View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PhilPapers

Finale Version in: Simon Lohse und Thomas Reydon (Hg.): Die Philosophien der Einzelwissenschaften. Meiner, Hamburg, 2017.

Philosophie der Neurowissenschaften

Holger Lyre
Lehrstuhl für Theoretische Philosophie & Center for Behavioral Brain Sciences
Universität Magdeburg

A. Einleitung

Neurowissenschaften und Philosophie

Mit dem Begriff „Neurowissenschaften" lassen sich all diejenigen Disziplinen umreißen, deren Ziel die Aufklärung der strukturellen und funktionellen Organisationsweise des Nervensystems ist. Hirnforschung und Neurowissenschaften im eigentlichen Sinne entwickelten sich erst ab der zweiten Hälfte des 19. Jahrhunderts, es handelt sich also um einen vergleichsweise jungen Wissenschaftsbereich. Wesentlicher Aspekt bei der Herausbildung der modernen Neurowissenschaften ist die historisch gewachsene Erkenntnis, dass das Gehirn bzw. das zentrale Nervensystem den zentralen Sitz des Kognitiven und Bewussten darstellt und insofern entscheidender Träger oder Vehikel mentaler Fähigkeiten und Eigenschaften ist. Dem wird in jüngerer Zeit dadurch Rechnung getragen, dass von den derart inhaltlich fokussierten Neurowissenschaften als *kognitiven Neurowissenschaften* die Rede ist. Genau hierauf soll auch der Fokus des vorliegenden Artikels liegen.

Spätestens seit den zur „Dekade des Gehirns" ausgerufenen 1990er Jahren erleben die Neurowissenschaften einen ungeahnten Aufwind. Hirnforschung ist zu einer der zentralen Forschungsfronten der Moderne herangewachsen, was sich nicht nur an einem hohen Aufkommen von Forschungsfinanzmitteln belegen lässt, sondern auch an der neumodischen Herausbildung einer großen Reihe von „Neuro-Bindestrich-Wissenschaften": klassische Disziplinen, die durch das Präfix „Neuro-" neuartige, sich den Neurowissenschaften anschließende oder doch wenigstens zu ihnen hin orientierte Disziplinen (oder Pseudodisziplinen) bilden wie zum Beispiel die Neuropsychologie, Neuroinformatik, Neurolinguistik, Neurophilosophie und Neuroethik bis hin zu Neurodidaktik, Neuroökonomie und sogar Neurotheologie. Der disziplinäre Status zumindest der drei letztgenannten ist dabei durchaus umstritten.

Neurowissenschaften und Philosophie berühren sich grundsätzlich sowohl im Bereich der theoretischen als auch der praktischen Philosophie. Letztere behandelt unter dem Titel Neuroethik zwei Fragestränge: zum einen Fragen der ethischen Dimension und Tragweite der Neurowissenschaften und ihrer lebenspraktischen Auswirkungen - Neuroethik in diesem Sinne ist weitestgehend eine Bereichsdisziplin der angewandten

Ethik. Zum anderen lässt sich unter Neuroethik auch der Versuch verstehen, ethische Fragen unter Rückgriff auf neurowissenschaftliche Ergebnisse und Methoden zu behandeln. Beide Gebiete sollen hier jedoch nicht zur Sprache kommen.

Für den Bereich der theoretischen Philosophie lässt sich eine systematische Unterscheidung treffen zwischen Neurophilosophie und Philosophie der Neurowissenschaften (wenngleich es keine kanonische Verwendung der Termini gibt). Neurophilosophie, wie insbesondere in Patricia Churchlands Klassiker „Neurophilosophy: Toward a Unified Science of the Mind-Brain" (1986) vertreten, zielt auf eine neurowissenschaftlich informierte Philosophie des Geistes ab; demgegenüber kann die Philosophie oder spezieller noch Wissenschaftstheorie der Neurowissenschaften als diejenige Disziplin angesehen werden, die die methodologischen, epistemologischen und ontologischen Probleme der Neurowissenschaften behandelt. In einem weiten Verständnis von Neurophilosophie ist letztere in ersterer enthalten. Der vorliegende Artikel fokussiert auf letztere, im Kern also auf die *Wissenschaftstheorie der kognitiven Neurowissenschaften.*

Disziplinentheoretisch besitzt die Philosophie (im Sinne einer Wissenschaftstheorie) der Neurowissenschaften in ihren Fragestellungen einen teilweise kontinuierlichen Übergang zu den Neurowissenschaften selbst – die nachfolgenden Inhalte demonstrieren dies. Zudem existieren bislang nur wenige Textsammlungen oder Überblicke (siehe Literaturempfehlungen). Dies zeigt, dass es sich um eine vergleichsweise junge Disziplin handelt, die noch keine relevante geschichtliche Dimension erkennen lässt.

Neuronale Komplexität und Stufenabfolge neurowissenschaftlicher Disziplinen

Es ist entscheidend richtig, von den Neurowissenschaften im Plural zu sprechen, da es sich hierbei nicht um eine homogene Gruppe, sondern ein heterogenes Gemenge von Fachgebieten handelt, das in verschiedenen Hinsichten disziplinär unterteilt werden kann. Mit Blick auf ihre praktische Ausübung gestatten die Neurowissenschaften eine Einteilung in theoretische, experimentelle und klinische Neurowissenschaften. Von letzteren, zu deren Kerndisziplinen die klinische Neurophysiologie, Neuroanatomie, Neurologie und Neurochirurgie gehören, wird in diesem Aufsatz abgesehen. Die in der Forschung und akademisch betriebenen Neurowissenschaften sind im hohen Maße experimentell und zeigen einen Mangel an langreichweitigen Theorien (also Theorien, Prinzipien und Gesetzen mit dem Anspruch auf weitgehend universelle Geltung anstelle nur eingeschränkt gültiger, lokaler Modelle und Mechanismen). Auf diesen Umstand wird im Laufe des Artikels noch gelegentlich eingegangen.

Neben einer vertikalen, an der Praxis orientierten Einteilung sind die Neurowissenschaften auch von einer horizontalen Einteilung bzw. Stufenabfolge neurowissenschaftlicher Teildisziplinen geprägt. Dies ergibt sich unmittelbar aus dem Forschungsgegenstand, also der Tatsache, dass Struktur und Funktion des neuronalen Systems auf unterschiedlichen Größenskalen und Organisationsebenen betrachtet werden können. Es lassen sich grob folgende Ebenen unterscheiden:

1. die molekulare Ebene
2. die zelluläre Ebene (des einzelnen Neurons)
3. die Netzwerk-Ebene neuronaler Verbände (neuronale Netze)

4. die systemische Ebene großräumiger neuronaler Verbände und Strukturen
5. die kognitiv-behavioral-psychische Ebene einzelner kognitiver Individuen
6. die psychisch-soziale Ebene von Individuen als Teil sozialer Gemeinschaften

Die Ebenen folgen einer Hierarchie wachsender neuronaler Komplexität. Entsprechend lassen sich gemäß einer horizontalen Aufteilung schichtenspezifische neurowissenschaftliche Subdisziplinen unterscheiden:

- Molekulare Neurowissenschaft (Ebene 1)
- Zelluläre Neurowissenschaft (Ebene 2)
- Computationale Neurowissenschaft (Ebenen 3 und 4, teilweise auch ab 2)
- Systemische Neurowissenschaft (Ebene 4, teilweise auch 5 und 6)
- Soziale Neurowissenschaft (Ebenen 5 und 6)

Nach herrschender Ansicht setzt die *kognitive Funktionalität* ab der zellulären Ebene bzw. im Zusammenspiel mehrerer Nervenzellen ein. In diesem Sinne sind computationale, systemische und soziale Neurowissenschaften die eigentlichen Unterdisziplinen der kognitiven Neurowissenschaften, hinzu kommen die Verhaltens-Neurowissenschaft (behavioral neuroscience) sowie Bio- und Neuropsychologie.[1] Eine Sonderrolle nimmt die Neuroinformatik ein, die ursprünglich nicht von der computationalen Neurowissenschaft abgetrennt wurde, in jüngerer Zeit aber vermehrt als Unterdisziplin der Informatik angesehen wird, deren Frageinteresse die Entwicklung und der Einsatz künstlicher neuronaler Netze zur Erledigung ingenieurtechnischer Aufgabenstellungen ist, ohne dass dabei entscheidend wäre, inwieweit die verwendeten neurocomputationalen Modelle einem biologischen Vorbild entsprechen.

Computationale Neurowissenschaft, Konnektionismus und Kognition

Das Feld der computationalen Neurowissenschaft (computational neuroscience) nimmt eine zentrale Stellung unter den neurowissenschaftlichen Arbeitsfeldern ein, insofern man hier neuronale Systeme als dezidiert informationsverarbeitend und computational ansieht (und sich in diesem Sinne für deren mathematische Modellierung interessiert). Dies beginnt auf der Einzelzellebene mit dem Hodgkin-Huxley-Modell als frühem und allgemeinen Neuronenmodell zur Entstehung und Dynamik von Aktionspotentialen. Besonderes Interesse gilt der Modellierung synaptischer Plastizität und den grundlegenden zellulären Mechanismen von Lernen und Gedächtnis. Den Schlüsselgedanken neuronaler Plastizität formuliert die von Donald Hebb 1949 neurophysiologisch begründete und bestätigte *Hebbsche Lernregel*, derzufolge die gleichzeitige Aktivität zweier Neuronen deren synaptische Verbindung verstärkt (Hebb 1949, S. 62). Hierdurch eröffnet sich ein grundlegendes Verständnis neuronaler Systeme als lernend, speichernd, adaptiv und assoziativ.

Auf der Ebene mehrerer Zellen oder Zellverbände lassen sich nunmehr Dynamiken der Verbindung von Neuronen betrachten (Hebbian cell assemblies). Hebbs Grundgedanke steht somit Pate für den *Konnektionismus* als computationalem Paradigma neurokognitiver Systeme. Als Berechenbarkeits-Paradigma steht der Konnektionismus dem Symbolismus der klassischen Künstlichen Intelligenz (KI) gegenüber. Während grundlegende neuronale Netzwerk-Modelle wie die Lernmatrix oder das Perzeptron bereits in den

[1] Vgl. Kandel et al. (1995) zu Neurowissenschaften allgemein sowie Gazzaniga et al. (2014) und Karnath & Thier (2012) zu den kognitiven Neurowissenschaften.

1950er und 60er Jahren entwickelt wurden, kam es in den 1970er Jahren zu einer weitgehenden Verdrängung durch die Dominanz der symbolverarbeitenden KI. Die 1980er Jahre führten dann zu einem explosionsartigen Wiedererstarken des (Neo-) Konnektionismus mit einer Vielzahl neuer Modelle wie dem Hopfield-Modell, Backpropagation, Boltzmann-Maschine und selbstorgansierenden Netzen nach Willshaw-Malsburg und Kohonen. Ab den 1990er Jahren treten mit Dynamizismus und situierter Kognition (Embodiment, Embeddedness; weitergehend auch Extended Cognition und Enaktivismus) paradigmatische Weiterentwicklungen auf[2] (vgl. den Beitrag „Philosophie der Kognitionswissenschaft" in diesem Band). In den computationalen Neurowissenschaften gewinnen gepulste Netzwerke zunehmend an Bedeutung (vgl. Fußnote 9 in Abschnitt C2).

Neuronale Systeme in physikalischer, funktionaler und intentionaler Hinsicht

Mit der Stufenabfolge neuronaler Komplexität gehen auch kategorial unterschiedliche philosophische Fragestellungen einher. Die Fragen und Probleme der unteren Stufen betreffen vornehmlich Struktur und Funktion des neuronalen Systems, nicht aber die Frage, wie dieses System mentale Zustände, Geist und Bewusstsein hervorbringen kann. Da die spezifischen Fragen der Philosophie der Kognition und des Geistes nicht Gegenstand dieses Aufsatzes sind, soll hier der Fokus nur darauf liegen, inwiefern Annahmen und Erkenntnisse der Kognitionswissenschaften und der Philosophie des Geistes Implikationen für die kognitiven Neurowissenschaften haben.

Die verschiedenen Frageebenen lassen sich gut anhand der drei von Daniel Dennett eingeführten Zuschreibungsarten illustrieren, die wir einem System gegenüber einnehmen können (Dennett 1987). Als Teil der physikalischen Welt kann jedes System von einer physikalischen Einstellung aus beschrieben werden (*physical stance*). Nimmt die Komplexität des Systems zu, so ist es angemessen, das System von einer funktionalen Einstellung aus zu beschreiben (*design stance*). Die Beschreibung der neuronalen Netzwerkstruktur oder des Konnektoms[3] wäre ein Beispiel dafür, wie im Rahmen der computationalen und systemischen Neurowissenschaft das neuronale System zunächst von der physikalischen Einstellung aus in seiner physikalisch-biologischen Struktur und Organisation beschrieben wird. Geht man von der Struktur zur Funktion über, so wechselt die Beschreibungseinstellung (vom *physical* zum *design stance*). Neuronalen Systemen wird dann typischerweise eine computationale Funktion unterstellt: sie sind informationsverarbeitend und üben eine Form von Rechentätigkeit aus. Zeigt ein System darüber hinaus ein Verhalten, das absichtsvoll und planvoll erscheint, so ist es nach Dennett zwingend, ihm gegenüber eine intentionale Einstellung einzunehmen (*intentional stance*). Sie eröffnet das gesamte Arsenal mentalen Vokabulars zur angemessenen Beschreibung des Systems, und betrifft die Ebenen kognitiver, behavioraler, psychischer und sozialer Zuschreibungen, und damit die Schnittstellen, die die Neurowissenschaften zur Psychologie, Psychiatrie und sämtlichen Verhaltenswissenschaften besitzen. An diesen Schnittstellen hat das traditionelle Leib-Seele-Problem als Gehirn-Geist-Problem seinen modernen Sitz.

