaviour took place at all. If the explanation given were a pattern explanation, I would not learn this. Hence, by simple counterposition, this explanation is obviously a causal and not a pattern explanation.

Resorting to the logical connection argument does not help here. As early as in the 60s it was shown with a great number of arguments that the logical connection argument is neither correct nor based on true premises. Davidson, for instance, has demonstrated convincingly that logical

---

[7] Or, what seems even more absurd, by being moved to a different society.
[8] Mackie 1965, 245–246.
[9] See Lewis 1986.

relations do not obtain between events, but between descriptions of events, and that for any two events one can always find descriptions that let them appear logically dependent as well as descriptions that make them logically independent. If *c* is the cause of *e*, *c* can always be described by the expression ‚the cause of *e*', but of course this does not render the sentence „the cause of *e* is the cause of *e*" false. On the other hand, the considerations concerning the theoretical character of mental expressions have made especially clear that there is a connection that obtains between the mental states and the behaviour that they are meant to explain, but that this connection is by no means so close that it would preclude a causal relationship. And anyway: even if by definition the expression „streptococci infection" meant „infection caused by streptococci", this would not render false the sentence „Streptococci infections are caused by streptococci".

To sum up. It seems to me that in the dispute about the causal character of mental explanations at the end of the 60s and the beginning of the 70s the causalists clearly won the day. At least, this is how it appears among the academic public. Further, it seems to me that it was precisely this that has permitted the new orthodoxy to prevail so quickly over Wittgensteinianism. However, we should not forget to realise the consequences the most remarkable of which consists in the fact that the mind-body problem is on the agenda again. Maybe the question of whether there are mental things like souls or other spooky stuff is obsolete even now. But if mental properties are pretty normal and seemingly causally efficacious properties we cannot ignore the question of how these properties fit into an essentially physical world. And that is exactly the question which has been addressed by most of the work in the philosophy of mind in the last decades.

In my mind, we should also acknowledge that we really have made some progress here, at least in understanding the question itself. However, in the mid-70s something began to happen within the framework of the new orthodoxy which could well be regarded as a return to Cartesianism. The starting point was Thomas Nagel's seminal paper „What is it like to be a bat?". The considerations of Nagel, Jackson, Levine, Chalmers[10] and many others all seem to point to the same result, namely, that at least phenomenal states have characteristic features that in the last consequence are not public, since they are neither tied to typical behaviours nor to causal roles. The idea of the philosophical zombie was born – the idea of a being that in all situations says exactly the same as I say, and does exactly the same as I

[10] Nagel 1974, Jackson 1982, Levine 1983, 1993, Chalmers 1996.

do, but whose phenomenal states are – on this assumption – either connected with radically different qualia or with none at all.

It seems to me that those who claim that philosophical zombies are possible have strayed a step too far from Wittgensteinianism. For this assumption has a number of consequences that cast doubt on its coherence. One of these consequences has been very clearly spelled out by Levine himself in his 1997 paper „Recent Work on Consciousness". Suppose, there is a creature that has the same functional structure as me, but whose functional states are not connected with any qualia. Suppose this creature is a functional zombie. According to the assumption, there are states within this creature – let's call him ‚Zansgar' – that play the same causal role as my sensations; but not only that, in this creature there are even states that play the same causal role as the beliefs that I have with regard to my sensations. Let us call these states Z-beliefs. Obviously, notwithstanding the differences, there is a great deal of similarity between my beliefs and Zansgar's Z-beliefs. Zansgar's Z-beliefs for example will make him say about himself: „Of course my phenomenal states are accompanied by certain qualitative experiences", even if this is false according to the assumption. Let us assume further, that the states through which Zansgar's Z-sensations are realised, are one-by-one replaced by states which not only play the right causal roles, but also are accompanied by the qualia belonging to that state. Would Zansgar notice any difference? Or Z-notice a difference? This does not seem to be the case since there is no change in the causal roles of his states. But how can one say under these circumstances that there is a significant difference between Zansgar and me? After all, by the reverse process I could be changed into Zansgar – without noticing anything.

> What's going on? The very intuitive stance, the first-person point of view, that fuelled the pro-zombie intuition, seems now to be undermining it. On the one hand, from within I seem aware of a feature of mental life whose absence I can so clearly conceive of in another. Yet, allowing for that absence in another seems to open up the possibility that its presence or absence makes no discernible difference, and that includes no discernible difference to me. (Levine 1997, 385)

If this is the case and there is no discernible difference whether a certain functional state is accompanied by a certain qualitative experience or not, how could it make sense to distinguish between functional states with and without this experiential quality? This truly seems to be an example of the Wittgensteinian wheel that is not part of the mechanism.[11]

[11] See *PI* § 271.

So, there are good reasons for following the new orthodoxy insofar as it holds the view that speaking about the mental is not different as a matter of principle from speaking about the physical and that therefore mental explanations can be regarded as causal explanations. However, one should become sceptical if the Cartesian picture of the mind creeps in through the back door. Wittgenstein's Anticartesianism and that of his followers is an achievement we should not fall behind.

*References*

Beckermann, A. 2000: „The perennial problem of the reductive explainability of phenomenal consciousness – C. D. Broad on the explanatory gap". In: T. Metzinger (Hg.) *Neural Correlates of Consciousness – Empirical and Conceptual Questions*. Cambridge MA: MIT-Press, 41–55. (Dt. Fassung in diesem Band Beitrag 2)

– 2001: *Analytische Einführung in die Philosophie des Geistes*. 2nd ed., Berlin/New York: de Gruyter.

Broad, C. D. 1925: *The Mind and Its Place In Nature*. London: Kegan Paul, Trench, Turbner & Co.

Carnap, R. 1936/1937: „Testability and Meaning". *Philosophy of Science* 3, 419–471, and *Philosophy of Science* 4, 1–40.

Chalmers, D. 1996: *The Conscious Mind*. Oxford: Oxford University Press.

Feigl, H. 1950a: „The Mind-Body Problem in the Development of Logical Empiricism". *Revue Internationale de Philosophie*. Repr. in: H. Feigl & M. Brodbeck (Hg.) *Readings in the Philosophy of Science*. New York: Appleton-Century-Crofts 1953, 612–626.

– 1950b: „Existential Hypotheses: realistic versus phenomenalistic interpretations". *Philosophy of Science* 17, 35–62.

Gean, W. D. 1965/1966: „Reasons and Causes". *Review of Metaphysics* 19, 667–688.

Glock, H.-J. 1996: *A Wittgenstein Dictionary*. Oxford: Blackwell.

Hempel, C. G. 1952: *Fundamentals of Concept Formation in Empirical Science*. Chicago: The University of Chicago Press.

Jackson, F. 1982: „Epiphenomenal Qualia". *Philosophical Quarterly* 32, 127–136.

Kim, J. 1996: *Philosophy of Mind*. Boulder COL: Westview Press.

Levine, J. 1983: „Materialism and Qualia: The Explanatory Gap". *Pacific Philosophical Quarterly 64*, 354–361.

– 1993: „On Leaving Out What It's Like". In: M. Davies & G. W. Humphreys (Hg.) *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell, 121–136.

– 1997: „Recent Work on Consciousness“. *American Philosophical Quarterly* 34, 379–404.

Lewis, D. 1986: „Causation“. In: David Lewis, *Philosophical Papers. Volume 2*. Oxford: Oxford University Press, 159–172.

Nagel, T. 1974: „What is it like to be a bat?“ *Philosophical Review* 83, 435–450.

Mackie, J. L. 1965: „Causes and Conditions“. *American Philosophical Quarterly* 2, 245–255.

Ryle, G. 1949: *The Concept of Mind*. Hutchinson 1949 (Reprint: Penguin Books 1963).

Sellars, W.S. 1948: „Realism and the new way of words“. *Philosophy and Phenomenological Research* 8, 601–634 .

ter Hark, M. 1995: „Wittgenstein und Russell über Psychologie und Fremdpsychisches“. In: E. von Savigny & O. Scholz (Hg.) *Wittgenstein über die Seele*. Frankfurt/M.: Suhrkamp, 84–106.

von Savigny, E. 1994: *Wittgensteins „Philosophische Untersuchungen“. Ein Kommentar für Leser. Band I.* 2nd, completely revised and augmented ed., Frankfurt/M.: Klostermann.

Wittgenstein, L. 1953: *Philosophical Investigations*. Oxford: Blackwell.

– 1967: *Zettel*. Oxford: Blackwell.

– 1969: *Blue and Brown Books*. Oxford: Blackwell.

