Abklärung des möglichen Beitrags der Neurowissenschaft und der Verhaltensforschung zum Verständnis moralischer Orientierung

Pilotstudie im Rahmen des Forschungsprojekts "Grundlagen moralischer Orientierung" des Ethik-Zentrums der Universität Zürich

Markus Christen

November 2005

Überarbeitet: April 2006

Zürich, Biel
November 2005, April 2006

Verfasser:
Markus Christen, Dr. sc. ETH
Institut für Neuroinformatik
Winterthurerstrasse 190
8057 Zürich
markus@ini.phys.ethz.ch

Atelier Pantaris
Bözingenstrasse 5
2502 Biel
Tel: +41 32 342 65 46
Fax: +41 32 342 65 47
markus.christen@pantaris.ch

Zusammenfassung

Die Suche nach biologischen Grundlagen von moralischem Verhalten ist in jüngerer Zeit vermehrt Gegenstand der Neurowissenschaft und Verhaltensforschung geworden. Im Rahmen des Schwerpunktes Ethik der Universität Zürich ist das Projekt "Grundlagen moralischer Orientierung" lanciert worden, das unter anderem die Relevanz dieser Forschung für genuin ethische Fragen untersuchen soll. Die vorliegende Arbeit dient als Pilotstudie für dieses Projekt. Sie soll aufgrund einer Analyse der aktuellen Forschungsliteratur aufzeigen, wie Moral in den letzten Jahren Gegenstand der Neurowissenschaft und Verhaltensforschung geworden ist, welche konkreten Fragestellungen untersucht werden, welche Methoden dafür Anwendung finden und welche Ergebnisse bisher erzielt worden sind. Basierend auf dieser Zusammenstellung sollen Forschungsfragen und potentielle Kooperationspartner im Raum Zürich identifiziert werden, welche für das Projekt von Nutzen sein können.

Die Pilotstudie beginnt mit einer Schärfung der Kernbegriffe "Ethik" bzw. "Moral" im Kontext der zu untersuchenden Arbeiten und einer thematischen Abgrenzung der Literaturrecherche. Aufgrund der untersuchten empirischen Studien werden vier Aspekte genauer vorgestellt: Erstens stellt sich die Frage, welche Stimuli bei Experimenten als "moralisch" qualifiziert werden und aufgrund welcher Kriterien dies geschieht. Zweitens wird untersucht, welche Mechanismen postuliert werden, die dem *moral decision making* zugrunde liegen sollen, und mit welchen Methoden dieser Prozess analysiert werden soll. Drittens stellt sich die Frage nach der empirischen Erfassung einer "moralischen Handlung" im Rahmen eines Experiments bzw. nach den Kriterien, die eine solche Handlung als "moralisch" qualifizieren. Viertens wird analysiert, welchen Stellenwert zentrale Elemente der philosophischen Ethik wie Begründungen, normative Theorien etc. im Rahmen solcher Untersuchungen bzw der daraus folgenden Erklärungsmodelle haben. Kaum eine der untersuchten Studien bezieht explizit zu allen vier Aspekten Stellung. Vielmehr kann eine eher unscharfe Begriffsverwendung festgestellt werden.

Da im Zug eines anhaltenden Booms von *Imaging*-Studien komplexe Verhaltensweisen von Menschen zunehmend Gegenstand der Neurowissenschaft geworden sind, war eine thematische Abgrenzung unumgänglich. Die grosse Mehrheit der untersuchten Studien benutzten zu einem wesentlichen Teil *Imaging* – also Techniken zur Visualisierung neurobiologischer Vorgänge *in vivo*. Da mit dieser Technik eine Reihe methodischer Probleme verbunden sind, werden in dieser Pilotstudie die verschiedenen Varianten des *Imaging*, wie auch Methoden der experimentellen Ökonomie vorgestellt. Zeitlich wurde der Fokus auf Arbeiten der vergangenen fünf bis zehn Jahre gelegt. Gewiss war Moral bzw. moralisches Verhalten bereits Jahrzehnte zuvor in einer Reihe von naturwissenschaftlichen Gebieten – vorab in der Neurologie und der Soziobiologie – ein Forschungsgegenstand. Eine

historisch fundierte Studie konnte in dieser Pilotstudie nicht geleistet werden. Eine bibliometrische Untersuchung zeigt aber, dass die Frage nach den neurobiologischen Grundlagen komplexer Verhaltensweisen, wie eben moralisches Verhalten, erst in jüngster Zeit ein stärker diskutiertes Thema der Neurowissenschaft geworden ist. Umfassend abgedeckt wurden *Imaging*-Studien, welche explizit moralisches Verhalten als Forschungsgegenstand wählten. Nur teilweise abgedeckt wurden Forschungsfelder, welche im Umfeld solcher Studien angesiedelt werden, vorab die *social cognitive neuroscience*, die Emotionsforschung und die Erforschung von Spiegelneuronen. Im Rahmen dieser Übersicht werden auch Studien über die Untersuchung moralnaher Verhaltensweisen bzw. Fähigkeiten wie Religiosität, Empathie, Intuition, Bedauern, Enttäuschung, Lügen, Vertrauen, Kooperation und Altruismus vorgestellt. Basierend auf den oben dargestellten vier Aspekten einer empirischen Untersuchung der Moral werden dann Ansätze zu einem Modell der neuronalen Grundlagen der Moral präsentiert.

Die Resultate vieler dieser Studien zeigen sehr deutlich die Probleme auf, welche bei der empirischen Untersuchung eines komplexen sozialen Verhaltens wie "moralisch Handeln" auftreten. In den meisten Fällen wird ein recht einfacher Begriff von "Moral" verwendet. Die *Imaging*-Studien erreichen in der Regel kaum mehr als eine Lokalisierung der – erwartungsgemäss zahlreichen – Hirnregionen, welche bei *moral decision making* involviert sein sollen. Bei so mancher Studie treten auch grundlegende skeptische Fragen hinsichtlich des Studiendesigns und der Aussagekraft der erzielten Resultate auf. Interessanter erscheinen jene Studien, welche *Imaging* mit einem (spieltheoretischen) Verhaltensparadigma verknüpfen, um damit moralnahe Verhaltensweisen wie Vertrauen zu untersuchen, da dieses Verhalten damit besser quantifiziert werden kann. In theoretischer Hinsicht favorisieren die vorgeschlagenen Modelle einen Automatismus beim *moral decision making*, wonach unbewusste Prägungen moralischer Stimuli den Entscheidungsprozess vorspuren und die daraus folgende Handlung erst *post faktum* mit einer rationalen Erklärung unterlegt würden.

Im Zug der Literaturrecherchen hat sich ergeben, dass ethische Probleme der Neurowissenschaft im Zug der so genannten Neuroethik etwa ab 2002 vermehrt diskutiert wurden. Aus diesem Grund wurde der Neuroethik ein eigenes, zusätzliches Kapitel gewidmet, um eine Übersicht über den Diskussionsstand in diesem Bereich zu geben. Als zentrale ethische Problemfelder haben sich hierbei der Umgang mit *Imaging*-Daten und künftig mögliche Eingriffe in das Gehirn im Zug einer möglichen Renaissance der Psychochirurgie, der Fortschritte in der Neuroprothetik und den zunehmenden Möglichkeiten des *neural enhancement* ergeben. Auf die seit längerer Zeit laufende Debatte um Auswirkungen der Neurowissenschaft auf Recht und Philosophie wird nur am Rande eingegangen.

Die Studie schliesst mit einer Ausformulierung des Grundproblems, welches sich die Wissenschaftler im Bereich "Neurobiologie der Moralfähigkeit" stellen. Darauf aufbauend wird aufgezeigt, worin die Probleme in der bisherigen Forschung bestehen – vorab eine Art Blindheit gegenüber den kategorial verschiedenen Problemen, die in diesen Forschungskomplex involviert sind. Die Untersuchung der Grundlagen der moralischen Orientierung würden demnach nicht nur die Neurobiologie des *moral agent* umfassen. Benötigt werden auch eine Phänomenologie des *moral agent*, eine Theorie der Interaktion solcher *agents*, eine Ethologie des Moralverhaltens und den Einschluss einer kulturwissenschaftlichen Perspektive im Sinn einer Kulturgeschichte der Moral. Für all diese Bereiche werden im Raum Zürich interessierte Wissenschaftler genannt, welche bei einer umfassenderen Untersuchung der Grundlagen moralischer Orientierung beteiligt werden könnten.

Contents

Chapter 1

Einführung

1.1 Auftrag und Zielsetzung

Welche Rolle spielt die Intuition bei der Bildung und dem Gebrauch der moralischen Urteilsfähigkeit beim Menschen? Diese Frage steht am Beginn des Projektes "Grundlagen moralischer Orientierung", das im Rahmen des Forschungsschwerpunkts Ethik der Universität Zürich durchgeführt werden soll. Diese und weiterführende Fragestellungen, welche im Antrag für das angesprochene Projekt genauer ausgeführt werden [68], berühren Forschungsgegenstände von Neurowissenschaft, Verhaltensforschung, Moralpsychologie und Affektforschung. In all diesen Gebieten lässt sich ein (z.T. zunehmendes) Interesse an den biologischen Grundlagen der Moralfähigkeit des Menschen feststellen. Zwei Pilotstudien sollen aufzeigen, welche genauen Fragestellungen in den einzelnen Gebieten auf welche Weise untersucht werden. Die vorliegende Pilotstudie befasst sich mit den ersten beiden oben genannten Gebieten, die zweite Pilotstudie fokussiert die neueren Entwicklungen in Moralpsychologie und Affektforschung. Die Abgrenzungen der beiden Pilotstudien ist jedoch nicht immer klar, eine gewisse inhaltliche Überschneidung ist also wahrscheinlich. Im Einzelnen umfasst der Gegenstand des Auftrags folgende Punkte:

1. Die Durchführung einer Literaturrecherche, welche den aktuellen Stand der Forschung in der Neurowissenschaft und Verhaltensforschung bezüglich des Verständnisses moralischer Orientierung, insbesondere von deren intuitiver Komponente, ermittelt.

2. Die Auswertung dieser bereits existierenden Forschungsergebnisse.

3. Die Formulierung von Fragestellungen für weiterführende Forschungen im Kontext des Projekts "Grundlagen moralischer Orientierung".

4. Die Identifizierung geeigneter Projektpartner (bevorzugt im Bereich der Universität Zürich), die von ihrem Forschungsprofil her für eine Kooperation und für die Bearbeitung der weiterführenden Fragestellungen in Betracht kommen.

Der Begriff der "moralischen Orientierung" ist dabei im Auftragstext mit folgenden Fragen enger gefasst worden:

- Was bezeichnet der Begriff "moralische Intuition" und in welcher Beziehung steht er zu Begriffen wie "Gefühl" und "Affekt"?
- Auf welche biologischen Zusammenhänge lässt sich die intuitive Orientierung zurückführen?
- Wie wirken biologische und soziale Faktoren bei der Prägung der moralischen Orientierung?
- Wie greifen intuitive und rationale Orientierung ineinander?
- Welche Bedeutung kommt der intuitiven Komponente der Moral für die Moralerziehung zu?

Im Rahmen der Literaturrecherche hat sich ergeben, dass die begriffliche Unschärfe hinsichtlich von Konzepten wie "Moral/moralisch", "Intuition", "Gefühl/Emotion", etc. in den einzelnen untersuchten Studien erheblich ist. Weiter lässt sich in historischer Hinsicht feststellen, dass die Moralfähigkeit des Menschen bereits früher von verschiedenen Gebieten untersucht worden ist. So wurde beispielsweise die Frage nach den evolutionären Wurzeln moralrelevanter Fähigkeiten wie Altruismus oder Empathie bereits im 19. Jahrhundert thematisiert. Sie fand im Rahmen der sich entwickelnden Soziobiologie in den 1970er Jahren neue Relevanz und wurde danach in der Ethik im Kontext der evolutionären Ethik in den 1980er Jahren breit diskutiert. Die Frage nach den neuronalen Korrelaten moralischen Verhaltens wiederum ist vergleichsweise jung und hängt einerseits mit der Renaissance der Emotionen in der Hirnforschung in den 1990er Jahren zusammen (mehr dazu im Abschnitt 2.3.2), andererseits mit den neuen Möglichkeiten der bildgebenden Verfahren. Für die Umsetzung des Auftrags sind demnach folgende Präzisierungen notwendig geworden:

- Die begrifflichen Unschärfen verlangten nach einer vorgängigen Präzisierung des ethischen Begriffsfeldes. Diese orientiert sich an den in der heutigen Ethik gängigen Kategorien, setzt aber eigene Schwerpunkte aufgrund der Forschungsgegenstände in den einzelnen untersuchten Studien. Diese Präzisierung soll es erlauben, die untersuchten Studien besser einzuordnen, damit die Bezüge zu den in der philosophischen Ethik untersuchten Fragen deutlicher werden.
- Je nach gewähltem Suchkriterium (z.B. "Emotion") wächst die zu beachtende Literatur ins uferlose. Es wurde demnach nötig, die Literaturrecherche wie folgt einzugrenzen:
  - Die Untersuchung der neuronalen Korrelate von moralischem Verhalten mittels bildgebender Verfahren (*Imaging*) ist ein neues und überschaubares Forschungsgebiet, das im wesentlichen abgedeckt werden konnte. Da die Gültigkeit der von diesen Wissenschaftlern erzielten Resultate stark von den verwendeten Methoden abhängen, werden die Verfahren wie Grundprobleme des *Imaging* ebenfalls aufgezeigt.
  - Die Untersuchung "moralischer Pathologien" ist ein Klassiker der Neuropsychologie (mehr dazu in Abschnitt 3.4.1). Gemeint ist damit die In-Bezug-Setzung bzw.

Erklärung von moralisch abnormem Verhalten (was zweifellos nicht immer eindeutig bestimmbar ist) mit bestimmten Hirnschäden. Dieses Gebiet kann nicht umfassend abgehandelt werden, zumal hier auch eine Überschneidung mit der zweiten Pilotstudie zu erwarten ist. Im Rahmen dieser Studie wurden einige neuere Arbeiten seit dem Ende der 1990er Jahre ohne Anspruch auf Vollständigkeit untersucht.