---

[2] Vgl. Churchland & Sejnowski (1992), Dayan & Abbott (2001), Eliasmith & Anderson (2003), Gerstner et al. (2014) und Trappenberg (2010) zur computationalen Neurowissenschaft und Neuroinformatik sowie Bechtel and Abrahamsen (2002) zu Konnektionismus und Dynamizismus.

[3] Zur systemischen Neurowissenschaft, Netzwerktheorie und Konnektomie siehe B2.

B. Die Philosophie der kognitiven Neurowissenschaften

B1. Ontologische Fragestellungen

Multirealisierbarkeit und Reduktionismus

Zentrale ontologische Probleme der Neurowissenschaften ranken um die miteinander zusammenhängenden Themen Multirealisierbarkeit und Reduktionismus (siehe auch den Beitrag „Philosophie der Psychologie“ in diesem Band). Der Hinweis auf Multirealisierbarkeit (MR) dient traditionell als antireduktionistisches Argument. Gleichzeitig kann die Dominanz des MR-Arguments in weiten Teilen der Philosophie verbunden mit der Vorherrschaft des Funktionalismus dafür geltend gemacht werden, dass die Philosophie des Geistes den Neurowissenschaften über Jahrzehnte keine hinreichende Aufmerksamkeit geschenkt hat (trotz bereits beachtlicher empirischer Erfolge der Neurowissenschaften). Mentale Zustände, so das MR-Argument, lassen sich auf multiple, drastisch heterogene Weise physikalisch, auch neurobiologisch, realisieren. Daher besteht keine Identität zwischen mentalen und physischen Typen. Um Kognition und das Wesen des Mentalen zu verstehen, genügt es, die höherstufigen funktionalen und computationalen Eigenschaften kognitiver Wesen und Systeme zu untersuchen, die Details der neuronalen Realisierung sind ohne Belang.

Diese Sichtweise änderte sich erst allmählich mit dem Aufkommen des Neokonnektionismus in den 1980er Jahren. In Reaktion hierauf entstand (und entsteht) eine sich gegenüber der Philosophie des Geistes verselbstständigende Neurophilosophie und Philosophie der Neurowissenschaften erst ab den 1990er Jahren. Prominente Vorreiter der Neurophilosophie in den 1980er Jahren waren Patricia und Paul Churchland – bekannt für ihre radikale Position des eliminativen Materialismus. Nach dieser Position lässt sich unsere in der Alltagspsychologie sedimentierte Überzeugungs-Wunsch-Konzeption des Mentalen (*belief-desire folk psychology*) eines Tages durch neurowissenschaftlich fundiertes Vokabular verlustfrei ersetzen, so wie etwa unsere naive und in Teilen auch vorwissenschaftliche Konzeption dessen, was Licht ist, durch die Elektrodynamik und Redeweisen über elektromagnetische Strahlung abgelöst wurden. Gleichzeitig haben die Churchlands wesentlich dazu beigetragen, den in den 1980er Jahren aufkommenden Neokonnektionismus für die Philosophie des Geistes fruchtbar zu machen. In ihren Arbeiten (P. S. Churchland 1986, P. M. Churchland 1989) geht es beispielsweise darum, eine Art interne Funktionale-Rollen-Semantik für Systeme zu charakterisieren, deren Zustände nicht diejenigen eines klassischen symbolverarbeitenden Systems sind, sondern Zustände eines Klassifikations- oder Assoziativspeicher-Netzwerks, die aber gleichwohl als begriffliche oder propositionale Zustände ansehbar sind. Oder es geht um den Versuch, eine funktionalistisch-repräsentationalistische Auffassung von Qualia im Sinne sensorisch-neuronaler Zustände zu motivieren.

Einer der prononciertesten Vertreter eines molekular-neurobiologischen Reduktionismus ist John Bickle. Er folgt dabei Paul Churchlands Auffassungen zur Theorien-Reduktion. In Ernest Nagels klassischer Konzeption von Theorien-Reduktion wird von der Leitidee ausgegangen, dass eine niederstufige Theorie T1 eine höherstufige Theorie T2 genau dann reduziert, wenn es möglich ist, die Gesetze von T2 auch im Rahmen von T1 herzuleiten (Nagel 1961). Da höher- und niederstufige Theorien typischerweise

verschiedene Vokabularien verwenden, ist es ferner nötig, die theoretischen Terme von T2 und T1 mittels Korrespondenz-Regeln oder Brücken-Prinzipien miteinander zu verbinden. Cum grano salis lassen sich Brücken-Prinzipien als empirisch bestimmbare Identitäts-Relationen zwischen höher- und niederstufigen Eigenschaftstypen auffassen; denn Terme oder Begriffe (oder deren logische Kombinationen) dienen im Rahmen einer Theorie dazu, Mengen von Eigenschaften, also Eigenschaftstypen, in der Welt zu repräsentieren. Eine Gleichsetzung von Begriffen verschiedener Theorien über Brückenprinzipien entspricht also einer Gleichsetzung von Eigenschaftstypen verschiedener Stufe. Die Annahme strenger Typ-Typ-Identitäten führt jedoch zu den bekannten Schwierigkeiten der Nagelschen Theorie; denn im Zuge des wissenschaftlichen Theorienwandels und Fortschritts werden höherstufige Theorien häufig durch feinkörnigere, niederstufige Theorien ersetzt. Die höherstufigen Theorien erweisen sich dabei meist als ungenaue und strenggenommen falsche Näherungen, höherstufige Begriffe besitzen keine exakten Entsprechungen in der verbesserten niederstufigen Theorie, und strenge Typ-Typ-Identitäten sind in aller Regel nicht zu erwarten. Kenneth Schaffner (1967) und Clifford Hooker (1981) haben daher in Modifikation der ursprünglichen Nagelschen Konzeption vorgeschlagen, dass es hinreicht, die höherstufige Theorie T2 lediglich durch ein angenähertes, theoretisches Analogon TA zu ersetzen, das im Vokabular der reduzierenden Theorie T1 formuliert ist. Das Problem der Brücken-Prinzipien entfällt somit. Als neues Problem taucht allerdings die Frage auf, wann und in welcher Weise TA dem ursprünglichen T2 hinreichend analog ist. Bickle (1998, 2003) verwendet zusätzlich die modelltheoretische, semantische Konzeption von Theorien im Sinne des strukturalistischen Programms nach Sneed, Suppes und Stegmüller (vgl. Balzer et al. 1987), wonach Theorien nicht, wie in der syntaktischen Konzeption, als Mengen von Sätzen, sondern als Mengen von Modellen aufgefasst werden. Auf der Basis dieses Instrumentariums möchte er zeigen, dass sich beispielsweise die funktionale Charakterisierung von Gedächtnisleistungen mit einer funktionalen Charakterisierung der molekularen Grundlagen und Mechanismen von Gedächtnis gleichsetzen und insofern reduzieren lässt. Nach Bickle kann auf diese Weise eine „schonungslose" („ruthless") molekular- und zellbiologische Reduktion mentaler Eigenschaften vorgenommen werden.

Die Mehrheit der Autoren aus dem anti-reduktionistischen Lager ist von der Churchland-Bickle-Strategie nicht überzeugt, vor allem mit Hinweis auf das psychophysische MR-Argument. Seit den 2000er Jahren zeigt sich dabei als Trend in der Literatur, MR nicht mehr allgemein und abstrakt, sondern mit konkretem Bezug auf die Neurowissenschaften zu diskutieren. Dennoch bleiben die Lager gespalten: Aizawa und Gillett (2009) verteidigen die massive multiple Realisierbarkeit psychologischer Typen auf allen Komplexitätsstufen der Neurowissenschaften. Zuvor hatten Bechtel und Mundale (1999) argumentiert, dass Neurowissenschaftler oftmals Hirnzustände als typidentisch in verschiedenen neuronalen Organisationsformen und Spezies ansehen, dass sie dort funktional gleiche Mechanismen aufweisen und dass das MR-Argument zudem von einer Unausgeglichenheit zwischen grobkörnig individuierten psychologischen Zuständen und feinkörnig individuierten Hirnzuständen unzulässig profitiert. Immer häufiger wird der empirische Charakter der MR-Fragestellung in der Diskussion betont. Während Shapiro (2008) die methodischen Voraussetzungen der empirischen Testbarkeit von MR evaluiert, versucht Figdor (2010) zu zeigen, dass MR als positive These innerhalb der kognitiven Neurowissenschaften selbst eine Rolle spielt.

Ein genereller und viel beachteter Einwand gegen das MR-Argument stammt von Lawrence Shapiro (2000): In MR-Szenarios sind typischerweise nicht alle, sondern nur bestimmte Eigenschaften der Realisierer

relevant. Hieraus folgt nach Shapiro ein Dilemma: falls die Realisierer viele relevante Eigenschaften teilen, sind sie nicht typverschieden, falls sie nur wenige Eigenschaften teilen, lassen sie nur wenige und uninteressante höherstufige Generalisierungen zu (vgl. Lyre 2013a).

Netzwerk-Ontologien

Die generellen Fragen einer Ontologie der Neurowissenschaften lauten, welches die grundlegenden Bausteine, Einheiten und Entitäten neuronaler Systeme sind und wie sie sich identifizieren und individuieren lassen. Offenkundig hängen Antworten auf diese Fragen auch von der jeweiligen Ebene neuronaler Komplexität ab. Sie hängen zugleich von den Erklärungszielen und dem epistemisch-methodischen Zugriff auf das neuronale System ab, so dass auch die Diskussion in den Abschnitten B2 und B3 hier relevant ist. Eine notorische Fragestellung ist, inwieweit die Ontologie neuronaler Systeme vornehmlich über die *Struktur* oder die *Funktion* zu bestimmen ist, ob wir also in Dennettscher Terminologie die Ontologie eher vom physical, design oder intentional stance aus zu entwickeln suchen.

Die Nervenzelle stellt nach herrschender Ansicht die kleinste Einheit im Bereich der computationalen und systemischen Neurowissenschaft dar. Mit aufsteigender Komplexität folgen (ohne Anspruch auf Vollständigkeit) neuronale Ensembles, kortikale Säulen, neuronale Karten, großräumige Netzwerk-Komponenten, Module und Hirnareale. Dabei ist fortgesetzt zu bedenken, dass im Übergang von der Ebene des Einzelneurons bis zum menschlichen Gehirn 9 bis 10, bei der Zahl der Synapsen sogar 13 bis 14 Größenordnungen zu überbrücken sind. Eine derart immense Komplexität gestattet prinzipiell eine übergroße Vielzahl an Zwischenstufen mit je eigenen Bausteinen.

Die Frage der neuronalen Ontologie ist eng verbunden mit der Debatte um den Lokalismus versus Holismus des Gehirns. Zwar haben Annahmen über Lokalisierung, funktionelle Spezialisierung und Modularisierung im Gehirn seit jeher die Oberhand, eindeutig entschieden ist diese Frage jedoch bei Weitem nicht. Der Lokalismus speist sich zunächst aus anatomischer Evidenz: neuronale Systeme zeigen anatomisch und physiologisch abgrenzbare Strukturen auf, die es nahelegen, diese Strukturen auch als funktional relevant anzusehen. Die klassischen Studien von Hubel und Wiesel (1962) zur Aufklärung der grundlegenden Mechanismen der visuellen Informationsverarbeitung in der Retina und den primären cortikalen Arealen stützen die Lokalisierungsstrategie auch von Seiten der neurocomputationalen Modellbildung. Auf der Retina finden sich rezeptive Felder, die aufrund ihrer Filtereigenschaften eine subsymbolische Verarbeitung visueller Merkmale (Kanten, Texturen, Farben etc.) im Gesichtsfeld ermöglichen. Diese gefilterten Informationen werden retinotop, also topologieerhaltend, von der Retina (über eine Schaltstation im seitlichen Kniehöcker) in den primären visuellen Cortex abgebildet, wobei den rezeptiven Feldern auf der Retina kortikale Säulen (genauer: Hyperkolumnen) im Cortex entsprechen; sie können als elementare Verarbeitungsmodule des primären visuellen Cortex angesehen werden.

Ein weiteres gewichtiges Argument für die kognitiv-funktionale Spezifizierung des Gehirns entspringt den vielfältigen Erfahrungen mit Läsionen, spezifischen Hirnerkrankungen und neuronalen Detektions- und vor allem Stimulationsexperimenten. Wie Hardcastle & Steward (2002) hervorheben, besteht hier jedoch ein erheblicher experimentell-methodischer Bias bezüglich des Lokalismus; die meisten Verfahren setzten die

Lokalisierungsthese schon qua Methodik voraus, insofern in gängigen Stimulations- und Bildgebungsverfahren vornehmlich auf eine Unterscheidung mehr oder weniger aktiver Hirnregionen mit definiter Funktion abgezielt wird.

Nach Jerry Fodors (1983) einflussreicher Analyse sprechen vor allem allgemein kognitiv-psychologische Gründe für die von ihm prominent vertretene Modularitätsthese: die Reichhaltigkeit unserer kognitiven Fähigkeiten lässt sich nur unter der Annahme informatorisch und computatorisch voneinander abgrenzbarer, domänenspezifischer kognitiv-psychologischer Systeme und Mechanismen verstehen. Die Definition dessen, was ein Modul ist und welche kognitiven Bereiche modular organisiert sind, ist Gegenstand von Kontroversen. Carruthers (2006) vertritt die These massiver Modularität, dies beinhaltet die Idee, dass jegliche kognitive Funktion modularisiert vorliegt, wobei Fodors strenge Kriterien an Module (z.B. angeboren und klar abgegrenzt zu sein) hin zu rein funktional bestimmten Systemen umgewandelt werden.