– 1980: *Remarks on the Philosophy of Psychology I*. Oxford: Blackwell.

# 18

# Darwin – Was, wenn der Mensch auch nur ein Tier ist?*


ABSTRACT

According to Darwin, humans, just like other organisms, are not created by any special act. All organisms arise by natural processes from inanimate matter. Humans are no exception. But can it really be the case that even humans are ‚only' animals – natural beings which (a) are completely made up of natural parts (in the end, of macro-molecules which themselves consist of atoms), and for which it is (b) true that all processes that occur within them are physico-chemical processes? In recent years some German philosophers (e. g. Habermas and Wingert) have argued that man „is elevated above nature by his capacity for deliberation and his ability to judge and comprehend meanings" (Singer and Wingert 2003, 11). Does this mean that humans are ‚supernatural' in some way or other? My aim is twofold. Firstly, I point out that even non-human animals to a certain extent have the capacity to deliberate and act for reasons. Secondly, I argue that one can also putforward theoretical considerations in favour of the thesis that entirely natural beings may have this capacity. Thus, my answer to the question of what parts of our accustomed views of humanity and the world must we abandon if even we humans are ‚only' animals is: nothing or at any rate very little. At least, even if we humans are ‚only' animals, we may have the capacity for deliberation and the ability to judge and comprehend meanings.


## *1. Darwins Evolutionstheorie als Kränkung der Eigenliebe des Menschen*

150 Jahre nach seinem Erscheinen stellt Darwins *On the Origin of Species* immer noch eine der ganz großen Herausforderungen für das Selbstverständnis des Menschen dar. Zumindest auf den ersten Blick hat dieses Werk etwas äußerst Beunruhigendes. So ist es kein Wunder, dass Freud die Evolutionstheorie zu den großen „Kränkungen ihrer naiven Eigenliebe" rechnet, die „die Menschheit im Laufe der Zeiten von der Wissenschaft [hat] erdulden müssen."

> Die erste, als sie erfuhr, daß unsere Erde nicht der Mittelpunkt des Weltalles ist, sondern ein winziges Teilchen eines in seiner Größe kaum vorstellbaren Weltsystems. Sie knüpft sich für uns an den Namen Kopernikus […]. Die zweite dann, als die biologische Forschung das angebliche Schöpfungsvorrecht des Menschen zunichte machte, ihn auf die Abstammung aus dem Tier-

---


* Deutsche Fassung von „Darwin – What if Man is Only an Animal, After All?" *Dialectica* 64 (2010), 467–482.

> reich und die Unvertilgbarkeit seiner animalischen Natur verwies. Diese Umwertung hat sich in unseren Tagen unter dem Einfluss von Ch. Darwin, Wallace und ihren Vorgängern nicht ohne das heftigste Sträuben der Zeitgenossen vollzogen. Die dritte und empfindlichste Kränkung aber soll die menschliche Größensucht durch die heutige psychologische Forschung erfahren, welche dem Ich nachweisen will, daß es nicht einmal Herr ist im eigenen Hause, sondern auf kärgliche Nachrichten angewiesen bleibt von dem, was unbewusst in seinem Seelenleben vorgeht. (Freud 1917, 294 f.)

Allerdings: Sind diese vermeintlichen Kränkungen der Eigenliebe der Menschheit tatsächlich so schwerwiegend? Warum z. B. sollten wir den Umbruch vom geo- zum heliozentrischen Weltbild überhaupt als Kränkung empfinden? Was ist so schlimm daran, dass die Erde nicht den Mittelpunkt der Welt bildet? Gut, wir stehen nicht mehr im Zentrum und verlieren vielleicht etwas an gefühlter Wichtigkeit. Aber ändert das etwas für unser Leben? Heißt es, dass wir jetzt keine Rechte und keine Würde mehr besitzen? Sicher nicht, denn unser ethischer Status und unsere Würde hängen auf keinen Fall davon ab, ob die Erde den Mittelpunkt der Welt bildet. Und: Man kann die Dinge auch ganz anders sehen. Man kann bewundernd auf das schier unendliche Universum schauen, auf die vielen Galaxien, die es enthält, und dann zu sich sagen: Hier in dieser Galaxie – der Milchstraße – gibt es eine Sonne, um die unsere Erde als ziemlich kleiner Planet kreist. Und auf dieser kleinen Erde gibt es uns Menschen, die wir uns das alles anschauen können. Ist das nicht wunderbar?

Der Fall Darwin ist anders. Jahrtausende haben sich Menschen etwas auf ihre Sonderstellung eingebildet. Sie fühlten sich der übrigen Natur weit überlegen; ja, das Christentum bescheinigte ihnen einen ganz besonderen Status, da Gott sie nach seinem Ebenbild erschaffen habe. Der Mensch sei durch seine Teilhabe am Göttlichen dem Reich des nur Natürlichen entrückt. In seiner Gottesebenbildlichkeit sei er anders als alle übrigen Naturdinge. Und in seiner Gottesebenbildlichkeit gründe auch sein ethischer Status – seine Würde, die ihm Rechte verleihe, die reine Naturdinge nicht haben.

Es gehört zu unserem platonischen Erbe, dass wir dazu neigen, in der Welt den Bereich des Geistigen streng vom Bereich des Körperlich-Natürlichen zu unterscheiden. Und seit Platon denken viele, dass wir Menschen zumindest mit unserer Seele dem Bereich des Geistigen zugehören. Dieses Bild wird durch Darwin bis in seine Grundfeste erschüttert. Darwin verankert den Menschen fest in der Natur. Von einem göttlichen Funken weit und breit keine Spur. Wie die anderen Lebewesen wird auch der Mensch nicht in einem besonderen Akt geschaffen. Vielmehr entwickeln sich Lebewesen in einem natürlichen Prozess aus unbelebter Materie. Und im Bereich der Lebewesen entwickeln sich die komplexeren und leistungsfähigeren Formen ebenfalls in einem natürlichen Prozess aus den einfacheren und

weniger komplexen. Der Mensch bildet keine Ausnahme. Es gibt keine – zumindest keine stichhaltigen – Anhaltspunkte dafür, dass wenigstens in unserem Fall ein Schöpfer in die natürliche Entwicklung der Dinge eingegriffen hat, um wenigstens uns mit einem göttlichen Funken, einer immateriellen Seele auszustatten, die uns über die übrigen Lebewesen hinaushebt.

Sicher, die Evolutionstheorie allein impliziert keinen Naturalismus. Schließlich ist es zumindest möglich, dass an irgendeinem Punkt der evolutionären Entwicklung das Seelische als völlig neues Phänomen emergiert – als Phänomen, das naturwissenschaftlich nicht erklärbar ist. Doch die Evolutionstheorie stützt – zusammen mit der Entdeckung, dass sich biologische Phänomene grundsätzlich auf chemische Vorgänge zurückführen lassen – die Annahme, dass in der natürlichen Welt alles mit rechten Dingen zugeht. So wie in der Entwicklung der Lebewesen keine Punkte erkennbar sind, an denen ein Schöpfer in diesen Prozess eingegriffen hat, gibt es auch keine Punkte, an denen wirklich Neues, naturwissenschaftlich nicht Erklärbares entstanden ist. Es ist genau, wie Freud gesagt hat: Die Evolutionstheorie zeigt dem Menschen seine „Abstammung aus dem Tierreich" und verweist ihn damit auf „die Unvertilgbarkeit seiner animalischen Natur". Mit anderen Worten: Die Evolutionstheorie ist ein wichtiges Indiz dafür, dass auch wir tatsächlich „nur" Tiere sind. Und diese Auffassung scheint in der Tat eine erhebliche Kränkung unserer Selbstliebe darzustellen.

Allerdings: Was bedeutet es tatsächlich, wenn auch wir „nur" Tiere sind? Was an unserem herkömmlichen Welt- und Menschenbild müssen wir aufgeben, wenn Darwin Recht hat?

Zunächst liefert die Evolutionstheorie sicher ein starkes Argument gegen den Platonisch-Cartesischen Dualismus. Es wäre schon sehr merkwürdig, sich die Evolution als einen Prozess vorzustellen, bei dem sich nach und nach aus komplizierten Makromolekülen immer komplexere Lebewesen entwickeln, dass aber Menschen erst entstehen, wenn den am höchsten entwickelten Lebewesen zusätzlich eine immaterielle Seele eingehaucht wird. Nichts spricht für diese Annahme. Menschen sind ebenfalls Produkte der Evolution; alles, was sie zu Menschen macht, hat eine rein biologische Grundlage. Doch damit wird die Frage nur verschoben. Was an unserem herkömmlichen Welt- und Menschenbild müssen wir aufgeben, wenn es zutrifft, dass wir keine Platonisch-Cartesische Seele besitzen?

