Philosophisches Seminar, AG Theoretische Philosophie
Hauptseminar Intentionalität und mentale Repräsentationen (PdG4)
Leitung: Prof. Thomas Metzinger
Datum: 15.09.2004

Schwache Kompositionalität in Constraint Satisfaction Networks

Philosophische Perspektiven konnektionistischer Kognition

von
Dominique Kaspar
Philosophie (10), Informatik (5), Anthropologie (3)

GLIEDERUNG

1. Einleitung
2. Konnektionismus vs. Symbolizismus
   - 2.1. Das symbolizistische Paradigma: die Sprache des Denkens
     - a) Komplexe mentale Repräsentationen
     - b) Struktursensitive Prozesse
     - c) Physikalische Realisierung
   - 2.2.Das konnektionistische Paradigma
     - a) Eigenschaften neuronaler Netze
     - b) Ebenen der Analyse
3. Fundamentale Eigenschaften der Kognition
   - 3.1. Produktivität / Systematizität
   - 3.2. Kompositionalität
   - 3.3. Repräsentationen
4. Die konnektionistische Theorie der “Constraint Satisfaction Networks” (CSN)
   - 4.1. Korrellative Semantik
   - 4.2. Schwache Kompositionalität
   - 4.3. Erfüllung von Randbedingungen als holistische Strategie
5. Philosophische Implikationen der CSN Theorie
   - 5.1. Eliminativismus?
   - 5.2. Internalismus? Externalismus? Holismus!
   - 5.3. Gibt es eine Ethik des konnektionistischen Selbsverständnisses?
6. Fazit

1. Einleitung

In der vorliegenden Arbeit beschäftige ich mich mit kognitionswissenschaftlichen Theorien und ihren philosophischen Aspekten. Die Kognitionswissenschaft ist eine relativ junge *Interdisziplin*, deren Ziel es ist, kognitive Phänomene wie Wahrnehmen, Denken, Handeln, Entscheiden, Sprechen oder Sprachverstehen wissenschaftlich zu untersuchen. Dabei bedient sie sich in einem integrativen Ansatz ganz unterschiedlicher Denkmodelle aus verschiedenen Wissenschaftszweigen, von der Philosophie über Psychologie, Linguistik, Neurologie, Anthropologie, Biologie bis hin zur komplexen Modellbildung im Rahmen von Forschungen zu künstlicher Intelligenz in der Informatik oder bei der Modellierung dynamischer Systeme in der Physik.

Hintergrundannahme bzw. Metatheorie der Kognitionswisseschaft ist dabei seit der Mitte des 20. Jahrhunderts die These des Komputationalismus, d.i. die Annahme, kognitive Systeme - wie der menschliche Geist - liessen sich am besten verstehen, wenn man sie als informationsverarbeitende Systeme modelliert. Der klassische Ansatz der Kognitionswissenschaft ist sehr stark von den frühen Paradigmen der Forschung zu künstlicher Intelligenz beeinflusst, die sich an deduktiven Systemen (z.B. zur automatischen Beweisführung) der Termersetzung orientierten. Formallogische, meist auf expliziten Regeln basierende Symbolsysteme sollten in diesen Ansätzen über zunehmende Komplexität ihrer Regelwerke zu höheren kognitiven Leistungen befähigt werden. Die Grundidee dieser auch „*Symbolizismus*“ genannten Hypothese wurde als erstes von Allen Newell und Herbert A. Simon in ihrer „Physical Symbol Systems Hypothesis“[1] formuliert: demnach hat ein physikalisches (also physikalisch implementierbares) Symbolsystem sowohl alle notwendigen wie auch hinreichenden Mittel zu intelligentem Verhalten. Auch wenn sich symbolorientierte kognitionswissenschaftliche Modelle hinsichtlich ihrer konkreten Ausgestaltung häufig widersprachen, so war ihre theoretische Dominanz innerhalb der Kognitionswissenschaft doch nie ernsthaft gefährdet: es gab schlichtweg keine funktionierende Alternative[2].

1982 erscheint in einer Arbeit von J. A. Feldman und D. H. Ballard[3] zum ersten Mal der Begriff „konnektionistisches Modell“, die darauf folgenden Jahre zeigen einen starken Anstieg der Publikationen zu konnektionistischen Themen. Auch wenn die diskutierten konnektionistischen Modelle sehr unterschiedlicher Natur waren, so gab es doch eine Gemeinsamkeit: die mit ihnen verbundenen Hintergrundannahmen unterscheiden sich in wichtigen Punkten radikal von den Prämissen, die die klassische Kognitionswissenschaft postuliert; so radikal, dass ein signifikanter Teil

---

1 Newell, A., and H. A. Simon. (1976). Computer science as empirical inquiry: Symbols and search. *Commun. Assoc. Comput. Machinery* **19**:111-26.

2 Vgl. Dennett, D. (1991). Mother Nature Versus the Walking Encyclopedia: A Western Drama. S. 25 in: Ramsey, W., Stich, S., Rumelhart, D. (Hrsg.): Philosophy and Connectionist Theory. Erlbaum

3 Feldman, J. A., Ballard, D. H. (1982). Connectionist Models And Their Properties. *Cognitive Science*, **6**, 205-224. auch: http://cognitrn.psych.indiana.edu/rgoldsto/cogsci/Feldman.pdf

der im Bereich der Kognition arbeitenden ForscherInnen von einer „konnektionistischen Revolution“ spricht.[4]

Die vorliegende Arbeit beschäftigt sich mit philosophischen Implikationen dieser Revolution. In einem ersten, deskriptiven Teil (Abschnitt 2) versuche ich die Unterschiedlichkeit des klassischen und des konnektionistischen Ansatzes in einem (notwendig skizzenhaften) Vergleich wichtiger Merkmale herauszuarbeiten.

Der zweite Teil der Arbeit (Abschnitt 3) erläutert die Hauptkritikpunkte von Vertretern der orthodoxen, symbolizistischen Kognitionswissenschaft am neuen Programm des Konnektionismus: Produktivität, Systematizität, und Kompositionalität sind laut J. Fodor essentielle Bestandteile jeder kognitionswissenschaftlichen Strukturhypothese – Bestandteile, die konnektionistische Modelle nicht erklären können. Zudem wird die Frage erörtert, welchen Begriff von Repräsentation eine kognitionswisseschaftliche Strukturtheorie anzunehmen braucht.

Der dritte Teil der Arbeit (Abschnitt 4) beschäftigt sich mit dem konkreten Vorschlag einer kognitionswissenschaftlichen Strukturtheorie. T. Goschke und D. Koppelberg versuchen in ihrem Modell der „Constraint Satisfaction Networks“ (CSN) das konnektionistische Paradigma als Strukturtheorie der Kognition durchzudeklinieren. In diesem Teil untersuche ich die Antworten, die die Theorie der CSN auf die in Abschnitt 3 aufgeworfenen Probleme zu geben vermag.

Im vierten Teil der Arbeit (Abschnitt 5) geht es mir um die spezifisch philosophischen Aspekte der CSN. Ich untersuche, ob man gezwungen ist, den Konnektionismus als Variante des eliminativen Materialismus zu verstehen und welche Hinweise die CSN für philosophische Probleme wie die Debatte um Internalismus/Externalismus geben kann. Zudem möchte ich einige – zugegebenermassen hochspekulative – Ausblicke auf mögliche ethisch-praktische Implikationen konnektionistischer Modelle in ihrer Funktion als anthropologische Modelle wagen.

Im Rahmen der Einleitung möchte ich auch klarstellen, was *nicht* Ziel dieser Arbeit ist. Die Literatur zu konnektionistischen Themen ist mittlerweile unüberschaubar, die Vielfalt (und Komplexität) der vorgeschlagenen Modelle kann und soll in dieser Arbeit keine besondere Beachtung finden. Ich beschränke meine Vorstellung des Konnektionismus (mit Ausnahme der theoretischen Ausführungen der CSN) daher auf grobe Vereinfachungen, die die wichtigsten Strukturmerkmale möglichst vieler konnektionistischer Modelle wiedergeben können.