Ein weiterer Argumentstrang für kognitive Module entspringt der evolutionären Psychologie. Demnach ist die evolutionäre Hervorbringung höherer Kognition nur in Form modularer Adaptationen verstehbar (Tooby & Cosmides 1995). Ein gängiges Argument für diese These ist, dass von Problem zu Problem je unterschiedliches Verhalten zur Erhöhung der Fortpflanzungswahrscheinlichkeit führt und daher kein universeller, non-modularer Mechanismus alle Probleme optimal lösen kann. Ferner würde die Komplexität der Probleme in der realen Welt jedes generalistische, non-modulare System am Problem der kombinatorischen Explosion scheitern lassen.

Seit gut zehn Jahren erwächst der theoretischen Neurowissenschaft ein neuer Zugang: die Netzwerk-Theorie. Sie bedient sich des Datenmaterials der Bildgebung, vornehmlich fMRT und DTI.[4] Die mathematische Netzwerktheorie bietet theoretische Modelle zur Erfassung der großräumigen Verbindungsstrukturen des Gehirns – vornehmlich unter Verwendung des mathematischen Werkzeugs der Graphentheorie (vgl. Sporns 2011). Großräumige Netzwerke der Hirnorganisation realisieren typischerweise Small-world-Topologien, also Netzwerkarchitekturen, die eine hohe lokale Clusterung bei gleichzeitig kurzer mittlerer Pfadlänge aufweisen. Clusterung und Modularisierung sind somit netzwerktheoretisch darstellbar. Wie Colombo (2013) hervorhebt, lässt sich die Modularitätsdebatte durch das mathematische Instrument der Netzwerktheorie gegenüber ihrem vormals vagen hypothetischen Charakter präzisieren und damit empirisch zugänglicher machen.

Um zu einer kognitiven Ontologie einzelner Hirnregionen als Elementen und Entitäten eines übergeordneten Netzwerks zu gelangen, benötigt man Kriterien, um die Funktionalität einer Region zu spezifizieren. Hirnregionen lassen sich dann als funktional individuierte Entitäten ansehen. Besitzt eine Hirnregion R eine spezifische Funktion F, so ist bei Abruf von F eine Aktivierung AR in R zu erwarten, die bedingte Wahrscheinlichkeit $p(AR|F)$ ist also hoch. In der funktionalen Bildgebung (siehe Abschnitt B3) schließt man umgekehrt von der Hirnaktivität auf die Funktion. Dies als reverse Inferenz bekannte Verfahren ist mit dem Problem behaftet, dass Hirnregionen in hohem Maße pluripotent, also typischerweise an zahlreichen kognitiven Aufgaben und Funktionen beteiligt sind. Aus der Aktivierung einer bestimmten Region kann man daher nicht unmittelbar darauf schließen, dass R eine definite Funktion erfüllt. Die

[4] Zu Fragen der Bildgebung siehe Abschnitt B3. Die Diffusions-Tensor-Bildgebung (DTI - diffusion tensor imaging) ist eine Variante der diffusionsgewichteten MRT, die geeignet ist, die Diffusionsbewegung von Wassermolekülen in Körpergewebe darzustellen. Hierdurch lassen sich die Verläufe größerer Nervenfaserbündel visualisieren (Traktografie, *fiber tracking*).

bedingte Wahrscheinlichkeit p(F|AR) ist niedrig, auch wenn p(AR|F) hoch ist.

Reverse Inferenz ist gängige Praxis in der funktionalen Bildgebung. Price und Friston (2005) schlagen vor, dass man das Problem der reversen Inferenz dadurch lösen kann, dass man den jeweiligen Abstraktions- oder Verallgemeinerungsgrad von F genügend anpasst (und zwar typischerweise erhöht). Das aber hat zur Konsequenz, dass vermehrt allgemeine und somit zunehmend uninteressante funktionale Attributionen verbleiben. Um bei Vorliegen regionaler Aktivität zu einer Verbesserung der Spezifität von F zu gelangen, sollte daher der Netzwerkkontext von R betrachtet werden. Nach Klein (2012) besteht so die Hoffnung, dass Netzwerkanalysen zukünftig auch einen Gewinn für die Frage kognitiver Ontologien darstellen.

B2. Epistemische und explanatorische Fragestellungen

Mechanistische versus dynamische Erklärungen

Das klassische Modell wissenschaftlicher Erklärungen ist das deduktiv-nomologische Modell, auch als DN-Schema oder *covering law*-Modell bekannt. Dem DN-Schema zufolge spielen allgemeingültige Gesetze in wissenschaftlichen Erklärungen eine zentrale Rolle, genauer: eine wissenschaftliche Erklärung ist ein deduktiver Schluss, der mindestens ein Naturgesetz sowie spezielle Antecedensbedingungen (typischerweise Anfangs- und Randbedingungen) als Prämissen enthält. Es wurde ab Mitte des 20. Jahrhunderts vor allem durch die Arbeiten von Carl Hempel bekannt, die dann einsetzende Diskussion um das DN-Schema förderte aber auch schon bald Defizite zu Tage. So zeigte sich, dass die bloße Subsumption unter Naturgesetze nicht hinreichend ist, um notorische Gegenbeispiele zum DN-Schema auszuräumen. Insbesondere spielt, so scheint es, die Kausalität und der Unterschied von Ursache und Wirkung auch für die Richtung einer Erklärung eine entscheidende Rolle.[5] Und schließlich scheint das DN-Schema mit seinem Fokus auf Naturgesetze die vielen Disziplinen nicht zu berücksichtigen, die theoriearm und damit arm an langreichweitigen Gesetzesmäßigkeiten sind. Die Neurowissenschaften und Lebenswissenschaften insgesamt bieten hierfür beredte Beispiele.

In der seit den 1990er Jahren in den Fokus der Debatte gerückten mechanistischen Auffassung von Erklärungen wird die im DN-Schema enthaltene Forderung aufgegeben, dass wissenschaftliche Erklärungen unter allgemeine Naturgesetze subsumieren. So ist der Hinweis auf ein gebrochenes Federrad eine anstandslose Erklärung dafür, warum der Uhrenwecker nicht mehr funktioniert. Hierzu bedarf es auf Seiten des Uhrmachers keinerlei gesonderter Kenntnisse der Grundgesetze der Newtonschen Physik, wohl aber des inneren Mechanismus des Weckers. Was ist dabei unter einem Mechanismus zu verstehen? In der Erklärungsdebatte werden mehrheitlich vier Kriterien zur Definition von Mechanismen benannt (vgl. Bechtel & Richardson 2010, Machamer et al. 2000, Craver 2007, Bechtel 2008). Unter einem Mechanismus

[5] Ein klassisches Beispiel zur Erklärungsasymmetrie lautet: Die Länge des Schattens eines Fahnenmastes S lässt sich mittels der Gesetze der Strahlenoptik, dem Stand der Sonne und der Mastlänge M erklären. Im Einklang mit dem DN-Schema lässt sich die Erklärung aber auch umdrehen: M wird durch S erklärt, was offenkundig unsinnig ist. Der Grund hierfür liegt darin, dass S eine Kausalfolge von M und nicht umgekehrt ist. Erst unter Berücksichtigung der kausalen Abfolge tritt die erwünschte Erklärungsasymmetrie zu Tage. Wesley Salmon ist daher dafür eingetreten, wissenschaftliche Erklärungen über das DN-Schema hinaus essentiell als Kausalerklärungen anzusehen (Salmon 1989).

versteht man

(1) ein mehr oder weniger komplexes *System* oder eine *organisierte Struktur*
(2) bestehend aus *Teilen*, *Komponenten* oder *Entitäten*,
(3) deren *Operationen* oder *Aktivitäten*
(4) dem zu erklärenden *Phänomen* oder *Systemverhalten* zugrundeliegen.

Es kann kein Zweifel bestehen, dass sowohl in der alltäglichen Erklärungspraxis als auch in nahezu allen komplexen Wissenschaften und speziell den Neurowissenschaften mechanistische Erklärungen eine dominante Rolle spielen. Die philosophische Debatte macht sich dabei an verschiedenen Punkten fest: Inwieweit grenzen sich mechanistische Erklärungen von rein kausalen oder funktionalen Erklärungen ab? Sind allgemeine Naturgesetze mechanistisch verzichtbar? Sind mechanistische Erklärungen reduktiv oder antireduktiv? Im Zusammenhang mit letzterer Frage haben verschiedene Autoren den Multi-Ebenen-Charakter mechanistischer Erklärungen hervorgehoben; denn typischerweise sind wenigstens zwei Ebenen involviert: die höherstufige Ebene des zu erklärenden Phänomens und die niederstufige Ebene der mechanistischen Organisation der Komponenten und ihrer Operationen. Da die Individuation des niederstufigen Mechanismus notwendig über das höherstufige Phänomen erfolgt, beharrt zum Beispiel Craver (2007) darauf, dass es sich um nichtreduktive Erklärungen handelt, während Bechtel (2008) für eine gemäßigt reduktionistische Position steht. Glauer (2012) argumentiert, dass mechanistische Erklärungen als reduktive funktionale Erklärungen im Sinne von Cummins (1983) angesehen werden können.

Zahlreiche Beispiele mechanistischer Erklärungen entstammen der molekularen und zellulären Ebene der Biologie einschließlich der Neurobiologie. Auf höheren, systemischen Stufen bedienen sich die Neurowissenschaften aber teilweise auch einer anderen Strategie: der Modellierung und Erklärung neuronaler Systemdynamiken im Rahmen der Theorie dynamischer Systeme. Unter einem dynamischen System versteht man sehr allgemein das zeitliche Entwicklungsmodell eines physikalischen Systems, das durch eine oder mehrere Zustandsgrößen beschrieben wird. Die Zustandsgrößen spannen einen Zustandsraum auf, die zeitliche Entwicklung des Systems ist als Trajektorie oder Orbit im Zustandsraum darstellbar. Neuronale Netze, speziell rekurrente neuronale Netze, stellen eine interessante Klasse dynamischer Systeme dar. Zwischen Mechanisten und Dynamizisten ist eine mitunter hitzig geführte Debatte um die Frage entbrannt, inwieweit dynamische Erklärungen von mechanistischen Erklärungen zu unterscheiden sind. Dynamizisten wie beispielsweise Stepp, Chemero und Turvey (2011) betrachten dynamische Erklärungen als in sich geschlossen und eigenständig. Dynamische Modelle sind Ihrer Meinung nach genuin explanatorisch. In ihrer Frontstellung gegenüber Mechanismen betonen Chemero und Silberstein (2008), dass dynamische Gesetze höherstufige, makroskopische Eigenschaften der betrachteten Systeme betreffen, was sie als Hinweis auf einen ontologischen Anti-Reduktionismus und explanatorischen Pluralismus werten. Das Lager der Verteidiger mechanistischer Erklärungen spaltet sich demgegenüber in wenigstens zwei Parteien auf: Kaplan und Craver (2011) sehen dynamische Erklärungen als eine spezielle Variante mechanistischer Erklärungen an (s.a. Zednik 2011), während Bechtel (2012) dynamische und mechanistische Erklärungen als einander ergänzend ansieht.

Worin aber liegen die Besonderheiten dynamischer Erklärungen gerade im Zusammenhang mit neuronalen Systemen? Wie schon Bechtel und Richardson (2010) betonen, müssen mechanistische Systeme in lokalisierbare Komponenten dekomponierbar sein. Bereits konnektionistische Systeme erfüllen diese

Bedingungen ihrer Meinung nach nicht (s.a. Bechtel und Abrahamsen 2002). Die komplexe und nichtlineare Dynamik von Nervennetzen gestattet keine sequentielle Analyse in Form definiter Netzwerkkomponenten, denn die Komponenten und Operationen eines dynamischen Netzwerks fluktuieren zeitlich und werden kontextuell moduliert. An die Stelle definiter Netzwerkkomponenten und deren Operationen treten dynamische Variablen (Bechtel 2012). Silberstein und Chemero (2013) werten dies als Hinweis auf das Versagen der mechanistischen Doktrin.

Funktionale Erklärungen der Neurocomputation

Die prädominante Erklärungsform der computationalen und systemischen Neurowissenschaft ist die funktionale bzw. computationale Erklärung. Ein kognitives System ist *computational* in dem Sinne, dass es Berechnungen ausführt. Es ist dann zugleich *funktional* im Sinne des mathematischen Funktionsbegriffs: gegebener Input wird in spezifischen Output überführt. Innerhalb des Systems stellt sich dies so dar, dass bestimmte systemische Zustände in funktionalen bzw. computationalen Relationen zu anderen Zuständen stehen. Cummins (1989) weist darauf hin, dass computationale Relationen über rein kausale Beziehungen hinausgehen: ein computationaler Zustand Z steht mit vielen anderen Zuständen des Systems in kausaler Beziehung, aber nur eine kleine Untermenge dieser Beziehungen erfüllt eine computationale Rolle in dem Sinne, dass Z gemäß einer Funktionsvorschrift in einen neuen Zustand Z' überführt wird. Ziel neurocomputationaler Erklärungen ist es, diejenige kausale Substruktur neuronaler Systeme herauszuheben, die als computationale Struktur aufgefasst werden kann (die also Zustände umfasst mit Relationen, die eine computationale Rolle spielen). In Dennetts Terminologie geht man dabei von der physikalischen zur funktionalen Einstellung über.

David Marr, Pionier im Bereich des maschinellen Sehens (*computer vision*), hat eine viel beachtete Darstellung der methodischen Vorgehensweise der computationalen Neurowissenschaft angegeben (Marr 1982, S. 25). Demnach muss die Analyse eines Systems, das eine computationale Aufgabe erledigen soll, auf drei Ebenen erfolgen (mit der ersten Ebene als Top-Level):

1. *Rechenebene („computational theory")*: Was ist das Ziel der Berechnung? Und was ist die Logik der Strategie ihrer Ausführung?
2. *Algorithmische Ebene („representation and algorithm")*: Durch welche algorithmischen Manipulationen und welche Input-/Output-Repräsentationen kann das Rechenziel erreicht werden?
3. *Physische Implementierungs-Ebene („hardware implementation")*: Wie wird der Algorithmus physisch realisiert?