- Die Erforschung der neuronalen Grundlagen des *decision-making* hat im Rahmen der sich entwickelnden *social cognitive neuroscience* in den vergangenen Jahren einen enormen Aufschwung erfahren. Dieses Forschungsgebiet kann nur in seinen Grundzügen skizziert werden, wobei einerseits aktuelle Übersichtsarbeiten verwendet wurden, andererseits Studien, in denen sich ein Bezug zu *moral decision-making* finden liess.
- Dieselbe Einschränkung gilt für die Erforschung von Emotionen. Erstaunlicherweise hat sich gezeigt, dass der Begriff der "Intuition" in den neurowissenschaftlichen Arbeiten nur selten explizit Verwendung findet. Meistens taucht er im Kontext der Emotionsforschung auf und bezeichnet eine Komponente in einer "vor-bewussten neuronalen Maschinerie", welche zu bestimmten Prädispositionen im betreffenden Subjekt führen (z.B. ein gewisses moralisches Urteil vorspuren). In der vorliegenden Studie wurde vorab jene Arbeiten untersucht, welche einen prägenden Einfluss vorbewusstlicher, emotions-gesteuerter neuronaler Prozesse auf vermeintlich rationale Entscheidungen behaupten.
- Die (unter anderem) in den Verhaltenswissenschaften aufgekommene Frage nach den evolutionären Ursprüngen von Moral ist in dieser Studie ebenfalls nicht umfassend abgehandelt worden. Es zeigte sich, dass in den vergangenen Jahren Zugänge unter Einbezug der experimentellen Ökonomie für die Klärung dieser Frage zunehmend an Bedeutung gewonnen haben. Untersucht wurde hier vorab die Entstehung von Kooperation und Vertrauen in menschlichen Gemeinschaften. Die vorliegende Pilotstudie konzentrierte sich auf diese neueren Arbeiten.
- Die vergleichende Verhaltenswissenschaft schliesslich stellt sich die Frage, ob sich Vorformen von Moral auch in Tiergemeinschaften – vorab Primaten – finden lassen. Das Interesse an solchen Fragen geht (mindestens) auf die 1960er Jahre zurück. Für diese Studie wurden neuere Arbeiten ohne Anspruch auf Vollständigkeit untersucht.

- Im Rahmen der Studie hat sich weiter ergeben, dass Fragen der Neuroethik – die mutmasslichen ethischen Probleme, die sich aus der weiteren Entwicklung der Neurowissenschaft ergeben können – in den vergangenen Jahren ein starkes Interesse gefunden haben. Nach Rücksprache mit dem Auftraggeber wurde deshalb beschlossen, eine kurze Übersicht zu den als relevant betrachteten Problemen der Neuroethik zu geben.

Das Verständnis der einzelnen Studien verlangt nach gewissen Grundkenntnissen in der Hirnanatomie und den in den Neuro- und Verhaltenswissenschaften verwendeten Methoden (insbesondere *Imaging*, experimentelle Ökonomie). Da diese Pilotstudie auch als Arbeitsinstrument im Rahmen des Projektes "Grundlagen moralischer Orientierung" nützlich sein soll, wird dazu eine kurze Einführung gegeben.

1.2 Methoden

Die vorliegende Pilotstudie stützt sich auf eine Literaturrecherche und einer Expertenbefragung. In folgenden Fachzeitschriften wurde eine Volltextanalyse für den Zeitraum 2000 bis 2005 durchgeführt.

Annual Review in Neuroscience	Nature Neuroscience
Behavioral and Brain Research	Nature Reviews: Neuroscience
Behavioral and Brain Science	NeuroImage
Behavioral Neuroscience	Neuron
Biology and Philosophy	Proceedings of the National Academy
Current Opinion in Neurobiology	of Sciences USA*
Human Brain Mapping	Progress in Neurobiology
International Journal for Psychophysiology	Science*
Journal of Cognitive Neuroscience	The Journal of Neuroscience
Journal of the History of Behavioral Sciences	Trends in Cognitive Sciences
Nature*	Trends in Neurosciences

Table 1.1: Untersuchte Zeitschriften; *: Nur Stichwortsuche.

Früher erschienene Ausgaben wurden mit Stichwortsuche (unter Benutzung der Begriffe "ethic", "ethical", "moral") in den Suchfunktionen der jeweiligen Online-Zeitschrift-Archive geprüft. Eine solche Stichwortsuche wurde auch in der *MedLine*[1] Datenbank der *National Institutes of Health* (USA) durchgeführt, wodurch auch Beiträge in anderen als den oben aufgelisteten Zeitschriften identifiziert wurden. Es wurde praktisch ausschliesslich englischsprachige Literatur ausgewertet. Im Rahmen der Pilotstudie wurde nicht systematisch untersucht, inwieweit die Erforschung der neurobiologischen Grundlagen der Moral innerhalb der philosophischen Ethik thematisiert wurde, denn dies hätte den Rahmen der Pilotstudie gesprengt. Für das Erstellen der Pilotstudie wurde zudem mit insgesamt zehn Experten unterschiedlicher Gebiete der Universität Zürich Vorgespräche geführt. Es handelt sich dabei um folgende Personen, denen an dieser Stelle auch herzlich für Ihre Teilnahme gedankt werden soll:

- Peter Brugger (Neuropsychologie, UniversitätsSpital Zürich)
- Urs Fischbacher (experimentelle Ökonomie, Universität Zürich)
- Alumit Ishai (Neuro-Imaging, Universität Zürich)
- Lutz Jänke (Neuropsychologie, Universität Zürich)
- Daniel Kiper (Sinnesphysiologie, Universität Zürich)
- Eric Kubli (Zoologie, Universität Zürich)
- Kevan Martin (Neurophysiologie, Universität Zürich)

[1] Siehe `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed`

- Marianne Regard (Neuropsychologie, UniversitätsSpital Zürich)
- Anton Valavanis (Neuroradiologie, UniversitätsSpital Zürich)
- Carel van Schaik (Anthropologie, Universität Zürich)

Einem Teil dieser Personen (Brugger, Fischbacher, Jänke, Regard, Valavanis, van Schaik) wurde der Entwurf dieser Studie zwecks Prüfung und weiterer Kommentierung zugestellt – das Design der Befragung orientiert sich also am Konzept der Delphi-Methode.[2] Mit diesem Verfahren wurden gleichzeitig mögliche Kooperationspartner (Punkt 4 des Auftrags) identifiziert.

1.3 Begriffliche Grundlagen

1.3.1 Ethik und Moral

Dieser Abschnitt liefert eine Übersicht zu zentralen Begriffen der Ethik, welche durch die untersuchten naturwissenschaftlichen Arbeiten angesprochen werden. Gewiss bestehen auch in der heutigen Ethik viele offene Fragen, so dass nicht beabsichtigt wird, eine allgemein gültige Übersicht zu erstellen. Für nachfolgende Erläuterungen werden die Beiträge des Handbuch Ethik [53] benutzt, das einen guten Überblick über den aktuellen Diskussionsstand in der philosophischen Ethik gibt. Ergänzt werden diese mit eigenen Überlegungen des Autors, die sich aufgrund der Literaturrecherche ergeben haben. Eine generelle Bemerkung betrifft die Begriffe “Ethik” und “Moral”. In den untersuchten naturwissenschaftlichen Arbeiten werden diese Begriffe zuweilen synonym verwendet, was hier aber nicht geschehen soll. Grundsätzlich soll in dieser Studie der Begriff “Moral” eine bestimmte Gesamtheit[3] normativ geprägter Sachverhalte[4] oder Verhaltensmuster bezeichnen, während “Ethik” jenes wissenschaftliche Gebiet bezeichnet, welche Moral zum Gegenstand hat.

Zunächst folgen einige allgemeine Bemerkungen zur **Ethik** als wissenschaftliches Gebiet. Ethik wird gemeinhin in die Bereiche *deskriptive Ethik*, *normative Ethik* und *Metaethik* unterteilt. Diese einzelnen Gebiete können wie folgt kurz umschrieben werden:

- Aufgabe der *deskriptiven Ethik* ist es, einen strukturierten Überblick über ein Moralsystem zu erhalten, das beispielsweise in einem bestimmten sozialen Kontext vorherrscht oder dass einer Vielzahl unterschiedlicher Gesellschaften gemeinsam ist. Es geht also um die Ermittlung der moralisch relevanten Sachverhalte und Verhaltensmuster für

[2] Die Delphi-Methode (auch Delphi-Studie oder Delphi-Befragung genannt) ist ein systematisches, mehrstufiges Interview-Verfahren, welches dazu dient, zukünftige Ereignisse, Trends, technische Entwicklungen und dergleichen möglichst gut einschätzen zu können. Dazu wird einer Gruppe von Experten ein Fragenkatalog des betreffenden Fachgebiets vorgelegt. Die Antworten werden zusammengefasst und den Fachleuten (in der Regel anonymisiert, was hier nicht erfolgt ist) erneut für eine weitere Diskussion, Klärung und Verfeinerung der Schätzungen vorgelegt. Dieser kontrollierte Prozess der Meinungsbildung kann über mehrere Stufen erfolgen. Das Endergebnis ist eine aufbereitete Gruppenmeinung. Quellen: Fraunhofer Institut für System- und Innovationsforschung (http://www.isi.fraunhofer.de/p/Projektbeschreibungen/Cu-delphi.html), Internet-Enzyklopädie http://www.powerwissen.com/.

[3] Für die Bestimmung dieser Gesamtheit werden Abgrenzungskriterien verwendet – beispielsweise solche, die Kulturräume voneinander abgrenzen.

[4] Ein Sachverhalt ist eine abstrakte Entität, die als Gegenstand von Akten des Glaubens, Wissens und dergleichen figuriert.

ein bestimmtes System miteinander interagierender moralischer Agenten[5]. Der Begriff "moralischer Agent" (*moral agent*, im deutschen oft auch "moralisches Subjekt") bezeichnet eine Entität, die über eine Reihe von Eigenschaften (z.B. Intentionalität, Autonomie) derart verfügt, dass bestimmte Aktivitäten der Agenten als "moralisch" qualifiziert werden können. Welche Eigenschaften genau erfüllt werden müssen, damit ein Agent ein moralischer Agent genannt werden kann, ist Gegenstand von Debatten, die durchaus vergleichbar sind mit der Suche nach Kriterien für die Bestimmung des Begriffs "Person". Hier stellt sich auch die Frage, wie Agenten genannt werden sollen, die nur einige der möglichen Eigenschaften eine moralischen Agenten haben (z.B. bestimmte Primaten – siehe Abschnitt 3.4.2). Wichtig ist, dass sich Moral immer in einem sozialen Kontext (also in einem System interagierender Agenten) ausbildet bzw. manifestiert. Anders gesagt, ein bestimmtes Moralsystem ist eine Lösung für Probleme, die sich aus solchen Interaktionen ergeben können. Mehrere Disziplinen betreiben deskriptive Ethik, so die Moralpsychologie, die Kulturgeschichte der Moral, Moralsoziologie und die Ethnologie.

- Die *normative Ethik* untersucht die Begründung bzw. Kritik von Moralsystemen. In diesem Bereich lassen sich zwei unterschiedliche Forschungsgegenstände unterscheiden. Zum einen geht darum, grundlegende Begründungsstrategien für Moralsysteme zu entwickeln und zu verteidigen. Die bekanntesten Strategien sind teleologische Ethiken, deontologische Ethiken und so genannt schwach normative bzw. kontextualistische Ansätze. Zum anderen werden – basierend auf solchen Begründungsstrategien – detaillierte Moralsysteme für konkrete soziale und politische Probleme ausgearbeitet mit dem Ziel, Grundlagen für (meist politische) Entscheidungen zu liefern. Die bekanntesten dieser so genannten Bereichsethiken sind die Bioethik, die Medizinethik, die Umweltethik und die Technikethik.

- Die *Metaethik* schliesslich untersucht Fragen, die sich im Rahmen der Konstruktion normativer Ethiken ergeben. Diese Fragen lassen sich in vier Bereiche gliedern: Zum ersten können sprachphilosophische Aspekte, beispielsweise hinsichtlich der Bedeutung normativer Aussagen, untersucht werden. Zum zweiten kann analysiert werden, was moralische Überzeugungen bzw. moralische Gefühle sind und welchen Stellenwert man diesen in einer bestimmten normativen Ethik einräumen will. Drittens können ontologische Fragen betreffend den Status moralischer Eigenschaften und der Existenz moralischer Tatsachen gestellt werden. Viertens schliesslich können epistemologische Fragen hinsichtlich der Rechtfertigung und Begründbarkeit moralischer Urteile untersucht werden. Eine scharfe Grenzziehung zwischen normativer Ethik und Metaethik hingegen ist nicht in jedem Fall möglich, da beispielsweise das Finden von Begründungsstrategien für eine bestimmte normative Ethik auch die Beantwortung metaethischer Fragen mit einschliesst.

[5]Der Begriff "Agent" (*agent*) stammt meines Wissens aus den Sozialwissenschaften. Er bezeichnet in seiner allgemeinsten Form eine Entität, die aufgrund innerer Randbedingungen bzw. Regeln mit anderen Entitäten und einer Umwelt interagieren und auf diese (andere Agenten und Umwelt) einwirken kann. Im Kontext des *agent-based modelling* können Agenten im Rahmen eines Software-Programms genau spezifiziert werden. Derartige Simulationen eröffnen neue Zugänge zur Untersuchung sozialer Systeme, wobei aber die Aussagekraft solcher Analysen kontrovers diskutiert wird.

Anhand dieser Dreiteilung lassen sich erste Ansatzpunkte jener naturwissenschaftlicher Arbeiten aus den Bereichen Neurowissenschaft und Verhaltensforschung herausarbeiten, welche Beiträge für die Grundlagen moralischer Orientierungen liefern wollen. Die erste Anforderungen an solche Untersuchungen fallen in die deskriptive Ethik: Die Untersuchungen müssen eine präzise Vorstellung davon vermitteln, *welchen* moralischen Sachverhalt sie untersuchen wollen und in welchen grösseren Kontext dieser zu stellen ist. Dies bedeutet insbesondere die Spezifikation eines moralischen Stimulus für bestimmte Experimente oder die Operationalisierung einer moralischen Handlung, so dass diese quantitativ erfassbar wird. Im nachfolgenden Abschnitt wird ein strukturierter Überblick der in den verschiedenen Studien vorgeschlagenen moralischen Sachverhalte, welche Gegenstand empirischer Untersuchungen sein können, gegeben. Der zweite Ansatzpunkt betrifft die Frage, in welchem Sinn derartige Untersuchungen Argumente für oder gegen bestimmte Theorien der normativen Ethik liefern können. Unbestritten ist sicher, dass damit empirische Daten (beispielsweise ermittelt in soziologischen Studien) über die praktische Umsetzbarkeit bestimmter normativer Theorien gewonnen werden können. Fraglich ist aber, inwieweit damit bestimmte grundlegende Moralsysteme (z.B. die neurobiologische begründete Bevorzugung der aristotelischen Tugendethik gegenüber deontologischen oder teleologischen Ansätzen, wie dies von [32] vorgeschlagen wurde – siehe dazu Abschnitt 3.1) in grundsätzlichem Sinne gegenüber anderen ausgezeichnet werden können. Der dritte Ansatzpunkt schliesslich gehört in den Bereich der Metaethik und fällt in das Projekt der Naturalisierung von Ethik. Gemeint ist damit die Auffassung, dass moralische Tatsachen natürliche Tatsachen sind bzw. in einem zu bestimmenden Sinn durch natürliche Tatsachen konstituiert sind. Die Untersuchung der Frage, ob ein solches Projekt erfolgversprechend ist, hat in der Ethik eine lange Tradition und beinhaltet die Diskussion um den naturalistischen Fehlschluss bzw. um das MOORE'sche Argument der offenen Frage (siehe z.B. [53]: Kapitel 3). Einige der in dieser Pilotstudie untersuchten wissenschaftlichen Arbeiten haben dieses Ziel. Dennoch kann im Rahmen dieser Pilotstudie die Diskussion um die Naturalisierung von Ethik nicht vorgestellt werden.[6] Nachfolgend soll lediglich darauf hingewiesen werden, welche Autoren mit welchen Argumenten für eine solche Naturalisierung plädieren. Die Naturalisierung von Moral schliesslich sollte von einer Naturalisierung der Ethik unterschieden werden, denn erstere geht davon aus, dass der naturwissenschaftliche Begriffs- und Methodenapparat als der korrekter Ansatz für die genaue Bestimmung aller struktureller Komponenten moralischer Sachverhalte (siehe unten) mindestens für die Ebene von Einzelpersonen und wahrscheinlich auch für Kleingruppen angesehen wird. In diesem Sinn streben praktisch alle der untersuchten Studien eine Naturalisierung von Moral an. Nur wenige Studien behaupten aber, dass damit auch moralische Sachverhalte auf der gesellschaftlichen Ebene verstanden werden können. Grundgedanke ist hierbei, dass die naturwissenschaftliche Herangehensweise an das Problem der Moral nötig ist, um Moralität in Menschen zu verstehen, aber nicht ausreicht, um ethische Probleme auch gleich lösen zu können [136].