Als erstes ist wohl klar, dass damit die Hoffnung auf ein Weiterleben nach dem Tode einen erheblichen Dämpfer erhält. Natürlich können wir dieses Weiterleben auch an der Auferstehung des Fleisches oder an irgendeiner Form der Reinkarnation festmachen. Aber sehr viele, ich denke, die allermeisten von uns sind in dieser Frage Dualisten. Wir wissen, dass unser Körper sterben wird, denken oder hoffen aber trotzdem, dass damit noch

nicht alles vorbei ist. Ich finde es absolut verblüffend, dass wir aufgeklärten Mitteleuropäer im Kreationismusstreit meist eindeutig auf der Seite der Evolutionsbiologie stehen, andererseits aber ebenso hartnäckig an der Erwartung eines Lebens nach dem Tode festhalten. Sicher, diese Idee hat viel Attraktives. Auch ich würde gerne meine Eltern und andere liebe Menschen wiedersehen, die schon gestorben sind. (Es gibt aber auch einige, denen ich nicht gern noch einmal begegnen würde.) Allerdings glaube ich, dass es dafür wenig Hoffnung gibt. Meiner Meinung nach ist es unehrlich und inkonsequent, den Menschen einerseits als durch und durch biologisches Wesen zu verstehen, andererseits aber auf ein dualistisch verstandenes Leben nach dem Tode zu hoffen (siehe auch Beckermann 2012).

Zweitens ist immer wieder behauptet worden, dass wir unmöglich frei sein können, wenn wir keine Seele besitzen. Doch dies ist nichts weiter als ein gängiges Missverständnis. Richtig ist, dass unsere Handlungen weder auf ein immaterielles Ich noch auf ebenfalls immaterielle, selbst unverursachte Willensakte zurückgehen, wenn es keine Seele gibt. Und richtig ist auch, dass wir anders, als schon Platon dachte, nicht in der Lage sind, als Erstursachen Kausalketten spontan neu zu beginnen. Aber beides ist nicht nötig, um frei zu sein. Freiheit hängt ab von der Fähigkeit zur Handlungskontrolle; und diese Fähigkeit zu haben bedeutet, Impulskontrolle ausüben zu können, d. h., vor dem Handeln innehalten und überlegen zu können sowie nach dem Überlegen dem Ergebnis der Überlegung gemäß handeln zu können. Diese Fähigkeiten können wir auch besitzen, wenn wir rein biologische Wesen sind (vgl. Beckermann 2008, Kap. 3).

## *2. Nicht-menschliche Tiere und die Fähigkeit zu überlegen*

Die Willensfreiheitsdebatte der letzten Jahre ist aber noch aus einem anderen Grund interessant, wenn man versucht, eine Antwort auf die Frage zu finden, was wir am herkömmlichen Welt- und Menschenbild ändern müssen, wenn wir uns selbst als durch und durch natürliche Wesen begreifen. Von einigen deutschen Philosophen ist nämlich *gegen* Neurobiologen wie Roth und Singer ein Argument vorgebracht worden, das auch Implikationen für diese Frage hat. Vordergründig geht es bei diesem Argument um eine Kritik des Anspruchs der Naturwissenschaften auf ein umfassendes Deutungsmonopol der gesamten Welt. Im Hintergrund spielt aber auch die These ein Rolle, dass es unmöglich ist, den Menschen nur als Naturwesen zu sehen. Wie sieht dieses Argument aus?

In einem Streitgespräch mit Wolf Singer, das zuerst vor ein paar Jahren unter dem Titel „Wer deutet die Welt?“ in der ZEIT veröffentlicht wurde, sagt Lutz Wingert:

> Der Mensch hebt sich als ein Wesen, das zum Nachdenken, zum Beurteilen und zum Verstehen von Bedeutungen fähig ist, *aus der Natur heraus*. Ein Großteil seiner Welt besteht aus sinnhaft konstruierten Gegenständen – wie Aussagen, Zehnmarkscheine oder politische Verfassungen. Solche Gebilde sind traditionsgemäß der Gegenstand geisteswissenschaftlicher Disziplinen. Die Naturwissenschaft dagegen untersucht sinnfreie Gegenstände. (Singer/Wingert 2000, 11 – meine Hervorh.)

Und:

> Die Frage ist doch, ob wir den Menschen auch als ein urteilendes und wertendes Wesen – und nicht nur in seiner organischen Existenz – als Teil der Natur auffassen können. Können wir den Menschen komplett, wie andere Teile der Natur auch, allein mit den Mitteln der disziplinierten Naturbetrachtung beschreiben? Oder verschwindet unter dieser Beschreibung nicht doch ein zentrales Element von uns, nämlich all das, was mit der Fähigkeit zur Metarepräsentation und zur Selbstkritik zu tun hat? Ich glaube, ja. Gewiss, wir sind kein Kopf ohne Welt, wir sind mit unserem Denken und Handeln in der Welt. Aber die Wirklichkeit ist nicht bloß die Natur […]. (ebd., 17)

Und schließlich:

> Freiheit ist doch nicht bloß eine Vorstellung! Sie ist auch ein Zustand, in dem ich mich als fähig erfahre, zu sagen: Das war ich! Das tue ich! Das heißt, ich kann dann ein Verhalten, das ein Beobachter mir als Organismus kausal zuordnet, auch als mein Handeln anerkennen; und zwar deshalb, weil es aus Gründen erfolgt, die ich als meine – schlechten oder guten – Gründe erkenne, und nicht bloß aus Ursachen, die in mir liegen. Und ich kann mich auch täuschen, frei zu sein. Dann sagen die anderen: Du rationalisierst bloß! Aber so kann man nur reden, wenn der Unterschied von bloßen Ursachen für Verhalten und rechtfertigenden Gründen fürs Handeln bestehen bleibt. (ebd., 13)

Ich entdecke in diesen Ausführungen zwei Hauptmotive. Erstens: Ein Großteil der Welt des Menschen besteht aus sinnhaft konstruierten Gegenständen – darunter Aussagen mit einem semantischen Gehalt, der wahr oder falsch sein kann –, und solche Gegenstände sind keine natürlichen Gegenstände, sie können nicht mit naturwissenschaftlichen Mitteln untersucht werden. Zweitens: Für menschliches Handeln gibt es in aller Regel nicht nur Ursachen, sondern auch Gründe. Und auch dieses Reich der Gründe ist kein möglicher Gegenstand naturwissenschaftlicher Untersuchungen und Erklärungen. Beide Motive zusammen stützen, so Wingert, die These: „Der Mensch hebt sich als ein Wesen, das zum Nachdenken, zum Beurteilen und zum Verstehen von Bedeutungen fähig ist, aus der Natur heraus." „Die Wirklichkeit ist nicht bloß die Natur."

In einem neueren Aufsatz listet Wingert eine Reihe von lebensweltlichen Gewissheiten auf, in denen sich dieselben Motive wieder finden (Wingert 2008, 290 f.):

- […] Menschen können für ihr Tun so verantwortlich sein, dass sie dafür Lob und Tadel, Wertschätzung und Verachtung verdienen.
- Menschen haben bisweilen einen Spielraum effektiven praktischen Überlegens. Gründe sind wenigstens manchmal wirksam für Handlungen.
- Die Handlungswirksamkeit von Gründen hängt von einer Bejahung der Gründe durch den Handelnden ab. Gründe wirken nicht so wie Pillen, wenn sie denn wirken.
- Menschen haben einen Sinn für symbolische Zeichen. […]
- Menschen verhalten sich auch regelgeleitet und nicht bloß regelmäßig.
- Menschen können aufgefordert werden, etwas zu tun/zu unterlassen, und nicht bloß dazu gebracht werden, etwas zu tun bzw. zu unterlassen. Sie können den Status eines Adressaten von normativen Erwartungen oder eines Sollens erwerben.