Die Erklärungsrichtung in Marrs Ansatz ist top-down: die erste Ebene restringiert die zweite und diese die dritte Ebene. Marrs eigene Analyse der frühen visuellen Verarbeitung in neuronalen Systemen liefert ein Beispiel. Der Einfachheit halber sei hier nur die Fähigkeit der Kantendetektion als eines spezifischen Informationskanals des visuellen Systems betrachtet. Die computationale Theorie der frühen visuellen Verarbeitung besteht nun darin, diejenige mathematische Funktion zu spezifizieren, die die Kantenextraktion eines Bildes leistet und die damit das erzeugt, was Marr den „*raw primal sketch*" nennt (mathematisch geschieht dies etwa durch die Nulldurchgänge, „*zero crossings*", einer Laplacefilter-Transformation – genau dies legt Marr seinem Modell zugrunde). Auf der algorithmischen Ebene ist der

spezifische Algorithmus zu benennen, mit dessen Hilfe das System in der Lage ist, die mathematische Funktion der computationalen Ebene auszuführen. Hierzu ist zu beachten, dass im Allgemeinen viele unterschiedliche Algorithmen zur Berechnung einer mathematischen Funktion existieren, und dass es in Abhängigkeit vom jeweils gewählten Algorithmus entscheidend auf das Datenformat ankommt, in dem Input und Output (hier: Bild und Kantenbild) vorliegen. Schließlich muss, im biologischen Substrat, die neuronale Realisierung bzw., auf einer gegebenen Computerhardware, die spezifische Implementierung erfasst werden. Mit Blick auf das biologische Substrat konnte Marr auf die bedeutenden Arbeiten von Hubel und Wiesel zurückgreifen, die in ihren wegweisenden Arbeiten die rezeptiven Felder der Retina bei Katzen und Makakken ausgemessen hatten und dabei auf kantenselektive Rezeptorzellen gestoßen waren (s.a. Abschnitt B1).

Marrs Unterscheidung vor allem der beiden oberen Ebenen ist nicht immer ganz klar. Im Vergleich zur Denettschen Analyse wird dies deutlich: während die Implementierungs-Ebene von der physikalischen Einstellung aus erfasst werden kann, ist nicht ganz klar, ob schon auf der algorithmischen oder erst auf der computationalen Ebene die funktionale oder nicht doch sogar schon die intentionale Einstellung ins Spiel kommt (vgl. Shagrir 2010). Und während die untere Ebene wohl der Angabe eines Mechanismus entspricht, ist strittig, inwieweit dies auch die beiden höheren Ebenen mit einbezieht. Manche Autoren haben jüngst behauptet, dass computationale Erklärungen generell dem mechanistischen Erklärungstyp entsprechen (Kaplan 2011, Milkowski 2013, Piccinini 2015), andere widersprechen dem (Chirimuuta 2014).

Neurowissenschaftliche Erklärungen und das Verhältnis von Theorie und Experiment

Ein Charakteristikum der Neurowissenschaften ist ihr Mangel an Theoretizität. Sie teilen das Schicksal aller Lebenswissenschaften, weitestgehend theoriearm voranzuschreiten. Anders als etwa in der Physik, in der das Wechselspiel zwischen experimenteller und theoretischer Physik eine lange und eingespielte Tradition besitzt, liegt der Fokus der neurowissenschaftlichen Forschung in großem Maße auf der experimentellen Seite. Dies hat explanatorische und methodologische Konsequenzen, denn während die Physik wenigstens in der Grundlagenforschung dominant hypothesengetrieben voranschreitet, geht es in den Neurowissenschaften eher um die Etablierung lokaler Mechanismen und Modelle. Wissenschaftsstrukturell bilden die Neurowissenschaften ein großangelegtes Patchwork lokaler Modelle.

Aufgrund der Theorieferne und der damit verbundenen Abstinenz langreichweitiger Gesetzmäßigkeiten finden sich in den Neurowissenschaften bedeutend weniger deduktive oder nomologische Erklärungs- und Schlussweisen als etwa in der Physik. Auch induktive Erklärungen, insofern sie auf die Bestätigung gesetzesartiger Regularitäten abzielen, sind weniger verbreitet. Dies steht im Einklang mit der Beobachtung, dass mechanistische Erklärungen in den Lebenswissenschaften vorherrschend sind.

Gleichzeitig spielt das explorative Experimentieren eine zentrale Rolle. Hierunter versteht man Formen experimenteller Aktivität, bei denen die Etablierung neuer Hypothesen und Regularitäten, das tastende experimentelle Voranschreiten in theoretisch noch unverstandenes Neuland und die Entwicklung geeigneter Kategorien und Darstellungsmittel des Experimentierens selber im Vordergrund stehen. Experimente, Experimentierpraktiken und deren Darstellungen besitzen insofern einen autonomen Status innerhalb der

wissenschaftlichen Praxis. In der jüngeren Wissenschaftstheorie sind diese Überlegungen unter dem Schlagwort „Neuer Experimentalismus“ bekannt (Hacking 1983, Steinle 1997).

Stärker theoretisierte Disziplinen der Neurowissenschaften sind die computationale und systemische Neurowissenschaft. Doch die Theoriebildung muss hier zum Teil auch praktische Hindernisse überwinden. Es gibt keine einheitlichen Standards zur graphischen und notationellen Darstellung neuronaler Netzwerkmodelle unterschiedlicher Komplexität und Bandbreite. Nordlie et al. (2009) schlagen daher eine „good model description practice“ zur Darstellung und Simulation neuronaler Netzwerkmodelle vor.

Big Data, Konnektomik und neurowissenschaftliche Großforschung

Vor allem computationale und systemische Neurowissenschaft erleben in jüngster Zeit einen bemerkenswerten Wandel hin zu datenintensiven und datengetriebenen Wissenschaften. Das in diesem Zusammenhang einschlägige Schlagwort lautet „*Big data*“. Hierunter werden Datenmengen großen Volumens und großer Variabilität verstanden, die sich nur mittels hohem computationalen Rechen- und Speicheraufwand und mittels statistischer Algorithmen erfassen und bearbeiten lassen. *Big data* hat in Wissenschaftsgebieten wie der Genetik, Klimaforschung oder Hochenergiephysik zu erheblichen Transformation geführt, die Neurowissenschaften scheinen derzeit vor einem ähnlichen Wandel zu stehen (vgl. Kandel et al. 2013). Gut sichtbar ist dies durch die beiden Mega-Forschungsinitiativen des „Human Brain Project“ (HBP) (2013 durch die EU bewilligt mit einer Fördersumme von 1,19 Mrd. EUR) und der „BRAIN Initiative“ (2013 durch die US-Regierung initiiert mit anvisierten 300 Mio. Dollar pro Jahr bei einer Laufzeit von 10 Jahren). Hierbei zielt das HBP auf eine großangelegte Simulation, die BRAIN Initiative auf eine detaillierte Kartierung des menschlichen Gehirns. Die Vision einer umfassenden Kartierung der Verbindungsstrukturen des Gehirns hat unter dem Schlagwort Konnektomik („*connectomics*“, in Analogie zu Genomik oder Proteomik) zu einem neuen, boomenden Teilgebiet der Neurowissenschaften geführt.

Es liegt in der Natur der Sache, dass datenintensive Forschung weniger theoriegeleitet bzw. theoriegestützt und insofern weniger hypothesen- als datengetrieben ist. Genauer: mittels statistischer Analysemethoden und Datamining-Prozeduren werden aus den Daten Hypothesen über Korrelationen generiert, die dann direkt „getestet“ werden. Gegenstand datenintensiver Neurowissenschaft ist nicht mehr das primäre biologische Substrat, sondern sekundäre Daten. Der generelle Vorwurf lautet, dass Kausalzusammenhänge auf diese Weise weniger bis gar nicht in den Blick genommen werden. Pietsch (2015) verteidigt demgegenüber eine spezifische Form der Theoriebeladenheit der in der Datenanalyse eingesetzten algorithmischen Verfahren. Bedeutsam wird sein, wie Neurosystem- und Konnektom-Daten zukünftig gespeichert und zugänglich sind. Werden Kontext, spezielle Umstände der Datennahme sowie die Erfahrung der Experimentatoren mit erfasst? Und welche Spielräume explorativen Experimentierens gestatten die Daten? Insofern schließlich die gesammelten Daten zur Simulation großer Regionen des Gehirns oder des Gehirns als Ganzem genutzt werden sollen, lassen sich die typischen Fragestellungen der Wissenschaftstheorie der Simulation an die Hirnforschung herantragen: Welchen Erklärungswert haben Simulationen? Handelt es sich um dynamische Modelle, Fiktionen oder Formen des Experiments (vgl. Winsberg 2010)? Keine dieser Fragen wurde in der Wissenschaftsphilosophie der Neurowissenschaften bislang spezieller behandelt.

B3. Methodologische Fragestellungen der Neurowissenschaften

Die Neurowissenschaften bewegen sich auf unterschiedlichen Ebenen neuronaler Komplexität. Jede Ebene erfordert jeweils angepasste und spezifische Herangehensweisen, was zu einer Vielzahl heterogener Experimentiermethoden führt: anatomische Schneide- und Färbetechniken, elektrophysiologische Einzel- und Multi-Zellableitungen, Mikroskopie, Tiermodelle und Bildgebung, um nur einige wichtige zu nennen.

Zur Elektrophysiologie der Nervenzelle

Im Rahmen der molekularen und der zellulären Neurowissenschaft untersucht man sehr wesentlich Fragen der Signalleitung innerhalb von Nervenzellen (etwa mittels Ionenkanälen, Membran- und Aktionspotenzialen) sowie der sich daran anschließenden Signalübertragung zwischen Nervenzellen – speziell die synaptische Übertragung und Freisetzung von Neurotransmittern betreffend. Die hier auftretenden philosophischen Fragen werden weitgehend durch die Philosophie der Biologie, speziell der Molekular- und Zellbiologie, abgedeckt (vgl. Kapitel „Philosophie der Biologie“).

Hardcastle und Steward (2003) haben methodologische Probleme im Zusammenhang mit Einzelzell-Ableitungen untersucht. Sie beziehen sich dabei auf Bogen und Woodwards (1988) bekannte Unterscheidung von Phänomenen und Daten. Theorien erklären Phänomene, nicht Daten; die Phänomene selbst sind in der Regel unbeobachtbar und aus Daten abgeleitet. Als neurobiologische Phänomene können die Feuerungsrate und Amplitude von Neuronen angesehen werden, die Daten elektrophysiologischer Experimente liegen als Spannungsableitungen vor. Hardcastle und Steward zeigen auf, dass in der Neurophysiologie nahezu keine kodifizierten Standards oder Algorithmen existieren, ab wann ein Spannungssignal als Messdatum angesehen werden soll (man denke an überlagerte, verrauschte oder anderweitig kontaminierte Signale), die Experimentatoren schöpfen stattdessen aus ihrer erlernten Praxis. Zur notorischen Theoriebeladenheit der Beobachtung, einer in der Wissenschaftstheorie weithin anerkannten These, gesellt sich eine Praxis der Datengenerierung auf der eher vagen und intuitiven Basis geteilter experimenteller Erfahrung in der wissenschaftlichen Community.

Methodologische Probleme im Zusammenhang mit Neuro-Bildgebung

Keine Entwicklung hat die Neurowissenschaften in den letzten 20 Jahren derart vorangetrieben wie die Entwicklung bildgebender, nichtinvasiver Messmethoden der Hirntätigkeit. Hierzu zählen insbesondere die Elektroenzephalografie (EEG), Magnetenzephalografie (MEG), Positronen-Emissions-Tomographie (PET), Einzelphotonen-Emissionscomputertomographie (SPECT), Magnetresonanz- oder Kernspintomographie (MRT), funktionelle MRT (fMRT) und diffusionsgewichtete MRT (speziell DTI, siehe die Fußnote in B1). Bildgebungsverfahren dienen einerseits der strukturellen Untersuchung des Biosubstrats, andererseits aber auch der Aufklärung funktionaler Zusammenhänge. Hierbei ist fMRT das herausragende Verfahren, es gestattet die Darstellung neuronaler Aktivität über die Messung von Durchblutungsänderungen mittels

MRT (Details s.u.).[6]

Bildgebende Verfahren messen neuronale Aktivität immer nur indirekt. Der Evidenzstatus der daraus gewonnenen „Bilder" ist signifikant ein anderer als bei Fotographien. Doch nicht nur Laien, sondern auch Experten lassen sich hiervon täuschen (Roskies 2007), wozu auch eine häufig stark überzogene Datendarstellung beiträgt (durch Färbung und willkürliche Schwellwertlegung). Bei vielen Verfahren ist die „inferentielle Distanz" (Roskies 2010) zwischen demjenigen, was die Bilder tatsächlich zeigen, und demjenigen, was idealerweise abgebildet werden soll, sehr groß, vor allem bei fMRT. Funktionale Bilder sind weder Daten noch Phänomene im Sinne der Unterscheidung von Bogen und Woodward (siehe oben), sondern eine Art graphische Aufarbeitung oder Interpretation der Daten (Bogen 2002). Dabei existieren keinerlei allgemeingültige und verlässliche Maße zur Abschätzung des Abstands zwischen den aufbereiteten Bilderdaten und dem intendierten Phänomen, also kognitiver Aktivität. Ein wichtiger Gesichtspunkt ist, wie Stufflebeam & Bechtel (1997) bezüglich PET ausführen, die Ergebnisse einer speziellen Bildgebung mit anderen Verfahren abzugleichen. Auf den zusätzlich problematischen, von Hardcastle & Steward (2002) diagnostizierten methodischen Bias bezüglich der Lokalisierungsthese der Hirnfunktion wurde schon in Abschnitt B1 hingewiesen.