Die unter dem Begriff **Moral** subsummierten Sachverhalte werden in dieser Pilotstudie hinsichtlich ihrer Struktur in folgende vier Bereiche unterteilt: 1) in eine handlungsauslösende Komponente; 2) in eine Komponente des *decision making* (unterstützt von handlungsleitenden Komponenten); 3) in die dadurch verursachten Handlungen; 4) in die Kriterien für die

[6] Vgl. dazu beispielsweise die Diskussion in [110], wo verschiedene Strategien einer Naturalisierung von Ethik vorgeschlagen werden oder der Beitrag von ROTTSCHAEFER, der die Gültigkeit des Arguments des naturalistischen Fehlschlusses für die Kritik an der evolutionären Ethik zurückweist [145].

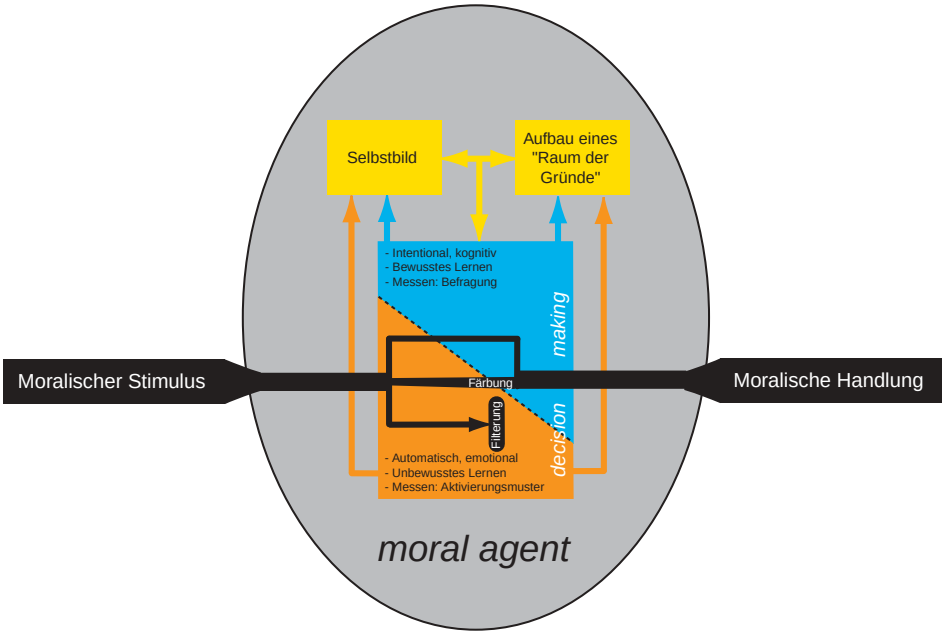

Figure 1.1: Arbeitsmodell eines *moral agent*.

Sicherung bzw. Stützung der Gültigkeit der handlungsleitenden Komponenten. Diese vier Bereiche, welche auch ein rudimentäres Modell eines *moral agent* mit einschliesst, lassen sich wie folgt genauer charakterisieren (vgl. mit Bild 1.1):

- Die erste Komponente ist ein für den betreffenden moralischen Agenten sinnlich erfahrbares raumzeitliches Ereignis – im Kontext eines Experiments beispielsweise ein moralischer Stimulus (zur Problematik dieses Begriffs siehe Abschnitt 3.2).
- Die zweite Komponente lässt sich in einen intentionalen und einen eher automatisch ablaufenden Prozess eines *decision making* trennen. Die genaue Charakterisierung und Unterscheidung dieser Prozesse ist Gegenstand von Untersuchungen (siehe Abschnitte 3.3 und 3.5). Beide Prozesse beinhalten (gelernte) Prädispositionen des moralischen Agenten (handlungsleitende Komponenten), welche die Bandbreite der möglichen Handlungen als Folge der sinnlichen Erfassung des moralischen Stimulus einschränken oder gar determinieren. Es sind also Prozesse wie “Filterung” oder “Färbung” von Wahrnehmungsaspekten zu erwarten. Im Rahmen einer naturwissenschaftlichen Erfassung des Phänomens geht man davon aus, dass diese Prozesse zumindest prinzipiell messbar sind – etwa im Sinn eines charakeristischen Erregungsmusters im Gehirn des betreffenden Agenten, das sich nach einem moralischen Stimulus aufbaut.
- Die dritte Komponente umfasst das Wirken des Agenten in der Raumzeit als Folge der handlungsauslösenden Komponenten und des von handlungsleitenden Komponenten begleiteten *decision makings*. Damit ist auch gesagt, dass die (empirisch erfassbaren)

moralischen Sachverhalte immer eine Handlung – sei dies ein tatsächlich körperliches Handel oder ein Sprachhandeln, etwa im Sinn einer beurteilenden Aussage – beinhalten. Dies ist nicht weiter erstaunlich, da die reine, nicht gegen aussen kommunizierte Introspektion bezüglich eines als moralisch erkannten Problems durch ein bestimmtes Individuum nicht Gegenstand einer empirischen Untersuchung sein kann.[7] Hier muss angefügt werden, dass der Begriff einer Entscheidung (*decision making*), der bei der empirischen Untersuchung moralischer Sachverhalte oft auch Anwendung findet, ebenfalls im Sinn einer Handlung zu verstehen ist, da das Kundtun einer Entscheidung in einem experimentellen Kontext (sei dies nun verbal oder mittels Knopfdruck) eine Handlung beinhaltet.[8] Weiter stellt sich die Frage, wie moralische von nichtmoralischen Handlungen unterschieden werden können. Bei den untersuchten Studien hat sich gezeigt, dass solche Fragen meist kaum nähere Beachtung fanden, bzw. im Kontext des Experiments wird davon ausgegangen, dass die Reaktionen der Versuchspersonen moralische Handlungen im oben erwähnten Sinn sind.

- Die vierte Komponente ist die empirisch wohl am schwierigsten fassbare Komponente eines moralischen Sachverhalts. Gemeint ist damit der Komplex von Begründungen, welche der betreffende moralische Agent für die Rechtfertigung seiner Handlung angibt (bzw. angeben würde). In der philosophischen Ethik basiert ein moralisches *decision making* auf dem selbstständigen Verhalten im logischen Raum der Gründe – ein von Wilfrid Sellars geprägter Begriff [147]. Hier würde man also die vierte Komponente mit einem intentionalen *decision making* identifizieren. In diesem Fall würde man eine rationalistische Auffassung von moralischem Handeln vertreten, wonach es letztlich die einem moralischen Agenten bewusst zugänglichen Gründe sind, welche zu einer moralisch (guten) Handlung führen. Eine Gegenposition, welche im Rahmen dieser Pilotstudie vorgestellt wird, wäre, dass vorab unbewusste, automatisierte Vorgänge eine moralische Handlung leiten und Rechtfertigungen erst *post fakto* (falls die Situation dies verlangt) durch den betreffenden Agenten erzeugt werden (erzeugt im Sinn einer bestimmten Sequenz von Aussagen für die Rechtfertigung der Handlung gegenüber Anderen). Empirisch würde diese vierte Komponente wohl mit Fragebögen, Interviews etc. erfasst werden.

Eine empirische Untersuchung moralischer Sachverhalte müsste grundsätzlich in der Lage sein, alle vier Komponente genau zu spezifizieren und deutlich zu machen, wie man diese quantifizieren will. In den untersuchten Studien geschieht dies in der Regel nicht, so dass Schwachstellen in der nachfolgenden Untersuchung identifiziert werden müssen. Bei der Messbarkeit einiger der oben erwähnten Komponenten stellt sich vom Standpunkt der philosophischen Ethik zudem ein grundlegendes Problem – vorab bei der vierten Komponente. Auch wenn man davon ausgeht, dass der Prozess des Suchen, Bewerten und

[7] Hier liesse sich einwenden, dass der Akt der Introspektion beispielsweise durch *Imaging* beobachtet werden könnte. Doch auch ein solches Experiment ist eingebunden in einen Handlungskontext, in dem beispielsweise das Experiment erklärt wird und die Versuchsperson mitteilen muss, wann die Introspektion beginnt und wann sie endet.

[8] Die Unterscheidung zwischen den Begriffen "Entscheidung" und "Handlung" – beispielsweise im Sinn, das nur jene Akte als Handlungen aufgefasst werden sollen, denen eine Entscheidung vorangeht [128] – soll damit aber nicht untergraben werden. Zu diesem Problem gibt es ausführliche Untersuchungen in der Handlungstheorie, welche für diese Pilotstudie aber nicht Berücksichtigung fanden.

Abwägen von Gründen auf bestimmten neuronalen Prozessen beruht, ist es unklar, aufgrund welcher Kriterien man derart komplexe psychische Entitäten mit physiologischen Entitäten in Beziehung setzen will. Dies mag mit ein Grund sein, warum das Modell entwickelt wurde, dass vorab unbewusste, automatisierte und oftmals auch emotionsgeladene Vorgänge eine moralische Handlung leiten. Die mit diesen automatisieren Vorgänge verbundene psychischen Entitäten haben mutmasslich eine einfachere Struktur und sind in der Emotionsforschung besser untersucht, was einen empirischen Zugang erleichtert.

Nebst ihrer Struktur lassen sich moralische Sachverhalte auch hinsichtlich des Anwendungskontexts unterteilen. Aufgrund der untersuchten Studien erweist sich folgende Dreiteilung als sinnvoll (vgl. mit Bild 1.2):

- So kann erstens ein einzelner *moral agent* Gegenstand der Untersuchung sein. Dies ist das übliche Vorgehen bei *Imaging* Experimente, in welchen beispielsweise eine Versuchsperson mit moralischen Dilemmas konfrontiert wird. Stimuli werden in solchen Experimenten praktisch immer visuell präsentiert. Was sind handlungsleitende Komponenten in diesem Kontext? Auf der intentionalen Ebene nennt man diese Überzeugungen oder Präferenzen. Diese stehen in einem grösseren Begründungskontext (die vierte Komponente), welche sich beispielsweise im Selbstbild der betroffenen Person ausdrücken. Die Frage ist natürlich, wie man sinnvolle neuronale Korrelate für diese Komponenten definiert.

- Zweitens kann eine Gruppe direkt interagierender moralischer Agenten Gegenstand der Untersuchung sein. "Direkt interagieren" bedeutet dabei, dass sich die beteiligten Partner über längere Zeiträume oft begegnen, so dass Beziehungen entstehen und die einzelnen Agenten einen Ruf erwerben und Vorstellungen über die jeweils anderen Agenten entwickeln können. Solche Interaktionen können beispielsweise im Rahmen der experimentellen Ökonomie untersucht werden. In diesem Kontext werden weitere Charakteristika von Kleingruppen genannt [25]: So existiert ein *public good*, dessen Verteilung geregelt werden muss, wobei keine zentralisierten Strukturen bestehen und die einzelnen Agenten keinen durch soziale Institutionen definierten Status haben. Eventuell würde auch eine Gemeinschaft von Primaten in diese Kategorie fallen, wenn man diesen den Status eines *moral agent* zubilligen möchte. Die handlungsleitenden Komponenten auf dieser Ebene kann man Gruppennormen nennen, welche durch Belohnungen und/oder Bestrafungen gestützt werden (eine mögliche vierte Komponente).

- Drittens kann man eine (grosse) Gruppe anonym interagierender moralischer Agenten untersuchen, die sich in Form von Institutionen organisiert haben. Viele praktische ethischen Probleme fallen in diesen Bereich. Sie können durch sozialwissenschaftliche Methoden (Befragungen etc.) aber auch durch ökonomische Experimente untersucht werden. Handlungsleitende Komponenten auf dieser Ebene können Werte genannt werden, die ihre Stützung durch bestimmte politische oder rechtliche Verfahrensregeln bzw. Gesetze erhalten (vierte Komponente).