So wird z.B. der in jüngster Zeit häufig in Konkurrenz bzw. Abgrenzung zum theoretischen Programm des Konnektionismus genannte „Dynamizismus“ bzw. die Theorie Dynamischer Systeme (Dynamical Systems Theory, DST) von mir nicht weiter diskutiert, da Ihr Status in Abgrenzung zu

---

[4] Medler, D. A. (1998). A Brief History of Connectionism. *Neural Computing Surveys* **1**(1), 61-10. auch: http://www.icsi.berkeley.edu/~jagota/NCS

konnektionistischen Theorien alles andere als eindeutig ist[5]. Auch das Programm des Symbolizismus hat eine Vielzahl an Strukturmodellen hervorgebracht, die in dieser Arbeit nicht diskutiert werden. Vielmehr will ich anhand einer Struktur, die den Verlauf der Diskussion – holzschnittartig – nachzeichnet darlegen, wie sich das konnektionistische Paradigma in Abgrenzung zur orthodoxen Theorie des Symbolizismus entwickelt hat.

2. Konnektionismus vs. Symbolizismus

2.1. Das symbolizistische Paradigma: Die Sprache des Denkens

In der Geschichte der Kognitionswissenschaft ist die klassische strukturelle Grundannahme die These, man könne kognitive Leistungen – insbesondere die höheren kognitiven Leistungen des menschlichen Geistes wie Sprache, logisches Schlussfolgern oder mathematisches Verständnis – am besten verstehen (bzw. modellieren), indem man kognitive Systeme allgemein als symbolorientierte Systeme versteht. Aus der Forschung zu künstlicher Intelligenz kommend sind diese Systeme durch ihre Fähigkeit charakterisiert, über strukturierte Operationen auf eindeutigen Symbolen intelligentes Verhalten qua Termersetzungsverfahren hervorzubringen. Übertragen auf kognitionswissenschaftliche Zusammenhänge nennt man die Symbole, auf denen die untersuchten kognitiven Systeme operieren, "mentale Repräsentationen".

Der Rahmen des symbolizistischen Paradigmas umfasst jedoch nicht nur "linguistisch infizierte Zustände"[6], auch das unbewusste, vorsprachliche Verhalten wie Wahrnehmen, motorische Steuerung, etc. wird als regelgeleitet, (meist) sequentiell und auf Symbolen mit kombinatorischem Charakter operierend verstanden. Wichtig ist hierbei vor allem die Tatsache, dass die Berechnung semantischen Gehalts, des *Inhalts* der mentalen Repräsentationen, auf der Ebene der Symbole stattfindet. Die wohl berühmteste klassische Metatheorie der Kognitionswissenschaft, die diese Anforderungen vereint, ist die der Theorie einer "Sprache des Denkens" (Language of Thought, LOT) von Jerry Fodor. Sie ist vor allem durch zwei Prämissen[7] (sowie eine Hypothese zur Realisierung) gekennzeichnet, die im Folgenden näher erläutert werden sollen:

(a) Komplexe mentale Repräsentationen

Alle mentalen Repräsentationen haben eine kombinatorische Syntax und Semantik, d.h. es existieren

---

[5] Vgl. Eliasmith, C. (1996). The third contender: A critical examination of the dynamicist theory of cognition. *Philosophical Psychology*. Vol. **9** No. 4 pp. 441-463. Reprinted in P. Thagard (Hrsg.) 1998. Mind Readings: Introductory Selections in Cognitive Science. MIT Press.

[6] Dennett, D. (1991). Mother Nature Versus the Walking Encyclopedia: A Western Drama. S. 26 in: Ramsey, W., Stich, S., Rumelhart, D. (Hrsg.): Philosophy and Connectionist Theory. Erlbaum

[7] Vgl. Abschnitt "The nature of the dispute", in: Fodor, J. A., Plyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28**, 3-71, S.12f

strukturell einfache ("atomic") und komplexe ("molecular") mentale Repräsentationen. Komplexe mentale Repräsentationen, sozusagen die Sätze der Sprache des Geistes, bestehen aus syntaktischen Konstituenten[8], die ihrerseits in ihrer Struktur entweder einfacher oder komplexer Natur sein können. Der semantische Gehalt – also die Bedeutung – komplexer mentaler Repräsentationen errechnet sich als Funktion über den semantischen Gehalt seiner syntaktischen Elemente wie auch aus der Struktur seiner Konstituenten. Der semantische Gehalt einer einfachen mentalen Repräsentation ist kontextunabhängig[9], d.h. einfache mentale Repräsentationen müssen in jedem Kontext, in den sie eingebunden werden, genau die gleiche Bedeutung eintragen – nur so ist gewährleistet, dass die (algorithmische, vollständige) Berechnung der Bedeutung des komplexen Ausdrucks über die Bedeutung der Konstituenten funktioniert.

(b) Strukturensitive Prozesse

Die Operationen bzw. Prozesse, die auf den mentalen Repräsentationen arbeiten und so von Inputstimulation zu errechnetem Output kommen, sind über die Struktur der mentalen Repäsentationen definiert. Aufgrund der kombinatorischen Struktur – Synatx und Semantik – mentaler Repräsentationen können mentale Prozesse auf mentalen Repräsentationen *qua Strukturinformation* operieren, d.h. mentale Prozesse benötigen Strukturinformationen sowohl um die Ausgangsrepräsentationen zu "finden" wie auch um sie in die Zielrepräsentation umzuwandeln. Das klassische kognitionswissenschaftliche Modell nimmt hier meist Formen der logischen Inferenz (Schlussfolgerung) an[10].

(c) Physikalische Realisierung

Die LOT Hypothese ist nicht nur eine repräsentationale Theorie des Geistes, in der Interpretation von Jerry Fodor ist sie auch im wörtlichen Sinne physikalisch instantiiert. Die mentalen Repräsentationen, deren kombinatorische Strukturmerkmale die kognitiven Fähigkeiten eines Systems definieren, sind als funktionale Module mit diesen Eigenschaften in den physikalischen Gesetzmässigkeiten des Systems zu finden, also als neuronale Aktivierungsmuster im Gehirn, Adressenbelegung in einer van Neumann Architektur, etc[11]. Das bedeutet, dass die LOT Hypothese eine *empirische* Hypothese darstellt: ihre Verfechter gehen davon aus, die kombinatorischen Strukturen einer LOT in den physikalischen

---

[8] Konstituenten sind Wortgruppen (bzw. Satzglieder), die in ihrer Gesamtheit untereinander austauschbar, verschiebbar oder ersetzbar sind (Umstellprobe). Sätze bestehen also nicht direkt aus Wörtern, sondern zunächst aus Konstituenten. Vgl. Schoebe, G. (1988) Elementargrammatik und Rechtschreibung. 2. Aufl. Oldenburg Verlag, München

[9] Fodor, J. A., Plyshyn, Z. W. (1988), a.a.O., S. 42

[10] Vgl. Beispiel von Fodor & Plyshyn: "[...]for example, in a model of inference, one might recognize an operation that applies to any representation of the form P & Q and transforms it into a representation of the form P.", Fodor, J. A., Plyshyn, Z. W., a.a.O., S.13, auch: S. 46f

[11] Fodor, J. A., Plyshyn, Z. W. (1988), a.a.O., S. 15

Gesetzmässigkeiten des jeweils untersuchten kognitiven Systems zu finden. Das ist bei Maschinen mit Turing oder van Neumann Architektur logisch gegeben, bei biologisch implementierten kognitiven Systemen wie dem menschlichen Gehirn jedoch zumindest fraglich[12].