Wingert interessiert sich primär für den epistemischen Status dieser Aussagen und besonders für die Frage, wie sich diese Gewissheiten zu den Ergebnissen wissenschaftlicher Forschung verhalten. Das ist hier nicht mein Thema; mir geht es allein um die Frage, ob diese Aussagen auch dann wahr sein können, wenn Menschen doch „nur" Tiere, doch „nur" durch und durch natürliche Wesen sind. Um die Antwort vorwegzunehmen: Meiner Meinung nach lautet die Antwort eindeutig Ja. Um dieses Antwort zu begründen, muss ich etwas ausholen.

a) In der Natur begegnen uns nicht nur Elementarteilchen, auch nicht nur physische Körper und chemische Verbindungen. In der Natur begegnen uns auch Lebewesen. Sicher, auch Lebewesen sind „nur" große Aggregate chemischer Verbindungen; aber diese Aggregate *als Lebewesen zu verstehen*, heißt zu erkennen, dass sie über Eigenschaften und Fähigkeiten verfügen, die andere physische Gegenstände nicht besitzen. Zentral sind die Fähigkeiten zur Selbsterhaltung und zur Reproduktion. Selbsterhaltung ist deshalb zentral, weil es sich bei Lebewesen – im Gegensatz zu den meisten Maschinen – um physische Strukturen mit einer *Tendenz zum spontanen Zerfall* handelt. Man kann einen Hund nicht – wie einen Staubsauger – in den Schrank stellen und nach vier Wochen wieder herausholen. Lebewesen müssen ständig gegen die Tendenz zum spontanen Zerfall anarbeiten. Sie sind in der Tat *autopoietische* Systeme, die immerwährend aktiv ihre Struktur erhalten, da sie sonst zugrunde gehen. Wenn die Mechanismen der aktiven Strukturerhaltung versagen, ist der Tod die unausweichliche Folge. Für das Funktionieren der aktiven Strukturerhaltung benötigen Lebewesen einen Stoffwechsel, der es ihnen erlaubt, aus der Umwelt Stoffe aufzunehmen, die für die Strukturerhaltung nötig sind. Für diesen Vorgang wiederum benötigen sie Energie, diese gewinnen sie in vielen Fällen aus Verbren-

nungsvorgängen, für die wiederum Nahrung und Sauerstoff erforderlich sind. Alle diese Tatsachen sind bekannt.

Mir ist hier Folgendes wichtig. Lebewesen, zumindest alle nicht-menschlichen Lebewesen, sind nach unserem heutigen Kenntnisstand durch und durch natürliche Wesen. Was bedeutet hier „natürlich“? Ein natürliches Wesen ist (a) ein Wesen, das vollständig aus natürlichen Teilen besteht – letzten Endes aus Makromolekülen, die ihrerseits aus Atomen aufgebaut sind. Ein natürliches Wesen ist (b) ein Wesen, bei dem alle in ihm ablaufenden Porzesse physikalisch-chemische Prozesse sind. Dass alle (nicht-menschlichen) Lebewesen durch und durch natürliche Wesen sind, bedeutet deshalb, dass sie nur aus natürlichen Teilen bestehen und dass alle für Lebewesen charakteristischen Prozesse – wie Nahrungsaufnahme, Verdauung, Atmung, Photosynthese, Fortpflanzung usw. – physikalisch-chemische Prozesse oder durch physikalisch-chemische Prozesse realisiert sind. Trotzdem: Einen physischen Gegenstand als Lebewesen zu verstehen, heißt die Funktion dieser Vorgänge zu erkennen, zu erkennen, dass sie alle der aktiven Strukturerhaltung dienen. In der Natur gibt es also nicht einfach nur chemische Prozesse; es gibt auch chemische Prozesse, die zugleich Prozesse der aktiven Strukturerhaltung sind; aber das ändert nichts am natürlichen Charakter dieser Prozesse.

Weil dies für meine Überlegungen von großer Bedeutung ist, möchte ich diesen Punkt noch einmal kurz zusammenfassen. (Nicht-menschliche) Lebewesen sind physikalisch-chemische Systeme, deren Struktur die folgenden Merkmale aufweist:

(i) Die Struktur hat eine Tendenz zum spontanen Zerfall; wenn nichts passiert, löst sich die Struktur eher früher als später auf.

(ii) Die Teile der Struktur wirken so zusammen, dass die Teile selbst und ihr Zusammenhang immer wieder repariert und erneuert werden; das System hat die Fähigkeit zur aktiven Selbsterhaltung; diese Fähigkeit beruht auch darauf, dass das System in der Lage ist, Stoffe aus der Umwelt aufzunehmen und in die Umwelt abzugeben.

(iii) Das System hat die Fähigkeit zur Fortpflanzung, d. h., es kann „Ableger“ erzeugen, die sich in der Interaktion mit der Umwelt von sich aus zu einem vollständigen System derselben Art entwickeln.

Dass ein Lebewesen ein physikalisch-chemisches System ist, soll heißen: Das System enthält nur physikalisch-chemische Teile. Die in dem System ablaufenden physikalisch-chemischen Prozesse folgen allein den grundlegenden Naturgesetzen. Alle höherstufigen Prozesse, Eigenschaften und Fähigkeiten lassen sich vollständig auf die niederstufigen physikalischen und chemischen Prozesse zurückführen.

b) Für die meisten Lebewesen sind drei Dinge zentral – Nahrung, Gefahrenvermeidung, Partnersuche. Wenn die Nahrung nicht da ist, wo das Lebewesen ist, ist es offenbar nützlich, wenn sich das Lebewesen bewegen kann – wenn es sich *zielgerichtet* bewegen kann; denn es will ja nicht irgendwo hin, sondern dorthin, wo sich die Nahrung befindet. Zielgerichtete Bewegung wiederum setzt zumindest rudimentäre Wahrnehmung voraus. Wenn ich nicht erkennen kann, wo sich die Nahrung befindet, hilft mir meine Bewegungsfähigkeit gar nichts. Der Zusammenhang zwischen Wahrnehmung und Bewegung beruht am Anfang auf einfachen Reiz-Reaktions-Mechanismen: helles Licht – Flucht, Berührung – Zuschnappen. Aber das reicht natürlich nicht, wenn ein Eichenhörnchen den Baum wiederfinden will, unter dem es einige Haselnüsse vergraben hatte. Die Evolution hat deshalb dafür gesorgt, dass sich Lebewesen ein von der aktuellen Wahrnehmung unabhängiges dauerhaftes „Bild" ihrer Umgebung machen können. Unter anderem dafür benötigen sie ein Gehirn. Offenbar gibt es in diesem Gehirn Strukturen, in denen zumindest einige Informationen über die jeweilige Umwelt gespeichert sind. Wir wissen nicht wirklich, wie diese Strukturen aussehen. Aber aus dem Verhalten von Lebewesen können wir schließen, dass es sie gibt.

Auch das ist alles nicht spektakulär. Aber vielleicht spürt man doch, dass wir jetzt schon ganz dicht dran sind an dem Punkt, an dem Bedeutungen und semantische Gehalte ins Spiel kommen. Wenn es Strukturen in den Gehirnen von Lebewesen gibt, die Merkmale ihrer Umwelt repräsentieren, dann haben diese Strukturen einen semantischen Inhalt. Und natürlich kann sich das Eichhörnchen den falschen Baum merken. D. h., schon auf dieser Ebene kommen die Begriffe wahr und falsch ins Spiel. Nicht für das Eichhörnchen selbst. Aber objektiv gesehen, kann das Eichhörnchen seine Umwelt richtig, aber auch falsch repräsentieren. Übrigens: Das Verhalten von Lebewesen kann schon auf dieser Stufe als ‚intentional' charakterisiert werden. Denn Lebewesen bewegen sich nicht einfach; sie bewegen sich zielgerichtet – sie laufen zur Wasserstelle, fliehen ins Unterholz, sie laufen dem Partner hinterher usw. Mit anderen Worten: Sie führen zwar eine physische Bewegung aus; aber entscheidend ist nicht der physikalische Charakter dieser Bewegung, sondern dass sie zu einem bestimmten Ziel führt. Schon das Greifen einer Banane ist eine zielgerichtete Bewegung.