Diese Unterscheidungen sind als heuristisch zu verstehen und soll nachfolgend nicht im Detail begründet werden. Sie sollen vielmehr dazu dienen, die verschiedenen Untersuchungen, welche im Bereich Neurowissenschaft (z.B. *Imaging*-Experimente und Auswertungen der Folge bestimmter Hirnläsionen) und Verhaltensforschung (z.B. Experimente mit

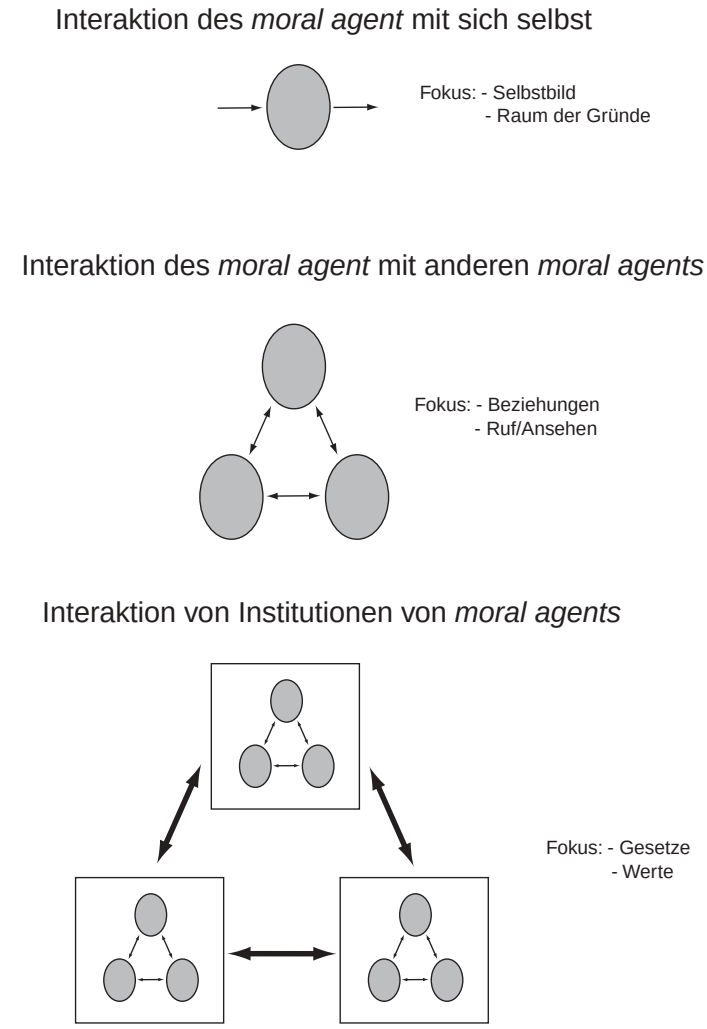

Figure 1.2: Ebenen der Interaktion von *moral agents*.

ökonomischen Spielen, Beobachtungen in der Ethologie) richtig einordnen zu können. Hinsichtlich des Anwendungskontextes wird übrigens auch angestrebt, Mischformen zu untersuchen – etwa die Analyse der Hirnvorgänge bei der direkten Interaktion verschiedener Personen (Multi-Scanner-Experimente). Diesbezüglich sind aber heute noch beträchtliche methodische Schwierigkeiten zu überwinden.

Eine weiterer Untersuchungsaspekt betrifft schliesslich die zeitliche Dimension hinsichtlich der Entstehung bzw. Veränderung der verschiedenen strukturellen Aspekte moralischer Sachverhalte. Eine für die Pädagogik relevante Frage ist beispielsweise die Ontogenese der handlungsleitenden Komponente moralischer Sachverhalte auf der Ebene von Einzelperso-

nen auf der Zeitskala von zehn bis zwanzig Jahren. Die Politikwissenschaften wiederum sind an der Frage interessiert, wie sich die handlungsleitenden Komponenten auf der Ebene der Interaktion von Institutionen (also Werte) auf der Zeitskala von mehreren Generationen verändern. Die evolutionsbiologische Perspektive schliesslich interessiert sich für die Frage nach der Entstehung von Normensystemen in bestimmten Arten über viele Jahrtausende (das Themengebiet der evolutionären Ethik). Diese unterschiedlichen Zeitskalen werden in einer Reihe der untersuchten Studien angesprochen. Studien, die sich über derart lange Zeitskalen erstrecken, finden sich aber nur selten.[9]

1.3.2 Neurowissenschaft und Verhaltensforschung

Nebst den begrifflichen Problemen, welche die naturwissenschaftliche Untersuchung des moralischen Urteils mit sich bringen, finden sich eine Reihe inhärenter begrifflicher Probleme der Neurowissenschaften und der Verhaltensforschung. Diese haben bei der nachfolgenden Analyse eine Bedeutung und sollen deshalb kurz skizziert werden. Natürlich kann an dieser Stelle keine eigentliche Kritik der philosophischen Probleme der Neurowissenschaft und Verhaltensforschung geleistet werden (zu ersterem siehe beispielsweise [19]). Es soll lediglich aufgezeigt werden, welche naturwissenschaftlichen Begriffe in den nachfolgend untersuchten Studien problematischer sind, als sie den Anschein machen.

Zu Beginn soll deutlich gemacht werden, was mit "**Neurowissenschaft**" und "**Verhaltensforschung**" gemeint ist. Der Terminus "Neurowissenschaft" (*neuroscience*) ist verhältnismässig neu und ist vermutlich durch den amerikanischen Forscher Ralph Gerard Ende der 1950er Jahre eingeführt worden (siehe [2]:Preface). Er beinhaltet insbesondere jene Forschungsanstrengungen, welche die funktionale Organisation neuronaler Systeme untersuchen, um damit die Verhaltensleistungen der entsprechenden Organismen erklären zu können.[10] Auch die Erforschung von Fehlleistungen neuronaler System – etwa in Folge degenerativer Erkrankungen des Nervensystems – fallen in den Bereich der Neurowissenschaft. Insofern besteht auch eine gewisse Überlappung mit der Neurologie, wobei aber beispielsweise Neurochirurgie wohl eher nicht zu den Neurowissenschaften gezählt wird. In den Bereich der Neurowissenschaft fallen unter anderem die (zelluläre) Neurophysiologie, die (oder zumindest Teile der) Kognitionswissenschaft, die Neuroinformatik und die Neuropharmakologie. Die modernen Verhaltensforschung (Ethologie) findet ihren Ursprung in der Instinktforschung – zu nennen sind hier unter anderem die Arbeiten von Oskar Heinroth und dessen Schüler Konrad Lorenz. Gegenstand der Verhaltensforschung sind in der Regel Tiere bzw. Tierpopulationen, während die Erforschung des Verhaltens von Menschen eher in den Bereich der Psychologie und teilweise auch der Ökonomie fällt. Wie die obige grobe Definition von Neurowissenschaft deutlich macht, bestehen heute natürlich auch enge Bezüge zwischen der Ethologie und der Neurowissenschaft, was zur Wortschöpfung der *behavioral neuroscience* geführt hat. Im weiteren hat sich gezeigt, dass in jüngerer

[9] Natürlich gibt es moralpsychologische Untersuchungen – etwa die Studien von Piaget oder auch das Modell der Moralentwicklung von Kohlberg [103]– wo längere Zeitskalen eine Rolle spielen. Es handelt sich hier um psychologische Studien, welche den methodischen Apparat der heutigen Neurowissenschaft nicht nutzen. In der Verhaltensforschung wiederum gibt es Feldstudien von Primatengemeinschaften, welche deren soziale Dynamik über viele Jahre erfassen. In Simulationen des *agent-based modelling* schliesslich können ebenfalls – da die Zeitachse beschleunigt werden kann – länger dauernde Prozesse untersucht werden.

[10] In einem Standardwerk wird die Aufgabe von *neuroscience* wie folgt umschrieben: "to explain behavior in terms of the activities of the brain" [100]:5

Zeit auch aus der Spieltheorie stammende Methoden Anwendung für Fragen der Verhaltensforschung gefunden haben – sowohl Tiere wie Menschen betreffend. Feststellbar ist beispielsweise ein verstärktes Interesse an experimenteller Ökonomie, wo Berührungspunkte zwischen Verhaltensforschung, Psychologie und neuerdings auch Neurowissenschaft (die so genannte Neuroökonomik, siehe [66] für eine kurze Übersicht) entstanden sind. Diese Forschungen finden für diese Pilotstudie ebenfalls Beachtung. Zu nennen ist schliesslich noch die Primatologie, deren Forschungsfragen zumindest teilweise mit der Verhaltensforschung zusammenfallen, und eventuell die Soziobiologie, welche aber hier nicht weiter vorgestellt wird. Diese grobe Charakterisierung zeigt, dass der Gegenstand dieser Studie in mehrere Gebiete fällt. Der Begriff Neurowissenschaft (*neuroscience*) ist sicher am umfassendsten, zumal hier der generelle Zusammenhang zwischen Verhalten und neuronalen Prozessen untersucht werden soll. Hier haben sich eine Reihe von Forschungsgebiete entwickelt, die Fragen der Grundlagen moralischer Orientierung betreffen (siehe Abschnitt 2.3). Viele klassische Bereiche der Verhaltensforschung hingegen fokussieren Tierverhalten und befassen sich demnach kaum mit Fragen nach moralischem Verhalten.

Nachfolgend sollen nun einige problematische naturwissenschaftliche Begriffe erläutert werden. So ist zum ersten in den untersuchten Studien oft die Rede von einem **neuronalen Zustand**, in dem sich ein gewisses System befindet. Es ist aber in verschiedener Hinsicht unklar (bzw. es wird nicht präzisiert), worin dieser Zustand bestehen soll. Folgende offene Probleme bestehen:

- Ein Aspekt ist der **relevante Parameter**, an dem sich dieser Zustand und dessen Änderung ablesen lassen soll, denn neuronale Aktivität manifestiert sich in verschiedener Weise: a) Im Energieverbrauch in Form von Sauerstoff-Verbrauch (das im fMRI gemessene Signal) oder in Form von umgesetztem ATP. b) In der elektrischen Aktivität wie beispielsweise Nervenimpulse (*spikes*), dem elektrischen Feldpotential (das mutmasslich von elektrischen Strömen in Dendriten herrührt) oder der elektrischen Aktivität ganzer Neuronengruppen (welche durch das EEG gemessen wird). c) In Änderungen des intrazellulären Metabolismus. d) In Änderungen der synaptischen Verschaltung. e) In Änderungen des chemischen Millieus, in denen sich die Zellen befinden. f) In möglichen wichtigen Beiträgen der Gliazellen. Auch heute ist nicht gesichert, welcher dieser Parameter (bzw. welche Kombination) als bedeutsam gelten soll. In der Regel bestimmt die gerade verwendete Methode den relevanten Parameter.

- Ein weiterer Aspekt betrifft die Frage, ob es ein **coarse graining** dieses relevanten Parameters derart gibt, dass die Rede von einem Zustand überhaupt Sinn macht. Dies ist ein wichtiges Problem für ein System, das (im Gegensatz etwa zu einem herkömmlichen Computerchip) nicht zentral getaktet ist und das vermutlich eine Art einer 'verteilten Berechnung' (*distributed computing*) benutzt. Demnach müsste man auch ein Kriterium finden, welches Regionen des Gehirns, die sich in einem Zustand A befinden sollen, von Regionen, die sich im Zustand B befinden sollen, abgrenzt.

- Ein dritter Aspekt betrifft die Frage nach der Zeitskala, innerhalb dessen ein Zustand als konstant angesehen soll. Je nach gewähltem Parameter unterscheiden sich die Vorschläge um Grössenordnungen: von Millisekunden (hinsichtlich der elektrischen Aktivität von Einzelzellen), über Sekunden (wenn gewisse Periodizitäten im EEG als

relevant gelten sollen), bis zu vielen Minuten (wenn strukturelle Änderungen der synaptischen Verschaltung als relevant gelten sollen).

Diese Aufzählung ist vermutlich nicht vollständig. Sie soll aber deutlich machen, dass die oft verwendete Rede von einem "neuronalen Zustand" meist eher metaphorisch verwendet wird, ohne dass präzise gemacht wird, worin dieser Zustand eigentlich genau besteht.

Ein weiteres Problem ist mit der Rede von **neuronalen Korrelaten** bestimmter psychischer Aspekte verbunden [34]. Die Rede von einem neuronalen Korrelat verlangt die Lösung dreier Probleme: Zum ersten muss hinreichend präzise geklärt werden, für welches psychische Phänomen man ein Korrelat finden will. Zum zweiten muss eine befriedigende Definition eines neuronalen Zustandes gefunden werden, der messbar ist und als gewünschtes Korrelat dienen kann. Drittens muss eine (optimalerweise kausale) Verkettung zwischen den beiden Phänomenen zumindest plausibel gemacht werden. So ist beispielsweise oft die Rede davon, dass ein gewisser neuronaler Zustand ein gewisses psychisches Phänomen (z.B. ein Sinneseindruck) "repräsentiert". Ist diese Repräsentation aber im Sinn einer Kodierung zu verstehen? Worin besteht der Informationsgehalt dieser Repräsentation? Ist es korrekt, die Veränderungen solcher Repräsentationen als eine Form von Berechnung (*computation*) aufzufassen? Was heisst "Berechnung" überhaupt im Kontext eines biologischen neuronalen Systems [153]? Diese sehr grundlegenden Probleme sind bei weitem nicht gelöst. Im weiteren ist nicht klar, ob es eine Komplexitätsschwelle bestimmter psychischer Phänomene derart gibt, dass die Rede von einem "neuronalen Korrelat" sinnlos wird, weil damit die Aktivität des gesamten Gehirns gemeint ist. Gerade psychische Phänomene, die man mit moralischen Sachverhalten in Verbindung bringen könnte (z.B. Emotionen wie Scham und Schuld) könnten derart komplex sein, dass eine Lokalisation des entsprechenden neuronalen Korrelats sinnlos wird (mehr dazu im Abschnitt 3.3.2). Zu nennen ist in diesem Zusammenhang auch die oft zu findende Behauptung, man suche nach den neuronalen "Grundlagen" moralischen Verhaltens. Dies impliziert eine hierarchische Gliederung derart, dass Prozesse auf einer Basisstufe (dem neuronalen System) die Prozesse auf einer höheren Stufe (den psychischen Phänomenen) in einem zu bestimmenden Sinn determinieren. Wie genau das aber zu verstehen ist, rüttelt an einem der Grundprobleme des Leib-Seele-Problems, das gemeinhin unter dem Begriff der Emergenz abgehandelt wird [95].

Ein weiterer, die Verhaltensforschung betreffender Problemkomplex betrifft die Frage, ab wann man beobachtete Regelmässigkeiten bei der Interaktion von Agenten als **Normen** betrachten will. Dieses Problem gewinnt an Schärfe, wenn man die evolutionäre Perspektive einbringen will und beispielsweise nach Vorformen von Moral sucht. Interessant ist insbesondere die Frage, ob das Auftreten des Normative als ein sprunghaftes Element zu verstehen ist (in der Art eines Phasenübergangs in der Physik) oder eher als gradueller Prozess gelten soll. Wenn ersteres der Fall ist: Anhand welcher Marker soll man einen solchen sprunghaften Wechsel erkennen können? Es wurde nicht untersucht, wie dieses Problem in der Verhaltensforschung angegangen wird. Aufgrund der analysierten Studien bleibt aber der Eindruck, dass diese wichtige Frage nicht hinreichend geklärt ist.