Im Folgenden sei vor allem fMRT betrachtet. Messgröße der fMRT ist das BOLD-Signal, das dem Sauerstoffgehalt im Blut entspricht (die eigentlichen Messdaten sind damit einhergehende Magnetresonanzen von Wasserstoffkernen). Das BOLD-Signal ist also ein indirekter Indikator für neuronale Aktivität, welche wiederum Indikator für kognitive Aktivität ist. Die technischen Limitationen des fMRT-Verfahrens haben Logothetis (2008) und Stelzer et al. (2014) von neurowissenschaftlicher Seite beschrieben. fMRT besitzt eine vergleichsweise gute räumliche, aber schlechte zeitliche Auflösung (im Bereich von Sekunden). Die räumliche Auflösung wird durch die Voxelgröße vorgegeben: ein Scandaten-Volumenelement mit Kantenlänge im Millimeterbereich, das je nach Scanner, Magnetfeldstärke und Gewebe größenordnungsmäßig 10^4 bis 10^6 Neuronen enthält. Das fMRI-Signal kann zwischen erregender und hemmender neuronaler Aktivität prinzipiell nicht unterscheiden, ebenso wenig unterscheidet es zwischen den hierarchisch differenten bottom-up– (also sensorisch aufsteigenden) und top-down-Prozessen (also rückgekoppelten Signalen aus höheren in niedere Areale). Schließlich wird sehr grob über individuelle anatomische Unterschiede der Probanden durch Abbildung auf ein Standardkoordinatensystem hinweg geglättet (Talairach-Atlas).

Von besonderer Bedeutung ist der Umgang mit der Auswertungsstatistik. In ihrer viel beachteten Kritik zum „Non-Independence Error" haben Edward Vul und Nancy Kanwisher nachgewiesen, dass viele Bildgebungsstudien elementare statistische Fehler enthalten; denn häufig werden erst nachträglich diejenigen Regionen selektiert, die als „interessant" gelten und mit weiteren Testvariablen korreliert sein sollen. Ein derartiger Selektionsbias der ROI's (*regions of interest*) führt jedoch zu massiv überschätzten Korrelationen (vgl. Vul & Kanwisher 2010). In ähnlicher Weise hat Colin Klein von wissenschaftstheoretischer Seite darauf hingewiesen, dass viele fMRT-Studien unsauber sind, da sie die ihnen zugrundeliegende Hypothese nicht gegenüber der Nullhypothese prüfen, was gegen elementare Bayesianische Einsichten verstößt (Klein 2010, Mole und Klein 2010; siehe aber auch Machery 2014).

[6] Vgl. Walter (2005) für einen Überblick zur funktionellen Bildgebung und Hanson & Bunzl (2010) zu den wissenschaftstheoretischen Grundlagen.

Soziale Neurowissenschaft

Die Entwicklung bildgebender Verfahren, vor allem fMRT, hat auch die Entstehung und Entwicklung der sozialen Neurowissenschaft stark vorangetrieben. Hier geht es unter anderem darum, die neuronalen Mechanismen sozialer Kognition zu klären. Das Feld der sozialen Kognition umfasst die gesamte Bandbreite der Interaktion und des Austauschs von Artgenossen untereinander – speziell beim Menschen. Hierbei spielt die Fähigkeit, mentale Zustände im Gegenüber zu erkennen, eine zentrale Rolle. Diese Fähigkeit firmiert unter den Begriffen Mindreading, Mentalisierung oder Theorie des Geistes (ToM- *Theory of Mind*; in Dennetts Terminologie die Fähigkeit, eine intentionale Einstellung einzunehmen). Zahlreiche konkurrierende Ansätze wurden entwickelt, um Mindreading-Fähigkeiten theoretisch zu erfassen. Hierzu zählen insbesondere die Theorie-Theorie, die Simulationstheorie sowie daraus entwickelte Hybride (vgl. Goldman 2012).

Die Theorie-Theorie sieht unsere Alltagspsychologie als theorieartig strukturiert an. Alltagspsychologische Generalisierungen über mentale Zustände und Eigenschaften anderer stehen in Analogie zu den gesetzartigen Aussagen einer naturwissenschaftlichen Theorie, deren theoretische Terme (z.B. Elektron, schwarzes Loch oder Gen) in gleicher Weise über die direkte Beobachtung hinausgehen wie das Fremdpsychische und Fremdkognitive (z.B. Überzeugungen, Absichten oder qualitative Erlebniszustände anderer). Demgegenüber ist die Simulationstheorie ein fähigkeitsbasierter Ansatz, die den Ursprung unserer ToM-Fähigkeiten nicht in speziellen Rationalitätsannahmen sieht, sondern in der Fähigkeit, sich in die Lage des anderen zu versetzen, ihn insofern „offline" zu simulieren. Simulationstheoretiker wie Gallese & Goldman (1998) haben die These vorgetragen, dass die in den 1990er Jahren entdeckten Spiegelneuronen die neuronale Basis der Mindreading-Fähigkeiten bilden. Jacob (2008) diskutiert einige konzeptionelle Probleme dieser These, insbesondere die Unterbestimmtheit der eigentlichen Handlungsabsicht (prior intention) durch Körpermotorik, sodass die Spiegelthese allenfalls auf die unmittelbare Motorintention einzuschränken wäre, was nicht hinreichend für Mindreading ist.

Experimente in der Tierkognition

Zahlreiche Experimente der sozialen Neurowissenschaft zielen darauf ab, die Ursprünge unserer ToM-Fähigkeiten zu ergründen. Sie bewegen sich daher in einem nicht unerheblichen Maße in den Bereichen der Entwicklungspsychologie und der kognitiven Ethologie, speziell der Primatenkognition. Dort wird unter anderem die Frage behandelt, ob und in welchem Maße Menschenaffen über eine ToM verfügen. Das experimentelle Paradigma hierzu sind „false belief tasks", bei denen es darum geht zu erkennen, dass andere Überzeugungen haben können, von denen man selber weiß, dass sie falsch sind (Call & Tomasello 2008).

Bei Experimenten im Bereich der Tierkognition ergeben sich höchst interessante wissenschaftstheoretische Fragestellungen (Hurley & Nudds 2006, Lurz 2009). Experimente zur Primatenkognition sind, wie Povinelli und Vonk (2004) argumentieren, von einem „logischen Problem" gekennzeichnet, das man als eine Instanz des Problems der Theorienunterbestimmtheit ansehen kann. Derzeitige experimentelle Resultate sind

sowohl im Einklang mit einer theoretischen Interpretation, die Menschenaffen Mindreading-Fähigkeiten unterstellt, als auch einer kognitiv weitaus weniger anspruchsvollen Variante, bei der sich evolutionär erfolg- und trickreiche „Behavior-Reading"-Fähigkeiten ausgebildet haben. Lurz (2011) legt ein experimentelles Design vor, anhand dessen sich die Frage, ob Primaten Behavior-Reader oder genuine Mindreader sind, entscheiden lassen soll. Die Idee ist, in geeigneter Weise auf die Unterscheidung von Erscheinung und Realität zurückzugreifen: ein Tier kann gegebenenfalls lernen, dass ein Unterschied darin besteht, wie ein Stimulus ihm erscheint, oder wie er real vorliegt, aber nur ein Mindreader kann diese Unterscheidungsfähigkeit auch einem Artgenossen attribuieren.

(Neue) Interventionsmethoden

Als letztes wichtiges Instrumentarium der Neurowissenschaften sei auf Interventionsmethoden verwiesen. Während Ableitung und Bildgebung lediglich detektieren, sind Interventionsmethoden geeignet, intendiert in neuronale Funktionskreise und Mechanismen einzugreifen und diese zu manipulieren. Erst auf diese Weise lassen sich Kausalzusammenhänge im Einklang mit der Interventionstheorie der Kausalität in zuverlässiger Weise aufdecken und bestimmen. Der Interventionstheorie zufolge ist U eine Ursache der Wirkung W, falls durch eine Manipulation von U nicht nur U, sondern auch W verändert wird (vgl. Woodward 2008 für eine Darstellung des kausalen Interventionismus mit Blick auf neuronale Systeme). Zu den wichtigen neurowissenschaftlichen Interventionsmethoden zählen seit jeher Läsionsstudien und Elektrostimulation, aber auch pharmakologische Interventionen. In neuerer Zeit kommen vor allem transkranielle Magnetstimulation (TMS), transkranielle Gleichstrom- (tDCS) und Wechselstromstimulation (tACS), tiefe Hirnstimulation (DBS) und Optogenetik hinzu (bei letzterer handelt es sich um eine noch relativ junge Methode zur Kontrolle und Manipulation genetisch modifizierter Zellen, also auch Nervenzellen, mittels Licht).

TMS erlaubt die kurzfristige Stimulation oder Hemmung kortikaler Regionen auf nichtinvasivem Wege durch äußere Magnetfelder. Aufgrund der beschränkten Eindringtiefe können jedoch nur nahe des Schädels gelegene Hirnpartien direkt beeinflusst werden, ein weiteres Problem ist die diffuse räumliche Auflösung (wenige Quadratzentimeter Schädelfläche abhängig von der Magnetfeldstärke und des Spulendesigns bei guter zeitlicher Auflösung). Dennoch ist TMS ein wichtiger methodischer Schritt in Richtung auf funktionale und kognitive Ontologien. Dies ist von wissenschaftstheoretischer und philosophischer Seite bislang nur unzureichend wahrgenommen worden. Als eine der wenigen diesbezüglichen Arbeiten weisen Schutter et al. (2004) im Zusammenhang mit TMS auf die neuen Möglichkeiten zur Lokalisierbarkeit „psycho-neuraler Entitäten" hin.

Im Sinne des Entitätenrealismus von Ian Hacking (1983) lässt sich folgende Überlegung anstellen: eine Hirnregion besitzt eine kognitive Funktion, falls es gelingt, die Wirkungen dieser Funktion intendiert zu manipulieren und dann entsprechend zu detektieren (etwa mithilfe von Bildgebung). Auf diese Weise lässt sich der Kreis aus Detektion und Intervention schließen. Für Hacking ist dies hinreichend für eine realistische Interpretation einer ansonsten nicht direkt beobachtbaren Entität, wie sein bekanntes Diktum über den Realstatus von Elektronen illustriert: „If you can spray them, they are real". Diese Überlegung ließe

sich gleichermaßen auf eine qua Interventionsmethoden begründete Ontologie funktional oder kognitiv individuierter Hirnregionen übertragen. Dies gilt für TMS ebenso für die Möglichkeiten von DBS und Optogenetik. Während DBS vornehmlich in der Klinik Anwendung findet, hat gerade die Optogenetik ein hohes forschungsrelevantes Potential. Für die Wissenschaftstheorie stellt die Analyse neurowissenschaftlicher Interventionsmethoden ein Forschungsdesiderat dar (siehe jedoch Craver, im Druck).

C. Schlussteil

C1. Weitere Probleme und Fragestellungen der Philosophie der Neurowissenschaften

Viele Probleme und Zukunftsfragen der Philosophie der Neurowissenschaften sind eng verbunden mit offenen Fragen der Neurowissenschaften selbst. Letztere betreffen nach Bekunden von Neurowissenschaftlern nahezu alle großen Themen (vgl. Hemmen & Sejnowski 2006): Wie funktionieren (multimodale) Wahrnehmung, Bewegungssteuerung, Lernen und Gedächtnis? Wie entwickelt sich das Gehirn – sowohl evolutionär als auch ontogenetisch? Welche Beiträge leisten Genetik und Umgebung? Wie funktionieren Sprache, Entscheidungsfindung und Handlungskontrolle? Was sind Schlaf, freier Wille und Bewusstsein? Wie hängen Gehirn- und psychische Störungen zusammen? Es folgt eine vermischte Auswahl bislang noch unbehandelter Themen, die einerseits mit offenen Fragestellungen der Neurowissenschaften zusammenhängen, andererseits aber auch in philosophischen Debatten Aufmerksamkeit gefunden haben.

Die Willensfreiheitsdebatte

Die Frage nach der Willensfreiheit hat im deutschsprachigen Raum in den 2000er Jahren bemerkenswert hohe Wellen geschlagen – bis hinein in die Medien. Wesentlicher Auslöser waren freiheitsskeptische und neurodeterministische Aussagen prominenter Neurowissenschaftler (vgl. Prinz, Singer und Roth in Geyer 2004). Im Hintergrund steht ein von Benjamin Libet bereits Ende der 1970er Jahre durchgeführtes Experiment, bei dem er zeigen konnte, dass eine zeitliche Verzögerung von mehreren hundert Millisekunden zwischen dem einer Handlung vorausgehenden neuronalen Bereitschaftspotential und dem Bewusstwerden der Handlungsentscheidung auftritt (Libet 1985).

Unter Philosophen bestehen zum Teil große Meinungsverschiedenheiten zur Frage der Willensfreiheit. Die beiden dominanten Positionen, der Kompatibilismus und der Libertarianismus, gehen von sehr verschiedenen Grundannahmen aus. Kompatibilisten sind der Ansicht, dass Freiheit und Determinismus grundsätzlich vereinbar sind. Libertarianer sind nicht nur Inkompatibilisten, sondern nehmen an, dass der Determinismus falsch sein muss, da Willensfreiheit besteht. Der Freiheitsbegriff des Libertariers lässt sich durch die Annahme charakterisieren, dass man anders hätte handeln können – auch unter ansonsten völlig unveränderten physikalischen Umständen. Dies impliziert Indeterminismus.