Dass auch nicht-menschliche Tiere zu Wahrnehmung und zielgerichtetem Verhalten fähig sind, bedeutet offenbar nicht, dass hier irgendetwas Übernatürliches im Spiel wäre. In den Gehirnen dieser Lebewesen laufen elektro-chemische Prozesse ab. Aber wir verstehen nicht wirklich, was vorgeht, wenn wir nicht sehen, dass es sich bei diesen Prozessen z. B. um Wahrnehmungsprozesse handelt – neuronale Prozesse, die zu Strukturen führen, die als Repräsentationen von Aspekten der Umwelt dienen. (Wahr-

nehmung besteht allerdings nicht allein in der Schaffung von Strukturen, die Teile der Umgebung des Lebewesens repräsentieren; diese Strukturen müssen auch bei der Leitung des zielgerichteten Verhaltens des Lebewesens eine entscheidende Rolle spielen.)

c) Wenn wir nur einen kleinen Schritt weitergehen, kommen zum ersten Mal Gründe ins Spiel. *Epistemische* Gründe sind Umstände, die dafür sprechen, dass bestimmte Annahmen wahr oder falsch sind. Denken Sie etwa an das Beispiel eines jungen Affen, der mitten auf dem Futterplatz eine Banane sieht. Soll er sich den Leckerbissen holen und verspeisen? Ganz so einfach ist die Sache nicht; in unmittelbarer Nähe befindet sich ein älteres, ranghöheres Männchen. Wird dieses Männchen zulassen, dass sich der jüngere Affe die Banane holt? Oder wird es die Banane für sich reklamieren und den jungen Affen ganz furchtbar vermöbeln, wenn der so dreist sein sollte, die Banane für sich zu beanspruchen? Der junge Affe muss abschätzen, was passieren wird. Wie macht er das? Er beobachtet das ältere Männchen. Schläft es? Ist es abgelenkt? Wohin blickt es? Macht es einen aggressiven Eindruck? Oder döst es ganz ruhig vor sich hin? Aufgrund dessen, was er beobachtet, kommt der junge Affe zu einer Einschätzung der Lage: Es ist gefährlich oder eben nicht gefährlich, wenn ich mir jetzt die Banane hole. Er nutzt das, was er beobachtet, als Grund für diese Einschätzung. Wieder gilt natürlich: Der Affe weiß nicht, was ein Grund ist. Dennoch nutzt er Informationen, die ihm zur Verfügung stehen, als Gründe für seine Beurteilung der Situation.

d) Wenn nicht-menschliche Lebewesen, intelligente Lebewesen wie Affen, in Gruppen leben, wird der Prozess der kognitiven Entwicklung einen wichtigen Schritt vorangetrieben. Denn für solche Lebwesen wird es, wie wir eben gesehen haben, geradezu lebenswichtig, nicht nur ihre physische Umwelt, sondern auch ihre Mitlebewesen in dieser Umwelt zu repräsentieren. Und diese Mitlebewesen zu repräsentieren, bedeutet insbesondere auch, ihre psychische Verfassung zu repräsentieren. Sind sie aggressiv? Oder gelangweilt? Wohin schauen sie? Was können sie sehen? Hier geht es um eine wie auch immer rudimentäre *theory of mind*, die zumindest Folgendes umfasst: Emotionen und Stimmungen, von denen das Verhalten entscheidend abhängt; Wahrnehmungen (Affen können recht gut einschätzen, was ein anderer Affe sieht und was nicht); und schließlich Erwartungen.

Fellpflege ist bei Affen sehr beliebt. Aber dazu braucht man einen Partner. Nichts liegt also näher, als sich diesem Partner zu nähern und ihm möglichst eindeutig zu verstehen zu geben, dass man gerne „gelaust“ werden möchte. Die Partner verstehen das auch; sie „wissen“, was der andere Affe von ihnen will; sie „wissen“, was er erwartet. Dasselbe gilt für Sex, wobei es, wie sich herausgestellt hat, einen interessanten Zusammenhang

gibt. Bei Javaneraffen in Indonesien wurde beobachtet, dass sich die Weibchen 1,5-mal pro Stunde mit einem der Männchen paarten. Nach Perioden ausgiebiger Fellpflege durch die Männchen stieg die Sexrate auf 3,5 Mal pro Stunde. Dabei boten die Weibchen vor allem denjenigen Männchen Sex an, von denen sie die Fellpflege erhalten hatten (Barras 2008). Wie dem auch sei, Affen haben manchmal Erwartungen an ihre Mitaffen und sie drücken diese Erwartungen auch aus. Schon auf dieser Stufe gibt es Aufforderungen. Diesen Aufforderungen muss der Adressat nicht nachkommen – er kann dies tun, er kann es aber auch lassen – in vielen Fällen, ohne unangenehme Konsequenzen befürchten zu müssen.

Ich beende diesen kurzen Abriss der Entwicklung kognitiver Fähigkeiten mit einem Zwischenfazit: Vieles von dem, was Lutz Wingert für den Menschen reklamiert, findet sich auch schon früher im Tierreich. Wir Menschen sind nicht so einzigartig, wie wir glauben. Aber ich höre natürlich schon den Einwand: Bei nicht-menschlichen Tieren gibt es doch gar keine wirkliche Intentionalität und nicht-menschliche Tiere handeln auch nicht wirklich aus Gründen. Erst beim Menschen treten diese Phänomene im Vollsinn auf. Ohne Zweifel gibt es im Hinblick auf die Fähigkeit, aus Gründen zu handeln, Unterschiede zwischen Menschen und nichtmenschlichen Tieren. Doch was folgt aus diesen Unterschieden? Ich will mich hier auf das Handeln aus Gründen beschränken. Meine Ausgangsfrage lautete, ob auch Menschen Tiere sind, d. h. ob auch Menschen durch und durch natürliche Wesen sind (vgl. oben S. 335). Aus Gründen Handeln scheint das Merkmal zu sein, an dem sich diese Fragen entscheidet. Kann ein Wesen in der Weise aus Gründen handeln, wie wir Menschen es tun, wenn es „nur“ rein natürliches Wesen ist? Oder zeigt unsere Fähigkeit, aus Gründen zu handeln, dass wir Menschen „aus der Natur herausgehoben sind“? Diese Frage scheint mir zumindest angedeutet in Formulierungen wie der Habermasschen vom „eigentümlich zwanglosen Zwang des besseren Arguments“ oder auch in der dritten lebensweltlichen Gewissheit Wingerts: „Die Handlungswirksamkeit von Gründen hängt von einer Bejahung der Gründe durch den Handelnden ab. Gründe wirken nicht so wie Pillen, wenn sie denn wirken“ (Wingert 2008, 290). Was steckt hinter diesen Bemerkungen? Wahrscheinlich die folgende Überlegung.

Naturalisten, das scheint hinter diesen Thesen zu stehen, behaupten, dass in der Natur ein durchgängiger Kausalzusammenhang herrscht. Wenn alles Natur wäre, hätten Gründe daher keinen Platz in der Welt. Denn das Reich der Ursachen ist kategorial verschieden vom Reich der Gründe. Wenn wir jemanden mit Gründen konfrontieren, *fordern wir ihn auf*, etwas Bestimmtes zu tun oder zu unterlassen; aber wir versuchen nicht, *ihn kausal dazu zu bringen*. Gründe wirken nicht wie Ursachen; aber wir können mit Gründen den Gang der Welt beeinflussen. Also kann in der Natur kein durchgängi-

ger Kausalzusammenhang herrschen. Wer den Unterschied zwischen Gründen und Ursachen leugnet, leugnet auch den Unterschied zwischen Tun und Geschehen sowie zwischen Überzeugen und Überreden. Reden wir also über Gründe und das Handeln aus Gründen.

### *3. Natürliche Wesen und die Fähigkeit, aus Gründen zu handeln*

In letzter Zeit ist gegen eine verbreitete Auffassung in der Handlungstheorie des Öfteren betont worden, Gründe seien *keine mentalen Zustände*, sondern *Umstände draußen in der Welt* (vgl. etwa Bittner 2005). *Dass dunkle Wolken aufziehen*, ist ein Grund dafür, den Regenschirm mitzunehmen; *dass Wasser den Durst löscht*, ist ein Grund dafür, Wasser zu trinken, wenn man Durst hat. Das ist richtig. *Epistemische* Gründe sind Umstände, die *prima facie* dafür sprechen, dass eine Überzeugung wahr ist; *praktische* Gründe sind Umstände, die *prima facie* dafür sprechen, eine Handlung aufzuführen. Aber: So verstandene Gründe rechtfertigen, sie erklären nicht. *Dass dunkle Wolken aufziehen*, führt nicht dazu, dass ich den Regenschirm mitnehme, wenn ich es nicht bemerke. Erst die entsprechende Überzeugung führt – zusammen mit anderen Bedingungen – zur Handlung; und diese Überzeugung führt zur selben Handlung, auch wenn sie nicht wahr ist, wenn also keine dunklen Wolken aufziehen. Wenn ich den Regenschirm mitnehme, dann also deshalb, weil ich *glaube*, dass dunkle Wolken aufziehen, und nicht, weil tatsächlich dunkle Wolken aufziehen. Man muss unterscheiden zwischen *Gründen* und dem *Haben von Gründen*.[1] Gründe *rechtfertigen*, das Haben von Gründen *erklärt*. Deshalb lässt sich als erstes festhalten: Jemand handelt aus Gründen, wenn er handelt, weil er davon überzeugt ist, dass bestimmte Umstände vorliegen, und wenn diese Umstände dafür sprechen, diese Handlung auszuführen.

