Erscheint in: *Zeitschrift für philosophische Forschung*.

Erweiterte Kognition und mentaler Externalismus

Holger Lyre
Institut für Philosophie
Universität Magdeburg
lyre@ovgu.de

5. Oktober 2009

Inhalt

Zusammenfassung
Die jüngeren Entwicklungen unter den Schlagwörtern Dynamizismus, Embodiment und situierte Kognition legen die Auffassung nahe, dass kognitive Systeme nicht auf das neuronale System beschränkt sind, sondern sich über die traditionellen Systemgrenzen hinaus in die Welt erstrecken. Dies ist die Grundthese der erweiterten Kognition. Eine derartige Erweiterung der kognitiven Vehikel führt auf einen neuartigen Gehalts-Externalismus, der als aktiver Externalismus bezeichnet wird. Der Aufsatz

verfolgt dreierlei Ziele: Erstens die Thesen der erweiterten Kognition und des aktiven Externalismus herauszuarbeiten und begrifflich voneinander abzugrenzen. Zweitens den aktiven Externalismus von seinen verschiedenen passiv-externalistischen Vorläufern in Form des physikalischen, historischen und sozialen Externalismus zu unterscheiden und in seiner Sonderstellung zu untersuchen, was auf eine sehr umfangreiche Diskussion und Schwachstellenanalyse aller genannten Spielarten des mentalen Externalismus hinausläuft. Und drittens zu zeigen, dass der soziale Externalismus im Gegensatz zum physikalischen und historischen Externalismus einen graduellen Übergang vom passiven zum aktiven mentalen Externalismus gestattet.

Abstract
Recent developments under the keywords dynamicism, embodiment and situated cognition suggest the view that cognitive systems are not confined to the neural system but leak into the world beyond the traditional system boundaries. This is the thesis of extended cognition. Such an extension of the cognitive vehicles leads to a new kind of content externalism, known as active externalism. The essay pursues three objectives: firstly, to distinguish the theses of extended cognitive and active externalism. Secondly, to delineate active externalism from its various passive-externalist precursors in the form of physical, historical and social externalism and to scrutinize its special position, which amounts to a comprehensive discussion and analysis of all variants of mental externalism. And thirdly, to show that social externalism as opposed to physical and historical externalism allows for a gradual transition from passive to active mental externalism.

0. Einleitung

Nach traditioneller Auffassung sind kognitive Prozesse als exklusive Aktivitäten unseres Gehirns anzusehen. Die Entwicklungen innerhalb der letzten Dekade unter den Schlagwörtern Dynamizismus, Embodiment oder situierte Kognition lassen jedoch zunehmend Zweifel an dieser Auffassung aufkommen. Demnach sind kognitive Systeme keineswegs auf den lokalen Verarbeitungsapparat, die neuronale Maschinerie, beschränkt, sondern erstrecken sich über die traditionellen Systemgrenzen hinaus in den Körper, die Umgebung und externe kognitive Werkzeuge. Dies ist die Grundthese der erweiterten Kognition (extended cognition, extended mind).

Die Erweiterungsthese lässt sich einerseits auf die physische Realisierung kognitiver Systeme, die *Vehikel* der kognitiven Aktivität, beziehen, sie kann sich andererseits aber auch auf die mit den physisch realisierten kognitiven Prozessen einhergehenden oder auf ihnen supervenierenden mentalen *Gehalte* beziehen. Für die erste Auffassung wird im Folgenden die Bezeichnung *Erweiterte Kognition*, für die zweite Auffassung die Bezeichnung *Externalismus* verwendet. Die Grundidee der Erweiterten

Kognition ist im angelsächsischen Raum unter dem Schlagwort „extended mind“ vor allem durch Andy Clark bekannt gemacht worden. Die spezifische Variante desjenigen Externalismus, der speziell mit der Extended mind-These verbunden ist, wird aus Gründen, die später erläutert werden, als *aktiver Externalismus* bezeichnet. In der angelsächsischen Debatte ist gelegentlich auch, Susan Hurley folgend, die Unterscheidung zwischen *Vehikel-* und *Gehalts-Externalismus* gebräuchlich. Da sich die Bezeichnung Externalismus in der sprach- und geistphilosophischen Tradition jedoch ausschließlich auf Gehalte bezieht, soll hier an diese Terminologie anschließend besser von einer *Erweiterung der Vehikel* in Abgrenzung zum *Externalismus der Gehalte* gesprochen werden.

Der vorliegende Aufsatz verfolgt dreierlei Ziele: Erstens die Thesen der erweiterten Kognition und des aktiven Externalismus begrifflich voneinander abzugrenzen und zu diskutieren. Zweitens den aktiven Externalismus von seinen verschiedenen passiv-externalistischen Vorläufern in Form des physikalischen, historischen und sozialen Externalismus abzugrenzen und in seiner Sonderstellung zu untersuchen, was auf eine sehr umfangreiche Diskussion und Schwachstellenanalyse aller genannten Spielarten des mentalen Externalismus hinausläuft. Und drittens zu zeigen, dass der soziale Externalismus im Gegensatz zum physikalischen und historischen Externalismus einen graduellen Übergang vom passiven zum aktiven mentalen Externalismus gestattet.

1. Erweiterte Kognition

Die Idee der erweiterten Kognition geht einher mit dem Aufschwung der neueren KI-Forschungstrends unter den Schlagwörtern Dynamizismus, Embodiment (verkörperlichte Kognition), Embeddedness und situierte Kognition. Man könnte dieses unscharfe Sammelsurium als „Neuere KI“ (NKI) bezeichnen. Die NKI kann als drittes Paradigma der KI-Forschung in der Nachfolge von Symbolismus und Konnektionismus angesehen werden (vgl. Lyre 2002, Kap. 4.1) und stellt gleichzeitig eine natürliche Fortentwicklung vor allem des Konnektionismus dar. Um zu einem angemessenen Verständnis der Idee erweiterter Kognition zu gelangen, sei hier zunächst eine knappe, thesenhafte Charakterisierung der NKI gegeben.

1.1 „Neuere KI“

Wir beginnen mit dem *Dynamizismus*. Dynamizisten machen sich wichtige Eigenschaften neuronaler Netze zunutze wie etwa diejenige, dass *„... rekurrente neuronale Netze und dynamische Systeme mit reellwertigen Systemgrößen ... im wesentlichen aufeinander abbildbar*“ sind (Jäger 1996, 3.3). Aufgrund der Ähnlichkeit bzw. Isomorphie der mathematischen Modelle des Konnektionismus und der Theorie dynamischer Sys-

teme hat sich der Dynamizismus zu einer der vorherrschenden Arbeits- und Modellierungsgrundlagen der Neuroinformatik etabliert.

> *These des Dynamizismus*: Kognitive Systeme sind eine spezielle Klasse dynamischer Systeme.

Als dynamisches System kann zunächst jedwedes genügend komplexe, d.h. mit einer Vielzahl an Freiheitsgraden und einer hinreichenden Menge an Kopplungen zwischen den einzelnen Systemkomponenten ausgestattetes System angesehen werden. Natürliche neuronale Systeme lassen sich als hochkomplexe Netzwerke verkoppelter neuronaler Oszillatoren und somit als paradigmatische Fälle dynamischer Systeme (vgl. Bechtel und Abrahamsen 1991, Eliasmith und Anderson 2002, Thelen und Smith 1994, Port und van Gelder 1996).

Das für den Dynamizismus wesentliche Motiv der Kopplung des kognitiven Systems an seine Umgebung macht es unerlässlich, die konkrete physische Realisation des Systems, seine körperliche Erscheinungsform, als integralen Teil seiner kognitiven Gesamtarchitektur anzusehen. Tim van Gelder (1995) betrachtet beispielhaft den Wattschen Fliehkraftregler. Dabei handelt es sich um ein simples rückgekoppeltes Regelungssystem, dass seine Regelungsaufgabe (die Steuerung einer Dampfmaschine) nicht auf computationalem Wege erbringt, sondern unter Zuhilfenahme der dynamischen Rückkopplung des Fliehkraftreglers mit dem zu regelnden System. Van Gelder sieht hierin, wie viele Dynamizisten (speziell Brooks 1991), einen Beleg für einen radikalen Anti-Repräsentationalismus. Die sich hier anbindende philosophische Kontroverse wollen wir jedoch nicht verfolgen. Das Beispiel des Fliehkraftreglers zeigt aber, dass es entscheidend auf die physische Realisierung des Regelungssystems ankommt. Überträgt man dies auf natürliche kognitive Systeme, so ergibt sich ein wesentliches Motiv von *Embodiment*. Dies ist verallgemeinerbar: Der Dynamizismus führt natürlicherweise auf die Betonung der physischen Realisierung eines kognitiven Systems, also auf Embodiment. Dynamizismus impliziert insofern Embodiment.