Chapter 2

Empirische Erforschung der Moral

2.1 Moral als Thema: eine Trendanalyse

Um einen ersten Eindruck über die Forschungstätigkeit in den für diese Pilotstudie relevanten Gebiete zu erhalten, wurde eine kurze bibliometrische Untersuchung basierend auf den Einträgen in der *MedLine*-Datenbank durchgeführt. Untersucht wurde, wie sich der Anteil der Forschung mit bildgebenden Verfahren, der *social cognitive neuroscience*, der Emotionsforschung und der Untersuchung moralischer/ethischer Aspekte der Neurowissenschaft relativ zur Gesamtzahl der neurowissenschaftlichen Arbeiten entwickelte. Fokussiert wurde die Publikationstätigkeit der vergangenen 30 Jahre – also ab 1975. In einem ersten Schritt wurde die pro Jahr publizierte Zahl an neurowissenschaftlichen Arbeiten mit dem folgenden booleschen Suchausdruck abgeschätzt: (neuron OR neuronal OR neural OR neuroscience OR brain) AND year[Publication date]. Das Resultat der Untersuchung findet sich in Abbildung 2.1.a: Es zeigt sich der erwartete stetige Anstieg der Publikationstätigkeit. Verglichen mit 1975 wurden im Jahr 2004 gut drei mal mehr Arbeiten publiziert. Die mit dieser Suche erhaltenen Menge an Publikationen dienen als Referenz für die nachfolgenden Analysen.

Der Anteil der *Imaging*-Arbeiten wurde mit folgendem Suchausdruck abgeschätzt: (imaging OR fMRI OR MRI OR PET) AND *Referenz*.[1] Hier zeigt sich (Abbildung 2.1.b, hellgraue Kurve) die stetig wachsende Bedeutung dieser Methode für die Neurowissenschaft. So finden sich in über zwanzig Prozent aller heute publizierten Arbeiten entsprechende Stichworte – ein erstaunlich hoher Anteil, der einem einzigen methodischen Ansatz zugeordnet werden kann. Die Zahl der Arbeiten, die der Emotionsforschung bzw. der *social cognitive neuroscience* zugeordnet werden können, wurden mit folgenden Ausdrücken abgeschätzt: (emotion OR emotional OR emotionally) AND *Referenz* / (social OR socially) AND *Referenz*. Hier zeigt sich bei beiden Forschungsgebieten ein markanter Anstieg ab etwa Mitte der 1990-er Jahre (Abbildung 2.1.b, schwarze und dunkelgraue Kurven). Die beiden Gebiete erreichen einen Anteil von gut 2.5 bzw. knapp zwei Prozent an der gesamten Publikationstätigkeit im Bereich

[1] Das EEG wurde in dieser Untersuchung nicht als eine Methode des *Imaging* bezeichnet, obgleich gewisse Darstellungsformen des EEG durchaus einen bildgebenden Charakter haben.

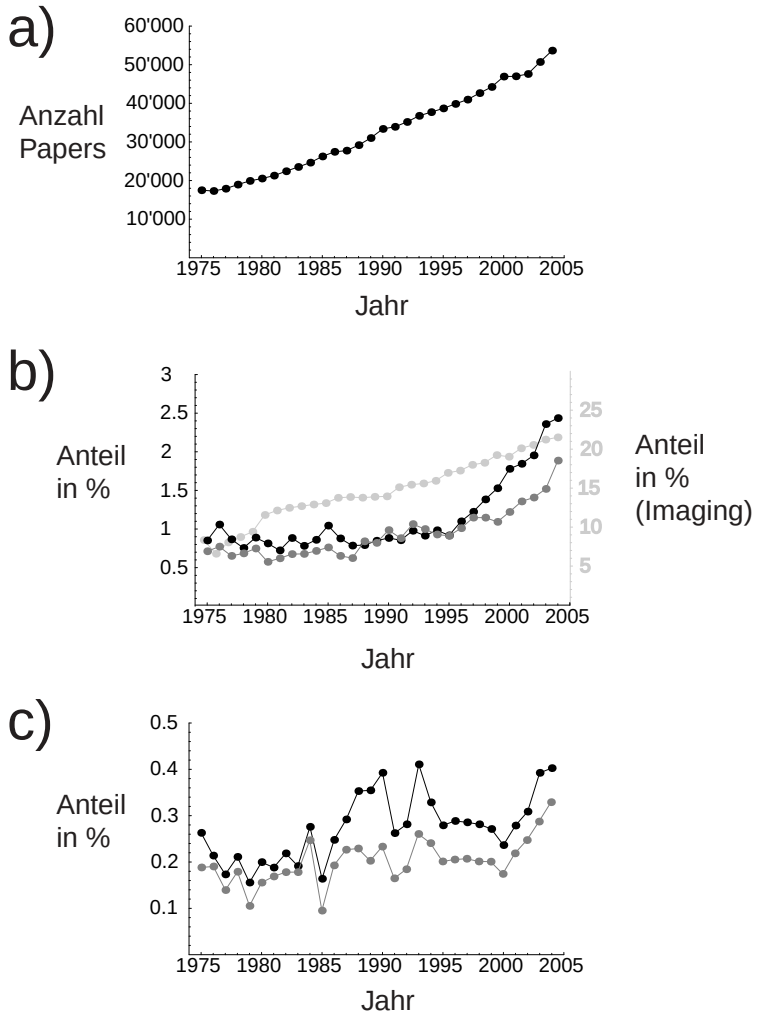

Figure 2.1: Bibliometrische Untersuchung neurowissenschaftlicher Arbeiten zu komplexen Verhaltensweisen: a) Generelle Publikationstätigkeit in der Neurowissenschaft. b) Anteil von Arbeiten aus den Bereichen *Imaging* (hellgrau, linke Ordinate), Emotion (schwarz, rechte Ordinate) und *social cognitive neuroscience* (dunkelgrau, rechte Ordinate). c) Anteil der Arbeiten aus den Bereichen Moral/Ethik (grau: ohne Arbeiten über Transplantation von Hirngewebe).

Neurowissenschaft. Die Zahl der Arbeiten über ethische und moralische Aspekte von Neurowissenschaft wurde schliesslich mit folgendem Ausdruck abgeschätzt: (moral OR morally OR ethics OR ethical) AND *Referenz*. Hier zeigt sich ein uneinheitlicheres Bild (Abbildung 2.1.c, schwarze Kurve): Offenbar lassen sich zwei Phasen mit einer verstärkten Publikationstätigkeit identifizieren: Zum einen Ende der 1980er und Beginn der 1990er (mit einem

Rückgang 1991 und 1992), zum anderen ab 2000. Da die absolute Zahl der Publikationen pro Jahr überschaubar war, konnten diese genauer untersucht werden. Hier zeigte sich, dass die erste Phase aus der verstärkten Diskussion der ethischen Probleme der Transplantation von foetalem Hirngewebe bei Parkinsonpatienten resultiert. Zieht man diese Arbeiten (Suchkriterium: (fetal OR transplantation) AND (moral OR morally OR ethics OR ethical) AND *Referenz*) ab, so zeigt die resultierende Kurve (Abbildung 2.1.c, graue Kurve) nur noch eine Phase verstärkter Publikationstätigkeit – jene ab 2000. Hier handelt es sich um Arbeiten, die entweder Moral als Gegenstand haben, oder der Neuroethik zugeordnet werden können.

Zusammengefasst bestätigt die bibliometrische Untersuchung den Eindruck, der sich aus der Lektüre der untersuchten Arbeiten ergeben hat. So ist das Interesse an komplexen Aspekten menschlichen Verhaltens (Emotionen, Sozialverhalten und eben auch moralisches Verhalten) in der Neurowissenschaft erst in den letzten Jahren auf ein zunehmendes Interesse gestossen. Vergleichbare bibliometrische Untersuchungen bestätigen diesen Befund: So finden Illes et al. ab den 1990er Jahren ein stark gestiegener Anteil von Arbeiten, welche fMRI verwenden (von weniger als 100 pro Jahr auf fast 1000 pro Jahr nach der Jahrtausendwende) [97]. So genannte *higher-order cognition* und Emotionen wiederum bilden einen zunehmenden Anteil in diesen fMRI-Studien (ihr Anteil beträgt gegen 20% im Jahr 2001). Eine auf psychologische Literatur beschränkte Untersuchung von Haidt basierend auf der PsycINFO-Datenbank zeigte ebenfalls ein teilweise enorm gesteigertes Interesse an der Erforschung bestimmter Arten von Emotionen im Zeitraum 1975 bis 1999 [84] – vorab solcher Emotionen, welche von Haidt als "moralisch" qualifiziert werden (siehe dazu Abschnitt 2.3.2). Die empirische Erforschung der Moral in der Neurowissenschaft ist also ein vergleichsweise junges Phänomen.

2.2 Anatomische und methodische Grundlagen

2.2.1 Grundlagen der Hirnanatomie

Viele der untersuchten Studien – vorab im Bereich *Imaging* – streben eine Lokalisation jener neuronalen Prozesse an, die bei moralischen Entscheidungen bzw. Handlungen aktiv sein sollen. Zu diesem Zweck soll nachfolgend eine kurze Einführung in die wichtigsten anatomischen Grundlagen des menschlichen Gehirns gegeben werden (vgl. dazu [81, 100]).

Für die **generelle räumliche Orientierung** im Körper eines Menschen bzw. Tieres werden in der Anatomie die folgenden Begriffe verwendet: Entlang der durch die Wirbelsäule vorgegebenen Achse wird die Richtung gegen den Kopf hin als *rostral* und die Richtung gegen den Schwanz / das Steissbein hin als *caudal* bezeichnet. Senkrecht zu dieser Achse wird die Richtung zum Brustbein hin als *ventral* und die Gegenrichtung als *dorsal* bezeichnet. Bei der dritten Richtung, die senkrecht zu diesen beiden steht, werden folgende Bezeichnungen verwendet. Orte nahe des Nullpunkts werden als *medial* bezeichnet, Orte fern des Nullpunkts als *lateral*. Auf das menschliche Gehirn bezogen (man stelle sich das Gehirn in einer Position eines gerade vor einem sitzenden Menschen vor) sehen die Bezeichnungen wie folgt aus: Die vorderen Hirnregionen (Stirnbereich) sind die rostralen Hirnbereiche, die hinteren Regionen (Hinterkopf) die caudalen Bereiche. Der obere Hirnbereich (oberes Kopfende) ist der dorsale Bereich, die unteren Regionen (gegen den Rachen hin) sind die ventralen Bereiche. Die mittlere Hirnregion schliesslich ist der mediale Bereich, während die Seitenbereiche (links

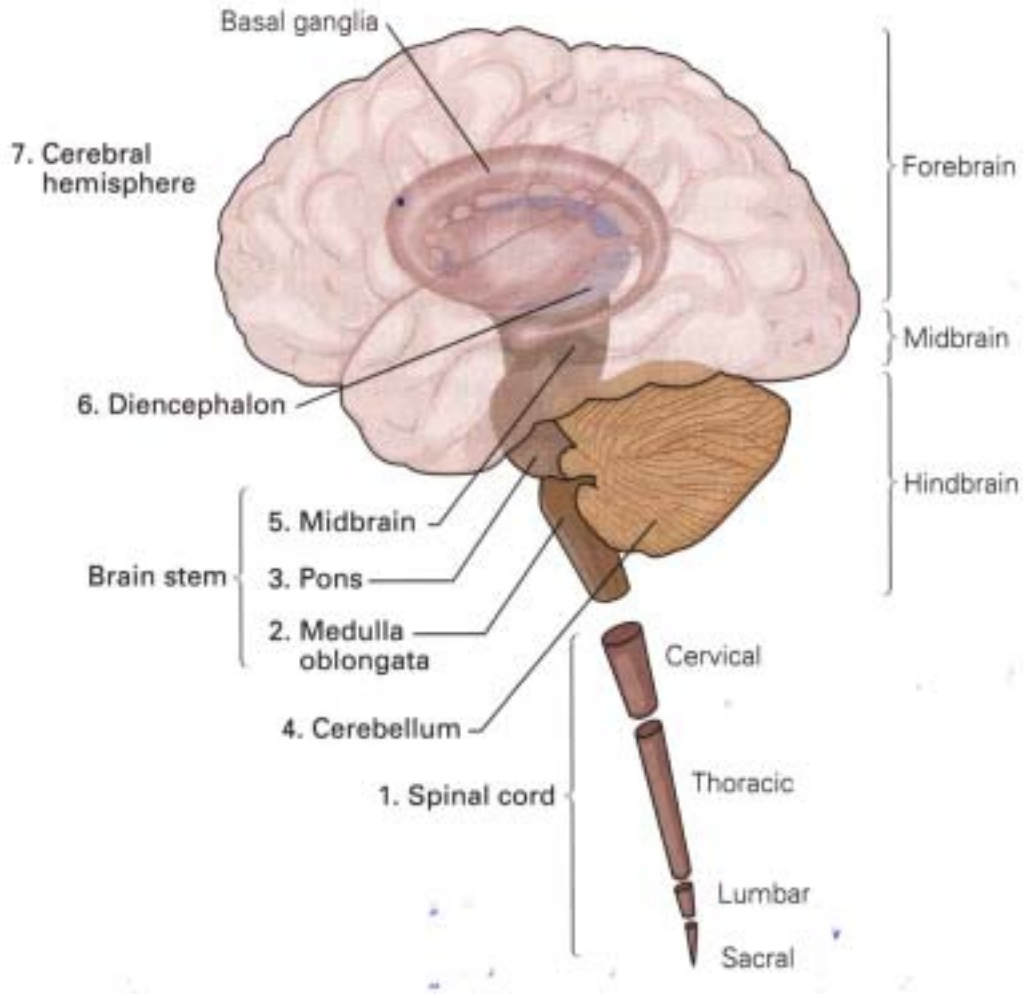

Figure 2.2: Anatomie des Zentralnervensystems (Abbildung aus [100]).

oder rechts) die lateralen Bereiche sind. Für die **Schnittebenen** (sowohl für anatomische *post mortem* Analysen wie auch beim *Imaging*) werden folgende Bezeichnungen verwendet. Man stelle sich den Kopf eines aufrecht sitzenden Menschen vor: Die Schnittebene parallel zum Boden ist die *horizontale* Schnittebene. Die senkrecht dazu stehende Ebene, die parallel zur Gesichtsfläche ist, nennt man die *coronale* Schnittebene. Die dritte Ebene senkrecht zu den beiden anderen ist die *sagittale* Schnittebene.