Nehmen wir noch einmal epistemische Gründe. Eines kann man meines Erachtens auf jeden Fall sagen: Es gibt *rein physische* Systeme – ich nenne sie kognitive Systeme –, die Informationen aus der Umwelt aufnehmen und auf der Grundlage dieser Informationen ihre Umwelt repräsentieren. (Ich denke hier nicht an Thermostate und Kameras, sondern an einige nichtmenschliche Lebewesen, aber auch an kleine Roberoter wie die, die an den RoboCup Turnieren teilnehmen.) Allerdings reicht das noch nicht. Jedes kognitive System steht vor der Aufgabe, seine Repräsentationen auf dem Laufenden zu halten. Es kann nicht einfach immer neue Repräsentationen erzeugen und zu den alten hinzufügen. Auf diese Weise würde ein inkohärentes und zum großen Teil veraltetes Bild seiner Umwelt entstehen. Wenn es eine neue Information aufnimmt, muss das System deshalb prüfen, wel-

[1] Zu dieser Unterscheidung vgl. Beckermann (1977), Abschn. 8.4.

che alten Repräsentationen aufgrund der neuen Information revidiert werden müssen. (Im Prinzip kann natürlich auch die neue Information verworfen werden.) Zur Lösung der Aufgabe, die jeweils richtigen Revisionen durchzuführen, gibt es inzwischen ausgeklügelte und sehr erfolgreiche Algorithmen, die sich ebenso wie andere Algorithmen physisch realisieren lassen.

Diese Zusammenhänge kann man meiner Meinung nach so beschreiben: Wenn ein kognitives System eine neue Information erhält und dies dazu führt, dass die bisherigen Repräsentationen so revidiert werden, dass sie die Umweltsituation weiterhin korrekt darstellen, dann behandelt das System die neue Information als *epistemischen Grund*. Ein epistemischer Grund ist, wie gesagt, ein Umstand, der dafür spricht, dass bestimmte Überzeugungen oder Repräsentationen wahr sind. Wenn ein kognitives System die Information erhält, dass Umstände vorliegen, die dafür sprechen, dass der Sachverhalt *A* vorliegt, ist es für das System rational, eine Repräsentation mit dem Inhalt *A* zu erzeugen (oder zu erhalten). Wenn ein System auf eine eingehende Information so reagiert, dass es die Repräsentationen erzeugt, für deren Wahrheit die Information spricht, und die Repräsentationen eliminiert, für deren Falschheit sie spricht, verhält es sich epistemisch rational. Es reagiert auf epistemische Gründe, denn es benutzt eingehende Informationen zu einer rationalen Revision der bestehenden Repräsentationen. All dies ist auch für rein physische Systeme möglich. Und dasselbe gilt in meinen Augen *mutatis mutandis* für *Handlungsgründe*. Ich will das hier aber nicht weiter erläutern. Denn ich höre schon den Einwand: Ja, aber wenn jemand *wirklich* aus Gründen handelt, dann führt er eine Handlung nicht nur aus, weil er etwas glaubt, das *de facto* für die Ausführung dieser Handlung spricht. Dann *weiß* er auch, dass das, was er glaubt, für diese Handlung spricht, dass es ein Grund dafür ist, so zu handeln. Handeln aus Gründen setzt voraus, dass man Gründe *als Gründe* erkennt. Und ist das bei rein physischen Systemen möglich?

Warum nicht? Ich hatte schon gesagt, dass kognitive Wesen nicht nur ihre physische Umwelt, sondern auch die Mitwesen in ihrer Umgebung repräsentieren müssen. Um ihre physische Umwelt repräsentieren zu können, müssen sie in der Lage sein zu erkennen, welche Objekte es in ihrer Umgebung gibt; d. h., sie müssen überhaupt Objekte von der Umgebung unterscheiden und diese Objekte in irgendeiner Weise kategorisieren können: Da ist ein Strauch, dort ein Baum, dort ein Fluss, dahinter eine Katze. Ich will mich hier nicht auf die Diskussion einlassen, ob solche Wesen über die entsprechenden Begriffe verfügen; aber dass auch nicht-menschliche Tiere verschiedene Dinge und verschiedene Arten von Dingen in ihrer Umgebung unterscheiden können, steht außer Frage. Mitwesen werden zunächst auch als Objekte in der Umgebung repräsentiert, aber als Objekte mit be-

sonderen Eigenschaften: Sie können handeln, sie haben Stimmungen, sie nehmen selbst die Umgebung wahr und reagieren auf das, was sie wahrnehmen. All dies führt, wie gesagt, in natürlicher Weise zu einer *theory of mind*. Kognitive Wesen repräsentieren ihre Mitwesen als Wesen, die Wünsche und Überzeugungen haben; sie bilden Metarepräsentationen.

Wir wissen, dass Affen die Fähigkeit besitzen, ihre Kumpane zu täuschen; sie können sie dazu bringen, etwas zu glauben, was nicht der Fall ist. Setzt das voraus, dass sie generell zwischen wahren und falschen Überzeugungen unterscheiden können? Das ist nicht recht klar. Aber was spricht dagegen, dass Wesen, die Objekte unterscheiden und kategorisieren können, die ihren Mitwesen Wünsche und Überzeugungen zuschreiben können, dass diese Wesen auch in der Lage sind zu lernen, wahre von falschen Überzeugungen zu unterscheiden? Von der Fähigkeit, Metarepräsentationen zu bilden, zur Fähigkeit, den repräsentierten Repräsentationen Eigenschaften wie Wahrheit und Falschheit zuzuschreiben, scheint es mir nur ein kleiner Schritt zu sein. Und wenn diese Stufe erreicht ist, ist auch die nächste Stufe nicht mehr weit, bei der es um Relationen zwischen Repräsentationen im Hinblick auf ihre Wahrheit geht: Die Wahrheit welcher Repräsentation spricht für die Wahrheit welcher anderen Repräsentationen, usw.? Wenn solche Relationen in den Blick kommen, kann man aber davon reden, dass manche Repräsentationen *als Gründe* für andere repräsentiert werden.[2]

Und wo bleibt das Normative? Wenn jemand einen Grund für eine Handlung hat, heißt das nicht nur, dass er ein Motiv für diese Handlung hat – also etwas, was ihnen dazu bringen kann, diese Handlung auszuführen; es heißt auch, dass es für ihn *rational* wäre, diese Handlung auszuführen, dass er diese Handlung *prima facie* ausführen *soll*. Gründe haben also immer auch einen normativen Aspekt, und wie kann eine naturalistische Sicht des Menschen dem gerecht werden?

Hinter diesem Einwand scheint mir folgende Idee zu stecken. Aus Gründen handeln heißt nicht einfach, etwas tun, weil man davon überzeugt ist, dass bestimmte Umstände vorliegen, die *de facto* für die Ausführung dieser Handlung sprechen. Es heißt auch nicht nur, eine Handlung ausführen, weil man erkennt, dass bestimmte vorliegende Umstände für diese Handlung sprechen. Es heißt, eine Handlung ausführen, weil man erkennt, dass man sie ausführen *soll*. Und wie sollen rein natürliche Wesen erkennen können, dass sie etwas sollen? Wie sollen Normen aber überhaupt im Bereich der Natur wirksam werden können?

---

[2] Wenn ich davon rede, dass Repräsentationen als Gründe behandelt oder erkannt werden, meine ich natürlich, dass ihre *Inhalte* als Gründe behandelt oder erkannt werden.

Zunächst noch einmal: Gründe bewirken nichts, auch Normen bewirken nichts. Es ist jeweils die *Überzeugung*, dass Gründe vorliegen, oder die *Überzeugung*, dass man etwas tun soll, die kausale Wirkungen zeitigt. Die Frage ist also: Können natürliche Wesen erkennen oder überzeugt sein, dass sie etwas tun sollen? Meine Antwort lautet wieder: Warum nicht? Zur Erläuterung vier Bemerkungen.

a) Kognitive Wesen sind nicht nur Wesen, die die Umwelt repräsentieren; sie sind primär Wesen, die in der Umwelt *handeln*. So wie die Fähigkeit zu handeln ohne Wahrnehmung sinnlos ist, hat umgekehrt Wahrnehmung ohne die Möglichkeit zu Handeln wenig Sinn. Der Zusammenhang zwischen Wahrnehmung und Handeln kann, das hatten wir schon gesehen, in einfachen Reiz-Reaktions-Mechanismen bestehen. Aber offensichtlich ist es im Laufe der evolutionären Entwicklung nicht bei diesen Mechanismen geblieben. Schon viele nicht-menschliche Tiere haben in vielen Situationen eine ganze Reihe von Handlungsoptionen. Und das bedeutet: Sie müssen sich entscheiden. Oder zunächst einfacher: Es muss in ihnen Prozesse geben, die aus den verschiedenen möglichen Handlungen eine auswählen und dafür sorgen, dass diese Handlung ausgeführt wird. Diese Prozesse können sehr einfach sein, sie können aber, wenn wir etwa an das Beispiel des jungen Affen denken, auch das Abwägen von Gründen beinhalten.