> *Embodiment-These*: Zustände und Prozesse der körperlichen Realisierung eines kognitiven Systems tragen wesentlich zu dessen kognitiver Aktivität bei und sind insofern integraler Bestandteil dieses Systems.

Dabei kann es durchaus so sein, dass auch die repräsentationalen Eigenschaften eines kognitiven Systems in die Verkörperlichung eingebunden oder eingeschrieben sind. Hierzu ein alltägliches Beispiel: manche Menschen merken sich die PIN-Geheimnummer ihrer Kredit- oder Geldkarte an einem Automaten nicht in Form der eigentlichen Zahlenkombination, sondern in Form desjenigen Bewegungsmusters, das sie vollführen müssen, wenn sie mit der Tastatur des Geldautomaten interagieren. Konfrontiert mit einer Tastatur, die von der üblichen Anordnung der Zifferntasten abweicht, sind sie dann oft nicht oder erst nach einer simulierten Rekonstruktion

ihrer Bewegung in der Lage, die korrekte Nummer einzugeben. Die Nummer war insofern nicht arithmetisch, sondern als dynamisches Bewegungsmuster, also verkörperlicht, gespeichert. Ein Beispiel für genuin verkörperlichten mentalen Gehalt stellt etwa unser Wissen um rechts und links dar (Lyre 2008). Nennenswerte philosophische Autoren im Zusammenhang mit Embodiment sind Clark (1997), Gallagher (2005), Lakoff und Johnson (1999), Rowlands (1999, 2006) sowie Varela, Thompson und Rosch (1991).

Die *Situiertheit* kognitiver Systeme lässt sich als weitere Folge des Dynamizismus sowie der verkörperlichten Kognition ansehen, denn nicht nur die physischen Gegebenheiten eine kognitiven Systems sind von Relevanz, sondern auch dessen spezifische Kopplungen mit dem jeweiligen Umweltkontext. Neben dem Begriff Situiertheit („situatedness“) sind in der angelsächsischen Literatur auch die Begriffe „embeddedness“ oder „environmentalism“ üblich.

> *These der kognitiven Situiertheit*: Der Ablauf kognitiver Zustände und Prozesse hängt wesentlich von der spezifischen Umgebungssituation und Einbettung des Systems in seine Umgebung ab.

Der Aspekt der Situiertheit hat nicht unerhebliche Auswirkungen auf unser Verständnis mentaler Repräsentationen, denn das kognitive Geschehen hängt nun nicht mehr von den Bedingungen der kognitiven Maschinerie alleine, sie mag mit dynamizistischen Mitteln beschrieben werden oder nicht, oder vom neuronalen System plus Körper, sondern von der äußeren physischen Umgebung ab. Der Zusammenhang zwischen Dynamizismus, Embodiment und Situiertheit ist dabei wie folgt: als ein wesentlich dynamisches System ist die kognitive Maschinerie in ihrer Aktivität von den Kopplungen zur systemexternen Umgebung abhängig und getrieben. Um aber überhaupt mit der Umgebung dynamisch verkoppelt zu sein, benötigt das kognitive System eine physische Verkörperung, deren individuelle Struktur wesentlich in die Art der Dynamik eingeht (vgl. Robbins und Aydede 2009).

1.2 Die EC-These

Clark (2005) drückt die Grundidee der erweiterten Kognition wie folgt aus: *„the human mind need not be in the human head… the material vehicles of cognition can be spread out across brain, body and certain aspects of the physical environment itself."* In verfeinerter Anlehnung daran sei hier folgende These formuliert:

> *These der erweiterten Kognition (EC: "extended cognition"):* Kognitive Systeme umfassen über das interne neuronale System hinaus all diejenigen Teile des Körpers, der Umgebung, externer kognitiver Hilfsmittel und Werkzeuge sowie sozialer Gemeinschaften, die zur Durchführung, Auf-

> rechterhaltung und Stabilisierung kognitiver Fähigkeiten und Aktivitäten benötigt werden.

Die EC-These wurde zunächst unter dem Schlagwort „extended mind“ durch einen gleichnamigen Aufsatz von Andy Clark und David Chalmers (1998) bekannt. Sie wird seither vor allem von Clark (1997, 2003, 2008) massiv weiterverfolgt. Weitere Autoren im näheren Umfeld der EC-These sind Susan Hurley (1998a, b, im Druck), Richard Menary (2007; im Druck), Alva Noë (2009), Mark Rowlands (1999, 2006), Robert Wilson (1994, 2004) und Michael Wheeler (2005).

Die EC-These lässt sich schon der Form nach als eine empirische These ansehen: die kognitiven Neurowissenschaften sind es, die bestimmen müssen, welche „Komponenten … zur Durchführung, Aufrechterhaltung und Stabilisierung kognitiver Fähigkeiten und Aktivitäten benötigt werden“. Die These bedarf also des Nachweises durch die entsprechende empirische Forschung und Modellbildung. Falls sie sich behauptet – wofür eine Menge spricht – führt sie zu einer erheblichen Abänderung unserer Auffassung von kognitiven Systemen und ihrer Dynamik. In ihrer obigen Formulierung enthält die EC-These vier Unterthesen, die den vier grundsätzlichen Arten kognitiver Erweiterungen entsprechen:

1. *Kognitive Systeme erstrecken sich in den Körper.*
2. *Kognitive Systeme erstrecken sich in die Umgebung.*
3. *Kognitive Systeme erstrecken sich in externe Hilfsmittel und Werkzeuge.*
4. *Kognitive Systeme erstrecken sich in soziale Gemeinschaften.*

Die vier Unterarten oder auch „Dimensionen“ kognitiver Erweiterungen lassen sich grob den verschiedenen Trends der NKI zuordnen: Embodiment (1), Dynamizismus (2), situierte Kognition (1-4) und, noch hinzutretend, soziale Kognition (4). Bereits die Konzeption der EC-These beinhaltet also, dass genau diejenigen Anwendungsfälle Belege der These sind, die auch Belege der Thesen des Dynamizismus, des Embodiments oder der Situiertheit sind.

Das wichtigste *indirekte* Argument für EC könnte man als „Argument des biologischen Chauvinismus“ bezeichnen. Denn es erhebt sich die Frage an den Vertreter einer konservativen internalistischen Position, was denn das biologische Vehikel und seine Grenzen so besonders macht und auszeichnet. Würde, wer am traditionellen neuronalen Kernsystem als kognitivem Vehikel festhält, nicht trotzdem zugestehen, dass Teile des neuronalen Systems künstlich, etwa durch neuronale Implantate, ersetzbar sind? Und würde man derartige Implantate, falls sie funktional einwandfrei in das neuronale System integrierbar sind, nicht auch als systemzugehörig ansehen? Gibt man dies aber zu, so verliert die Grenzziehung zwischen neuronalen und künstlichen Vehikeln ihren kategorischen Charakter. In einem zweiten Schritt lässt sich dann folgern, dass sich künstliche Vehikel auch ebenso gut außerhalb der traditionellen neuronalen Systemgrenzen befinden können.

Beispiele kognitiver Erweiterungen führen zu einer Liste etwa der folgenden Art:

Sozial-kognitive Werkzeuge
  - Sprache
  - Gestik
  - Sprachgemeinschaften, Teamarbeit
  - Schwarm-Intelligenz
- Verkörperlichte kognitive Werkzeuge
  - Teile des Körpers oder gesamter Körper
  - Sensorische Erweiterungen
    - Brillen, Hörgeräte
    - Mikroskope, Teleskope
  - Motorische Erweiterungen
    - Spazierstock, Balancierstock
    - Sportgeräte
    - Prothesen
- Dynamische Kopplungen
  - Sensomotorische Rückkopplungsschleifen
- Kognitive Werkzeuge (konventionelle Techniken)
  - Bücher
  - Notizen, Taschenkalender
  - Karteikästen
- Kognitive Werkzeuge (elektronisch)
  - Computer-Notebooks
  - Pocket-Organizer, Mobiltelefone
  - Internet
  - Virtuelle Realität
- Kognitive Werkzeuge (Hybridtechniken, Enhancement)
  - Biotronik
  - Neurobionik
  - Bio-/Neuropharmakologie
  - Implantate

1.3 Kritik an der EC-These und ihre Verteidigung

Die Liste kognitiver Erweiterungen lädt zu einer nahe liegenden Kritik an der EC-These ein: Lässt sich überhaupt noch sinnvoll zwischen EC- und Nicht-EC-Komponenten trennen, oder haben wir es im Falle der EC-These mit einer inflationären und damit sich selbst trivialisierenden These zu tun? Es droht ein Dammbruch. Andererseits scheint aber klar zu sein, dass selbst das kognitive Ich eines passionierten Computerhackers nicht beliebig über das gesamte Internet verstreut ist. Wo also liegen die Grenzen?