2.2. Das konnektionistische Paradigma

Um das konnektionistische in Abgrenzung zum symbolizistischen Paradigma zu kennzeichnen ist es interessanterweise nicht hinreichend, die syntaktische Struktur, also die implementatorischen Kapazitäten der jeweiligen Modelle zu charakterisieren.[13] Sowohl Turing und van Neumann Architekturen wie auch Architekturen auf der Basis von neuronalen Netzen sind *universal*, d.h. es ist sowohl möglich, ein neuronales Netz in einer Turing oder van Neumann Architektur (wie z.B. der PC Architektur) zu implementieren, wie es auch möglich ist, auf der Basis einer neuronales-Netz-Architektur eine "klassische" symbolorientierte Maschine zu implementieren. Die Unterschiede liegen in der Art, wie die Semantik, d.h. der Inhalt der mentalen Repräsentationen implementiert wird. Um das konnektionistische Paradigma genauer zu kennzeichnen sind jedoch trotzdem verschiedene Vorüberlegungen zur Struktur hilfreich.

a) *Eigenschaften neuronaler Netze*

Neuronale Netze funktionieren, indem sie einen Input über ein Netz einzelner "Einheiten" in ein Output überführen. Diese Einheiten bestehen meist aus einer einfachen Input-Berechnung-Output Struktur, die bei ihrer Aktivierung eine Berechnung – meist die Summe der Inputstimuli – als Outputstimulation an andere Einheiten des Netzes weitergeben. Die Vernetzung dieser Einheiten ist gewichtet, d.h. das Netz der untereinander verbundenen Einheiten ist an verschiedenen Stellen unterschiedlich stark verbunden. Über diese Gewichtung können neuronale Netze Strukturen abbilden. Die Gewichtung kann unterschiedlichen Ursprungs sein: so gibt es Netze, die Ihre anfangs komplett neutrale Gewichtung nach statistischen Methoden adaptiv dem gegebenen Input anpassen, wie auch Netze, deren Gewichtung bereits eine vorgeformte Struktur besitzt. Sogenannte "backpropagation" Modelle erlauben, das Netz anhand von Fehlerhäufigkeiten in Bezug auf gewünschte Fähigkeiten zu trainieren: ein Feedback über den Grad des Erfolges des bisherigen Outputs ist jeweils Teil der neueren Inputstimuli und kann so korrektive Funktionen annehmen. Es ist Charakteristikum dieser neuronalen Netze, dass sie ihre Struktur in Abhängigkeit von der Häufigkeit bzw. Stärke der auf die einzelnen

---

[12] Vgl. Dennett, D. C. (1986) The logical geography of computational approaches: A view from the East Pole. S. 66f in: Harnish, M., Brand, M. (Hrsg.) Problems in the representation of knowledge, S. 59-79, Tuscon, AZ, University of Arizona Press. zit. nach: Dennett, D. (1991). Mother Nature Versus the Walking Encyclopedia: A Western Drama. a.a.O., S. 22

[13] Vgl. Abschnitt 3.3 in: Chalmers, D. (1992) Subsymbolic Computation and the Chinese Room. In: Dinsmore, J. (Hrsg.) *The Symbolic and Connectionist Paradigms: Closing the Gap*, S. 25-48. Hillsdale, NJ: Lawrence Erlbaum. auch: http://jamaica.u.arizona.edu/~chalmers/papers/subsymbolic.pdf

Einheiten jeweils eingehenden Aktivierungen verändern.
Neuronale Netze sind meist in mehreren Schichten aufgebaut, die untereinander komplexe Verbindungsmuster eingehen, so gibt es sowohl einfache "feed-forward" Strukturen, die den Stimulus von den Inputeinheiten zu den Outputeinheiten des Netzes weitertragen, wie auch "feed-back" Strukturen, die bereits berechnete, interne Stimuli wieder an vorhergehende Einheiten zurücktragen. Diese sind nicht direkt an globalen Input- oder Outputrelationen beteiligt sondern repräsentieren als sog. versteckte Einheiten die interne Konnektivität des Netzes. So sind z.B. rekursive Relationen möglich, die bereits berechnete Output-Zustände innerhalb der globalen Berechnung wieder als Inputstimulationan an die vorhergehenden Einheiten zurückgeben und so verstärkend bzw. hemmend wirken. Auch verschiedene Formen von "Gedächtnis" können so modelliert werden, da Aktivierungsstrukturen über diese Schleifen signifikante Zeiträume "im Netz" wirksam bleiben – auch wenn sie nicht in expliziten Speicherstrukturen gesichert sind.
Zusammenfassend lässt sich sagen, dass in neuronalen Netzen die Gewichtung oder Verbindungsstärke zwischen den einzelnen Einheiten die Rolle des Programms in klassischen Modellen übernimmt[14]. Der große Vorteil neuronaler Netze als kognitive Modellannahmen liegt in ihrer Fähigkeit, sich selbst zu programmieren: d.h. der globale Zustandsraum des Netzes wird nicht von aussen festgelegt, er entwickelt sich vielmehr aus der Kombination von strukturellen Anforderungen der Inputstimuli und innerer Konnektivität. Diese Fähigkeit zur Selbstorganisation in Verbindung mit distribuierten Repräsentationen (sowie einem gewissen Grad an Redundanz[15]) ermöglichen es dem neuronalen Netz, Schäden in seiner Architektur durch neue adaptiv gewonnene Konnektivität abzumildern oder gar zu beheben. Interessanterweise zeigen neuronale Netze bereits ab einer architektonischen Dichte, die nur leicht über dem Trivialen liegt, eine Komplexität der inneren Vorgänge, die dem Menschen das exakte Nachvollziehen der auf gegebene Inputstimuli produzierten Lösungen unmöglich macht.[16]

b) Ebenen der Analyse

In der klassischen Sicht der Kognitionswissenschaft gibt es nur eine Ebene, auf der es sinnvoll ist, von Kognition zu sprechen: nämlich diejenige, auf der die struktursensitiven Prozesse auf mentalen Repräsentationen arbeiten, die symbolische (auch: konzeptuelle) Ebene. Die konnektionistische Sichtweise widerspricht dem radikal: Wenn Kognition in konnektionistischen Modellen als neuronales

---

[14] Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**:1-23. S. 1

[15] Vgl. Foster, M. Saidel, E. (1994) Connectionism and the fate of folk psychology: a reply to Ramsey, Stitch and Garon. *Philosophical Psychology Vol* **7**, No. 4, S. 437-452, auch: http://philosophy.wisc.edu/forster/papers/Connectionism.HTM, S. 447

[16] Vgl. Eraßme, R.: Der Mensch und die 'Künstliche Intelligenz'. Eine Profilierung und kritische Bewertung der unterschiedlichen Grundauffassungen vom Standpunkt des gemäßigten Realismus. Diss. Aachen, 2002, S. 21. Im Gegensatz zur dort vertretenen Meinung, dass dies ein Nachteil sei, gehe ich jedoch davon aus, dass hier eine der großen Chancen der konnektionistischen Modellbildung liegt.

Netz implementiert wird, so ist die komplette, formale und vollständige Beschreibung des Phänomens der Kognition nicht auf der Ebene der Symbole zu suchen, da sie dort nicht berechnet werden kann. Was unterscheidet also die Beschreibung kognitiver Leistungen im konnektionistischen Paradigma von Beschreibungen innerhalb des symbolizistischen Paradigmas? Im symbolizistischen Paradigma haben Symbole nicht nur eine klar definierte Syntax, sie haben, damit einhergehend, eine klar definierte Semantik. Die Bedeutung komplexer Phänomene (wie sie die Kognitionswissenschaft untersucht) ergibt sich aus der klar definierten, kontextunabhängigen Semantik der einzelnen mentalen Repräsentationen und ihrer syntaktischen Beziehungen, da auf dieser Ebene die komplexen Leistungen "berechnet" werden[17].

In konnektionistischer Sichtweise hingegen sind mentale Repräsentationen implizit in vielen einzelnen, untereinander verbundenen Einheiten des neuronalen Netzes distribuiert als Strukturinformation vorhanden. Die Bedeutung bzw. Semantik einzelner mentaler Symbole ist auf der Basis mentaler Repräsentationen nie komplett formal und präzise vorherzusagen, da ihre konkrete Implementation immer auch strukturell mit der aktuellen Implementation des globalen Zustandsraumes des kognitiven Systems verbunden ist. Somit ist es *unmöglich*, komplexe Phänomene der Kognition auf der Ebene der mentalen Symbole *formal vollständig* zu beschreiben, ihre (im Sinne einer erfolgreichen kognitionswissenschaftlichen Deutung notwenige!) Explikation als mentale Symbole erfahren sie stets nur durch Analyse, Zuschreibung und Annäherung. Die formale und vollständige Berechnung mentaler Repräsentationen im konnektionistischen Paradigma kann nur auf subsymbolischer bzw. subkonzeptueller Ebene stattfinden[18].

3. Fundamentale Eigenschaften der Kognition

Wie die Skizze des konnektionistischen und des symbolizistischen Paradigmas im vorhergehenden Abschnitt gezeigt hat sind die beiden Ansätze in einigen Punkten nicht ohne weiteres miteinander vereinbar. Im folgenden Abschnitt werde ich einige Kritikpunkte aufgreifen, die von Jerry A. Fodor und Zenon W. Plyshyn[19] als Vertretern des Symbolizismus gegenüber konnektionistischen Modellen angeführt wurden. Es geht mir nicht darum, diese Kritikpunkte in extenso zu wiederholen, vielmehr möchte ich über eine kurze Darstellung der Kritik des konnektionistischen Paradigmas darlegen, welche Fragen sich einer generellen Strukturtheorie der Kognition stellen.