Die Willensfreiheitsfrage hat aber nicht nur mit kontroversen Auffassungen zum Determinismus zu tun, sondern ist zudem mit zahlreichen weiteren philosophischen Problematiken verwoben wie mentaler Verursachung, Kausalität, dem Handlungsbegriff sowie der Unterscheidung von Ursachen und Gründen. Beim Problem der Willensfreiheit handelt es sich also in gewisser Weise um eine ganze Klasse von Problemen, was nicht nur die Diskussion zwischen Neurowissenschaftlern und Philosophen, sondern auch von Philosophen untereinander erschwert. Dennoch lässt sich feststellen, dass der Kompatibilismus heute die Mehrheitsposition unter Philosophen darstellt. Dies hängt auch an einer scheinbar unvermeidlichen Konsequenz des Libertarianismus, dass nämlich Willensfreiheit auf die Zufälligkeit und damit Beliebigkeit des Handlungsverlaufs führt. Nach Mehrheitsauffassung ist freies Entscheiden und Handeln aber ein Entscheiden und Handeln nach Gründen (ohne äußere Zwänge). Um hiervon ausgehend eine naturalistische Theorie der Willensfreiheit zu entwickeln, muss man den Graben zwischen Ursachen und Gründen überwinden. Aber auch hier setzt sich die Überzeugung durch, dass dies prinzipiell möglich ist, indem die handlungsrelevante Wirksamkeit von Gründen mit neuronalen Realisierungen geeignet verknüpft wird (vgl. Beckermann 2008, Pauen und Roth 2008).

Unter der Maßgabe unseres weiterhin unvollständigen neurowissenschaftlichen Wissens der Hirnfunktion erscheint ein strenger Neurodeterminismus eher als Chimäre (und viele Kompatibilisten sind ohnehin der Auffassung, dass die Frage, ob die Welt deterministisch oder indeterministisch ist, für die Frage der Willensfreiheit irrelevant ist). Interessanter ist die generelle Frage, inwiefern empirische Befunde einen Beitrag für eine kompatibilistisch verstandene Willensfreiheit leisten. Wie beispielsweise Sven Walter (2016) ausführt, zeigen sozialpsychologische Studien, dass ein nicht unerheblicher Teil unserer Handlungssteuerung unbewusst abläuft. Dies unterminiert unsere rationale Kontrollfähigkeit, wir sind in bestimmten Grenzen manipulierbar. Freiheit besteht demnach nicht uneingeschränkt, sondern nur graduell.

Eine freiheitsskeptische Konsequenz ganz anderer Art zieht Carruthers (2007). Sie steht im Kontext seiner Auffassung, dass Metakognition das Resultat auf uns selbst gerichteter Mindreading-Fähigkeiten ist (wodurch Fälle von Konfabulationen erklärbar werden). Das Selbst ist eine beständige Form nachträglicher Selbstzuschreibung und Selbstnarration. Nach Carruthers neigen wir fälschlich dazu, unsere inneren Verbalisierungen (etwa der Form „Ich werde jetzt dies tun“) als transparent anzusehen und nicht als reine Selbstinterpretation. Der freie Wille ist demnach eine Illusion.

Das Bindungsproblem

Die subsymbolische Verarbeitungsstrategie des Konnektionismus bringt, bei allen Vorteilen, auch Probleme mit sich. Unsere Sinnesorgane zeigen spezifische Rezeptivität für ausgewählte Merkmale des sensorischen Inputs. Eine gegebene Szene, etwa das Gesichtsfeld im Falle der visuellen Wahrnehmung, wird in eine Vielzahl von einzelnen Merkmalen („features“) zerlegt (wie Kantenorientierung, -stärke, Texturen, Farben etc.). Auf den frühen neuronalen Verarbeitungsstufen entsteht so das Problem, zusammengehörige Merkmale geeignet zu verbinden und zu einem einheitlichen Wahrnehmungserlebnis zu integrieren, wobei zusätzliche Mehrdeutigkeiten im Merkmalsraum zu überwinden sind. Liegt etwa ein grünes Dreieck und ein rotes Quadrat im Gesichtsfeld vor, so sind diejenigen Neurone aktiv, die für die Merkmale rot, grün,

dreieckig und quadratisch codieren.[7] Allerdings rufen ein rotes Dreieck und ein grünes Quadrat dieselbe Aktivierung hervor, es kommt zur „Superpositions-Katastrophe“ (Malsburg 1999). Um ihr zu entgehen, benötigt man einen geeigneten, neuronal plausiblen Bindungsmechanismus. Bereits Anfang der 1980er Jahre hat Christoph von der Malsburg vorgeschlagen, Neurone auf der Basis der Korrelationen ihrer zeitlichen Aktivitäten zu großen Ensembles zu verbinden (im Effekt handelt es sich um eine Hebbsche Bindungsregel auf einer schnellen, dynamischen Zeitskala). Diese Lösung des Bindungsproblems durch neuronale Synchronizität hat aufgrund der experimentellen Funde vor allem durch Singer und seine Mitarbeiter in den 1990er Jahren große Beachtung gefunden (vgl. Malsburg et al. 2010; für eine kritische Diskussion der experimentellen Evidenz neocortikaler Rhythmen siehe den dortigen Text von Singer).

Bennett und Hacker (2003, Kap. 4.2.3) halten ein derartiges Verständnis des Bindungsproblems für konfus. Das Problem entspringt ihrer Meinung nach der falschen Vorstellung und Redeweise, dass das Gehirn aus verschiedenen Merkmalen der Sinneswahrnehmung ein kohärentes Bild konstruiert. Aber die Fähigkeit eines Wesens zu sehen heißt nicht, interne Bilder zu konstruieren. Konziser ist die Kritik von Hardcastle (1994) und Garson (2001), die die Annahme vertreten, dass zwischen wenigstens zwei Varianten von Bindung zu unterscheiden ist, die man als funktionale und phänomenale Bindung bezeichnen kann. Funktionale bzw. perzeptionelle Bindung findet relativ unkontrovers im Bereich der Wahrnehmung und entsprechend auf den frühen Verarbeitungsstufen merkmalscodierender neuronaler Aktivitäten statt. Die funktionale Bindung von Stimulimerkmalen folgt dabei über weite Strecken den Gesetzen der klassischen Gestaltpsychologie, wie in Wahrnehmungsexperimenten demonstrierbar ist (Treisman & Gelade 1980). Kontrovers bleibt die Frage, ob es im Sinne phänomenaler Bindung darüber hinaus nötig und begrifflich sinnvoll ist, das subjektive Erleben der Einheit der Wahrnehmung und des Bewusstseins durch spezifische neuronale Mechanismen der Synchronisation zu erklären.

Bewusstseinsforschung

Erst ab den 1980er Jahren setzt allmählich eine nennenswerte neurowissenschaftliche Bewusstseinsforschung ein. Francis Crick und Christof Koch (1990) postulieren das Phänomen synchroner 40 Hz-Oszillationen als neuronales Korrelat von Bewusstsein. Die Kritik an phänomenaler Bindung des vorangehenden Abschnitts bezieht sich hierauf, steht aber auch im Zusammenhang mit dem notorischen Qualia-Problem. Mittlerweile sind eine ganze Reihe theoretischer Ansätze zu den neuronalen Grundlagen von Bewusstsein vorgelegt worden, aber keiner dieser Ansätze kann den spezifisch qualitativen Charakter phänomenalen Bewusstseins zufriedenstellend erfassen. Ein früher Ansatz ist die Theorie des „*global workspace*“ von Baars (1988), eines global verfügbaren Arbeitsspeichers für kurzlebige, wechselnde Repäsentationen. Tononi (2004) hält einen hohen Grad an Informations-Integration für entscheidend. Thomas Metzinger (2003) vertritt eine Selbstmodell-Theorie, derzufolge repräsentationale Modelle der Außenwelt in ein Selbstmodell eingebettet werden, und zwar so, dass Weltmodell und Selbstmodell im selben neuronalen Format vorliegen und das Selbstmodell aufgrund seiner semantischen Transparenz vom System nicht mehr als Modell erkannt wird.

Einen Ansatz besonderer Art stellt die Neurophänomenologie dar. Drei Stränge sollen nach Ansicht von Shaun Gallagher und Dan Zahavi (2008) in diesem Gebiet zusammenlaufen: neurowissenschaftliche

[7] Es ist für die Didaktik des Beispiels unerheblich, ob derartige Merkmale tatsächlich neuronal codiert sind, was vermutlich nicht der Fall ist.

Bewusstseinsforschung, verkörperlichte Kognition („*embodiment*"; vgl. Varela, Thompson & Rosch 1991, Lyre 2013b) und die phänomenologische Tradition. Vor allem in Hinblick auf letzteren Punkt betonen die Vertreter der Neurophänomenologie die Akzeptanz und Bedeutung von Erster- (und neuerdings auch Zweiter-) Person-Perspektive zur wissenschaftlichen Untersuchung von Bewusstsein, ganz im Gegensatz etwa zu Dennett (1991), der mit seiner „Heterophänomenologie" eine strenge Dritte-Person-Methodologie vertritt.

Menschenbild und kritische Neurowissenschaft

Neurowissenschaften und Hirnforschung erleben seit einem guten Vierteljahrhundert einen ungebrochenen Boom. Dabei geht es auch, dem Selbstverständnis und den Erklärungsansprüchen vieler Neurowissenschaftler nach, um die Etablierung eines wissenschaftlich fundierten, neuen Bildes vom Menschen. Nicht selten werden diese Ansprüche in den Medien von großen Ankündigungen und Versprechungen begleitet (vgl. Elger et al. 2004). Die Fachphilosophie ist dem zum Teil entgegengetreten. Unter Rückgriff auf wittgensteinianische Überlegungen zeigen Bennett und Hacker sowohl systematisch (Bennett & Hacker 2003) als auch historisch (Bennett & Hacker 2008) einen weitverbreiteten „mereologischen Fehlschluss" auf, der in der gängigen neurowissenschaftlichen Praxis besteht, mentale und psychologische Prädikate nicht Personen als Ganzen, sondern dem Gehirn oder auch Teilen des Gehirns zuzuschreiben. Die zum Teil polemischen Ausführungen der beiden Autoren sind aber auch unter Philosophen nicht unumstritten. Wie beispielsweise Dennett (in Bennett et al. 2007, S. 85ff) anmerkt, muss die Frage, welche Sprachgebrauchsregeln etabliert sind, empirisch entschieden werden (und nicht durch Sprachverbote).

Neurowissenschaftliche Deutungsansprüche markieren aber nicht nur das wissenschaftliche Selbstverständnis, sondern auch Hoheitsansprüche gegenüber benachbarten Disziplinen, und zielen zum Teil bewusst auf Öffentlichkeit und Politik. Projekte wie die „Kritische Neurowissenschaft" treten mit dem Versuch an, die Neurowissenschaften vor überzogenen, voreiligen Schlüssen und Interpretationen zu bewahren (Choudhury & Slaby 2011). Im Fokus stehen Konzepte wie Freiheit, Selbst, Subjektivität oder der psychische Krankheitsbegriff. Dennoch ist eine kritische Reflektion der Neurowissenschaften nicht automatisch gleichbedeutend mit einer Verabschiedung des naturalistischen Menschenbildes (vgl. Beckermann 2008). Mit Blick auf gesellschaftliche Konsequenzen der zunehmenden Entschlüsselung der Neurobiologie des Bewusstseins zielt Metzinger (2009) auf eine neue Bewusstseinsethik.

C2. Zukünftige Entwicklungen und Herausforderungen der Philosophie der Neurowissenschaften

Es ist nicht leicht, die zukünftigen Entwicklungen eines Gebiets auszumachen, das einen so raschen Wandel erfahren hat und auch absehbar erfahren wird wie die Neurowissenschaften (vgl. Marcus & Freeman 2014 für einen Blick auf die neurowissenschaftliche Zukunft). Entsprechend sind auch die Herausforderungen für die begleitende Wissenschaftsphilosophie schwer abzuschätzen. Ein allgemeiner Punkt sei noch einmal wiederholt: die Neurowissenschaften sind vergleichsweise theoriearm und experimentorientiert. Eine umfassende, vereinheitlichte Theorie des Gehirns ist nicht in Sicht. Eine visionäre Hoffnung wäre es,

wenigstens einige grundsätzliche Organisations- und Stilprinzipien des Gehirns benennen zu können. Umrisse und Ansätze hierzu sind immerhin vorhanden. Ein wichtiges Prinzip ist zweifellos Hebbsche Plastizität, auch dynamische Selbstorganisation ist in der ein oder anderen Weise relevant. Beides zusammengenommen führt in den Ideenkreis von Synchronisation und dynamischer Koordination (siehe Abschnitt C1). Auf vielen Skalen und in vielen Bereichen findet sich das Grundmotiv lokaler Aktivierung bei lateraler Inhibition. Friston (2010) postuliert ein „*free energy principle*", wonach kognitive biologische Systeme grundsätzlich danach streben, die freie Energie ihrer sensorischen Zustände (interpretiert als statistisches, informationstheoretisches Maß für deren Überraschungsgrad) zu minimieren. Ein verwandter Ansatz ist die generelle Idee des „*predictive coding*", wonach das Gehirn keine rein passive Repräsentationsmaschinerie darstellt, sondern aktiv und beständig neue Vorhersagen oder Weltmodelle auf der Basis bekannter Zustände generiert, um sensorischen Input vorherzusagen bzw. mit der Vorhersage in Abgleich zu bringen (vgl. Clark 2013, Hohwy 2013). Beide Konzepte, *free energy principle* und *predictive coding*, sind über die wachsende Bedeutung bayesianischer Ansätze in den Neurowissenschaften verbunden (Schlagwort „*Bayesian brain*", vgl. Knill & Pouget 2004). Wie Colombo & Seriès (2012) argumentieren, handelt es sich bei diesen Ansätzen nicht um mechanistische Erklärungsformen, man sollte daher Instrumentalist bezüglich der Frage sein, ob das Gehirn eine Bayesianische Maschinerie darstellt.