Um einem Dammbruch zu begegnen, ist es wichtig, limitierende Kriterien zu benen-

nen, denen die externen EC-Komponenten genügen müssen. Hier ein Vorschlag in Anlehnung an Clark und Chalmers (1998):

1. *Zugänglichkeit:* Die externen Komponenten müssen jederzeit direkt und unmittelbar für das kognitive System zugänglich sein.
2. *Stabilität:* Die externen Komponenten müssen, bezogen auf die Zeitskala der Aufgabenstellung, stabil und robust im Zugriff sein.
3. *Zuverlässigkeit:* Die externen Komponenten müssen zuverlässig und valide belastbar sein.

Bei dieser Auflistung handelt es sich nicht um eine strenge und kategorische Liste, die Anwendung der verschiedenen Kriterien (und vielleicht noch weiterer, weniger zentraler Kriterien) erfolgt graduell und gestattet in Folge dessen auch lediglich eine unscharfe Individuation der externen EC-Komponenten. Es scheint eine Folge der EC-These zu sein, dass das erweiterte kognitive System keine scharfen Grenzen zur Umgebung besitzt.

Ein alternatives Individuationskriterium wird von Clark (2005) als *Paritätsprinzip* („parity principle") vorgeschlagen: Diejenigen Prozesse und Komponenten können als Erweiterungen angesehen werden, die, wären sie „im Kopf" verortet, unkontrovers dem System zugeschlagen würden. An diesem Prinzip und seinen superfunktionalistischen Implikationen haben vor allem Fred Adams und Kenneth Aizawa ihre Kritik festgemacht (Adams und Aizawa 2001, 2008; aber auch Rupert 2004). Ihr Hauptargument fassen die beiden Autoren unter dem Schlagwort ‚coupling-constitution fallacy'. Demnach unterliegen die Vertreter der EC-These einem Fehlschluss, da sie von den rein kausalen Kopplungen eines kognitiven Systems mit gewissen externen Komponenten auf deren konstitutive Rolle für das kognitive System schließen. Den Ursprung des Fehlschlusses verorten Adams und Aizawa darin, dass EC-Proponenten keine tragfähige Theorie des ‚mark of the cognitive' anzubieten haben, d.h. keine Theorie darüber, was es eigentlich heißt, dass bestimmte Prozesse kognitive Prozesse sind. Nur wenn man über eine solche Theorie verfügt, kann man sagen, was konstitutiv für kognitive Prozesse ist. EC-Vertreter, so der Einwand, können dies nicht.

Es ist allerdings fraglich, ob sich hiermit überhaupt ein irgendwie gearteter Einwand gegen die erweiterte Kognition verbindet. Denn die EC-These beansprucht ja nicht *aus sich heraus*, eine Definition dessen zu geben, was Kognition ihrem Wesen nach ist. Für die Anwendung des Paritätsprinzips geht man von einem bereits gegebenen kognitiven System aus und fragt danach, ob ein bestimmter kognitiver Zustand oder Prozess auch auf einem anderen als dem traditionell internalistischen Wege realisiert werden kann. Man bezieht sich also direkt auf die Möglichkeit der multiplen Realisierung kognitiver Zustände und Prozesse. Da das Charakteristikum der Multirealisierbarkeit eine Besonderheit funktional definierter Zustände ist, liegt es nahe, EC-Vertreter als Funktionalisten einzuordnen. Es scheint sogar, als ob das Konzept erweiterter Kognition nichts anderes als einen konsequent zu Ende gedachten Funktionalismus darstellt. EC ist eine Form von Super-Funktionalismus, Clark (2008) und

Wheeler (im Druck) sprechen neuerdings von "extended functionalism".

Die Frage bleibt, ob EC-Proponenten eine Theorie über das Wesen des Kognitiven („mark of the cognitive“ à la Adams und Aizawa) vorlegen müssen. Robert Rupert (2009) spricht vom Problem der Demarkation des Kognitiven. Das Paritätsprinzip leistet eine derartige Demarkation nicht, welche Alternativen bieten sich? Es wäre denkbar, dass Kognition eine natürliche Art bezeichnet. Diese Option scheint aber, auch nach Rupert, nicht sehr überzeugend. Kognitive Systeme sind durch eine Vielzahl von Merkmalen wie beispielsweise Sensorik, Motorik, Gedächtnis, Informationsverarbeitung etc. gekennzeichnet. Es müssen aber nicht immer sämtliche Merkmale zugleich zutreffen. Andererseits scheint jedes Merkmal nur für sich genommen zu einer zu weitherzigen Charakterisierung kognitiver Systeme zu führen. Die Klasse kognitiver Systeme scheint viel eher über Familienähnlichkeiten zusammenzuhängen. Es stellt sich die Frage, inwieweit dies ein spezielles Problem des Gegenstandsbereichs Kognition ist. Sind wir in anderen Bereichen etwa eher in der Lage, die Gegenstände in natürlicher Weise herauszugreifen? Gelingt es uns, ein „mark of the physical” oder ein “mark of the biological” anzugeben? Selbst wenn es sich als problematisch erweist, kognitive Systeme im Unterschied zu nicht-kognitiven Systemen zu individuieren, so reiht sich diese Problematik in die viel generellere Frage nach natürlichen Arten ein – keinesfalls aber handelt es sich um eine Fragestellung, die für erweiterte Kognition spezifisch ist.

Es existiert eine weitere nennenswerte Option zur Demarkation des Kognitiven, die sich bereits versteckt in der obigen Auflistung limitierender Kriterien finden lässt. Vor allem mit Blick auf die Praxis der theoretischen und modellbildenden Kognitions- und Neurowissenschaften scheint es plausibel zu sein, kognitive Systeme systemtheoretisch als Klassen von miteinander integriert zusammenhängenden Mechanismen zu demarkieren (vgl. auch die derzeitige Debatte über Mechanismen in der Biologie und den Neurowissenschaften; z.B. Bechtel 2008 und Craver 2007). Diese systemtheoretisch-mechanistische Option wird auch von Rupert (2009) favorisiert, er versucht aber zu zeigen, dass sie in Hinblick auf erweiterte kognitive Systeme empirisch unplausibel ist, da die geforderte Integration der Systemmechanismen seiner Meinung nach nicht gezeigt werden kann (demgegenüber ist er, wie schon in Rupert 2004, der Ansicht, dass verkörperlichte Kognition durchaus empirische Plausibilität besitzt). Die Vielzahl EC-freundlicher Beispiele etwa bei Clark (2003, 2008) macht aber deutlich, dass es sehr wohl zuhauf empirische plausible Instanzen für kognitive Erweiterungen auf verschiedensten Stufen gibt. Es zeigt sich einmal mehr, dass diese Diskussion letztlich nicht auf philosophische, sondern auf empirisch entscheidbare Fragen hinausläuft.

Kehren wir abschließend nochmals zur Frage einer Theorie des „mark of the cognitive“ zurück. Wir haben dafür argumentiert, dass EC-Proponenten auf diese Frage *qua EC-These* keine Antwort anbieten müssen. Darüber hinaus steht das Konzept erweiterter Kognition dem Funktionalismus nahe. Es zeigt sich, dass der tiefere Grund

zahlreicher Attacken auf EC darin zu suchen ist, dass EC-Opponenten häufig Anti-Funktionalisten sind und letztlich eine Charakterisierung des Kognitiven darüber zu geben versuchen, dass kognitive Zustände *gehaltvolle* Zustände sind (im Sinne semantisch-repräsentationalen oder qualitativen Gehalts oder beidem). Die Diskussion um das Wesen des Kognitiven wird dann zu einer Diskussion um das Wesen mentalen Gehalts – und hierbei ist in den meisten Fällen und ganz ausdrücklich und notorisch bei Jerry Fodor (2009) „originärer", d.h. nicht-abgeleiteter, intrinsischer mentaler Gehalt gemeint bzw. die intrinsischen repräsentationalen Kräfte „echt" kognitiver Systeme. Dann aber verlagert sich die Diskussion von der Betrachtung kognitiver Zustände als reiner Vehikel zu den sie tragenden mentalen Gehalten. Unter EC-Gesichtspunkten markiert dies den Übergang von „extended cognition" zu „extended mind", von Vehikel- zu Gehalts-Externalismus.

2. Semantischer Externalismus und naturalisierte Semantik

Mentale Zustände sind repräsentational, sie scheinen semantisch gehaltvolle Zustände zu sein. Im Rahmen naturalistischer Theorien des Geistes interessiert man sich für die Frage, unter welchen physischen Bedingungen mentale Repräsentationen gehaltvoll sind. Dabei ist es grob möglich, den bekanntesten Kandidaten naturalistischer Repräsentationstheorien eine je spezifische Variante eines Gehalts-Externalismus zuzuordnen. Externalistischen Positionen ist gemeinsam, dass zur vollständigen Fixierung und Individuation mentalen Gehalts system-externe Komponenten beitragen. Die verschiedenen Varianten des Externalismus unterscheiden sich dann je nach Art der relevanten externen Komponenten. Als eine Art Lackmus-Test auf die externe Natur von Gehalten können Zwillingserde-Gedankenexperimente angesehen werden. Hierbei betrachtet man mögliche Duplikat-Welten, in denen die internen Komponenten und Prozesse eines kognitiven Systems invariant gehalten werden, während lediglich die spezifischen externen Komponenten variieren. Für einen Externalisten ändern sich dabei die Gehalte weltenüberquerend, für einen Internalisten nicht.