3.1. Produktivität / Systematizität

Unter der Produktivität des Denkens versteht man klassischerweise die Hypothese, dass ein kognitives

[17] Die "Berechnung" im klassischen Ansatz ist eher der logischen Schlussfolgerung gleichzusetzen, Vgl. Fußnote 10
[18] Vgl. Smolensky, P. (1988), a.a.O., S. 6f
[19] Fodor, J. A., Plyshyn, Z. W. (1988) a.a.O.

System unter idealen Bedingungen in der Lage ist, mit seinen durch die physikalische Konstitution vorgegebenen, endlichen Mitteln unendlich viele Aussagen bzw. Propositionen zu formen. Um in einem System diese Fähigkeit zu modellieren benötigt man eine Möglichkeit, unendlich viele komplexe mentale Repräsentationen aus einem endlichen Set an mentalen Kapazitäten zu formen: ein ideales Argument für die Annahme einer "Sprache des Denkens".

Die Annahme einer derart ungebundenen expressiven Fähigkeit mentaler Systeme ist allerdings ein a priori Argument: in der Realität gibt es stets eine finite Reihe von Propositionen, die ein System formen kann. Bei Theorien, die die ungebundene Produktivität des Geistes annehmen, wird demnach zwischen aktualer Performanz und ungebundener, hypothetischer Kompetenz gesprochen[20]. Die Notwendigkeit, die Produktivität des Denkens – ich kann zu jedem gedachten Gedanken einen neuen denken – als konstituives Element der Kognition zu verstehen hängt damit davon ab, ob man die a priori Annahme der ungebundenen expressiven Fähigkeit menschlichen Denkens akzeptiert.

Die Systematizität des Denkens besteht hingegen in der weniger grundlegenden Annahme, dass die Fähigkeit des Menschen (bzw. eines kognitiven Systems) zur Bildung verschiedener Gedanken instrinsisch mit der Fähigkeit zur Bildung anderer Gedanken verbunden ist. Das klassische Beispiel lautet: wenn ich den Gedanken "John liebt Mary" verstehen kann, so ist es mir unmöglich, den Gedanken "Mary liebt John" nicht zu verstehen[21]. Die Argumentationsstrategie ist hier, aufzuzeigen, dass man um die Systematizität mentalen Geschehens zu erklären eine *Konstituentenstruktur* des Denkens benötigt: wenn alle Gedanken Einzelfälle sind, d.h. nicht aus untereinander austauschbaren, zumindest jedoch strukturell verbundenen Konstituenten bestehen, so ist es unerklärlich, wieso kognitive Systeme derartige Fähigkeiten aufweisen.

3.2. Kompositionalität

Systematizität ist also die Eigenschaft komplexer mentaler Repräsentationen, aufgrund ihrer syntaktischen Struktur systematische Relationen zu anderen mentalen Repräsentationen zu erklären. Diese Eigenschaft kann jedoch nicht auf strukturelle – also syntaktische – Eigenschaften mentaler Repräsentationen beschränkt bleiben. Es ist darüber hinaus notwendig, dass mentale Repräsentationen als Konstituenten komplexer mentaler Repräsentationen in ihrem Inhalt (ihrer Semantik) kontextunabhängig sind – nur über eine solche kombinatorische Struktur kann die Fähigkeit, Produktivität und Systematizität des Denkens zu erklären, konstruiert werden. Man benötigt, zusätzlich

[20] Vgl. Chomsky, N. (1968): Language and Mind. New York: Harcourt, Brace and World. zit. nach: Fodor, J. A., Plyshyn, Z. W., a.a.O., S. 34

[21] Dabei wird von den Vertretern des Symbolizismus gefordert, dass *alle* kognitiven Systeme diese Fähigkeit zeigen. Das ist allerdings mehr als unwahrscheinlich, vgl. folgendes anschauliches Beispiel: "There are organisms of which one would say with little hesitation that they think a lion wants to eat them, but there is no reason at all to think they could 'frame the thought' that they want to eat the lion!" (Dennett, D. (1991). Mother Nature Versus the Walking Encyclopedia: A Western Drama. a.a.O., S. 27)

zur syntaktischen Systematizität, eine *semantische Systematizität*. Ein Beispiel[22] mag das verdeutlichen:

> "Der Gedanke, dass ein Mann mit einem kalten Eisen in der Hand sich fürchtet, ein heisses Blech anzufassen."
> "Der Gedanke, dass ein Mann mit einem kalten Blech in der Hand sich fürchtet, ein heisses Eisen anzufassen."

Die syntaktische Austauschbarkeit der Satzkonstituenten ist nicht hinreichend, um die semantische Nicht-Austauschbarkeit der Satzkonstituenten vollständig zu beschreiben. Man benötigt eine Theorie des semantischen Gehalts der Konstituenten, da "heisses Eisen" idiomatisch ist.
Hier wird deutlich, welche starken Prämissen die "Sprache des Denkens" voraussetzt: mentale Repräsentationen müssen, um, wenn schon nicht Produktivität, so doch zumindest Systematizität erklären zu können, eine syntaktisch strukturierte Konstituentenstruktur aufweisen, die aus auch semantisch individuierten, einfachen mentalen Repräsentationen besteht. Man benötigt, um es salopp zu formulieren, mentale Repräsentationen als Bedeutungsatome im Kopf, sozusagen einen Inhalts-Essentialismus in Bezug auf mentale Repräsentationen[23].

3.3. Repräsentation

Es ist die These dieser Arbeit, dass eine umfassende Strukturtheorie der Kognition nur dann als erfolgreich gelten kann, wenn sie in Bezug auf mentale Repräsentationen eine realistische Position einnimmt. Der Begriff der Repräsentation ist zentral für die Erklärung kognitiver Zielphänomene wie Wahrnehmung, Entscheidung, Handlung etc. - ohne die Annahme einer wie auch immer gearteten repräsentationalistischen Theorieebene fehlt der Kognitionswissenschaft ihr eigentlicher Untersuchungsgegenstand. Nach Jerry Fodor kann man sowohl symbolizistische wie auch konnektionistische Strukturtheorien der Kognition als repräsentationalistische Theorien des Geistes verstehen[24]. Was jedoch unter dem Begriff der Repräsentation genau zu verstehen sei, ist umstritten. Im folgenden möchte ich ein Repräsentationskonzept vorstellen, um notwendige Charakteristika des Begriffs der Repräsentation näher zu spezifizieren.
Als Minimalexplikation des Begriffs der Repräsentation soll folgende Definition dienen:
Repräsentation ist

- eine dreistellige Relation a**R**b*S* zwischen
  - dem zu repräsentierenden (Repräsentanda) a

---

[22] aus: Werning, Markus: Compositionality and the Basis of Mental Concepts, Colloquium of the Philosophy of Science, Jan. 2001, Erfurt, Germany. Presentation: http://service.phil-fak.uni-duesseldorf.de/ezpublish/index.php/filemanager/download/216/Kompositionaltit%E4t%20und%20die%20Basis%20mentaler%20Begriffe.ppt

[23] Vgl. An Interview with Daniel Dennett, in: The Dualist, Stanfords Undergraduate Journal of Philosophy (2001), Volume **8**, http://www.stanford.edu/group/dualist/vol9/pdfs/dennett.pdf, S. 78

[24] Vgl. Fodor, J. A., Plyshyn, Z. W. (1988) a.a.O., S. 7f

- dem repäsentierten (Repräsentat) b
- dem individuellen System

- wobei die Relation a**R**b asymetrisch ist:
  - die Identität von a und b ist ausgeschlossen
  - die konverse Relation b**R**a ist nicht identisch mit a**R**b
  - die Relation ist nicht transitiv, d.h. sie überträgt sich nicht
  - die Relation entspricht einer Stellvertreter- bzw. Abbildungsfunktion

Dieses Konzept der Repräsentation muss um eine zusätzliche Spezifikation des Repräsentats b als mental ergänzt werden, um zum Begriff der mentalen Repräsentation zu kommen. Eine mögliche Definition mentaler Repräsentation[25]:

- **REP$_M$** (S, X, Y), wobei
  - S = ein individuelles informationsverarbeitendes System
  - Y = ein Aspekt des gegenwärtigen Zustands der Welt
  - X = als funktional individuierter, interner Systemzustand das Representat von Y für S

Es ist hier wichtig, sich die verschiedenen Teilaspekte von X zu vergegenwärtigen: X ist ein funktional individuierter interner Systemzustand(1), der einen Aspekt der Welt(2) für das System(3) repräsentiert. Die symbolizistische Modellannahme hat mit diesem Repräsentationsbegriff keine Probleme: ihr Konzept von mentalen Repräsentationen, die a priori sowohl syntaktisch wie auch semantisch individuiert sind, erfüllt alle Randbedingungen einer representationalistischen Theorie des Geistes. Die Frage ist, welchen Status mentale Repräsentationen in konnektionistischen Modellen kognitiver Systeme besitzen.