So wie man wahre und falsche Repräsentationen unterscheiden kann, kann man auch richtige und falsche Entscheidungen unterscheiden. Wenn man einem Affen eine Bananenattrappe aus Kunststoff hinhält, wird er sie zuerst für eine Banane halten; aber spätestens, wenn er versucht, sie zu schälen und zu essen, bemerkt er seinen Irrtum. Wenn solche Attrappen häufiger in seiner Umgebung vorkommen, wird er lernen, auf der Hut zu sein, und er wird lernen, dass er sich bei seiner Klassifizierung irren kann. Bei Entscheidungen gilt ganz Ähnliches. Wenn sich unser junger Affe – nach einigem Überlegen und Abwägen – entscheidet, sich die Banane zu holen, dann aber, weil er die Gemütslage des älteren Männchens falsch eingeschätzt hat, fürchterlich vermöbelt wird, hat er die falsche Entscheidung getroffen. Und er wird sich hüten, noch einmal einer solchen Fehleinschätzung zu unterliegen.

Was uns Menschen von unseren nächsten Verwandten unterscheidet, ist allerdings eine größere Reflektiertheit. Wir nehmen nicht nur wahr und überlegen, wir *wissen* auch, dass wir wahrnehmen und überlegen; wir entscheiden nicht nur aus Gründen, wir *wissen* auch, dass wir das tun. Das ermöglicht uns eine sehr viel größere Distanz zu uns selbst und zu anderen. Und es ermöglicht uns, uns selbst und andere jederzeit zu hinterfragen. Ist das auch richtig, was der andere glaubt und tut? Ist das auch richtig, was ich selbst glaube und tue? Dass wir diese Fähigkeiten besitzen, steht außer Frage. Es kann sein, dass diese Stufe der kognitiven Entwicklung nur er-

reicht werden kann, wenn sich zuvor die Fähigkeiten zu Selbstbewusstsein und zu sprachlicher Verständigung entwickelt haben.[3] Das kann ich hier nicht vertiefen. Allerdings sehe ich nicht, warum nicht auch die Entwicklung dieser Fähigkeiten in einem vollständig natürlichen Rahmen stattfinden kann.

b) Auf der Tagung *Naturalismus und Menschenbild* des Forums für Philosophie, auf der Lutz Wingert seinen Vortrag „Lebensweltliche Gewissheit versus wissenschaftliches Wissen?“ hielt, sprach Carl-Friedrich Gethmann zu dem Thema „Warum sollen wir überhaupt etwas und nicht vielmehr nichts?“. Und er gab auf diese Frage eine interessante Antwort: Das Sollen kommt vom Wollen. „Wir sollen etwas tun, weil und soweit jemand von uns etwas will“ (Gethmann 2008, 138). Das Sollen beruht auf den Erwartungen unserer Mitmenschen.

Nun gibt es zwei Arten von Erwartungen – faktische und normative. Faktische Erwartungen haben zum Inhalt, dass etwas passieren wird. Normative Erwartungen dagegen beinhalten Aufforderungen an andere, etwas zu tun oder zu unterlassen. Vielleicht erscheint das Adjektiv „normativ“ hier nicht in allen Fällen passend. Wenn etwa der ältere Affe erwartet, dass der jüngere ihm die Banane überlässt, wird man nicht ohne Weiteres sagen, dass es sich hier um eine Norm handelt. Auch wenn es um die Erwartung geht, gelaust zu werden, sprechen wir nicht von Normen. Allerdings, wir sagen schon: „Heinz, Du *sollst* zum Chef kommen.“ Soziale Normen (auch sprachliche Normen) beruhen aber einzig und allein auf Erwartungen. Die Geltung der Norm, dass man grüßt, wenn man ein Zimmer betritt, beruht darauf, dass dies in der Regel geschieht, dass Nichtgrüßen sanktioniert wird und dass solche Sanktionen in der Regel akzeptiert werden.[4] Es gibt eine von allen Mitgliedern der Gemeinschaft geteilte Erwartung. Und genau dies ist es, was die Norm ausmacht.

Wenn sich unsere faktischen Erwartungen nicht erfüllen, sind wir (im Wortsinn) enttäuscht. Wenn sich unsere normativen Erwartungen nicht erfüllen, sind wir auch enttäuscht, aber in einem anderen Sinn. Wir halten es dem anderen vor, dass er nicht so gehandelt hat, wie wir wollten, wir nehmen es ihm übel, wir machen ihm Vorwürfe und manchmal greifen wir zu Sanktionen.

c) Auch moralische Urteile sind präskriptiv; auch sie drücken Aufforderungen, Erwartungen aus – allerdings Erwartungen, die in besonderer Weise begründungsbedürftig sind. Man kann ja auf Erwartungen nicht nur reagieren, indem man sie befolgt oder nicht befolgt; man kann sie auch in Frage

---

[3] Zur Entwicklung von Selbstbewusstsein vgl. Beckermann (2005; 2008, Kap. 2).

[4] Ich beziehe mich hier auf den Normbegriff von H.L.A. Hart (1961, 54 ff.).

stellen. „Wie kannst Du das von mir erwarten?“ Überhaupt kann man auf jede Aufforderung „Tu das!“ antworten „Warum?“ Unsere nächsten Verwandten im Tierreich können das nicht – allein schon, weil ihnen die Sprache fehlt. Aber so etwas wie Entrüstung, wenn die Dinge nicht den normativen Erwartungen gemäß verlaufen, gibt es auch schon bei Affen.[5]

d) Erwartungen sind im Übrigen auch entscheidend, wenn es um „sinnhaft konstruierte Dinge“ wie Zehn-Euro-Scheine geht. Ein Zehn-Euro-Schein ist das, was er ist – ein Zahlungsmittel mit einem bestimmten Wert –, weil es die Erwartung derer gibt, die einen solchen Schein besitzen, dass Händler ihnen beliebige Waren in einer gewissen Menge geben, wenn sie ihnen den Schein übergeben, und umgekehrt auch die Bereitschaft der Händler, genau dies zu tun.

Alles in allem: Wenn der Kern des Normativen in Erwartungen besteht, gibt es keinen Grund, daran zu zweifeln, dass auch rein natürliche Wesen etwas glauben können, weil sie es glauben sollen, und etwas tun können, weil sie es tun sollen.

Und wie steht es mit der Aussage, dass Gründe nicht so wirken wie Ursachen? Viele Menschen glauben, dass sich Computer bzw. Roboter von Menschen dadurch unterscheiden, dass Computer immer nur tun, was ihr Programm ihnen vorschreibt; dass sie keine Spielräume haben und insofern nur rein mechanisch handeln. Informatiker selbst unterscheiden allerdings zwischen „herkömmlichen“ und so genannten „autonomen“ Systemen. Diese Unterscheidung hat ihren Sinn im Wesentlichen bei Systemen, deren Aufgabe es ist, in verschiedenen Situationen unterschiedliche Handlungen durchzuführen, um bestimmte Ziele zu erreichen. Herkömmliche Systeme sind so programmiert, dass man ihnen für jeden möglichen Situationstyp genau eine Verhaltensanweisung gibt. Sie müssen also nur feststellen, in welcher Art von Situation sie sich befinden, und dann genau das tun, was das Programm ihnen vorschreibt. Es ist klar, dass diese Art von Programmierung nicht sehr erfolgreich ist, wenn ein System mit unvorhergesehenen Situationen konfrontiert wird. Autonome Systeme sind deshalb anders aufgebaut. Sie verfügen über mehrere Komponenten. Eine erste Komponente erlaubt ihnen, sich Informationen über die Situation zu verschaffen, in der sie sich befinden. Was sind die wesentlichen Merkmale der Situation? Wo befinden sich die für meine Aufgabe relevanten Objekte? Welche Eigenschaften haben diese Objekte? Usw. Eine zweite Komponente erlaubt diesen Systemen, auf der Grundlage der erhobenen Informationen eigene Handlungspläne zu entwickeln, also eigene Antworten auf die Frage zu finden „Wie gehe ich am besten vor, um in der gegebenen Situation meine

[5] Vgl. die Untersuchungen von Brosnan und de Waal an der Emory-Universität über den Gerechtigkeitssinn von Kapuzineraffen.