2.1 Internalismus

Der Internalismus geht von der These aus, dass die Gehalte eines kognitiven Systems nur von dessen internen Zuständen und keinerlei externen Komponenten abhängen. Gehalt wird als intrinsische Eigenschaft eines kognitiven Systems konzipiert. Innerhalb einer sehr groben Klassifizierung lässt sich auch hier eine Bedeutungstheorie zuordnen, nämlich die klassische Ähnlichkeits- oder Abbildtheorie. Wenngleich der Internalismus auf diese Theorie nicht festgelegt ist, wurde und wird sie aber faktisch von vielen Internalisten vertreten. Der problematischste Punkt dabei ist nicht einmal die schon für sich bestehende Schwierigkeit, den Begriff der Ähnlichkeit zu präzisie-

ren. Nochmals problematischer ist es, angesichts der intrinsischen Natur des Gehalts plausibel zu machen, wie es zu der nach Putnam (1981) „magischen“ Bezugnahme mentaler Repräsentationen auf die äußere Welt kommen kann. Internalisten haben hierauf im Wesentlichen zwei Erklärungsvorschläge gegeben: Mentale Repräsentationen bzw. die ihnen subvenierenden internen kognitiven Vehikel haben aus sich heraus, intrinsisch, die Kraft, repräsentational zu sein (diese Kraft könnte etwa nur biologischem Substrat vorbehalten sein) oder, näher an der funktionalistischen Auffassung orientiert, eine „Funktionale Rollen-Semantik“ interner kognitiver Vehikel (hiergegen wendet sich John Searles Gedankenexperiment des Chinesischen Zimmers).

In Reaktion auf die Schwierigkeiten internalistischer Konzeptionen von Gehalt und als bekannteste externalistische Alternativen haben sich die kausale Theorie, die Teleosemantik und die Gebrauchstheorie hervorgetan. Ihnen lassen sich der physikalische, der historische und der soziale Externalismus zuordnen.

2.2 Kausale Theorie der Bedeutung

Im Rahmen einer Kausaltheorie der Bedeutung geht man von der Annahme aus, dass die Referenz durch eine Kausalkette bestimmt ist. Hieran anbindend hat Putnam (1975) erstmalig eine Form des Externalismus deutlich gemacht. Anhand eines „Twin earth"-Experiments, dem Ahnvater aller späteren Zwillingserde-Experimente, versucht er zu zeigen, dass Bedeutungen nicht (allein) *„im Kopf“* sind, sondern (auch) von der Natur der Referenzobjekte abhängen. Er betrachtet hierzu den physikalischen Doppelgänger eines Menschen auf einer gedachten Zwillingserde, die sich von unserer Erde lediglich dadurch unterscheidet, dass auf ihr die Natur von Wasser nicht H_2O, sondern XYZ ist. Der Kausaltheorie der Referenz zufolge muss nun den Wasser-Gedanken des Doppelgängers ein anderer Gehalt zugesprochen werden.

In der Vergangenheit sind vor allem zwei Varianten kausaler Theorien der Repräsentation hervorgetreten. Nach Jerry Fodors Theorie kausaler Kovarianz (Fodor 1987) gilt vereinfacht:

> *Eine Repräsentation R hat dann und nur dann den Gehalt „p”, falls R durch ein Vorkommnis von p verursacht wurde.*

Der auf ein vergleichbares Ergebnis hinauslaufende informationale Ansatz von Fred Dretske (1981) besagt in Kürze:

> *Ein Signal R trägt die Information „p”, wenn p eine notwendige Bedingung für das Auftreten von R ist.*

Kausale Theorien - wenigstens in ihrer vereinfachten, kruden Form - binden dem-

nach den Gehalt einer Repräsentation starr an entsprechende ursächliche Vorkommnisse. Dies führt jedoch zu beträchtlichen Problemen. Zunächst lässt sich einwenden, dass einfache kausale Kovarianz zu unspezifisch ist, um Gehalt zu fixieren, da nicht jegliche kausale Kovariation semantisch gehaltvoll scheint. Das gravierendere Problem kausaler Theorien besteht jedoch nach allgemeiner Überzeugung in der mangelnden Erklärbarkeit von Fehlrepräsentationen, denn nur ein Repräsentationsbegriff, der auch das im Alltag ja hinlänglich bekannte Versagen von Repräsentationsleistungen erfasst, wäre begrifflich zufriedenstellend. Wegen der strengen kausalen Bindung von Repräsentat und Repräsentandum in der (kruden) Kausaltheorie kann es dort jedoch per definitionem zu keinerlei Fehlrepräsentation kommen. Dem klassischen Beispiel zufolge können nur Kühe Kuh-Repräsentationen hervorrufen. Dann aber bleibt unerklärlich, dass auch gelegentlich Pferde fehlerhaft Kuh-Repräsentation hervorrufen. Allenfalls könnte gelten, dass Kuh-Repräsentationen sich auf die disjunktive Klasse >Kuh oder Pferd< beziehen. Dann aber wäre der Gehalt von „Kuh" nicht spezifisch bestimmt – es ergibt sich ein Disjunktionsproblem. In der Literatur besteht weitgehend Übereinkunft, dass die Fehlrepräsentations-Problematik eine ernsthafte begriffliche Schwierigkeit für kausale Theorien darstellt und im Rahmen rein kausaler Theorien nicht befriedigend lösbar ist.

Fodor hat auf diese Einwände reagiert, indem er versucht hat, die krude symmetrische Kausalbedingung durch eine asymmetrische, kontrafaktische Bedingung zu ersetzen: Falls Kühe keine Kuh-Repräsentationen hervorrufen, rufen auch Pferde keine Kuh-Repräsentationen hervor, es ist aber möglich, dass Pferde keine Kuh-Repräsentationen hervorrufen, jedoch Kühe dies trotzdem tun. Pferde-Repräsentationen sind daher parasitär gegenüber Kuh-Repräsentationen. Wie aber lässt sich eine derartige asymmetrische Beziehung zwischen Repräsentation und Repräsentat naturalisieren? Es scheint klar zu sein, dass dies durch Rekurs auf rein kausale Beziehungen allein nicht gelingen kann, die kontrafaktische Bedingung wurde ja gerade zur Verfeinerung der Schwächen kruder Kausalbedingungen eingeführt. Die von diesem Befund aus nahe liegendste naturalistische Konzeption scheint zu sein, auf die Lern- oder Adaptationsgeschichte zur Hervorbringung von Repräsentationen zu rekurrieren. Verfolgt man diesen Weg konsequent, so geht die Kausaltheorie letztlich in eine teleosemantische Theorie über.

2.3 Teleosemantik

Die Teleo- oder Biosemantik bildet die zweite große Klasse von Naturalisierungsprogrammen der Semantik. Die Grundidee dieses vor allem auf Arbeiten von Ruth Millikan (1984) und Dretske (1995) zurückgehenden Ansatzes besteht vereinfacht in Folgendem:

> *Eine Repräsentation R besitzt den Gehalt „p", wenn R die Funktion hat, das Vorliegen von p anzuzeigen.*

Die Grundidee ist, den Repräsentationsbegriff zunächst auf den Funktionsbegriff zurückzuführen. Denn die Anzeigefunktion einer Repräsentation lässt sich, so die entscheidende Annahme der Teleosemantik, insofern einwandfrei naturalisieren, als sie durch die jeweilige evolutionäre Selektions-, Adaptations- oder auch Lerngeschichte eines kognitiven Systems eindeutig festgelegt ist. Offenkundig führt auch dies auf eine Form von Gehalts-Externalismus, diesmal im diachronen Sinne eines historischen Externalismus: es ist die adaptive Historie des Systems, die dessen Gehalte individuiert.

Doch auch die Teleosemantik ist nicht frei von Problemen, die in erster Linie mit der Bestimmbarkeit von Funktionen zusammenhängen. So reden wir im Rahmen biologischen Vokabulars nahezu immer von Funktionen, ohne dass uns die genaue Selektions- oder Lerngeschichte bekannt wäre. Hieran hängt eine bekannte Schwachstelle des Adaptationismus. Adaptationistisches Denken ist in der Praxis häufig auf geschicktes, rein auf Plausibilität abzielendes „Geschichtenerzählen" angewiesen. In nahezu allen Fällen sind evolutionäre Erklärungen Schlüsse auf die beste Erklärung. Durch die Unterbestimmtheit der Adaptationsgeschichte bleibt auch die durch sie individuierte Funktion unterbestimmt.