Die **grobe Anatomie** des menschlichen Zentralnervensystems und Gehirns umfasst folgende Bereiche (von caudaler in rostraler Richtung, siehe Abbildung 2.2): Das (noch nicht zum Gehirn gehörende) Rückenmark, der Hirnstamm mit den Unterbereichen *Medulla*,

Pons und Mittelhirn, das Kleinhirn (*cerebellum*), das Diencephalon mit den Unterbereichen Thalamus und Hypothalamus (und weiteren Regionen) und schliesslich das Telencephalon, welches den cerebralen Kortex und subkortikale Zentren umfasst. Da die überwiegende Mehrzahl der Studien sich auf Regionen im Bereich des Telencephalon beziehen, wird dieser Bereich noch genauer vorgestellt: Die subkortikalen Regionen werden üblicherweise in die Regionen der Basalganglien, des Hippocampus und der verschiedenen Kerne der Amygdala unterschieden (dazu mehr unten). Für den cerebralen Kortex existieren unterschiedliche Arten für eine weitere Differenzierung: Hinsichtlich der Struktur des biologischen Gewebes wird zwischen *grey matter* und *white matter* unterschieden. Erstere besteht primär aus den Zellkörpern von Nervenzellen[2] (und Gliazellen[3]) und weist eine Schichtenstruktur auf (sechs Schichten, mit unterschiedlicher Ausprägung je nach Ort), letztere besteht primär aus myelinisierten[4] Axonen von Nervenzellen, welche verschiedene Felder des Kortex miteinander verbinden. Eine weitere grundlegende Differenzierung betrifft jene zwischen der linken und rechten Hemisphäre und den Nervenfasertrakten (Kommissuren), welche die Hemisphären verbinden. Es gibt vier solche Kommissuren, wobei der Balken (das *corpus callosum*) die wichtigste ist. Die Hemisphären wiederum lassen sich in vier so genannte Lappen unterteilen (siehe Abbildung 2.3): Den Frontallappen (rostral), den Occipitallappen (caudal), den Temporallappen (lateral) und den Parietallappen.

Die **kortikale Anatomie** lässt sich wie folgt weiter ausdifferenzieren: Die menschliche Hirnrinde weist eine ausgesprochene Furchung auf. Die "Hügel" dieser Furchen werden *Gyri* (Einzahl: *Gyrus*) genannt, während die "Täler" *Sulci* (Einzahl: *Sulcus*) genannt werden. Die wichtigsten Sucli trennen die einzelnen Lappen des Kortex: Der zentrale Sulcus trennt den Frontallappen vom Parietallappen. Der laterale Sulcus trennt den Temporallappen vom Parietal- und Frontallappen. Der parietal-occipitale Sulcus trennt die gleichnamigen Lappen. Aufgrund der Furchung der Hirnrinde sind einzelne Regionen des Kortex von Aussen nicht sichtbar. Die wichtigsten dieser inneren Regionen sind der insuläre Kortex (Insula) und das Cingulum (*cingulate gyrus*). Im Verlauf einer sich über viele Jahrzehnte erstreckenden Lokalisationsforschung hat man einzelnen Regionen des Kortex (Rindenfelder)

[2]Nervenzellen (Neuronen) weisen folgenden grundlegenden Bauplan auf: Im Zellkörper (Soma) befinden sich der Zellkern und die überwiegende Mehrzahl der Zellorganellen, welche für die Synthese von Proteinen, Neurotransmittern, etc., für die Energieversorgung und für die Erfüllung weiterer Zellfunktionen sorgen. Die Dendriten sind eine weitverzweigte Struktur, über welche die Nervenzelle Signale anderer Nervenzelle aufnimmt. Diese Signalaufnahme geschieht via chemische oder elektrische Synapsen, welche von anderen Nervenzellen ausgehen. Aktivität in diesen Synapsen bewirken eine Änderung des Membranpotentials in den Dentriten (oder auch im Soma, das ebenfalls von Synapsen kontaktiert werden kann) und damit zu Stromflüssen innerhalb der Zelle. Diese können bei einer speziellen Struktur des Somas – dem Axon-Hügel (*axon hillock*) – ein aktiv erzeugtes elektrisches Potential (der Nervenimpuls oder *spike*) auslösen, das über das Axon der Nervenzelle zu anderen Nervenzellen weitergeleitet wird. Diese Axone verzweigen sich in der Regel und enden in so genannten Synapsen – den Verbindungsstellen zu andere Nervenzellen (oder Muskelzellen). In morphologischer Hinsicht wie auch in physiologischer Hinsicht lassen sich eine Vielzahl von Neuronentypen unterscheiden. Beeindruckend ist die Konnektivität innerhalb des cerebralen Kortex: Jede Nervenzelle wird in der Grössenordung von 10'000 andern Nervenzellen kontaktiert. Die Ermittlung der genauen Anatomie dieser lokalen Verschaltung (*micro-circuits*) ist ein aktuelles Forschungsgebiet.

[3]Verschiedene Unterarten von Gliazellen bilden die grosse Mehrheit der Zellen des zentralen Nervensystems. Ursprünglich wurden Gliazellen nur als eine Art "Hilfszellen" der Neuronen aufgefasst. Neuere Untersuchungen zeigen aber einen bedeutenden Beitrag von Gliazellen für die Signalverarbeitung im zentralen Nervensystem. Die genaue Funktion der einzelnen Arten von Gliazellen ist Gegenstand von Forschungen.

[4]Bestimmte Arten von Gliazellen haben die Funktion, die Axone von Neuronen mit einer Art elektrischer Isolationsschicht (Myelinschicht) zu umgeben. Dies beschleunigt die Weiterleitung von Nervenimpulsen.

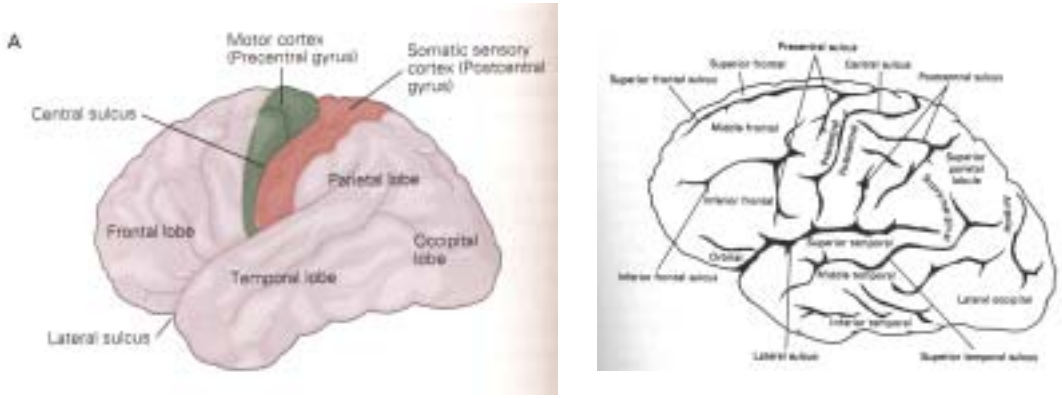

Figure 2.3: Anatomie des Kortex (linke Abbildung aus [100], rechte Abbildung aus [159]).

bestimmte Funktionen zuordnen können. So sind jene Regionen, welche periphere Informationen der Sinnesorgane aufnehmen (sensorische Rindenfelder wie der primäre visuelle Kortex) bzw. Signale an das Bewegungssystem abgeben (motorische Rindenfelder) gut bekannt und erforscht. Weit problematischer ist die Lokalisation "höherer" Leistungen des zentralen Nervensystems. Im Rahmen der Erforschung der "Neuroanatomie der Moral" sind in den für diese Pilotstudie untersuchten Arbeiten folgende Regionen genauer untersucht worden, welche nachfolgend aufgelistet sind. Hierbei muss auf das Problem hingewiesen werden, dass dieselben Hirnregionen in unterschiedlichen wissenschaftlichen Gebieten (Neuroanatomie, Sinnesphysiologie, *Imaging*) unterschiedlich bezeichnet werden – was die Vergleichbarkeit von Studien erschwert. Ein allgemeiner, von allen Neurowissenschaftlern akzeptierter Hirnatlas existiert nicht (Kevan Martin: persönliche Mitteilung). Es ist demnach nicht auszuschliessen, dass in der nachfolgenden Liste gewisse Begriffe dieselbe Hirnregion bezeichnen. Der für diese Pilotstudie beschränkte Zeitrahmen hat eine entsprechende Kontrolle verunmöglicht, so dass auf eine Darstellung dieser Hirnregionen in einer Abbildung verzichtet wird. Die Liste gibt vielmehr einen Hinweis auf die Vielzahl der Hirnregionen, welche bei moralischem bzw. moralnahem Verhalten involviert sein sollen.

Amygdala: [4, 5, 122, 150]
(Bilaterale) Insula: [134, 150]
Bilateraler ventromedialer präfrontaler Kortex: [91]
Bilaterale *temporal poles*: [91]
Caudatum (*Caudate nucleus*): [141]
Cingulum (*Cingulate cortices*): [4, 5]
Dorsaler medialer präfrontaler Kortex: [134]
Dorsales *Striatum*: [48, 102]
Dorsolateraler präfrontaler Kortex: [134]
Fusiformer Gyrus: [150]
Medialer präfrontaler Kortex: [134]
Lateraler orbitaler Gyrus: [122]

Linguale Gyri: [122]
Linker anteriorer mittlerer temporale Gyrus: [61]
Linker lateraler präfrontaler Kortex: [91]
Linker posterior superior temporale *Sulcus*: [91]
Linker superiorer frontaler *Gyrus*: [61]
Medialer orbitofrontaler Kortex: [122]
Medialer präfrontaler Kortex: [121]
Mittlerer temporaler Gyrus: [91]
Nucleus accumbens: [141]
Occipitalen Kortex: [134]
Orbitaler präfrontaler Kortex: [121]
Orbitofrontaler Kortex: [4, 5, 141]
Orbitofrontaler Gyrus: [61]
Orbitofrontaler-striataler *circuit*: [123]
Posteriorer Gyrus des Cingulum: [61]
Rechter somatosensorischer Kortex: [4, 5]
Rostraler anteriorer Cingulum: [141]
Subcallosal cinculate: [134]
Superiorer temporaler Sulcus: [121, 122, 132, 134, 150]
Temporaler Pol: [122]
Unterer Pränucleus: [61]
Ventrolateraler präfrontaler Kortex: [134]
Ventromedial-frontaler Kortex: [141]
Vorderes Cingulum (*Anterior cingulate*): [134]

2.2.2 Methoden der Neurowissenschaft

Die modernen Neurowissenschaften können auf ein ganzes Arsenal an Methoden für Messungen und Eingriffe in neuronale Systeme zurückgreifen. Die überwiegende Mehrzahl der für die Pilotstudie relevanten Studien verwenden verschiedene Varianten bildgebender Verfahren (*Imaging*), welche in einem gesonderten Abschnitt vorgestellt werden. Hier sollen kurz weitere, hauptsächlich *in vivo* eingesetzte Methoden vorgestellt werden.

Wichtige Erkenntnisse wurden – vorab in frühen anatomischen Untersuchungen – durch ***post mortem* Untersuchungen** gewonnen. Durch entsprechende Färbetechniken ist es beispielsweise gelungen, die verschiedenen Arten von Neuronen morphologisch zu charakterisieren und die wesentlichen Nervenfasertrakte im Gehirn zu erkennen. *Post mortem* Untersuchungen bilden auch bei heutigen Tierexperimenten eine wichtige Rolle – etwa um den Ort einer Mess- oder/und Stimulationselektrode anatomisch korrekt zu identifizieren.

Ein weiteres wichtiges methodisches Instrument bildet die **Läsionsforschung**. Vorab an Tieren wurde diese Methode seit dem Aufkommen der neuzeitlichen Hirnforschung Ende des 18. Jahrhunderts oft angewandt. Bei Menschen waren solche Experimente natürlich – abgesehen von Nebenfolgen chirurgischer Eingriffe – nicht möglich. Hingegen haben Läsionen hervorgerufen durch Verletzungen oder Hirnblutungen die Möglichkeit eröffnet, den funktionellen Beitrag bestimmter Hirnregionen zu gewissen Verhaltensweisen zu ermitteln. Die Auswertung von Läsionsexperimenten ist mit zwei Hauptschwierigkeiten konfrontiert [159]: Zum einen ist das Verursachen einer Läsion in einem Versuchstier mit traumatischen Schäden

am Hirngewebe verbunden, deren Auswirkungen nie ganz erfasst werden können. So können beispielsweise unbeabsichtigt Faserverbindungen zerstört werden, so dass der beobachtete Funktionsausfall gar nicht auf die Läsion der betreffenden Hirnregion zurückgeführt werden kann, sondern auf die Unterbrechung der Verbindung zweier sonst nicht betroffener Regionen. Zum anderen besitzt das Gehirn eine erstaunliche Regenerationsfähigkeit (Plastizität). Da aufgrund der Verletzungen, die aus einer Läsion resultieren, Verhaltensexperimente in der Regel nicht gleich nach dem Verursachen der Läsion durchgeführt werden können, könnten eine unbemerkte (teilweise) Regeneration der durch die Läsion beeinträchtigten Funktion stattgefunden haben. Dies würde die aus der Läsion gezogenen Schlussfolgerungen verfälschen. In jüngerer Zeit finden in Tierexperimenten auch Methoden einer reversiblen Läsion Anwendung. So kann durch Kühlung von Hirnregionen, durch Injektion bestimmter Chemikalien oder durch transkraniale magnetische Stimulation (siehe unten) die Funktion dieser Region zeitweise ausgeschaltet werden.

Ein für das Aufkommen der modernen Neurowissenschaft zentrales methodisches Instrument ist die **Messung bzw. Induktion elektrischer Phänomene** im Gehirn durch geeignete Elektroden. Eine bekannte, nichtinvasive Methode ist das in den 1920er Jahren entwickelte Elektro-Encephalogramm (EEG), das später auch durch die Methode der evozierten Potentiale (EEG-Signale die mit bestimmten Stimuli in Bezug gesetzt werden können) ergänzt wurde [24]. Über lange Zeit war das EEG das einzige nichtinvasive Instrument für die Messung der elektrischen Hirnaktivität und es erlangte insbesondere für diagnostische Zwecke Bedeutung (z.B. für die Feststellung epileptischer Anfälle oder für die Charakterisierung verschiedener Schlafphasen). Von grosser Bedeutung war die in der ersten Hälfte des 20. Jahrhunderts entwickelte Technik der Ableitung der elektrischen Aktivität einzelner Neuronen bzw. Nervenfasern. Diese invasive Technik ermöglichte nicht nur die Entschlüsselung der genauen biophysikalischen Vorgänge, die bei der Erzeugung von Nervenimpulsen auftreten. Sie erlaubten auch eine gezielte Stimulation des Nervengewebes und damit eine Analyse der funktionellen Aspekte. Keine andere Methode erreicht eine vergleichbare räumliche Auflösung. In jüngerer Zeit finden auch *arrays* von Elektroden Anwendung, welche die Ableitung bzw. Stimulation ganzer Neuronenpopulationen erlauben. Auch die chronische Implantierung von Elektroden oder Elektrodenarrays wird untersucht. Dies unter anderem mit dem Ziel, neuartige *brain-machine-interfaces* zu kreieren, mit deren Hilfe beispielsweise dereinst Prothesen gesteuert werden sollen [6]. Vorab die chronische Implantierung von Elektroden ist mit Problemen verbunden, da durch das Eindringen der Elektrode das Gewebe verletzt wird und zudem eine Gegenreaktion des Körpers erfolgt (Bildung von "Narbengewebe" um die Elektroden, was deren Mess- bzw. Stimulationseigenschaften beeinträchtigt). Mittels transkranialer elektrischer Stimulation besteht schliesslich auch die Möglichkeit einer nichtinvasiven elektrischen Stimulation [146]. Dazu werden auf der Kopfhaut Elektroden derart angebracht, dass der Stromfluss zwischen den Elektroden auch durch das Gehirngewebe fliesst, was Neuronen zum Feuern bringt. Da mit dieser Methode auch die Gesichtsmuskeln gereizt werden können, was nicht schmerzlos ist, wird diese Methode nur selten eingesetzt.