4. Die konnektionistische Theorie der "Constraint Satisfaction Networks"

Im folgenden soll eine konnektionistische Strukturtheorie untersucht werden, die sich auf interessante Weise mit dem philosophischen Problem des Inhalts mentaler Repräsentationen und ihrer unterstellten semantischen Kompositionalität widmet. Dabei ist anzumerken, dass die Autoren der untersuchten Hypothesen explizit darauf hinweisen, mit ihren Vorschlägen "keine formale konnektionistische Theorie der Repräsentation und Verarbeitung von Konzepten"[26] aufzustellen. Vielmehr geht es ihnen um das Aufstellen eines "vielversprechenden theoretischen Rahmen[s]", in dem sich Lösungsmöglichkeiten für die o.g. Probleme suchen lassen. Meine Bezeichnung "Constraint Satisfaction Networks" für den Komplex der Überlegungen von Thomas Goschke und Dirk Koppelberg

[25] Die folgende Definition ist aus meiner Mitschrift zur Vorlesung PdG IV – Intentionalität und mentale Repräsentationen (SS 2004) von T. Metzinger an der Universität Mainz

[26] Goschke, T. & Koppelberg, D. (1993). Konnektionistische Repräsentation, semantische Kompositionalität und die Kontextabhängigkeit von Konzepten. In H. Hildebrandt & E. Scheerer (Hrsg.), *Interdisziplinäre Perspektiven der Kognitionsforschung*. Frankfurt a.M., Peter Lang, S. 90

postuliert daher möglicherweise einen formaltheoretischen Zusammenhang bzw. Systemcharakter, der in den ursprünglichen Überlegungen so nicht vorhanden ist, die philosophische Arbeit aber erleichtert, indem er die begriffliche Vorsicht der Forscher[27] zugunsten einer eindeutigen Position aufgibt.

4.1. Korrelationale Sematik

Der erste Pfeiler der Theorie entsteht aus Überlegungen zu der Frage, inwiefern man innere Zustände konnektionistischer Modelle als (mentale) Repräsentationen eines bestimmten Inhalts verstehen kann. Hierbei spielen die in Abschnitt 2.2.a vorgestellten "versteckten Einheiten" neuronaler Netze eine wichtige Rolle. In Experimenten, in denen (mittels "back-propagation" Algorithmen trainierte) neuronale Netze digitalisierten menschlichen Sprachlauten passende Phoneme zuordnen sollten, hatten sich im Testverlauf in den versteckten Einheiten "Detektoren" herauskristallisiert, die zwischen Vokalen und Konsonanten unterscheiden konnten[28]. Dieses Unterscheiden-können wurde durch das Beobachten einer stabilen Korrelation zwischen Eingabemustern (Vokal vs. Konsonant) und der Aktivierung der jeweiligen (Knoten von) versteckten Einheiten festgestellt. Die Interpretation von Aktivierungsmustern als in Korrelation zu Eingabemustern stehend kann dabei von einzelnen, direkt beobachteten, inputabhängig aktivierten Einheiten bis hin zu komplexen, multivariaten Clusteranalysen etabliert werden, in denen die *Ähnlichkeit* von Aktivierungsmustern in Abhängigkeit von Eingabevektoren als klassifikatorisches Merkmal dienen muss. Die "Detektoren" sind also nicht notwendig einzelne Einheiten, vielmehr sind es oft Gruppen von versteckten Einheiten, die als komplexe innere Aktivierungsmuster äussere Zustände kodieren.

Bedeutung bzw. Inhalt kommt mentalen Repräsentationen nach diesen Vorstellungen somit qua *Indikatorfunktion* zu: insofern sie verlässlich mit äusseren Zuständen der Welt korrelieren, repräsentieren sie diese Zustände für das System (Vgl. Abschnitt 3.3).

Dieser Ansatz unterscheidet sich in den Grundannahmen stark von denen, die das klassische, symbolizistische Paradigma als semantische Theorie annimmt. Der Hauptunterschied liegt in der Art und Weise, wie die korrelationale Semantik die Arbitrarität der Bedeutungsrelation der klassischen Bedeutungstheorien vermindert. Auch in der korrelationalen Semantik muss die Bedeutung mentaler Repräsentationen letztlich interpretiert, d.h. zugeschrieben werden: ein Aktivierungsmuster im neuronalen Netz als solches hat keine intrinsische, d.h. extrarepräsentationale Bedeutung. Allerdings ist die Zuschreibung von Bedeutung zu mentalen Repräsentationen (die, wir erinnern uns, als Aktivierungsmuster des Netzes verstanden werden) *weniger arbiträr* als in klassischen, symbolizistischen Modellen, da die inter-repräsentationale Bedeutung, d.h. die Korrelation von

---

[27] Vgl. Goschke, T. & Koppelberg, D. (1993), a.a.O., S. 91
[28] Vgl. Goschke, T. & Koppelberg, D. (1993), a.a.O., S. 68

Aktivierungsmustern in Abhängigkeit von Eingabemustern mit der Ähnlichkeit der Eingabemuster kovariiert[29]. Während im Symbolizismus Symbole als Setzung letztlich vollständig arbiträr zugeschriebene Weltbedeutung ("Inhaltsessenzen", Vgl. Abschnitt 3.3) besitzen müssen, wird in der korrelationalen Semantik Bedeutung nicht als arbiträre extrarepräsentationale Äquivalenz, sondern funktional als "nützliche", intrarepräsentationale Ähnlichkeit verstanden. Die Komplexität der Indikatorfunktion in Modellen kognitiver Systeme mit mentalen Repräsentationen kann dabei so hoch sein, dass deren semantischer Gehalt für den Beobachter "weitgehend intransparent bleibt"[30].

4.2. Kontextsensitive Konzepte

Der zweite theoretische Pfeiler von CSN als kognitiver Strukturtheorie entsteht bei der Betrachtung des Problems, wie die Bedeutung komplexer mentaler Strukturen aus der Kombination von einfachen mentalen, repräsentationalen Konzepten hergestellt werden kann. Die damit zusammenhängenden, zu erklärenden Phänomene wurden bereits in Abschnitt 3.1 und 3.2 vorgestellt (Produktivität, Systematizität und Kompositionalität). Nach der klassischen symbolizistischen Modellannahme ist die Kontextunabhängigkeit von basalen mentalen Repräsentationen Grundbedingung für die Erklärung der kognitiven Zielphänomene[31]. Wir erinnern uns zudem an den Abschnitt 2.1.c, in dem deutlich wurde, dass die Autoren der LOT Hypothese ihre Theorie als physikalisch realistische und damit empirische Theorie verstehen.

Gerade die empirische Psychologie ist es aber, die zahlreiche Befunde[32] für eine starke Kontext*sensitivität* von basalen mentalen Konzeptstrukturen nahelegt: so werden Fragen nach Typikalitätsurteilen, mit denen diese Konzeptstrukturen ermittelt werden können, sowohl von unterschiedlichen Populationen (z.B. StudentInnen vs. ProfessorInnen), als auch von unterschiedlichen Versuchspersonen gleicher Populationen, wie auch von der gleichen Versuchsperson an unterschiedlichen Tagen(!) unterschiedlich beantwortet. Auch die innere strukturelle Verbindung der basalen mentalen Repräsentationen von Konzepten ist keine invariante Beziehung: die wahrgenommene Ähnlichkeit von Konzepten korreliert "untereinander nur mässig[...]", der "[...] Eindruck der Stabilität von Konzeptstrukturen dürfte vielmehr ein Artefakt der Mittelwertbildung über viele Personen, Kontexte und Aufgaben sein."[33].