Man kann dies als rein epistemischen Einwand abtun. In einem systemischen Ganzen lassen sich allenfalls proximate, mit der direkten kausalen Rolle einer Systemkomponente verbundene Funktionen bestimmen. Die dahinter stehenden ultimaten Funktionen der evolutionären Historie sind uns häufig epistemisch versperrt oder doch wenigstens nicht vollumfänglich zugänglich, können aber dennoch, so die Teleosemantik, ontologisch verteidigt werden, insofern jedes evolutionäre Vorkommnis eine eindeutige Herkunftsgeschichte, ob bekannt oder unbekannt, besitzt.

In ontologischer Hinsicht wären ultimate Funktionen dann bedroht, wenn die evolutionäre Geschichte zur Individuation ultimater Funktionen nicht hinreichend ist. Dies ist aber sehr wahrscheinlich genau der Fall. Ist etwa der Fangmechanismus eines Frosches selektiv auf Fliegen oder schwarze Schatten? Insofern er zum Fliegenfang evolutionär hervorgebracht wurde, ist ihm nach teleosemantischer Auffassung die Fliegenselektivität als „eigentliche Funktion" (proper function) zuzusprechen. Demnach müsste beispielsweise das Schnappen eines Frosches nach künstlichen Pillen, die seine Gesundheit verbessern, streng genommen als dysfunktional angesehen werden. Man darf in diesem Fall höchstens von einer abgeleiteten, nicht ursprünglichen und lediglich durch Interpretation zugesprochenen Funktion reden. Auch wenn dies üblichen Intuitionen und Sprechweisen eher zuwiderlaufen dürfte, wird der Teleosemantiker dennoch an seiner Behauptung festhalten, dass die „eigentliche" ultimate Funktion in diesem Szenario tatsächlich verfehlt wird. Allerdings muss auch im Rahmen der Teleosemantik die Möglichkeit bestehen, dass ultimate Funktionen sich mit der Zeit verschieben, insofern sich äußere Selektionszwecke ändern können. Nach einer Vielzahl von Froschgenerationen, die sich in einer Umwelt mit

fliegenden Gesundheitspillen ihren Schnappmechanismus bewahrt haben, ist eben das Schnappen nach derartigen Pillen zur eigentlichen Funktion geworden. Aber nach wie vielen Generationen? Und was wäre, wenn wir uns eine mögliche Welt denken, in der wahlweise Fliegen und Gesundheitspillen herumfliegen? Besteht die Funktion dann in der Selektivität auf die disjunktive Klasse von Fliegen und Pillen? In der Vagheit und Unbestimmtheit der teleosemantischen Antwort liegt die Unbestimmtheit oder mangelnde Feinkörnigkeit der ultimaten Teleofunktion begründet. Man muss bezweifeln, dass evolutionäre Historien Funktionen eindeutig auszeichnen, ja vielleicht sogar, dass sie überhaupt Funktionen auszeichnen.

Eine Variation des Beispiels zeigt eine weitere Facette des Problems, denn unser Beispiel lässt sich ja auch so lesen, dass eben ein und derselbe Kausalmechanismus in Lebewesen auf unterschiedlichen Selektionswegen herbeigeführt werden kann. In unserem Beispiel der Schnappmechanismus für Fliegen- und Pillenfang-Frösche. Gegenüber unserem konstruierten Beispiel lassen sich für diesen Umstand in der Tat auch zahlreiche biologisch plausible Beispiele angeben (so wurde etwa das Auge bzw. der Sehmechanismus auf zahllosen verschiedenen Wegen in unterschiedlichsten Lebensformen durch die Evolution „entdeckt"). Doch trotz kausal-funktionaler Äquivalenz ist die Teleofunktion nicht dieselbe, insofern die externe historische Komponente der Funktionsindividuation nicht dieselbe ist. Dies ist reichlich kontraintuitiv.

Im Rahmen von Zwillingserde-Szenarien lassen sich diese kontraintuitiven Konsequenzen des historischen Externalismus nochmals verdeutlichen. Ein besonders drastisches de facto-Zwillingserde-Szenario wurde erstmals von Davidson im Rahmen seiner bekannten Swampman-Überlegung angestellt (vgl. Detel 2001). Der Swampman steht hier für einen internalistisch invarianten, gegenüber seinem Original jedoch auf der Zwillingserde spontan kreierten Doppelgänger, dem der Teleosemantiker aufgrund seiner fehlenden evolutionären oder sonstigen Geschichte keinerlei Gehalte zusprechen kann. Der Swampman erweist sich, wenigstens im Moment seiner Erschaffung, als radikal gehaltloser Zombie. Und dies ist, da zugestanden wird, dass ein solcher Doppelgänger keinerlei Verhaltensdifferenzen gegenüber seinem Original zeigt, eine speziell für Naturalisten schwerlich akzeptable Konsequenz.

Man sieht deutlich, dass die Teleofunktion bzw. die ihr unterliegende evolutionäre Historie als gehaltsindividuierende externe Komponente nicht das beobachtbare Verhalten betrifft. Von außen lässt sich der Swampman sehr wohl als ein kognitives System auffassen, das dieselben funktionalen Rollen erfüllt wie das Original. Er darf insofern als „funktional äquivalent" bezeichnet werden. Hier liegt jedoch eine gefährliche Äquivokation im Begriff der Funktion vor, die leicht zu Missverständnissen und begrifflicher Verwirrung führen kann. Denn diese Art funktionaler Äquivalenz ist genauer eine Äquivalenz bezüglich kausaler Struktur und Mechanismen. Die entscheidende Frage ist, ob es sich um bloß zugeschriebene, abgeleitete oder originäre, eigentliche Funktionalität handelt. Man kann auch sagen, dass die zugeschriebene

Funktionalität nur die proximaten Funktionen betrifft. Demgegenüber ist die Teleofunktion eine genuine, nicht bloß zugeschriebene, ultimate Funktion aufgrund der evolutionären Historie.

2.4 Gebrauchstheorie der Bedeutung

Die Gebrauchstheorie zählt nicht per se zu den Naturalisierungsprogrammen von Semantik, lässt aber eine naturalistische Flanke und speziell eine Schnittstelle zur Neuen KI offen. *„Man kann für eine große Klasse von Fällen der Benützung des Wortes „Bedeutung" – wenn auch nicht für alle Fälle seiner Benützung – dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache"*, so Wittgenstein in einer bekannten Passage aus den Philosophischen Untersuchungen (Wittgenstein 1953, § 43). Dem zweiten Teil des Satzes lässt sich die rohe Charakterisierung einer Gebrauchstheorie mentaler Gehalte entnehmen:

> *Der Gehalt einer mentalen Repräsentation entspricht ihrem Gebrauch in einer Sprachgemeinschaft.*

Im Unterschied zur Kausaltheorie und Teleosemantik wird der Gehalt nun nicht länger referentiell mittels p analysiert. Konzentrieren wir uns wieder auf den externalistischen Aspekt: Die Gebrauchstheorie führt auf einen sozialen Externalismus, denn die Gehalte hängen in systematischer Weise von den Gebrauchsweisen in der Sprachgemeinschaft ab, sie supervenieren über (Teilen) der Gemeinschaft und sind insofern holistisch, nicht individualistisch fixiert.

Eine entsprechende Zwillingserde-Überlegung stammt von Tyler Burge (1979). Aufgrund seiner Schmerzen im Oberschenkel ist Otto der Überzeugung, Arthritis zu haben. Dies ist aber falsch, da nur bei Gelenkerkrankungen von Arthritis gesprochen wird. Auf der Zwillingserde werden auch Knochenentzündungen Arthritis genannt. Zwottos Überzeugung, Arthritis zu haben, ist daher korrekt. Die Gebrauchsbedeutung von Arthritis auf der Zwillingserde (zur Vermeidung von Konfusionen in der Erdsprache besser mit einem neuen Wort, Zwarthritis, belegt) ist also different von Arthritis auf der Erde.

Durch den Sprachgebrauch legt die Sprachgemeinschaft Bedeutungsnormen fest. Durch diese „eingebaute" Normativität gelingt es in der Gebrauchstheorie grundsätzlich, das Problem der Fehlrepräsentation gar nicht erst aufkommen zu lassen. Fehlrepräsentationen lassen sich als Abweichungen praktizierter Normen auffassen. Eine streng naturalisierte Gebrauchstheorie darf derartige Gebrauchsnormen jedoch nicht als genuin normativ, sondern lediglich als Zuschreibungen ansehen. Insofern unser Auftakt: die Gebrauchstheorie zählt nicht per se zu den Naturalisierungsprogrammen, lässt aber eine naturalistische Flanke offen.