Ziele zu erreichen?“ – Antworten, die nicht schon explizit durch ein Programm vorgegeben sind. Dieser letzte Schritt kann dadurch weiter flexibilisiert werden, dass man dem System darüber hinaus die Möglichkeit gibt zu entscheiden, welche Ziele vordringlich verfolgt werden sollen. Soll ich fortfahren, Bodenproben zu entnehmen? Soll ich zuerst dafür sorgen, dass meine Batterien aufgeladen werden? Oder gibt es eine unmittelbare Gefahr, der ich ausweichen muss?

Natürlich sind beide Komponenten – die Informationsbeschaffungs- und die Handlungsplanungskomponente – selbst programmiert; der Programmierer hat also festgelegt, wie die Informationsbeschaffung und die Handlungsplanung ablaufen. Allerdings sind auch autonome Systeme zweiter Stufe denkbar, die auf Grund der Erfahrungen, die sie machen, die Programme der Informationsbeschaffung und der Handlungsplanung selbst modifizieren können. Dafür ist natürlich eine zusätzliche Lernkomponente erforderlich, die ihrerseits wieder programmiert ist. Soweit wir die Autonomie auch treiben, am Ende gibt es doch so etwas wie ein unveränderliches Basisprogramm – ein Basisprogramm, das bei Wesen, die der Evolution unterliegen, aber durch Mutation und Selektion modifiziert werden kann.

Ich denke, dass wir selbst autonome Systeme der gerade geschilderten Art sind. Wenn wir mit Gründen konfrontiert werden, d. h., wenn wir die Information erhalten, dass Umstände vorliegen, die für eine bestimmte Handlung sprechen, reagieren wir nicht immer mechanisch auf dieselbe Weise. Unsere Reaktion hängt vielmehr davon ab, wie wir diese Information verarbeiten. Erstens entscheiden wir, ob wir diese Information überhaupt für wahr halten sollen; zweitens überlegen wir, ob die Umstände tatsächlich für die Ausführung der Handlung sprechen; drittens wägen wir ab, ob es nicht andere Umstände gibt, die dafür sprechen, eine andere Handlung auszuführen, usw. Diese Entscheidungs-, Überlegens- und Abwägungsprozesse laufen nach Mustern ab, die wir selbst verändern können. Aber auch bei uns gibt es an einem bestimmten Punkt angeborene genetisch bedingte Muster, die nur noch durch Mutation und Selektion modifiziert werden können. Das alles ändert aber nichts daran, dass wir die Fähigkeiten besitzen, vor dem Handeln innezuhalten und zu überlegen, bei der Handlungsplanung auf Gründe zu reagieren und mit sinnhaft konstituierten Gegenständen angemessen umzugehen.

## *4. Schluss*

Zurück zu Darwin. Darwins Evolutionstheorie legt zumindest die Annahme nahe, dass wir Menschen – ebenso wie Katzen, Hunde, Elefanten und Primaten – Tiere sind, rein natürliche Lebewesen. Meine Frage war, was wir

an unserem herkömmlichen Welt- und Menschenbild aufgeben müssen, wenn diese Annahme zutrifft. Die Antwort sollte inzwischen klar geworden sein: Nichts, oder doch zumindest nur wenig. Darwins Theorien haben die größten Konsequenzen dort, wo es um religiöse Überzeugungen geht. Sie sprechen ebenso gegen die Existenz einer immateriellen Seele wie gegen die Annahme, dass die Welt das Werk eines allmächtigen, allwissenden und allgütigen Schöpfers ist (vgl. Kitcher 2009, 152 ff.). Aber ansonsten lassen sie unser Menschenbild weitestgehend intakt. Oder anders: An der Wahrheit der lebensweltlichen Gewissheiten Wingerts ändert sich auch dann nichts, wenn wir tatsächlich „nur" Tiere sind.

Daniel Dennett hat einmal auf die Frage, ob er die Vorstellung, dass der Mensch nur eine Maschine sei, nicht deprimierend fände, geantwortet:

> Überhaupt nicht; denn ich denke, dass wir solche wundervollen Maschinen sind. Wenn man eine miese Vorstellung davon hat, was eine Maschine ist, wenn man glaubt, sie sei nichts weiter als ein aufgeblasener Toaster, wissen Sie, dann ist das nicht besonders aufregend. Aber wenn man sich klar macht, was die Maschinen, aus denen wir bestehen, alles können – sie können sich selbst reparieren, sie können Infektionen bekämpfen und sie können erstaunliche Berechnungen im Gehirn durchführen –, dann ist das gewaltig. http://www.pbs.org/saf/1103/features/dennett4.htm – Abruf 14. April 2007, 18:40 Uhr)

Und in dem Dialog „Ein Kaffeehaus-Gespräch über den Turing-Test" lässt Douglas Hofstadter Sandy am Schluss sagen:

> Wenn man mir eine Maschine zeigte, die Sachen kann, wie ich sie kann – den Turing-Test bestehen, meine ich –, dann würde ich mich nicht etwa gekränkt oder bedroht fühlen, sondern im Einklang mit dem Philosophen Raymond Smullyan sagen: „Wie wunderbar sind doch Maschinen!" (Hofstadter 1981, 551)

Auf unseren Kontext bezogen: Es ist nicht schlimm, ein Tier zu sein; dass wir Menschen Tiere sind, zeigt nur, was für wunderbare Tiere es gibt.

## Literatur

Barras, Colin (2008) „Macaque monkeys ‚pay' for sex". *New Scientist* 2637, 2.1.2008.

Beckermann, Ansgar (1977) *Gründe und Ursachen*. Kronberg/Ts.: Scriptor Verlag. (Elektronischer Nachdruck: http://phillister.ub.uni-bielefeld.de/publication/650)

Beckermann, Ansgar (2005) „Selbstbewusstsein in kognitiven Systemen". In: M. Peschl (Hg.) *Die Rolle der Seele in der Kognitionswissenschaft und Neurowissenschaft*. Königshausen & Neumann, 171–187; in diesem Band Beitrag 13.

Beckermann, Ansgar (2008) *Gehirn, Ich, Freiheit*. Paderborn: mentis.

Beckermann, Ansgar (2012) „Der Mensch als Tier und biologische Maschine. Anmerkungen eines Naturalisten zu den Aussichten, den biologischen Tod zu überleben“. In: K.-L. Koenen & J. Schuster SJ (Hg.) *Seele oder Hirn? Vom Leben und Überleben der Personen nach dem Tod*. Münster: Aschendorff, 29–48

Bittner, R. (2005) *Aus Gründen handeln*. Berlin: de Gruyter.

Freud, S. (1917) *Vorlesungen zur Einführung in die Psychoanalyse*. In: ders., *Gesammelte Werke, Band 11*, Frankfurt/M.: Fischer Verlag 1969.

Gethmann, Carl-Friedrich (2008) „Warum sollen wir überhaupt etwas und nicht vielmehr nichts?“ In: Peter Janich (Hg.) *Deutsches Jahrbuch Philosophie Band I: Naturalismus und Menschenbild*. Hamburg: Felix Meiner, 138–156.

Hart, H.L.A. (1961): *The Concept of Law*. Oxford.

Hofstadter, D. R. (1981) „Ein Kaffeehaus-Gespräch über den Turing-Test“. In: D. R. Hofstadter, *Metamagicum*. Stuttgart: Klett-Cotta 1991, 529–551.

Kitcher, Phillip (2009) *Mit Darwin leben*. Frankfurt am Main: Suhrkamp.

Singer, Wolf & Lutz Wingert (2000) „Wer deutet die Welt?“, DIE ZEIT, 50, 2000. (Wieder abgedruckt in Wolf Singer *Ein neues Menschenbild?* Frankfurt/M.: Suhrkamp 2003, 9–23; zitiert nach dem Wiederabdruck.)

Wingert, Lutz (2008) „Lebensweltliche Gewissheit versus wissenschaftliches Wissen?“ In: Peter Janich (Hg.) *Deutsches Jahrbuch Philosophie Band I: Naturalismus und Menschenbild*. Hamburg: Felix Meiner, 288–309.