Da transiente elektrische Ströme von Magnetfeldern begleitet sind bzw durch solche ausgelöst werden können, finden auch **magnetische Mess- und Stimulationsverfahren** Anwendung. Vor allem die transkraniale magnetische Stimulation (TMS) ist hier bedeutsam, da vermutet wird, dass magnetische Felder keine relevanten Nebenwirkungen auf die Versuchspersonen haben sollen [146]. Dazu wird ein starkes, transientes magnetisches Feld

aufgebaut, das im Innern des Gehirns elektrische Ströme induziert und Neuronen zum Feuern bringt. TMS hat zudem das Potential, die Aktivität gewisser Gehirnregionen zu unterbinden und kann damit temporäre Läsionen verursachen. Bei der mehrfachen Anwendung von TMS in der gleichen Hirnregion ist aber nicht auszuschliessen, dass dort bleibende strukturelle Veränderungen entstehen. Zudem ist die räumliche Auflösung von TMS nicht sehr gross (die *peak activity* erreicht eine Grössenskala von einem Zentimeter) und es ist unklar, in welchem Umfeld des Aktivitätspeaks Ströme induziert werden. So existiert eine Reihe anekdotischer Berichte über physiologische und verhaltensmässige Nebenwirkungen nach der Anwendung von TMS. Auch Ratten meiden starke Magnetfelder [52]:173, was gewisse Zweifel an der vermeintlichen Schadlosigkeit dieses Verfahrens nährt.

2.2.3 Bildgebende Verfahren

Die überwiegende Mehrzahl der in dieser Pilotstudie untersuchten Arbeiten aus dem Bereich Neurowissenschaft verwendet bildgebende Verfahren, so dass diese Methoden und die damit verbundenen Probleme genau vorgestellt werden. In den vergangenen Jahrzehnten wurden mehrere derartige Methoden in die Neurowissenschaft (und Neurologie) eingeführt. Insbesondere die funktionelle Magnetresonanz-Tomographie (fMRI) hat in jüngster Zeit starkes Interesse gefunden, weil diese Methode einen nichtinvasiven Einblick in die funktionelle Aktivität des Gehirns verspricht. Folgende *Imaging*-Verfahren sind in der Neurowissenschaft von Bedeutung [151, 159]:

- **Röntgenstrahlen-basierende Technologien:** Röntgenstrahlen sind seit Jahrzehnten in der Hirnforschung eingesetzt worden und sie bilden insbesondere in der Neurologie auch heute noch ein wichtiges diagnostisches Mittel. Das physikalische Grundprinzip dieser Methode beruht auf unterschiedlichen Durchdringungseigenschaften verschiedener biologischer Gewebe für Röntgenstrahlen, welche sich in einem für solche Strahlen empfindlichen Film entsprechend niederschlagen. So können also unterschiedliche Gewebedichten entsprechend abgebildet werden, der Kontrast ist aber eher gering. Durch Zugabe bestimmter Kontrastmittel kann aber beispielsweise das Blutgefässsystem des Gehirns dargestellt werden. Die klassische Röntgenaufnahme ist ein analoges Verfahren und es besteht keine Möglichkeit einer digitalen Nachbearbeitung des Ergebnisses. Zudem liefert das klassische Röntgenbild ein zweidimensionales Rissbild einer dreidimensionalen Struktur, was natürlich die Aussagekraft des erzielten Bildes weiter vermindert. Durch Computer-assistierte Röntgentomographie hingegen können auch dreidimensionale Strukturen mittels Röntgenstrahlen adäquat abgebildet werden. Dazu werden die Quelle der Röntgenstrahlen und der Detektor bewegt, so dass Aufnahmen aus verschiedenen Blickwinkeln möglich werden. Der Computer rekonstruiert aufgrund dieser Daten ein dreidimensionales Bild. Das damit erzeugte Bild enthält lediglich Informationen über anatomische Strukturen, nicht aber über funktionale Vorgänge im Gehirn. Die Methode führt zudem zu einer Belastung der Versuchsperson mit Röntgenstrahlen und kann deshalb nicht beliebig wiederholt werden. In den in dieser Pilotstudie untersuchten Arbeiten wird dieses Verfahren nicht verwendet.
- **Positron-Emissions Tomographie (PET):** PET ist ein weit verbreitetes *Imaging*-Verfahren, das im Unterschied zu röntgenbasierten Verfahren nebst anatomischer

Information auch solche über die funktionelle Aktivität der Gehirns liefern kann. Das physikalische Grundprinzip lässt sich kurz wie folgt erklären: In den Körper des Patienten werden kurzlebige radioaktive Substanzen injiziert, welche eine bestimmte Rolle im Stoffwechsel des Gehirns erfüllen (beispielsweise Fluordesoxyglucose mit radioaktivem Fluor) oder sonstwie durch das Gefässsystem im Gehirn verteilt werden (z.B. radioaktives Wasser). Diese Radionuclide (C^{11}, F^{18}, N^{13}, O^{15}) setzen bei ihrem Zerfall Positronen frei, das Antiteilchen des Elektrons. Ein so erzeugtes Positron trifft in unmittelbarer Nähe seines Entstehungsortes ein Elektron und zerfällt unter Freisetzung zweier Gamma-Quanten, welche in jeweils entgegengesetzte Richtung emittiert werden. Diese Gamma-Quanten durchdringen biologisches Material ohne Wechselwirkung und werden ausserhalb des Körpers durch einen Messapparat erfasst. Dieser Apparat bestimmt das gleichzeitige Eintreffen solcher Gammaquanten, welche aus demselben Zerfallsereignis herrühren und ermittelt dadurch den Ort des Zerfalls (Auflösung: 3-5 mm). Mit diesem Verfahren können – je nach verwendetem Radionuclid – verschiedene Informationen gewonnen werden: Die Verwendung von radioaktivem Wasser gibt Informationen über den Blutfluss im Gehirn. Durch die Verwendung von radioaktiver Glucose kann Information über erhöhte Stoffwechselaktivität gewonnen werden, da an diesen Orten vermehrt Glucose konsumiert wird. Durch die Verwendung radioaktiv markierter Rezeptorliganden kann Aufschluss über die räumliche Verteilung von ganz spezifischen Nervenzell-Rezeptoren erhalten werden. Kein anderes Verfahren erlaubt einen derart spezifischen Einblick in bestimmte metabolitische Prozesse im Gehirn. Diese Methode hat aber auch mehrere Nachteile: So unterwirft sich die Versuchsperson einer gewissen Strahlenbelastung, welche die mehrfache Wiederholung einer PET-Messung verbietet. Im weiteren ist die zeitliche Auflösung sehr ungenau und erlaubt kaum Aussagen über den zeitlichen Verlauf der untersuchten Prozesse. Das Verfahren ist auch teuer, weil die benötigten Radionucleide aufgrund ihrer Kurzlebigkeit vor Ort in einem Zyklotron hergestellt werden müssen. Das Verfahren ist schliesslich sehr sensitiv für Bewegungs-Artefakte. Ein alternatives Verfahren ist SPECT (single photon emission computerized tomography). Hier werden langlebige Radionuclide verwendet, welche direkt einzelne Photonen aussenden, die dann detektiert werden. Die räumliche Auflösung des Verfahrens ist aber tiefer als beim PET.

- **Magnetresonanztomographie (magnetic resonance imaging MRI):** Die Einführung von MRI in der Medizin (und in der Neurowissenschaft) gilt als eine der bedeutendsten wissenschaftlichen Innovationen der vergangenen Jahrzehnte. MRI ermöglicht die Gewinnung von Informationen über die Anatomie des untersuchten Systems in einer Auflösung von derzeit rund einem Kubikmillimeter. Das Volumenelement, welches diese Auflösung definiert, nennt man *Voxel* (parallel zum *Pixel* – der Auflösungsgrenze eines digitalen Bildes). Durch geeignete Wahl der Messparameter (Repetitionszeit und Echozeit – siehe unten) können unterschiedliche Gewebekontraste erreicht und damit unterschiedliche Aspekte hervorgehoben werden. Man kann auch Informationen über den Fett- und Proteinanteil von Geweben und damit über den Metabolismus erhalten. Neuere Geräte erlauben auch die Erzeugung von Filmen (z.B. Herzbeobachtung), vergleichbar mit Ultraschall-Bildgebung. Das physikalische Prinzip von MRI kann wie folgt skizziert werden: Bestimmte Atome haben einen (quantenmechanisch erklärbaren) Spin. In biologischen Geweben ist das Wasserstoffatom –

und damit Wasser – die dominierende Atomart mit dieser Eigenschaft. Insofern misst MRI im wesentlichen die strukturelle Verteilung von Wasser in einem Gewebe. Dies ist auch der Grund, warum Knochenstrukturen mit MRI nicht abgebildet werden können, weil diese kaum Wasser enthalten. Dieser Spin erzeugt einen magnetischen Dipol. Die Richtungen all dieser Dipole in biologischem Material ist zufällig verteilt. Durch Anlegen eines starken Magnetfeldes werden diese Dipole aber in eine Richtung ausgerichtet (z-Richtung). Durch einen weiteren magnetischen Puls, werden die Dipole in eine (zunächst phasengleiche) Präzession gebracht. Die Periode zwischen Anregungspulsen nennt man Repetitionszeit. Durch die Anregung erhält der Dipolvektor Komponenten in Richtung der x-y-Ebene. Die Frequenz der Präzession (Lamorfrequenz) ist proportional zur Stärke des Magnetfeldes, das die Spins ausrichtet. Dadurch, dass das Magnetfeld einen Gradienten besitzt, ist die Lamorfrequenz an unterschiedlichen Orten verschieden, was zur Lokalisierung des gemessenen Signals dient. Zwei Signale können gemessen werden. Erstens richten sich die Dipole nach dem magnetischen Puls wieder in z-Richtung aus. Dieser Vorgang ist mit einer Energieabgabe verbunden, der gemessen werden kann. Die Zeitkonstante dieses Vorgangs heisst T1-Zeit, sie hängt von der Stärke des Magnetfeldes wie auch von gewebetypischen Aspekten ab. Die T1-Zeit bestimmt gewissermassen, wie schnell sich die Dipole von der Anregung durch den magnetischen Puls erholen und wieder anregbar werden. Wird die Repetitionszeit kurz gewählt, so bestimmt im wesentlichen die T1-Komponente den Bildkontrast (T1-Wichtung). Zweitens desynchronisiert sich die durch den magnetischen Puls synchronisierte Präzession der Dipole ebenfalls auf eine charakteristische Weise. Auch dieses Signal kann gemessen werden, wobei hier die Dauer zwischen dem Zeitpunkt des magnetischen Pulses und der Messung – die Echozeit – entscheidend ist. Je länger die Echozeit ist, desto stärker erscheinen gewebetypische Unterschiede im T2-Signal (T2-Wichtung). Durch geeignete T1- und T2-Wichtung können so spezifische, auf die Fragestellung angepasste Bilder produziert werden und man erreicht mittels Computerunterstützung schliesslich ein dreidimensionales Bild der verschiedenen Gewebetypen im Messobjekt. Da MRI auf diesen magnetischen Eigenschaften des untersuchten biologischen Materials beruht, ist das Verfahren weit weniger problematisch für die Versuchspersonen als röntgenstrahlenbasierte Verfahren oder PET. Die Nachteile entsprechen jenen von funktioneller MRI und werden dort besprochen.

- **Funktionelle Magnetresonanztomographie (fMRI):** Eine wichtige Erweiterung erfuhr MRI zu Beginn der 1990er Jahre durch die funktionelle Magnetresonanztomographie, welche derzeit vor allem in der Variante des BOLD-MRI benutzt wird (BOLD steht für *blood oxygenation level dependent*). Dieses Verfahren beruht auf der Entdeckung, dass der Blufarbstoff Hämoglobin unterschiedliche magnetische Eigenschaften hat, je nachdem ob sich dieser in einem oxygenisierten Zustand (also Sauerstoff zur Zelle transportiert) bzw. desoxygenisierten Zustand (d.h. Sauerstoff an die Zelle abgegeben hat) befindet. Wird an einer bestimmten Stelle im Gehirn viel Sauerstoff verbraucht, so verändert sich das Verhältnis zwischen osygenisiertem und desoxygenisiertem Hämoglobin (die Hemodynamik). Das so genannte BOLD-Signal misst die Veränderung dieses Verhältnisses. Das Problem ist dabei aber, dass dieses Verhältnis durch drei unterschiedliche Prozesse bestimmt ist: durch das cerebrale Blutvolumen, durch den Blutfluss und durch die eigentliche Hemodynamik. Mit BOLD-MRI wird