3. Der Aktive Externalismus und seine Konsequenzen

Dem physikalischen, historischen und bis auf Weiteres (s. Abschnitt 4) auch dem sozialen Externalismus ist gemeinsam, dass die gehaltsfixierenden externen Komponenten aufgrund ihres distalen Charakters für das beobachtbare Verhalten des kognitiven Systems keine Rolle spielen. Umgekehrt besitzt das kognitive System keine Einflussmöglichkeit auf die jeweiligen externen Komponenten. Das kognitive System steht diesen Komponenten *passiv* gegenüber, die drei externalistischen Varianten können daher treffend als *passive Externalismen* charakterisiert werden. Gleichwohl geht von den jeweiligen distalen externen Komponenten einen gehaltsdeterminierender „Einfluss" aus. Im Unterschied dazu führt EC auf einen *aktiven Externalismus*, dessen Implikationen nun genauer analysiert werden sollen.

Eine Zusatzbemerkung: Unser Augenmerk soll ausschließlich auf repräsentationalem Gehalt liegen, die Externalismusdebatte lässt sich aber auch auf qualitative Erlebnisgehalte ausweiten. Überlegungen im Umfeld eines aktiven Externalismus qualitativer Gehalte werden insbesondere von Hurley (1998) und Noë (2004, 2009) unter dem Schlagwort Enaktivismus angestellt.

3.1 Aktiver Externalismus

Für das Folgende wollen wir die EC-These als zutreffend voraussetzen. Unter der gängigen Annahme, dass Gehalte über ihren physischen Vehikeln supervenieren, folgt aus EC eine neue Variante eines Externalismus, die systemexterne Komponenten wie dynamische Kopplungen mit der Umgebung, kognitive Werkzeuge, Kommunikationsgemeinschaften, künstliche neuronale Implantate oder sonstige Vehikel-Erweiterungen betrachtet. Die solcherart durch EC induzierten externen Komponenten sind jedoch nicht distaler, sondern unmittelbarer und proximater Natur. Infolgedessen erweist sich der durch EC induzierte Externalismus als aktiver Externalismus.

> *These des aktiven Externalismus (AE):* Mentale Gehalte hängen nicht nur von den internen Zuständen eines kognitiven Systems ab, sondern auch von extensionalen Komponenten – und zwar so, dass eine Variation der gehaltsfixierenden externen Komponenten prinzipielle Verhaltensrelevanz besitzt.

Im Gegensatz zur EC-These, die im Wesentlichen empirischen Charakter hat, ist AE eine dezidiert philosophische These. Während EC unabhängig von AE vertreten werden kann, ist AE hier unter Bezugnahme auf EC konzipiert, da die gängige Annahme der Supervenienz des Mentalen über dem Physischen es nahe legt, den aktiven Externalismus als eine direkte Folge der erweiterten Kognition anzusehen.

Ebenso wie die EC-These wurde auch die AE-These durch den Aufsatz von Clark und Chalmers (1998) prominent gemacht, allerdings präsentieren die Autoren beide Thesen in vermengter Form, genauer, sie führen zwar die zwei unterschiedlichen Begriffe „extended mind“ und „active externalism“ ein, sagen aber nicht, ob und inwiefern zwischen beiden Schlagwörtern unterschieden werden soll. Es ist das Verdienst von Susan Hurley (1998a, b), auf die begriffliche Trennung zweier Thesen hingewiesen zu haben, die sie als Vehikel- und Gehalts-Externalismus bezeichnet. Dieser Trennung folgend wollen wir statt von „extended mind“ präziser von „extended cognition“ als der EC-These im Sinne der Vehikel sprechen, die These des aktiven Externalismus (AE) aber terminologisch streng als eine philosophische These über Gehalte ansehen.

Inwiefern führt nun EC auf einen Gehalts-Externalismus? Clark und Chalmers (1998) haben hierzu folgerichtig ein Zwillingerde-Szenario betrachtet: Otto leidet an Alzheimer, kann aber seinen Alltag meistern, indem er die für ihn wichtigen Daten, Hinweise, Termine etc., an die er sich nicht mehr erinnern kann, in ein Notizbuch schreibt, das er ständig bei sich führt und auf das er in Folge dessen auch ständig zurückgreift. Nun möchte Inga sich mit ihm am Museum of Modern Art treffen, das sich in der 53. Straße befindet. Da Otto sich die Adresse des MOMA nicht merken kann, konsultiert er kurz vor dem Treffen sein Notizbuch und kann sich dann glücklich mit Inga treffen.

Auch auf der Zwillingserde will sich Zwotto mit Inga (identisch mit Zwinga) treffen. In Zwottos Notizbuch hat sich jedoch ein Fehler eingeschlichen: als Adresse des MOMA ist dort die 51. Straße notiert. In Folge dessen scheitert die Verabredung von Inga und Zwotto. Der veränderte Notizbucheintrag ist verhaltensrelevant, hierin zeigt sich die aktive Natur des AE-Gehaltsexternalismus. Man kann geltend machen, dass dadurch wesentliche naturalistische Intuitionen eingefangen werden – bzw. umgekehrt, dass die mit dem passiven Externalismus verknüpften kontraintuitiven Folgen vermieden werden. AE steht daher auch mit mentaler Verursachung in größerem Einklang (wie weiter unten ausgeführt).

Zur Plausibilität von AE ist vor allem das in der AE-These enthaltene „Master-Argument“ heranzuziehen: Falls EC richtig ist, dann folgt AE unter der (üblichen) Annahme der Gehalt-Vehikel-Supervenienz. Die Plausibilität von AE speist sich insofern aus EC. Dies bedeutet auch, dass sich inhaltlich ihrer Natur nach sehr verschiedene externe Komponenten für AE heranziehen lassen. Nicht nur Notizbücher als konventionelle von Menschen verwendete Hilfsmittel, sondern kognitive Werkzeuge jeglicher Art (auch elektronischer und biotronischer Natur) sowie dynamische Rückkopplungsschleifen unter Einbeziehung von Teilen der Umgebung (das Motiv der Situiertheit) sind hier denkbar. Bei Prismenbrillen-Experimenten beispielsweise superveniert der Wahrnehmungsgehalt über der gesamten sensomotorischen Rückkopplungsschleife, also über den Zuständen der Brille, den aktiv bewegten motorischen Effektoren (im Sinne des Reafferenzprinzips) und all denjenigen äußeren Sti-

mulusanteilen, die zur Herbeiführung und Stabilisierung des neuen Wahrnehmungsbildes nötig sind.

3.2 Aktiver Externalismus und multiple Realisierung

Allem Anschein nach sind mentale Gehalte multipel realisierbar. Multiple Realisierbarkeit impliziert, dass mentale Typen und physische Typen nicht identisch sind, sondern dass mentale Vorkommnisse durch vielerlei, gegebenenfalls stark heterogene physikalische Vorkommnisse instantiierbar sind. Ein wenig beachteter Umstand ist, dass die These der multiplen Realisierbarkeit nicht in jedem Fall die externen gehaltsfixierenden Komponenten des passiven Externalismus betrifft. Der Gehalt meiner „Wasser"-Gedanken wird durch die wahre Essenz von Wasser, etwa seine Natur als H_2O, fixiert. Aber natürlich ist die Essenz eines Dings ihrerseits nicht multipel realisierbar. Sofern die kausale Referenztheorie Recht hat, sind lediglich die systeminternen, engen Gehalte multipel realisierbar. In diesem Fall ließe sich multiple Realisierbarkeit geradezu als ein Kriterium zur Individuation enger Gehalte auffassen.

Auch die Teleosemantik gestattet keine Multirealisierbarkeit der externen Komponente. Dies mag zunächst überraschen. Ist es nicht denkbar, dass ein und derselbe Gehalt mittels verschiedener adaptiver Geschichten hervorgerufen wird? In der Tat ist dies sogar zu erwarten: Meine Lerngeschichte der Bedeutung eines gewöhnlichen deutschen Satzes wie „Das Haus ist groß" ist gewiss different von derjenigen anderer Sprecher, und dennoch darf man vermuten, dass für eine Mehrzahl der kompetenten Sprecher des Deutschen der Satz buchstäblich dasselbe bedeutet. Also scheint es doch, als ob die externe historische Komponente multipel realisierbar ist. Dies ist aber ein Missverständnis. Für den Teleosemantiker liegen die Dinge anders. Für ihn können die durch unterschiedliche Historien herbeigeführten Gehalte niemals identisch sein. Hierzu muss man sich die in Abschnitt 2.3 hervorgehobene Unterscheidung von proximaten und ultimaten Funktionen in Erinnerung rufen. Wie wir dort gesehen haben, ist die Hervorbringung funktional äquivalenter Mechanismen – also äquivalent hinsichtlich der proximaten Funktion – kein Garant für die Gleichheit der ultimaten und distalen Teleofunktion. Im Gegenteil: der Unterschied in der externen historischen Teleokomponente bedingt einen Unterschied im Gehalt, da Gehalte erst aufgrund ihrer Teleogeschichte individuiert werden. Die Verhaltensäquivalenz oder äquivalente Reaktion verschiedener Sprecher bezüglich gewöhnlicher deutscher Sätze wie „Das Haus ist groß" besagt nicht, dass der Teleogehalt ihrer entsprechenden mentalen Repräsentationen derselbe ist. Teleogehalte sind gerade nicht multipel realisierbar. Oder bezüglich Funktionalität gesprochen: Multirealisierbarkeit bezieht sich lediglich auf proximate, nicht auf ultimate Funktionen.