---

[29] Vgl. Christiansen, M. H., Chater, N. (1992) Connectionism, Lerning and Meaning. *Connection Science*, **4**, 227-252. Im Gegensatz zu den Autoren bin ich der Ansicht, dass die Tatsache, dass extrarepräsentationale Bedeutung *immer arbiträr* ist, nicht notwendig zu einer Abwertung der erstaunlichen Tatsache führt, dass n. Netze über inter-repräsentationale Strukturähnlichkeiten Strukturen der Welt modellieren können. Eine derart naturalisierte Epistemologie ist nicht notwendig eliminativ, sie verneint allerdings ihre radikale Autonomie, Vgl. Churchland, P. M. (1989) On the nature of theories: A neurocomputational perspective. *Minnesota Studies in the Philosophy of Science* **14**, S. 59-101

[30] Goschke,, T. & Koppelberg, D. (1993), a.a.O., S. 73

[31] Fodor, J. A., Plyshyn, Z. W. (1988), a.a.O., S. 42

[32] Vgl. Goschke,, T. & Koppelberg, D. (1993), a.a.O., S. 80ff

[33] Vgl. Goschke,, T. & Koppelberg, D. (1993), a.a.O., S. 83

Wenn sich herausstellt, dass die Bedeutung von mentalen Repräsentationen in Abhängigkeit zu situativen oder globalen Einflüssen steht, kann eine kontextunabhängige, den mentalen Repräsentationen intrinsische Semantik nicht kohärent gefordert werden. Sowohl die Strategie, Kontextsensitivität immer als echte Mehrdeutigkeit, also unterschiedliche lexikalische Einträge zu verstehen, als auch die Strategie, Kontextsensitivität als Phänomen synkategorematischer Ausdrücke (Ausdrücke die nur in Verbindung mit anderen Ausdrücken eine Bedeutung erlangen) zu verstehen, müssen als gescheitert gelten[34].

Die Kontextsensitivität von komplexen Begriffen lässt sich auch nicht formalisieren, indem man Schnittmengen über die Extensionen der basalen Begriffe bildet oder Merkmalsdimensionen formuliert, auf denen sich ihre Bedeutung ausdehnt, da in die Berechnung stets "allgemeines Weltwissen über Kovariationen zwischen Merkmalen" einfliesst[35].

Diese Befunde lassen eine starke Kompositionalität, wie sie der klassische symbolizistische Ansatz erfordert, unwahrscheinlich erscheinen. Um das Sprachverhalten von Sprechern/Hörern einer Sprache, das das mühelose Verstehen komplexer Begriffe beinhaltet, adäquat zu modellieren, benötigt man jedoch (irgend-)eine Theorie der Kompositionalität. Die Idee ist hier, dass eine Verbindung der im vorigen Abschnitt erörterten korrelationalen Semantik mit den Eigenschaften distribuierter Repräsentationen in neuronalen Netzen ein Modell der schwachen Kompositionalität ermöglicht, dass über den Dreiklang von Semantik, Pragmatik und Weltwissen die Kompositionalität strukturierter mentaler Repräsentationen erklären kann, ohne selbst aus strukturierten Symbolsystemen zu bestehen.

4.3. Constraint Satisfaction Networks als Strukturtheorie der Kognition

Im Rahmen des Modells von CSN als Strukturannahmen der Kognition stellen T. Goschke und D. Koppelberg folgende Analogie vor:

> "Für das Verständnis der Arbeitsweise von Constraint-Satisfaction-Modellen ist es hilfreich, Verarbeitungseinheiten als Hypothesen, Aktivierungsgrade als Bestätigungsgrade einer Hypothese und Konnektionen zwischen Einheiten als Randbedingungen zu interpretieren[...]"[36]

In einem neuronalen Netz implementieren die Randbedingungen die Beziehungen zwischen unterschiedlich gewichteten Räumen, die unterschiedliche Merkmalsdimensionen darstellen. Die Unterschiedlichkeit der Merkmalsdimensionen ist dabei das Produkt der strukturellen Eigenschaften von Eingabevektoren sowie der bereits bestehenden Randbedingungen, also der bestehenden Konnektivität des Netzes. Ein neuronales Netz in seiner Funktion als kognitives Modell hat also immer einen gewissen Grad an erfüllten Randbedingungen, an globaler Kohärenz, den es zu maximieren

---

[34] Goschke,, T. & Koppelberg, D. (1993), a.a.O., S. 87
[35] Goschke,, T. & Koppelberg, D. (1993), a.a.O., S. 89
[36] Goschke,, T. & Koppelberg, D. (1993), a.a.O., S.93

versucht[37].

In diesem Modell sind Eingabevektoren nichts anderes als weitere Randbedingungen des globalen Zustandsraums, da sie nur als Veränderung der Konnektivität aufgrund ihrer spezifischen Aktivierungsgeschichte auf den globalen Zustandsraum wirken.

So wird denn auch klar, wie CSN Modelle Produktivität und Systematizität erklären: als Ergebnis statistischer Kovariationsmuster über Merkmalsdimensionen, letztlich also über eine Ähnlichkeitsrelation, deren Bedeutung aus der Zuschreibung von semantischen Inhalten zu internen "Gravitationszentren der Aktivität" des Netzes resultiert[38].

Die Tatsache, dass Eingabevektoren in CSN Modellen als weitere Randbedingungen des Netzes auf ihre interne Verarbeitung sowie die existierende Konnektivität wirken, kann in plausibler Weise erklären, wie die Kontextsensitivität der basalen mentalen Repräsentationen zustande kommt: es existiert schlicht keine mentale Repräsentation, die nicht als distribuiertes Aktivierungsmuster von inneren und äußeren Kontextvektoren beeinflusst wäre[39]. CSN als Modelle einer kognitiven Strukturtheorie zu postulieren erlaubt also, einen holistischen Blick auf die Fragen einer korrelativen Semantik in Verbindung mit dem Konzept einer schwachen Kompositionalität zu werfen, die die Bedeutung komplexer mentaler Repräsentationen "mit dem Prozess identifiziert [...], durch den ein Netzwerk auf jenen Punkt im konzeptuellen Raum konvergiert, der möglichst viele der Randbedingungen erfüllt, die zugleich durch die aktuelle Eingabe und die innere Konnektivität des Netzes definiert werden."[40]

5. Philosophische Implikationen der CSN als Strukturtheorie der Kognition

5.1. Eliminativismus?

Eine der interessantesten Fragen, die man sich stellen muss, wenn man den Konnektionismus als kognitionswissenschaftliche Strukturtheorie annehmen will, ist die Frage nach dem Status propositionaler Einstellungen. Propositionale Einstellungen sind die terminologischen Grundlagen der von Philosophen häufig als "folk psychology" (FP) bezeichneten Theorie, sie bezeichnen Aussagen wie

---

[37] Vgl. Goschke,, T. & Koppelberg, D. (1993), a.a.O., S.94. Interessant ist in diesem Zusammenhang auch die Terminologie der "Energie" des Netzes: je höher das Maß globaler Kohärenz, desto niedriger die Energie des Netzes. Man ist versucht, hier weitläufige Parallelen zum zweiten Hauptsatz der Thermodynamik zu ziehen, auch gerade im Hinblick auf die externalistische Theorie der Bedeutung und die Nichtabgeschlossenheit kognitiver Systeme, vielleicht auch mit einem Seitenblick auf Vorschläge zu einem "aktiven Externalismus" (Vgl. Clark, A., Chalmers, D. (1998) The Extended Mind. *Analysis* **58** S. 10-23. auch: Grim, P. (Hrsg.) (1998) *The Philosopher's Annual*, vol XXI). Derartige Spekulationen bleiben in dieser Arbeit allerdings eine Fußnote.

[38] Ein Punkt den Fodor und seine Mitstreiter – wenig überraschend – nicht akzeptieren. Für ihre Argumentation vgl. Lepore, E., Fodor, J. (1999) All At Sea in Semantic Space: Churchland on Meaning Similarity. *the Journal of Philosophy,* **96***, S. 381-403*

[39] Goschke,, T. & Koppelberg, D. (1993), a.a.O., S.96

[40] Goschke,, T. & Koppelberg, D. (1993), a.a.O., S. 97

“glauben, dass P”, “hoffen, dass P”, “wünschen, dass P”, “meinen, dass P”. Der philosophische Diskurs um den Status der FP ist wichtig, da viele gesellschaftlich-kulturelle Institutionen, von juristischen Fragen zu Schuldvermögen über pädagogische Konzepte bis hin zu Fragen des Common-Sense Selbstverständnisses als freie, handelnde, denkende Wesen implizit die Wahrheit der FP annehmen, zumindest jedoch ihre Ersetzung durch ontologisch konservative Theorien erwarten[41].