Ein offensichtliches Hindernis auf dem Weg zu einem umfassenden Verständnis ist die schiere Komplexität des Gehirns. Man könnte dies das „Skalenproblem der kognitiven Neurowissenschaften" nennen: Angenommen, die Mechanismen und Dynamiken auf der Einzelzellebene und für überschaubare Netzwerke von bis zu 100 oder 1000 Neuronen seien halbwegs verstanden und modellierbar, wie ließen sich dann diese Mechanismen, Dynamiken und Modelle übertragen bzw. hochskalieren auf die Größenordnung der 10^{10} Neuronen des gesamten Gehirns? Immerhin müssen sieben bis acht Größenordnungen überbrückt werden, was zahlreiche praktische, technische und methodologische Fragen aufwirft, die Neurowissenschaften aber auch vor das epistemische Problem stellt, welche Art Zugriff der menschliche Forscherverstand auf diese Komplexitätsebenen haben kann.

Der Code des Gehirns

Zentrale Grundannahme der kognitiven Neurowissenschaften ist, dass das Gehirn informationsverarbeitend und repräsentational ist. In Strenge sind dies zwei miteinander verbundene Annahmen: um informationsverarbeitend zu sein, muss das Gehirn erstens Information in bestimmten Datenformaten kodieren und zweitens nach gewissen syntaktischen Regeln verarbeiten.[8] Es tritt dann die Annahme hinzu, dass die Syntax auf biologischen Vehikeln operiert, die semantisch bewertbar oder repräsentational sind. Man kann also eine syntaktische von einer semantischen Annahme unterscheiden.

Während die Philosophie des Geistes traditionell auf Fragen der Semantik und der mentalen Repräsentation abzielt, wird häufig übersehen, dass bereits das syntaktische Problem neuronaler Codierung mit eminenten wissenschaftstheoretischen Fragen verknüpft ist. Es ist eine offene Frage, welche biologischen Elemente und Strukturen im Gehirn überhaupt als informationsverarbeitend und codierend anzusehen sind (Vreeswijk 2006, Stanley 2013). Vermutlich sind mehr als nur die Eigenschaften von Nervenzellen relevant. In jüngerer

[8] Zum Begriff der Information siehe Lyre (2002).

Zeit rücken beispielsweise Gliazellen ins Interesse, deren häufigste Form, die Astrozyten, an der neuronalen Signalübertragung und Reizleitung in unterschiedlicher Form beteiligt sind (vgl. Fields 2010). Dabei hängt nicht nur die Grundlagenforschung, sondern beispielsweise auch die neurotechnologische Entwicklung an einer Aufklärung dieser Fragen, denn der Einsatz von neuronalen Prothesen, Implantaten oder Gehirn-Computer-Schnittstellen setzt ein Verständnis des cerebralen Codes zur Informationsübertragung voraus.

Doch auch mit dem Fokus auf rein neuronale Codes treten zahlreiche Fragen auf. Es ist davon auszugehen, dass neuronale Systeme keinen universellen Code verwenden, sondern in höchstem Maße kontextabhängig operieren. Differenzen hinsichtlich der neuronalen Codierung sind zwischen verschiedenen Spezies, Individuen, Hirnregionen und Typen neuronalen Gewebes zu erwarten. Sehr wahrscheinlich nutzt auch jede Modalität ein eigenes Datenformat (z.B. 1D auditorische Daten gegenüber 2D visuellen Daten). Neuronale Aktivität besitzt zwei augenfällige Charakteristika, die prinzipiell zur Codierung von Information dienen können: die (Feuerungs-) Rate der Aktivität eines Neurons pro Zeiteinheit (r*ate code, spike trains*) und die zeitliche Struktur neuronaler Antworten sowie deren Bezogenheit aufeinander (*temporal code*). Doch wie schon Hebb betonte, codieren nicht nur einzelne Zellen, sondern auch ganze Populationen (*population coding*). Die Betonung der zeitlichen Aspekte neuronaler Aktivität hat in der computationalen Neurowissenschaft zur bedeutsamen Kategorie der gepulsten Netzewerke (*spiking networks*) geführt.[9]

Man kann auch umgekehrt fragen: ist jede neuronale Aktivität codierend? Wie ist es mit neuronaler Aktivität im Schlaf oder im sogenannten Hirnruhezustand (*resting state, default mode network*; vgl. Raichle & Snyder 2007)? Wissenschaftstheoretisch bemerkenswert ist das Problem neuronaler Codierung als exklusiv empirisches Problem. Was kann und will man herausfinden, wenn man rein empirisch untersucht, wie das Gehirn Information verarbeitet und codiert? Und was kann als überzeugender empirischer Nachweis eines neuronalen Codes angesehen werden? Die übliche Vorgehensweise besteht darin, Korrelationen zwischen äußeren Stimuli und neuronalen Antworten aufzuzeigen. Auf diese Weise finden sich merkmalscodierende Neurone auf den frühen sensorischen Verarbeitungsstufen, Orts- und Gitterzellen im Hippocampus (Moser & Moser 2014, s.a. Bechtel 2014), Spiegelneurone und neuerdings auch „Jennifer Aniston-Zellen“ bzw. Konzeptzellen (eine abgemilderte Variante von Großmutterzellen, vgl. Quiroga et al. 2005, Quiroga 2012), um nur einige bekannte Beispiele zu nennen.

Dennoch dürften Korrelationen zur Bestätigung codierender Elemente letztlich nicht ausreichend sein. Die Parallele zum genetischen Code der Molekularbiologie drängt sich auf, greift aber zu kurz. Denn der genetische Code lässt sich anhand der durch ihn codierten und hervorgebrachten Endprodukte individuieren, nämlich Proteine in der Zelle. Analog bringt der neuronale Code bedeutungshafte mentale Repräsentationen hervor. Diese sind aber nicht in gleicher Weise empirisch fassbar und nachweisbar wie Proteine. Kann man sagen, wie das Gehirn Information verarbeitet und codiert, ohne dass man weiß, wie das Gehirn bedeutungshaft operiert und repräsentiert? Falls nein, kann das Problem nicht (jedenfalls nicht ausschließlich) bottom-up angegangen werden.

Da die Frage nach dem Wesen mentaler Repräsentation seit jeher im Zentrum der Philosophie des Geistes

[9] Nach Maas (1997) lassen sich Spiking Networks als dritte Generation neuronaler Netze verstehen. Zu den Netzen erster Generation zählen Netze mit digitalisierten Outputs wie etwa McCulloch-Pitts-Netzwerke. Netze zweiter Generation gestatten kontinuierliche neuronale Antwortfunktionen, Netze dritter Generation deren gepulsten, zeitlichen Charakter. Spiking Networks besitzen gegenüber klassischen neuronalen Netzen höhere biologische Plausibilität (Maass und Bishop 1999, Gerstner et al. 2014).

steht, findet die Wissenschaftstheorie der Neurowissenschaften hier ihren natürlichen Übergang zur Philosophie des Geistes. Eine top-down geleitete Erforschung des neuronalen Codes durch die Neurowissenschaften könnte in diesem Sinne durch Arbeiten im Bereich der Philosophie assistiert und gestützt werden. Dabei könnte es sich als wesentlich erweisen, dass in weiten Teilen der Philosophie des Geistes und der Sprachphilosophie Fragen nach Bedeutung mit der These des semantischen Externalismus verknüpft sind. Nach dieser These supervenieren Bedeutungen nicht über neuronalen Zuständen allein, sondern hängen auch von Faktoren der physischen und sozialen Umgebung ab. Es ist eine bis heute offene Frage, ob und wie dies mit Fragen der neuronalen Codierung zusammenhängt.

Die Neurowissenschaften und die Ebene des Behavioralen, Psychischen und Sozialen

Die wohl wichtigste Zukunftsaufgabe der Neurowissenschaften ist, den Anschluss zu den höherstufigen Wissenschaften wie der kognitiv und verhaltensorientierten Psychologie, der Psychiatrie und den Kognitionswissenschaften insgesamt herzustellen. In diesen Disziplinen nimmt man typischerweise Dennetts intentionale Einstellung ein. Eine Verbindung neurowissenschaftlicher Kategorien mit psychischen, kognitiven, behavioralen und sozialen Kategorien geht insofern mit einer Klärung der Fragen der Semantik und damit der Rückbindung an den vorigen Abschnitt einher. Um etwa zu entscheiden, wie neuronale und zerebrale Zustände und Aktivitäten mit psychischen Zuständen und Aktivitäten, auch mit psychischen Störungen, zusammenhängen, ist es gleichermaßen entscheidend, den Code des Gehirns zu verstehen, wie auch die psycho-neuronale Supervenienz zu überdenken. Die Berücksichtigung einer breiten transkraniellen Supervenienzbasis könnte, durchaus im Einklang mit einem geläuterten Naturalismus und Reduktionismus, eventuell dazu beitragen, die Zukunftsaufgaben der Neurowissenschaften und ihrer Philosophie anzugehen.

C3. Literaturempfehlungen

Bechtel, William, Peter Mandik, Jennifer Mundale & Robert S. Stufflebeam, Hg. (2001). Philosophy and the Neurosciences: A Reader. Oxford: Blackwell.

Bickle, John, Hg. (2009). Oxford Handbook of Philosophy and Neuroscience. New York: Oxford University Press.

Bickle, John, Peter Mandik & Anthony Landreth (2012). The Philosophy of Neuroscience". In Edward N. Zalta (ed.): The Stanford Encyclopedia of Philosophy. URL = <http://plato.stanford.edu/archives/sum2012/entries/neuroscience/>.

Bickle, John, Hg. (2004, 2005, 2006); Kenneth Aizawa & John Bickle, Hg. (2009); Gualtiero Piccinini, Hg. (2011, 2012, 2014). Special Issue: Neuroscience and its Philosophy. Synthese.

Brook, Andrew & Kathleen Akins, Hg. (2005). Cognition and the Brain: The Philosophy and Neuroscience Movement. Cambridge: Cambridge University Press.

Machamer, Peter, Rick Grush & Peter McLaughlin, Hg. (2001). Theory and Method in Neuroscience. Pittsburgh: University of Pitt Press.

Pauen, Michael & Gerhard Roth, Hg. (2001). Neurowissenschaften und Philosophie: Eine Einführung. München: Fink (UTB 2208).

Walter, Henrik (2013). Neurophilosophie und Philosophie der Neurowissenschaften. In A. Stephan & S. Walter (Hg.): Handbuch Kognitionswissenschaft. Stuttgart: Metzler.

Literatur

Aizawa, Kenneth & Carl Gillett (2009). Levels, individual variation, and massive multiple realization in neurobiology. In: Bickle (2009), 529–581.

Baars, Bernard (1988). A Cognitive Theory of Consciousness. New York: Cambridge University Press.

Balzer, Wolfgang, C. Ulises Moulines & Joseph D. Sneed (1987). An Architectonic for Science: The Structuralist Approach. Dordrecht: Reidel.

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

Bechtel, William (2012). Understanding endogenously active mechanisms: A scientific and philosophical challenge. European Journal for the Philosophy of Science 2: 233-248

Bechtel, William (2014). Investigating neural representations: the tale of place cells. Synthese, online, DOI 10.1007/s11229-014-0480-8.

Bechtel, William & Robert C. Richardson (1993, ^{2}2010). Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research. Cambridge, MA: MIT Press.

Bechtel, William & Adele Abrahamsen (1991, ^{2}2002). Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks. Oxford: Blackwell.

Bechtel, William & Jennifer Mundale (1999). Multiple realizability revisited: linking cognitive and neural states. Philosophy of Science 66: 175–207.

Beckermann, Ansgar (2008). Gehirn, Ich, Freiheit. Neurowissenschaften und Menschenbild. Paderborn: Mentis.

Bennett, Max, Daniel Dennett, Peter M. S. Hacker & John Searle (2007). Neuroscience and Philosophy. Brain, Mind and Language. New York: Columbia University Press.

Bennett, Max & Peter M. S. Hacker (2003). Philosophical Foundations of Neuroscience. Oxford: Blackwell.

Bennett, Max & Peter M. S. Hacker (2008). History of Cognitive Neuroscience. Oxford: Blackwell.

Bickle, John (1998). Psychoneural Reduction: The New Wave. Cambridge, MA: MIT Press.

Bickle, John (2003). Philosophy and Neuroscience: A Ruthlessly Reductive Account. Dordrecht: Kluwer.

Bogen, James & James Woodward (1988). Saving the phenomena. Philosophical Review 97: 303-352.

Bogen, James (2002). Epistemological custard pies from functional brain imaging. Philosophy of Science 69 (3): S59-S71.

Call, Josep & Michael Tomasello (2008). Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Science 12: 187-192.

Carruthers, Peter (2006). The Architecture of the Mind. Oxford: Oxford University Press.

Carruthers, Peter (2007). The illusion of conscious will. Synthese 96(2): 197-213.

Chemero, Anthony & Michael Silberstein (2008). After the philosophy of mind: replacing scholasticism with science. Philosophy of Science 75: 1-27.

Chirimuuta, Mazviita (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. Synthese 191: 127-153.

Choudhury, Suparna & Jan Slaby, Hg. (2011). Critical Neuroscience A Handbook of the Social and Cultural Contexts of Neuroscience. Chichester: Wiley-Blackwell.

Churchland, Patricia S. (1986). Neurophilosophy: Toward a Unified Science of the Mind-Brain. Cambridge, MA: MIT Press.

Churchland, Patricia S. & Terrence J. Sejnowski (1992). The Computational Brain. Cambridge, MA: MIT Press.