In der Gebrauchstheorie liegen die Dinge anders, wenigstens dann, wenn wir eine rein behavioral verstandene Theorie betrachten. Mentale Gehalte sind hier einzig durch Verhalten und Gebrauch der Sprecher individuiert, Gehalte supervenieren

über den Gebrauchsweisen von Wörtern in Sprechergemeinschaften. Gebrauchsweisen sind aber durchaus multipel realisierbar. Es ist einwandfrei denkbar, dass verschiedene Sprechergemeinschaften zu vergleichbaren oder genau den gleichen Gebrauchsweisen tendieren.

3.3 Aktiver Externalismus und mentale Verursachung

Im Rahmen unserer Alltagspsychologie schreiben wir uns selbst die Fähigkeit zu, autonom und auf der Basis rationaler Überlegungen entscheiden und handeln zu können. Kognitive Systeme einer gewissen Komplexitätsstufe sind zu kognitiver Agentenschaft fähig. Dabei ist es offenkundig der semantisch gehaltvolle Charakter unserer mentalen Zustände, der es macht, dass wir handelnd in die Welt eingreifen. Diese kausale Wirksamkeit unseres geistigen Innenlebens wird als mentale Verursachung bezeichnet. Mentale Verursachung und passiver Externalismus stehen vordergründig in einem seltsamen Spannungsverhältnis, denn wie können mentale Gehalte kausal wirksam sein, falls diese Gehalte von *distalen* externen Faktoren abhängen? Der springende Punkt ist, dass der Externalismus mit der gängigen Annahme der *lokalen* psychophysischen Supervenienz unverträglich zu sein scheint, da Verursachung gewöhnlich als ein lokales Phänomen angesehen wird. Kausale Wirkungen sollten daher nicht von raumzeitlich getrennten Entitäten hervorgerufen werden wie im Falle distaler externer Faktoren.

Man bringt gelegentlich dies auch dadurch zum Ausdruck, dass man den passiven Externalismus als Anti-Individualismus charakterisiert: mentale Gehalte sind nicht agenten-intrinsisch und lokal, sondern relational bestimmt. Mentale Verursachung scheint jedoch einen Individualismus vorauszusetzen, insofern kognitive Agenten aufgrund der „individuellen" internen Natur ihrer mentalen Zustände handeln. Mit „individuell" ist dabei für gewöhnlich zweierlei zugleich gemeint: intrinsisch und lokal. Für die Spannung zwischen Externalismus und mentaler Verursachung ist es aber weniger entscheidend, ob Gehalte intrinsisch sind, sondern ob Gehalt lokal superveniert. Und dies scheint für den Externalismus nicht erfüllt zu sein.

Der auf Fodor (1987) zurückgehende Standardvorschlag zur Auflösung dieser Spannung besteht darin, zwischen weitem und engem Gehalt zu unterscheiden. Lediglich enger Gehalt ist verhaltensrelevant und superveniert über den lokalen internen Zuständen eines kognitiven Agenten. Nimmt man externe Komponenten in die Supervenienzbasis auf, so erhält man weiten Gehalt, der jedoch nicht verhaltensrelevant ist. Analog haben wir von einem passiven Externalismus gesprochen. Man kann die Passivität der Komponenten im Sinne eines *genitivus subjectivus* oder *genitivus objectivus* verstehen: die externen Komponenten sind nicht verhaltensrelevant und eben darum passiv, umgekehrt stehen kognitive Systeme der Zuweisung weiten Gehalts passiv gegenüber, denn die über den engen Gehalt hinausgehende Gehaltszuweisung hängt ja nicht von den internen system-intrinsischen Zuständen ab.

Bemerkenswert ist, dass die beschriebene Spannung nur den passiven, nicht aber den aktiven Externalismus betrifft. Denn die unmittelbaren und proximaten externen Komponenten des AE führen durchaus zu Verhaltensänderungen und unterminieren somit per se nicht die Frage der Wirksamkeit mentaler Gehalte. Die von vielen ohnehin als unplausibel angesehene Unterscheidung von weitem und engem Gehalt ist dann nicht erforderlich. AE steht insofern in Einklang mit mentaler Verursachung als einem der vermeintlichen Grundcharakteristika des Mentalen. Die Konsequenz ist allerdings, dass es das im Sinne von EC erweiterte und vergrößerte System ist, dem wir Agentenschaft zubilligen müssen. Die entscheidende Besonderheit ist zudem die je aufgabenspezifische Variabilität des Systems. Wir haben es unter verschiedenen Umständen mit verschiedenen kognitiven Erweiterungen zu tun. Es ist absehbar, dass eine derartige Konzeption erhebliche Folgen für die Fragen nach Personen- und Urheberschaft nach sich zieht, die wir hier jedoch aus Platzmangel nicht ansprechen können.

4 Vom passiven zum aktiven Externalismus

Die Gebrauchstheorie wird üblicherweise als Form des passiven Externalismus rekonstruiert. Eine einfache Überlegung zeigt aber, dass dies in Strenge nicht zutrifft: die Gebrauchstheorie gestattet in der Tat einen graduellen Übergang zum aktiven Externalismus. Nach Voraussetzung supervenieren Gebrauchsbedeutungen auf der gesamten Sprechergemeinschaft. Da ein einzelner Sprecher wie Otto dem faktischen Gebrauch eines Wortes in einer großen Sprachgemeinschaft quasi passiv gegenübersteht, handelt es sich beim sozialen Externalismus um einen passiven Externalismus – für nahezu alle praktischen Belange. Es genügt aber, die Sprechergemeinschaft im Gedankenexperiment hinreichend klein zu machen, um zu sehen, dass Ottos Gebrauch des Wortes Arthritis durchaus auch zu einer Bedeutungsverschiebung innerhalb der Gemeinschaft führen kann. Wenn sozial einflussreiche Teile einer Gemeinschaft neue Gebrauchsweisen von Wörtern etablieren, so sind dies fortan die neuen Bedeutungen. Als aktiver Teil der Sprechergemeinschaft besitzt Otto daher durchaus eine aktive, wenn auch für große Sprachgemeinschaften faktisch nur marginale Einflussnahme auf den gemeinschaftlichen Sprachgebrauch. Die Unterscheidung von passivem und aktivem Externalismus ist letztlich graduell.

Dies steht in Einklang mit der in 3.2 gewonnenen Einsicht, dass die externen Komponenten der Gebrauchstheorie multipel realisierbar sind, nicht aber die externen Komponenten der Kausaltheorie und Teleosemantik. Der soziale Externalismus zeigt auch in dieser Hinsicht seine Affinität zur AE-These, denn auch die externen AE-Komponenten sind multirealisierbar. Die Erweiterung meiner mathematischen Denkleistungen lässt sich auf vielfache, funktional äquivalente (im Sinne kausaler Rollen) Arten und Weisen erreichen, etwa durch einen Rechenschieber, Abakus oder Taschenrechner. Und Ottos Notizbuch kann ohne jede performative Einbuße durch

eine Kopie ersetzt werden. Die prinzipielle Multirealisierbarkeit von AE-Komponenten blockiert die Möglichkeit der Erweiterung in die Kausaltheorie, steht jedoch in Einklang mit der prinzipiellen Erweiterbarkeit in die Gebrauchstheorie.

Auf die weiteren, offensichtlichen Anknüpfungsmöglichkeiten einer im Sinne eines aktiven Externalismus verstandenen Gebrauchstheorie sei hier abschließend nur kurz hingewiesen. In 1.2 wurde als vierte Komponente die kognitive Erstreckung in soziale Gemeinschaften, also im Sinne sozialer Kognition genannt. Dies bindet an jüngere Überlegungen zur Schwarmintelligenz, kollektiven Intelligenz sozialer Gruppen und zur „Shared Intentionality"-These an. Von herausragender Bedeutung sind etwa die Arbeiten und Experimente zur Roboter-Kommunikation und zum Ursprung der Sprache durch Luc Steels und seine Mitarbeiter (vgl. Steels 2003). Diese Forschungen sollen zeigen, dass die Gemeinschaft der Sprecher- oder Kommunikationsagenten als ein komplexes adaptives System aufzufassen ist, dass das Problem der Herausbildung eines Kommunikationssystems in kollektiver Weise löst. Die Gemeinschaft erreicht dabei schließlich Einigkeit über ein rudimentäres Vokabular und eine entsprechende Syntax, der bekannte Streit zwischen Symbolisten und Konnektionisten zum Verständnis und zur Herkunft einer vermeintlichen Ur-Syntax könnte so eine empirische Auflösung im Rahmen evolutionärer Linguistik finden.