Dieser Erwartung widersprechen einige Forscher[42]. Ihrer Meinung nach gilt folgender Konditionalsatz: wenn konnektionistische Modelle für kognitive Phänomene die geeignete Beschreibung sind, dann ist die FP eine *falsche* Theorie, da propositionale Einstellungen keine funktional diskreten, semantisch interpretierbaren Zustände mit einer kausalen Rolle in der Verhaltenssteuerung des kognitiven Systems darstellen.

Die Argumentationsstrategie ist auf folgenden Prämissen aufgebaut[43]: (1) In konnektionistischen Modellen sind viele Verbindungen und Einheiten gleichzeitig aktiv. (2) Jede dieser Einheiten kodiert gleichzeitig mehrere – wenn nicht alle – Repräsentationen des Systems. (3) Wenn die Einheiten und Verbindungen in *einer* Repräsentation kausal aktiv sind, dann sind *alle* Repräsentationen, die durch diese Einheiten und Verbindungen kodiert werden, kausal aktiv. Konklusion:(4) Es macht keinen Sinn davon zu sprechen, dass einige Repräsentationen kausal aktiv sind und andere nicht.

Diese argumentative Strategie schlägt jedoch fehl, da zwar sowohl (1) wie auch (2) in konnektionistischen Netzwerken mit distribuierten Repräsentationen gelten, (3) jedoch weder für kurzfristige noch für längerfristig bestehende Repräsentationen notwendig gilt. Ein Beispiel: sechs Personen sind in die Arbeit von zwei gemeinnützigen Organisationen A und B involviert. Alle sechs Personen arbeiten sowohl in Organisation A wie auch in Organisation B. Aus dieser Tatsache kann man jedoch nicht schliessen, dass Organisation A in allen Aktivitäten von Organisation B implizit aktiv ist[44]. Analog diesem hübschen Beispiel explizieren Foster und Saidel die Rolle distribuierter Repräsentationen in neuronalen Netzen. Die Tatsache, dass Repräsentationen bei genügend differenzierter Analyse funktional unterscheidbar sind, etabliert trotz der Implementation der Repräsentationen über die gleichen Einheiten und Verbindungen des neuronalen Netzes ihren kausalen Modulcharakter.

---

41 Vgl. Unterschied ontologisch konservativ vs. ontologisch radikal in: Metzinger, Thomas (1996) Anthropologie und Kognitionswissenschaft. in: Engel, A. & Gold, P. (Hrsg.), *Der Mensch in der Perspektive der Kognitionswissenschaft*. Frankfurt am Main: Suhrkamp. auch: http://www.philosophie.uni-mainz.de/metzinger/publikationen/1996q.html

42 Am prominentesten wahrscheinlich: Ramsey, W., Stich, S., Garon (1991). Connectionism, Eliminativism, and the Future of Folk Psychology. Auf sie rekurrierend argumentiert auch: Davies, M. (1991). Concepts, connectionism, and the language of thought. *beides in:* Ramsey W., Stich S., & Rumelhart D.E. (Hrsg.), *Philosophy and connectionist theory*. Hillsdale, NJ, Erlbaum.

43 Die klare Darlegung der Argumentationsstrategie von RSG wie auch das schlagende Argument dagegen verdanke ich Malcolm Foster und Eric Saidel, vgl. Foster, M. Saidel, E. (1994) Connectionism and the fate of folk psychology: a reply to Ramsey, Stitch and Garon, a.a.O.

44 Foster, M. Saidel, E. (1994) Connectionism and the fate of folk psychology: a reply to Ramsey, Stitch and Garon, a.a.O, S. 444

Nach diesen Überlegungen ist die FP keineswegs eine falsche Theorie, vielmehr bietet das konnektionistische Paradigma die mathematisch präzise beschreibbare Mikrostruktur, um die Generalisierungen der FP ontologisch konservativ zu explizieren. Die Beschreibungsebene der FP hat zudem den unschätzbaren Vorteil, das Verhalten kognitiver Systeme ohne Berücksichtigung komplexer implementatorischer Details korrekt vorauszusagen[45]. Bei dieser Betrachtung kann die These, konnektionistische Strukturtheorien der Kognition führten qua ihrer Struktur zu eliminativ-materialistischen Theorien nicht aufrecht gehalten werden.

5.2. Internalismus? Externalismus? Holismus!

Einer der interessantesten Aspekte der CSN Hypothese als strukturtheoretische Modellannahme der Kognition besteht in ihrem Potential, einem berüchtigten erkenntnistheoretischen Problem der Philosophie einen möglichen Lösungsrahmen vorzuschlagen: nämlich dem der Internalismus/Externalismus Debatte. Das philosophische Problem besteht traditionell in der Frage, wie epistemische Subjekte gerechtfertigt Meinungen haben bzw. bilden können. Woran kann ich eine wahre von einer falschen Meinung unterscheiden, was ist das Wahrheitskriterium für propositionale Einstellungen wie glauben, dass p, meinen, dass p, etc.? Während Externalisten die Wahrheit einer propositionalen Einstellung (z.B. “ich glaube, es regnet”) als von ihrer Realisierung in der Welt abhängig definieren, sehen Internalisten die Notwendigkeit, die Wahrheit einer propositionalen Einstellung von subjektinternen Faktoren (wie Argumente-dafür-haben, introspektiv-zugänglich-sein, etc.) abhängig zu formulieren. Eine Theorie, die in dieser Debatte fruchtbar argumentieren will, ist also stark von ihren Hintergrundannahmen zur Entstehung von Bedeutung bei mentalen Repräsentationen abhängig.

In Abschnitt 4.1 habe ich gezeigt, dass das konnektionistische Programm der CSN in Bezug auf die Bedeutung mentaler Repräsentationen weitgehend aus einer korrelativen Semantik besteht. Demnach entsteht Bedeutung aus der korrelativen Beziehung von internen Zuständen des Netzes zu in Inputvektoren kodierten Strukturmerkmalen der Welt. Diese Beziehung wird auch als Indikatorfunktion bezeichnet. Diese externalistische Postition hat Probleme zu bewältigen: so ist z.B. die Frage, wie Fehlrepräsentationen in diesem Modell möglich sind, schwierig zu beantworten. Ein Beispiel: ich stehe morgens auf und gehe in die Küche, um mir einen Kaffee zuzubereiten. Ich nehme das Kaffeepulver, koche es auf und trinke die schwarze Flüssigkeit. Ich habe nun die Überzeugung, durch mein Kaffeetrinken Koffein aufgenommen zu haben. Dummerweise hat mein Mitbewohner aber

[45] Vgl. Foster, M. Saidel, E. (1994) Connectionism and the fate of folk psychology: a reply to Ramsey, Stitch and Garon, a.a.O, S. 449. In diesem Punkt ähnelt der hier vertretene Standpunkt dem von Daniel Dennett und seiner Unterscheidung zwischen physikalischer, funktionaler und intentionaler Beschreibungsebene: die FP ist die klassische “intentional stance” des Menschen. Vgl. Dennett, D. C. (1971) Intentional Systems. *Journal of Philosophy* **68**, S. 87-106, auch: Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