Churchland, Paul M. (1989). A Neurocomputational Perspective. Cambridge, MA: MIT Press.

Clark, Andy (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral & Brain Sciences 36(3): 181–204.

Colombo, Matteo (2013). Moving forward and beyond the modularity debate: a network perspective. Philosophy of Science 80(2): 356–377.

Colombo, Matteo & Peggy Seriès (2012). Bayes in the brain — on Bayesian modelling in neuroscience. The British Journal for the Philosophy of Science 63(3): 697–723.

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

Craver, Carl (im Druck): Thinking about interventions: optogenetics and makers' knowledge of the brain. In K. Waters et. Al (Hg.): Causation in Biology and Philosophy. Minnesota Studies in Philosophy of Science. Minneapolis, MN, University of Minnesota Press.

Crick, Francis & Christof Koch (1990). Towards a neurobiological theory of consciousness. Seminars in Neuroscience 2, 263-275.

Cummins, Robert C. (1983). The Nature of Psychological Explanation. Cambridge, MA: MIT Press.

Cummins, Robert C. (1989). Meaning and Mental Representation. Cambridge, MA: MIT Press.

Dayan, Peter & L. F. Abbott (2001). Theoretical Neuroscience. Cambridge, MA: MIT Press.

Dennett, Daniel C. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

Dennett, Daniel C. (1991). Consciousness Explained. Boston, MA: Little, Brown and Co.

Elger, Christian E., Angela D. Friederici, Christof Koch, Heiko Luhmann, Christoph von der Malsburg, Randolf Menzel, Hannah Monyer, Frank Rösler, Gerhard Roth, Henning Scheich & Wolf Singer (2004). Das Manifest. Elf führende Neurowissenschaftler über Gegenwart und Zukunft der Hirnforschung. Gehirn und Geist 6: 30-36.

Eliasmith, Chris & Charles H. Anderson (2003). Neural Engineering. Cambridge, MA: MIT Press.

Fields, R. Douglas (2010). The Other Brain. New York: Simon & Schuster.

Figdor, Carrie (2010). Neuroscience and the multiple realization of cognitive functions. Philosophy of Science 77 (3).419-456.

Fodor, Jerry (1983): The Modularity of Mind: An Essay on Faculty Psychology. Cambridge, MA: MIT Press.

Friston, Karl (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11(2): 127-38.

Gallagher, Shaun & Dan Zahavi (2008). The Phenomenological Mind. An Introduction to Philosophy of Mind and Cognitive Science. London: Routledge.

Gallese, Vittorio & Alvin I. Goldman (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences 12: 493-501.

Garson, James W. (2001). (Dis)solving the binding problem. Philosophical Psychology 14(4): 381-392.

Gazzaniga, Michael S., Richard B. Ivry & George R. Mangun (2014, 4th edition). Cognitive Neuroscience: The Biology of the Mind. New York: W.W. Norton.

Gerstner, Wulfram, Werner M. Kistler, Richard Naud & Liam Paninski (2014). Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition. Cambridge: Cambridge University Press.

Geyer, Christian (2004). Hirnforschung und Willensfreiheit. Frankfurt a.M.: Suhrkamp.

Glauer, Ramiro (2012). Emergent mechanisms. Münster: mentis.

Goldman, Alvin I. (2012). Theory of mind. In: E. Margolis, R. Samuels & S. P. Stich (Hg.). Oxford Handbook of Philosophy of Cognitive Science. Oxford: Oxford University Press.

Hacking, Ian (1983). Representing and Intervening: Introductory Topics in the Philosophy of Natural Science. Cambridge: Cambridge University Press.

Hanson, Stephen José & Martin Bunzl, Hg. (2010): Foundational Issues in Human Brain Mapping. Cambridge, MA: MIT Press.

Hardcastle, Valerie G. (1994). The binding problem and possible solutions. Journal of Consciousness Studies 1: 66–90.

Hardcastle, Valerie G. & C. Matthew Steward (2002). What do brain data really show? Philosophy of Science 69(3): S72-S82.

Hardcastle, Valerie G. & C. Matthew Steward (2003). Neuroscience and the art of single cell recordings. Biology and Philosophy 18(1): 195-208.

Hebb, Donald (1949): The Organization of Behaviour. New York: Wiley.

Hemmen, J. Leo van & Terrence J. Sejnowski (2006). 23 Problems in Systems Neuroscience. Oxford: Oxford University Press.

Hohwy, Jakob (2013). The Predictive Mind. Oxford: Oxford University Press.

Hooker, Clifford A. (1981). Towards a general theory of reduction. Dialogue 20: 38-59, 201-236, 496-529.

Hubel, David H. & Torsten N. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cats visual cortex. Journal of Physiology 160: 106-154.

Hurley, Susan & Matthew Nudds, Hg. (2006). Rational Animals? Oxford: Oxford University Press.

Jacob, Pierre (2008). What do mirror neurons contribute to human social cognition? Mind and Language 23(2): 190-223.

Kandel Eric R., Henry Markram, Paul M. Matthews, Rafael Yuste & Christof Koch (2013). Neuroscience Thinks Big (and Collaboratively). Nature Reviews Neuroscience 14(9): 659-664.

Kandel, Eric R., James H. Schwartz & Thomas M. Jessel (1995). Neurowissenschaften. Eine Einführung. Heidelberg: Spektrum.

Kaplan, David M. (2011). Explanation and description in computational neuroscience. Synthese 183: 339–373.

Kaplan, David M. & Carl F. Craver (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. Philosophy of Science 78: 601-627.

Karnath, Hans-Otto & Peter Thier, Hg. (2003, ^{3}2012). Kognitive Neurowissenschaften. Berlin: Springer.

Klein, Colin (2010). Images are not the evidence in neuroimaging. British Journal for the Philosophy of Science 61(2): 265-278.

Klein, Colin (2012). Cognitive ontology and region- versus network-oriented analyses. Philosophy of Science 79: 952-960.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27(12): 712–719.

Libet, Benjamin (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. The Behavioral and Brain Sciences 8: 529-539.

Logothetis, Nikos K. (2008). What we can do and what we cannot do with fMRI. Nature 453: 869–78.

Lurz, Robert W. (2009): The Philosophy of Animal Minds. Cambridge: Cambridge University Press.

Lurz, Robert W. (2011). Mindreading Animals. Cambridge, MA: MIT Press.

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

Lyre, Holger (2013a). Reduktionismus, Multirealisierbarkeit und höherstufige Näherungen. In: J. Michel & G. Münster (Hg.): Die Suche nach dem Geist. Mentis, Münster, 55-80.

Lyre, Holger (2013b). Verkörperlichung und situative Einbettung. In: A. Stephan und S. Walter (Hg.): Handbuch Kognitionswissenschaft. Stuttgart: Metzler, 186-192.

Maass, Wolfgang (1996). Networks of Spiking Neurons: The Third Generation of Neural Network Models. Neural Networks 10(9) 1659-1671.

Maass, Wolfgang & Christopher M. Bishop (1999). Pulsed Neural Networks. Cambridge, MA: MIT Press.

Machery, Eduard (2014). Significance testing in neuroimagery. In M. Sprevak & J. Kallestrup (eds.): New Waves in the Philosophy of Mind. London: Palgrave Macmillan, S. 262-277.

Machamer, Peter, Lindley Darden & Carl F. Craver (2000). Thinking about mechanisms. Philosophy of Science 67(1): 1-25.

Malsburg, Christoph von der (1999). The what and why of binding: the modeler's perspective. Neuron 24: 95–104.

Malsburg, Christoph von der, William A. Phillips & Wolf Singer (2010). Dynamic Coordination and the Brain: From Neurons to Mind. Cambridge, MA: MIT Press.

Marcus, Gary & Jeremy Freeman, Hg. (2014). The Future of the Brain: Essays by the World's Leading Neuroscientists. Princeton: Princeton University Press.

Marr, David (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York: Freeman.

Metzinger, Thomas (2003). Being No One. The Self-Model Theory of Subjectivity. Cambridge, MA: MIT Press.

Metzinger, Thomas (2009). Der Ego-Tunnel. Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsethik. Berlin: Berlin Verlag.

Milkowski, Marcin (2013). Explaining the Computational Mind. Cambridge, MA: MIT Press.

Mole, Christopher & Colin Klein (2010). Confirmation, refutation, and the evidence of fMRI. In: Hanson & Bunzl (2010), 99-111.

Moser, May-Britt Moser & Edvard I. Moser (2014). Understanding the cortex through grid cells. In: Marcus & Freeman (2014), S. 67-77.

Nagel, Ernest (1961). The Structure of Science, New York: Harcourt, Brace, and World.

Nordlie, Eilen, Marc-Oliver Gewaltig & Hans Ekkehard (2009). Towards reproducible descriptions of neuronal network models. PLoS Computational Biology 5(8): e1000456.

Pauen, Michael & Gerhard Roth (2008). Freiheit, Schuld und Verantwortung. Grundzüge einer naturalistischen Theorie der Willensfreiheit. Frankfurt a.M.: Suhrkamp.

Piccinini, Gualtiero (2015). Physical Computation: A Mechanistic Account. Oxford: Oxford University Press.

Pietsch, Wolfgang (2015). Aspects of theory-ladenness in data-intensive science. Philosophy of Science 82: 905–916.

Povinelli, Daniel J. & Jennifer Vonk (2004). We don't need a microscope to explore the chimpanzee's nind. Mind and Language 19(1): 1-28. (In: Hurley & Nudds 2006, 1-28).

Price, Cathy & Karl Friston (2005). Functional ontologies for cognition: the systematic definition of structure and function. Cognitive Neuropsychology 22(3): 262–275.

Quiroga, Rodrigo (2012). Concept Cells: The Building Blocks of Declarative Memory Functions. Nature Reviews Neuroscience 13: 587-597.

Quiroga, Rodrigo, Leila Reddy, Gabriel Kreiman, Christof Koch & Itzhak Fried (2005). Invariant visual representation by single neurons in the human brain. Nature 435: 1102-1107.

Raichle, Marcus E. & Abraham Z. Snyder (2007). A default mode of brain function: a brief history of an evolving idea. Neuroimage 37(4): 1083-1090.

Roskies, Adina L. (2010). Neuroimaging and Inferential Distance: The Perils of Pictures. In: Hanson & Bunzl 2010, Chap. 16.

Roskies, Adina L. (2007). Are neuroimages like photographs of the brain? Philosophy of Science 74: 860–72.

Salmon, Wesley (1989). Four Decades of Scientific Explanation, Minneapolis: University of Minnesota Press.

Schaffner, Kenneth F. (1967). Approaches to reduction. Philosophy of Science 34(2): 137-147.

Schutter, Dennis J. L. G., Jack Van Honk & Jaak Panksepp (2004). Introducing transcranial magnetic stimulation (TMS) and its property of causal inference in investigating brain-function relationships. Synthese 141 (2): 155-173.

Shagrir, Oron (2010). Marr on computational-level theories. Philosophy of Science 77(4): 477-500.

Shapiro, Lawrence A. (2000). Multiple realizations. Journal of Philosophy 97(12): 635–654.

Shapiro, Lawrence A. (2008). How to test for multiple realization. Philosophy of Science: 75: 514–525.

Silberstein, Michael & Anthony Chemero (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. Philosophy of Science 80: 958–970.

Sporns, Olaf (2011). Networks of the Brain. Cambridge, MA: MIT Press.

Stanley, Garrett B. (2013). Reading and writing the neural code. Nature Neuroscience 16: 259-263.

Steinle, Friedrich (1997). Entering new fields: exploratory uses of experimentation. Philosophy of Science 64: S65-S74.

Stelzer, Johannes, Gabriele Lohmann, Karsten Mueller, Tilo Buschmann & Robert Turner (2014): Deficient approaches to human neuroimaging. Frontiers in Human Neuroscience 8: 462.

Stepp, Nigel, Anthony Chemero & Michael T. Turvey (2011). Philosophy for the rest of cognitive science. Topics in Cognitive Science 3: 425-437.

Stufflebeam, Robert S. & William P. Bechtel (1997). PET: Exploring the myth and the method. Philosophy of

Science 64(4): 95-106.

Tononi, Giulio 2004: An information integration theory of consciousness. BMC Neuroscience 5: 42.

Trappenberg, Thomas P. (2002, ^{2}2010). Fundamentals of Computational Neuroscience. Oxford: Oxford University Press.

Treisman, Anne & Garry Gelade (1980). A feature-integration theory of attention. Cognitive Psychology 12: 97-136.

Tooby, John & Leda Cosmides (1995). Mapping the evolved functional organization of mind and brain. In: M. Gazzaniga (Hg.): The cognitive neurosciences. Cambridge, MA: MIT Press.

Vreeswijk, Carl van (2006). What is the neural code? In: Hemmen & Sejnowski (2006), 143-159.

Varela, Francisco J., Evan Thompson & Eleanor Rosch (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: MIT Press.

Vul, Edward & Nancy Kanwisher (2010). Begging the question: The non-independence error in fMRI data analysis. In: Hanson & Bunzl (2010), 71-91.

Walter, Henrik, Hg. (2005). Funktionelle Bildgebung in Psychiatrie und Psychotherapie: Methodische Grundlagen und klinische Anwendungen. Stuttgart: Schattauer.

Walter, Sven (2016). Illusion freier Wille? Grenzen einer empirischen Annäherung an ein philosophisches Problem. Stuttgart: Metzler.

Winsberg, Eric (2010). Science in the Age of Computer Simulation. Chicago: University of Chicago Press.

Woodward, James (2008). Mental causation and neural mechanisms. In: J. Hohwy & J. Kallestrup (Hg.): Being Reduced: New Essays on Reduction, Explanation, and Causation. Oxford: Oxford University Press.

Zednik, Carlos (2011). The nature of dynamical explanation. Philosophy of Science 78: 238-263.