Inhaltlich benachbart sind auch die Arbeiten von Michael Tomasello (2008) aus dem Bereich der Primatenkognition. Hier geht es sehr wesentlich darum zu zeigen, dass die Fähigkeit des Menschen zur Intentionalität zweiter Stufe, nämlich der Erkennung eines intentionalen Gegenübers und der versuchten Deutung der entsprechenden Absichten („Mindreading"), zu den charakteristischen Fähigkeiten menschlicher Kognition im Sinne eines Alleinstellungsmerkmals gehört (vgl. auch Goldman 2006). Die auch in Primaten auffindbaren Spiegelneuronsysteme werden dabei bekanntlich als wichtiger funktionaler Baustein dieser Fähigkeiten angesehen. Geteilte Intentionalität ließe sich somit als eigene Variante erweiterter Kognition verstehen, denn ihr zufolge erstreckt sich das Mentale nicht nur in die Welt, es kann sich auch in andere kognitive Systeme erstrecken, so dass es zu mentalen Symbiosen kommt.

Ferner diskutiert bereits Hutchins (1995) in detaillierter Form, wie auf einem US-Kriegsschiff Entscheidungen gefällt werden, wobei er zeigt, dass die Mannschaft bzw. Mannschaftsteile in vielen Prozessen als Ganze agieren, eine Form von „distributed cognition" unter den Besatzungsmitgliedern. Hier deutet sich eine Anwendung der Ideen erweiterter Kognition auf nächsthöherer Stufe an: kognitive Systeme können bei entsprechender Kopplung kognitive Super-Systeme ausbilden, von denen sie selber nur einen Teil darstellen. Offensichtlich stellt dies eine extreme und äußerst radikale Anwendung der EC-These dar, deren Plausibilität hier auch nicht weiter verfolgt werden soll.

Für die Zwecke der vorliegenden Diskussion ist es allein wichtig festzuhalten, dass es sich bei allen angedeuteten Möglichkeiten sozialer Kognition zugleich um Erwei-

terungsstufen und Spielarten eines aktiven Externalismus handelt, so dass die Debatte um diese neuartige Form des Externalismus, gemessen an den konzeptionellen Innovationen an der derzeitigen empirischen Forschungsfront der sozialen kognitiven Neurowissenschaften erst am Anfang steht. Es war das Ziel dieses Aufsatzes, zu einer ersten begrifflichen Klärung der Grundlagen dieser Entwicklungen und ihrer Rückbindung an schon bekannte philosophische Diskussionen um Umfeld des mentalen Externalismus beizutragen.

Literatur

Adams, Frederick, und Kenneth Aizawa (2001): The Bounds of Cognition, in: *Philosophical Psychology* 14: 43-64.

Adams, Frederick, und Kenneth Aizawa (2008): *The Bounds of Cognition*. Oxford: Blackwell.

Bechtel, William (2008): *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bechtel, William, und Adele Abrahamsen (1991, 2nd 2002): *Connectionism and the Mind*. Oxford: Blackwell.

Block, Ned (1978): Troubles with functionalism. In C. W. Savage, ed., Minnesota Studies in the Philosophy of Science IX. Minneapolis: University of Minnesota Press.

Brooks, Rodney (1991): Intelligence Without Representation, in: *Artificial Intelligence* 47: 139-159.

Burge, Tyler (1979): Individualism and the Mental, in: *Midwest Studies in Philosophy* 4:73-121.

Clark, Andy (1997): *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA.

Clark, Andy (2003): *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, Oxford.

Clark, Andy (2005): Intrinsic Content, Active Memory and the Extended Mind, in: *Analysis* 65(285): 1-11.

Clark, Andy (2008): *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, New York.

Clark, Andy und David Chalmers (1998): The Extended Mind. Analysis 58(1): 7-19.

Craver, Carl (2007): *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Detel, Wolfgang (2001): Haben Frösche und Sumpfmenschen Gedanken? Einige Probleme der Teleosemantik, in: *Deutsche Zeitschrift für Philosophie* 49:601-626.

Dretske, Fred (1981): *Knowledge and the Flow of Information*. MIT Press, Cambridge, MA.

Dretske, Fred (1995): *Naturalizing the Mind*. MIT Press, Cambridge, MA.

Eliasmith, Chris und Charles H. Anderson (2002): *Neural Engineering. Computation, Representation, and Dynamics in Neurobiological Systems*. MIT Press, Cambridge, MA.

Fodor, Jerry A. (1987): *Psychosemantics*. MIT Press, Cambridge, MA.

Fodor, Jerry A. (2009): Where is My Mind?, in: *London Review of Books* 31 (3).

Gallagher, Shaun (2005): *How the Body Shapes the Mind*. Oxford University Press, Oxford.

Gelder, Tim v. (1995): What might cognition be if not computation?, in: *Journal of Philosophy* 92: 345-381.

Goldman, Alvin (2006): *Simulating Minds*. Oxford University Press, New York.

Hurley, Susan (1998a): *Consciousness in Action*. Harvard University Press, Cambridge, MA.

Hurley, Susan (1998b): Vehicles, contents, conceptual structure and externalism, in: *Analysis* 58(1): 1-6.

Hurley, Susan (im Druck): Varieties of Externalism. In: Menary (im Druck).

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Jäger, H. (1996): Dynamische Systeme in der Kognitionswissenschaft, in: *Kognitionswissenschaft* 5 (4): 151-174.

Lakoff, George, und Mark Johnson (1999): *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books, New York.

Lyre, Holger (2002): *Informationstheorie. Eine philosophisch-naturwissenschaftliche Einführung*. Fink, München.

Lyre, Holger (2008): Handedness, Self-Models and Embodied Cognitive Content, in: *Phenomenology and the Cognitive Sciences* 7(4): 529–538.

Menary, Richard (2007): *Cognitive Integration: Mind and Cognition Unbounded*. Palgrave Macmillan, Basingstoke.

Menary, Richard, Hg. (im Druck): *The Extended Mind*. MIT Press, Cambridge, MA.

Millikan, Ruth (1984): *Language, Thought and Other Biological Categories*. MIT Press, Cambridge, MA.

Noë, A. (2004): Action in Perception. MIT Press, Cambridge, MA.

Noë, Alva (2009): *Out of Our Heads*. Hill and Wang, New York.

Port, Robert F., und Tim van Gelder, Hg. (1996): *Mind as motion: explorations in the dynamics of cognition*. MIT Press, Cambridge, MA.

Putnam, Hilary (1975): The meaning of `meaning'. In: K. Gunderson (Hg.), *Language, Mind, and Knowledge*. University of Minnesota Press, Minneapolis.

Putnam, Hilary (1981): *Reason, Truth and History*. Cambridge University Press, Cambridge.

Robbins, Philip, und Murat Aydede, Hg. (2009): *The Cambridge Handbook of Situated Cognition*. Cambridge University Press, Cambridge.

Rowlands, Mark (1999): *The Body in Mind: Understanding Cognitive Processes.* Cambridge University Press, Cambridge.

Rowlands, Mark (2003): *Externalism: Putting Mind and World Back Together Again*. McGill-Queen's University Press, Montreal & Kingston.

Rowlands, Mark (2006): *Body Language: Representing in Action*. MIT Press, Cambridge, MA.

Rupert, Robert (2004): Challenges to the Hypothesis of Extended Cognition. *Journal of Philosophy* 101(8): 389-428.

Rupert, Robert (2009): *Cognitive Systems and the Extended Mind*. Oxford University Press, New York.

Steels, Luc (2003): Evolving grounded communication for robots, in: *Trends in Cognitive Science* 7(7): 308-312.

Thelen, Esther, und Linda Smith (1994): *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA.

Tomasello, Michael (2008): *Origins of Human Communication*. MIT Press, Cambridge, MA.

Varela, Francisco, Evan Thompson und Eleanor Rosch (1991): *The Embodied Mind*. MIT Press, Cambridge, MA.

Wheeler, Michael (2005): *Reconstructing the Cognitive World: the Next Step*. MIT Press, Cambridge, MA.

Wilson, Robert A. (1994): Wide computationalism, in: *Mind* 103: 351–372.

Wilson, Robert A. (2004): *Boundaries of the mind: the individual in the fragile sciences: cognition*. Cambridge University Press, New York.

Wittgenstein, Ludwig (1953): *Philosophical Investigations*. Blackwell, Oxford.