koffeinfreien Kaffee gekauft, ich habe also de facto eine Fehlrepräsentation des Kaffees als koffeinhaltig. Im Rahmen radikal-externalistischer Bedeutungstheorien ist nicht klar, wie ich überhaupt die Überzeugung, koffeinhaltigen Kaffee getrunken zu haben, formulieren könnte: man müsste hier sagen, dass ich diese Überzeugung nicht haben *kann*, da das Kaffeepulver kein Koffein enthielt.
Ein weiteres Problem der externalistischen Position ist die Tatsache, dass mentale Repräsentationen untereinander in begrifflichen Relationen zu stehen scheinen. Wenn die Bedeutung der mentalen Repräsentationen exklusiv aufgrund korrelativer Beziehungen zu äusseren Faktoren in der Welt zugeschrieben werden kann, ist zumindest nicht offensichtlich, wie strukturelle Verbindungen der mentalen Repräsentationen untereinander entstehen.
Die Theorie der CSN bietet für beide Probleme einen Rahmen, in dem die vermeintlichen Gegensätze zwischen Internalismus – der einen hilfreichen Hinweis auf interne strukturelle Verbindungen zwischen mentalen Repräsentationen, sowie eine einfache Möglichkeit zu Fehlrepräsentationen bietet – und Externalismus – der eine plausible naturalistisch inspirierte Theorie der Bedeutung mentaler Repräsentationen bietet – aufgelöst werden. Wie in Abschnitt 4.3 ausgeführt werden im Rahmen der Theorie der CSN Inputvektoren einfach als weitere Randbedingungen des Systems charakterisiert, innere Einheiten jedoch ebenfalls als aktivierende (oder hemmende) Randbedingungen involviert. Für die Frage nach der Semantik mentaler Repräsentationen ist es demnach schlicht irrelevant, ob die Randbedingung aufgrund von interner Konnektivität oder externer Inputvektoren erscheint: beide beeinflussen das Resultat[46]. Auch wenn man konstatieren muss, dass man ohne eine korrelative Beziehung zu (in Inputvektoren kodierten) strukturellen Merkmalen der Welt überhaupt nicht von Bedeutung sprechen könnte, so ist es doch ebenso klar, dass man, um die Bedeutung einer distribuierten Repräsentation verstehen zu können, die innere Konnektivität des Aktivierungsvektors berücksichtigen muss. In dieser holistischen Betrachtung scheint es unplausibel, eine prinzipielle Inkommensurabilität der beiden Faktoren interne Konnektivität und externem Inputvektor zu postulieren.

5.3. Gibt es eine Ethik des konnektionistischen Selbstverständnisses?

Es ist eine interessante Frage, inwiefern man aus einer deskriptiven Strukturtheorie des Geistes, die sich anschickt, eine naturalisierte Epistemologie aufzustellen, normative Feststellungen treffen kann[47]. Während primitiv-naturalistische Fehlschlüsse vom "Sein" der Dinge auf ihr "Sollen" sicherlich falsch sind, zeigt die Beschäftigung mit den im Rahmen der Diskussion des Konnektionismus als kognitiver

[46] Goschke, T. & Koppelberg, D. (1991). The concept of representation and the representation of concepts in connectionist models. S. 154 in: Ramsey W., Stich S., & Rumelhart D.E. (Hrsg.), *Philosophy and connectionist theory* S.129-162. Hillsdale, NJ, Erlbaum.

[47] Vgl. Churchland, P. M. (1989) On the nature of theories: A neurocomputational perspective. *Minnesota Studies in the Philosophy of Science* **14**, S. 101

Strukturtheorie aufgeworfenen Fragen doch auch, dass es direkte Implikationen für unser Selbstverständnis hat, welche Theorie der Kognition wir annehmen. So kann man z.B. an der Auseinandersetzung mit der Position des eliminativen Materialismus feststellen, dass sich die theoretische Stellung der Erkenntnistheorie als prima philosophia zugunsten einer differenzierteren Position ihrer Stellung aufgelöst hat.[48] Die Erkenntnis, dass keine Theorie ohne die ihr eigenen "Sprachspiele" verstanden werden kann, findet sich zwar bereits in Quines berühmten Gedankenexperiment zur Unbestimmtheit der Übersetzung, in konnektionistischen Zusammenhängen finden Philosophinnen nun vielleicht sogar empirische Bestätigung für derlei Hypothesen[49].
Man kann das Programm des Konnektionismus als Versuch auffassen, eine biologisch plausible, physikalistische Theorie mentaler Repräsentationen zu postulieren, ohne einen radikalen eliminativen Materialismus zu fordern. In diesem Zusammenhang betrachtet zeigen sich ganz unterschiedliche philosophische Projekte, so z.B. das der Frankfurter Schule in ihrem Bestreben, den "herrschaftsfreien Diskurs" als notwendigen ethischen Aspekt des soziologischen Geschehens zu formulieren, oder das des Dekonstruktivismus in seiner Ablehnung einer ultimativen Wahrheitsebene, in neuem Licht. So kann man in einer zugegebenermassen gewagten Spekulation die These formulieren, dass sich im Rahmen des konnektionistischen Paradigmas erste Anzeichen einer möglichen realen Integrierbarkeit normativen Wissens in deskriptive Modellannahmen zeigen, die ethischen Intuitionen, wie sie sich in aktuellen gesellschaftlichen Institutionen wie Demokratie, Minderheitenschutz, Emanzipierung oder kultureller Toleranz äussern, wie auch einem konsistenten, physikalistischen Weltbild die Chance ermöglichen, sich gegenseitig sinnvoll zu Befruchten.

6. Fazit

Die Konsequenzen dieser theoretischen Einsichten (Ansichten?) sind naturgemäss umstritten. Während pessimistische Stimmen im Zusammenhang mit konnektionistischen Vorstellungen des Menschen vor einem reduktionistischen Verfall unseres Selbstbildes warnen, in dem "keine Individualität und Einmaligkeit im strengen Sinne"[50] und demnach auch keine "vollwertige" Ethik mehr möglich sei, feiern optimistische Ausblicke auf die "kognitive Revolution" enthusiastisch die geglückte naturalistische Situierung bislang nur dualistisch "begründeter", mentaler Zustände im Rahmen der Wissenschaften als die wichtigste Entwicklung der letzten 300 Jahre und preisen deren Potential zur Etablierung menschlicher Werte, der Bekämpfung globaler Mißstände und dem Erhalt eines hohen

---

[48] Vgl. Abschnitt "Eliminativism Then And Now" in: Rockwell, Teed (2004) Eliminativism. in: Dictionary of Philosophy of Mind, C. Eliasmith (Hrsg.) 12.09.2004 http://www.artsci.wustl.edu/~philos/MindDict/eliminativism.html

[49] Vgl. Goschke,, T. & Koppelberg, D. (1991), a.a.O., S. 155

[50] Eraßme, R.: Der Mensch und die 'Künstliche Intelligenz'. Eine Profilierung und kritische Bewertung der unterschiedlichen Grundauffassungen vom Standpunkt des gemäßigten Realismus. Diss. Aachen, 2002, S. 104

qualitativen Lebensstandards[51]. Wieder andere sehen in der Zurückweisung der Erkenntnistheorie als prima philosophia (und damit der Entwicklung eines relativierten Wahrheitsbegriffs) eine Entwicklung, die es vor allem den Geisteswissenschaften ermöglicht, ihr heuristisches Potential mit neuem Selbstbewusstsein zu postulieren[52].

Auch wenn man die optimistischen Ausblicke nicht teilt, festzuhalten ist in jedem Fall, dass es dem Konnektionismus als Mitbewerber um eine umfassende kognitive Strukturtheorie gelungen ist, Vorstellungen kognitiver Leistungen in vielerlei Hinsicht zu präzisieren und Überlegungen zum Status des "Geistes" im Rahmen der philosophischen Leib/Seele Problematik in verdienstvoller Weise anzureichern. In den meisten Zusammenhängen ist eine primitive komputationalistische Annahme des Menschen als starre "syntactic engine" schlicht keine zutreffende Metapher[53]. Welche weiteren Verdienste der konnektionistischen Modellbildung, gerade auch in Verbindung mit der geisteswissenschaftlichen und philosophischen Theoriebildung zukommt, wird die Zukunft zeigen.

---

[51] Vgl. Sperry R. W. (1993) The impact and promise of the cognitive revolution. *American Psychologist*, **48** (3), S. 878-885. Auch wenn in diesem Artikel nicht explizit die Rede von konnektionistischen Modellen, sondern etwas schwammig von der "kognitiven Revolution" ist, so kann der Ausblick meiner Meinung nach dennoch auf die in konnektionistischen Modellannahmen formulierten Zusammenhänge gelten.

[52] Vgl. Rorty, R. (2000) The Decline of redemptive truth and the rise of literary culture. 12.09.2004, http://www.stanford.edu/~rrorty/decline.htm

[53] Vgl. Metzinger, Thomas (1996) Anthropologie und Kognitionswissenschaft. in: Engel, A. & Gold, P. (Hrsg.), *Der Mensch in der Perspektive der Kognitionswissenschaft*. Frankfurt am Main: Suhrkamp. auch: http://www.philosophie.uni-mainz.de/metzinger/publikationen/1996q.html

E R K L Ä R U N G

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Mainz, 21.09.2004 _____

(Dominique Kaspar)