also letztlich der Energieverbrauch des Gehirns bestimmt, wobei neuere Untersuchungen darauf hinweisen, dass das BOLD Signal tatsächlich mit der Aktivität der Neuronen korreliert – aber nicht wie ursprünglich vermutet mit der Feuerrate [89], sondern eher mit dendritischen Prozessen [113, 125, 157]. Diese Studien weisen aber auch darauf hin, dass das Signal/Rausch-Verhältnis der fMRI Signals wesentlich kleiner ist als jenes, das man mit Ableitungselektroden ermittelt. Demnach würde das fMRI-Signal die tatsächliche Aktivierung bestimmter Bereiche des Gehirns unterschätzen. Die genaue Dynamik wie auch die zugrunde liegenden Mechanismen des BOLD-Signals ist weiterhin nicht geklärt. Die physiologischen Prozesse während welchen die (venöse) Hemodynamik in den "nichtaktivierten" Zustand zurückfällt erstreckt sich über viele Sekunden [146]:22. Selbstverständlich benötigt während einer BOLD-MRI Messung das gesamte Gehirn Sauerstoff. Will man also eine Aussage darüber gewinnen, ob eine spezifische Aktivität den Energieverbrauch an einem spezifischen Ort erhöht, muss das BOLD-Signal immer gegenüber einer Referenzmessung (*baseline*) bestimmt werden. Man muss also ein experimentelles Paradigma derart wählen, dass die Versuchsperson im fMRI-Scanner (die Messapparatur) zwei ausreichend ähnliche *tasks* durchführt, die sich lediglich in der zu untersuchenden Eigenschaft unterscheiden (die damit verbundenen Probleme werden nachfolgend ausführlich diskutiert). Die räumliche Auflösung von fMRI ist – wie jene von MRI – gut und erstreckt sich in der Grössenordnung von einem Kubikmillimeter. Höhere Magnetfeldstärken erlauben im Prinzip eine feinere Auflösung. Die zeitliche Auflösung hat eine Grössenordnung von einer Sekunde und ist damit zu grob, um die genaue neuronale Dynamik abzubilden (die zeitliche Auflösung des EEG ist im Vergleich im Bereich einer Millisekunde). Theoretisch sind jedoch Auflösungen in der Grössenordnung von 0.5 Millimeter bzw. 50 Millisekunden möglich. Die Erzeugung eines fMRI-Bildes ist schliesslich mit einem erheblichen statistischen Aufwand verbunden, weil sich die gemessenen Aktivitäten im Testfall vergleichen mit dem Referenzfall teilweise um weniger als ein Prozent unterscheiden. fMRI ist damit eine anspruchsvolle Methode, bei welcher eine Reihe von Problemen beachtet werden müssen. Diese werden nachfolgend genau erläutert. Die attraktive Grundidee des BOLD-MRI ist, dass man die raumzeitlichen Veränderungen der neuronalen Aktivität erfassen will. Es geht nicht nur um ein Abbilden des Gehirns, sondern um eine Art "Parsing" [51]. Man will gewissermassen die Buchstaben und Wörter eines "Hirnsatzes" erkennen – man beachte die linguistische Konnotation, die durch den Ausdruck des *Parsing* hervorgerufen wird. Es wäre demnach falsch, fMRI lediglich als eine Methode anzusehen, welche eine neue Form von Lokalisation von Funktionen erlaubt, denn man will das mit einer Funktion verbundene zeitliche Aktivitätsmuster im Gehirn ermitteln, nicht nur die aktiven Regionen. Zum Schluss sei noch erwähnt, dass nebst der BOLD-MRI auch andere Varianten von fMRI existieren. So kann beispielsweise auch die Bewegung der Wassermoleküle erfasst werden. Wassermoleküle können sich im Gehirn nicht in beliebige Richtungen bewegen, sondern deren Bewegung wird durch die Faserstruktur etc. in bestimmte Richtungen begrenzt. Durch entsprechende statistische Auswertungen dieser Bewegungen lassen sich detaillierte Informationen über die Faserstruktur des Gehirns ermitteln. Weitere Varianten von Bildgebung sind derzeit in Entwicklung.

- **Optische Methoden:** Der Vollständigkeit halber soll noch auf optische Methoden

des Imaging hingewiesen werden. Hierbei handelt es sich in meisten Fällen um invasive Verfahren, die am offenen Gehirn (und demnach in Tierversuchen) durchgeführt werden. Optische Methoden können einerseits ebenfalls das BOLD-Signal erfassen, weil oxygenisiertes bzw. desoxygenisiertes Hämoglobin von unterschiedlicher Farbe ist. Andererseits sind Farbstoffe entwickelt worden, welche ihre Farbe in Abhängigkeit vom elektrischem Potential ändern (*voltage-sensitive dyes*). Mit solchen Farbstoffen lässt sich die elektrische Aktivität einer Vielzahl einzelner Neuronen bestimmen. Sie bilden also eine Alternative zu Multi-Elektroden *arrays*.

Die funktionelle Magnetresonanztomographie hat bei der Erforschung der neuronalen Grundlagen von komplexem menschlichen Verhalten – zu welchem sicherlich auch moralische Entscheidungen bzw. Handlungen zu zählen sind – einen grossen Stellenwert eingenommen. Zielsetzung dieser Studien ist in meisten Fällen eine Form von Lokalisierung: Man will jene Hirnregionen identifizieren, welche bei einem bestimmten Verhalten übermässig aktiv sind. Die Qualität dieser Studien ist aber teilweise mangelhaft, wie in den Gesprächen mit Fachpersonen des Imaging an der Universität Zürich (Alumit Ishai und Anton Valavanis) deutlich geworden ist. Alumit Ishai vom Institut für Neuroradiologie betont, dass die Problematik der richtigen Wahl der *baseline* die Aussagekraft einer allfälligen Lokalisierung erschwert. Viel relevanter ist ihrer Ansicht nach die Ermittlung der zeitlichen Dynamik des Aktivitätsmusters, da bei praktisch allen Formen von Verhalten das gesamte Gehirn Aktivität aufweist. Studien, welche komplexes Verhalten auf die übermässige Aktivität einer einzelnen oder einiger weniger Regionen zurückführen wollen, sollten demnach mit grosser Vorsicht genossen werden. Anton Valavanis, ebenfalls vom Institut für Neuroradiologie, ist diesbezüglich noch weit kritischer. Er bezweifelt bei der Mehrzahl der Forschenden, die *Imaging* anwenden, deren methodische Kompetenz. So läuft an seinem Institut derzeit eine Auswertung aller im Zeitraum 2001-2004 publizierten *Imaging*-Studien. Vorläufige Resultate lassen darauf schliessen, dass in rund 60 Prozent aller Abbildungen von fMRI-Messungen die anatomischen Bezeichnungen inkorrekt sind. Dies mag auch mit dem erwähnten Fehlen eines einheitlichen Hirn-Atlas zusammenhängen. Im weiteren werde seiner Ansicht nach in vielen Studien die Anatomie und funktionalen Eigenschaften des Gefäss-Systems (beispielsweise die Dichte von Blutgefässen in bestimmten Regionen und Änderungen des Blutdrucks im Verlauf von Experimenten) nicht berücksichtigt, was die Aussagekraft der Resultate weiter vermindere. Valavanis befürchtet, dass gerade im Gebiet der *social cognitive neuroscience*, wo neuronale Grundlagen komplexer menschlicher Verhaltensweisen mittels *Imaging* ermittelt werden sollen, sich viele Forschende den methodischen Problemen des *Imaging* nicht bewusst seien. Er beobachte ein "unkontrolliertes, euphorisches Vorgehen". Valavanis wies schliesslich auf einen weiteren problematischen Aspekt der heutigen *Imaging*-Technologie hin: Demnach erlaube die digitale, computergestützte Visualisierung eine bis hin zur Manipulation und Fälschung reichende Veränderung wissenschaftlicher Daten. So habe er an medizinischen Kongressen selbst erlebt, wie Referenten vor ihrem Vortrag Bildmaterial entsprechend veränderten, um die Ergebnisse eines neurochirurgischen Eingriffs zu beschönigen. Seiner Ansicht nach erlaubten die neuen Technologien nicht mehr nur ein *Imaging* (d.h. ein Abbilden) neuronaler Strukturen und Prozesse, sondern eine eigentliche *Visualization* (also ein aktives Verändern von Bildern) der Vorgänge im Nervensystem. Die damit einhergehenden Probleme reichen in den Problemkreis der Neuroethik und werden dort weiter besprochen (Abschnitt 4.2). Im Folgenden soll nun anhand des BOLD-MRI –

der in den für diese Pilotstudie untersuchten neurowissenschaftlichen Studien am häufigsten verwendeten Methode – aufgezeigt werden, welche methodischen Probleme mit *Imaging* Studien verbunden sein können (hauptsächlich verwendete Quellen: [30, 146, 159]):

- **Beeinträchtigungen der Versuchspersonen:** Der nichtinvasive Charakter von fMRI soll nicht darüber hinwegtäuschen, dass Versuchspersonen durchaus gewisse Beeinträchtigungen erfahren, welche auf das Ergebnis bestimmter Experimente rückwirken können. So ist der Raum, in dem sich die Personen befinden, sehr eng und laut (bis 120 Dezibel), was entsprechende psychische Folgen haben kann. Im weiteren gilt festzuhalten, dass die verwendeten magnetischen Felder zwar nichtionisierend sind, hingegen Wärme erzeugen. Sehr starke Felder können auch Ströme induzieren, deren möglichen Auswirkungen beispielsweise auf das Kreislaufsystem offenbar noch nicht hinreichend untersucht worden sind (siehe auch die Bemerkungen zur TMS). Schliesslich gilt noch anzumerken, dass die Anwesenheit magnetischer Gegenstände in der Nähe der Scanner gravierende Auswirkungen haben können, da diese vom Scanner angezogen werden und unter Umständen in den von der Versuchsperson besetzten Messbereich kommen können. In Einzelfällen ist es zum Tod von Versuchspersonen gekommen [106]. Diese Risiken sind aber in der Regel gut kontrollierbar. Bisher (Ende 2005) wurden weltweit gegen 200 Millionen MRI und fMRI Untersuchungen durchgeführt. Bekannt (publiziert) sind 14 Todesfälle (vorab wegen Herzschrittmachern, die inaktiviert wurden) und gegen 100 Verletzte) z.B. wegen unentdeckter Metallsplitter in den Augen).

- **Probleme im Messprozess:** Bei der eigentlichen Messung müssen eine Reihe von Fehlerquellen berücksichtigt werden: Bewegungen von Versuchspersonen (etwa hervorgerufen durch Sprechen) beeinflussen die Uniformität des magnetischen Feldes und damit die Auswertung der Daten. Bei bestimmten Regionen im Schädel (Stirnhöhlen) bestehen Luft-Gewebe-Grenzen, was die Messung gerade in möglicherweise für Studien der *social cognitive neuroscience* interessanten Regionen (präfrontaler Kortex) erschwert. Zahnimplantate, Haarspangen etc. können zu (leicht erkennbaren) Bildartefakten führen.

- **Das Problem der Variabilität:** Bisherige Studien lassen vermuten, dass sowohl die *inter-trial* Variabilität (also die an derselben Versuchsperson vorgenommenen Messungen bei Wiederholung eines bestimmten Experiments), wie auch die individuelle Variabilität sowohl in anatomischer wie auch in funktioneller Hinsicht gross sind. So können anatomisch gleiche Regionen bei verschiedenen Menschen unterschiedlich gross sein und auch bei der Wiederholung des (scheinbar) gleichen *tasks* können jeweils unterschiedliche Regionen maximale Aktivität aufzeigen. Eine Mittelbildung über verschiedene Personen könnte dabei zu einer nur vermeintlichen Lokalisierung führen. Zudem ist es möglich, dass die Anatomie verschiedener Hirnregionen eine unterschiedliche Varianz aufweist, was die Mittelbildung weiter erschwert.

- **Probleme der experimentellen Methode:** Damit sind vor allem zwei Probleme angesprochen. Zum einen stellt sich das Problem der Wiederholbarkeit: Um statistische Aussagekraft zu erreichen, müssen Versuchspersonen mehrfach denselben Stimuli ausgesetzt werden. Dies wird dann zu einem Problem, wenn ein Stimulus verwendet wird, der eine emotionale Reaktion bei der Versuchsperson auslösen soll. Es

ist zu erwarten, dass ein Gewöhnungseffekt auftritt (bzw. der Stimulus erreicht mit mehrfacher Präsentation eine immer geringere emotionale Reaktion), der weit stärker ist als beispielsweise bei der Untersuchung der optischen Informationsverarbeitung. Zum anderen stellt sich das Problem der *baseline*: Es hat sich gezeigt, dass unterschiedliche Kontrollbedingungen unterschiedliche Aktivierungsmuster erzeugen können [30, 152]. Als Lösung für dieses Problem wird das so genannte *mixed design* von *Imaging*-Studien vorgeschlagen, wonach man zuerst die möglichen interessanten Regionen erhöhter neuronaler Aktivität identifizieren soll und danach die zeitliche Dynamik des Signals in diesen Regionen misst [51]. Im weiteren muss man sich der Natur eines Subtraktionsverfahrens bewusst sein – bei der Erstellung eines fMRI-Bildes wird ja schliesslich immer eine Versuchsbedingung mit einer Referenzbedingung verglichen bzw. die jeweils erzielten Werte der Aktivität werden Voxel für Voxel voneinander subtrahiert. Mathematisch gesehen wird eine solche Abbildung immer ein Maximum und ein Minimum aufweisen, d.h. man findet immer Regionen erhöhter Aktivität.

- **Das Problem der Korrelation psychischer und physiologischer Entitäten:** Dies betrifft ein begriffliches Grundproblem des *Imaging*, das letztlich an das Leib-Seele-Problem anknüpft: Kann man das zu untersuchende psychischen Phänomene hinreichend genau definieren, um es mit den gemessen Aktivitätsmustern in Beziehung setzen zu können? Hier sind Zweifel angebracht, weil viele dieser Phänomene vielleicht gar keine gar keine psychobiologischen Entitäten sind, sondern Manifestationen der verwendeten experimentellen Methoden und Theorien. Zudem sind solche Prozesse in ihrer Gesamtheit nicht zugänglich, weil nur subjektiv erfahrbar. Die durch fMRI festgestellte, erhöhte Aktivität bestimmter Regionen könnte zudem von unbewussten Hirnaktivitäten herrühren, die dem Stimulusparadigma gar nicht zugänglich sind – was die entsprechende Korrelation zusätzlich erschwert. Gemäss Uttal sind alle bisherigen Ansätze zur Erstellung einer Taxonomie mentaler Aspekte aufgrund solcher gescheitert, so dass eine Lokalisierung von Verhaltensphänomenen im Sinn, dass diese mit der erhöhten Aktivität einiger weniger Hirnregionen in Beziehung gesetzt werden können, nicht funktionieren könne – abgesehen von der Lokalisation in den primären sensorischen und motorischen Bereichen des Kortex.

- **Statistische Probleme:** Ein fMRI-Bild ist das Resultat ausgefeilter statistischer Analysen. Hier stellt sich das Problem des Schwellenwertes, anhand dessen sich Regionen einer statistisch signifikant erhöhten Aktivität identifizieren lassen. So zeigt sich, dass unterschiedliche – aber allesamt hohe Signifikanz anzeigende – Schwellenwerte zu unterschiedlichen Bildern führen. Ausserdem können Untersuchungen bei gleichem experimentellen Paradigma, gleichen Versuchspersonen und gleichem Schwellenwert, die mit einem Scanner mit stärkeren Magnetfeldern durchgeführt werden, ebenfalls unterschiedliche Bilder erzeugen. Dies ist eine Folge des statistischen Paradox von Paul Meehl, wonach die Frage 'Unterscheidet sich Gruppe A von Gruppe B hinsichtlich der Eigenschaft X statistisch signifikant?' je eher mit Ja beantwortet wird, desto stärker der verwendete statistische Test ist (d.h. genaueres Messen, grössere Grundgesamtheit, längere Messzeit) für *beliebige* A,B,X. [146]. Zudem müssen wegen der statistischen Natur des Resultats zwei Arten von Fehler beachtet werden: Typ I Fehler (*false positive*) und Typ II Fehler (*false negative*). Optimierung in eine Richtung führt zu einem erhöhten Risiko, Fehler des jeweils anderen Typs in Kauf zu nehmen. Schliesslich kann

es bei *Imaging*-Studien vorkommen, dass man sich von Anfang an nur auf bestimmte Regionen des Gehirns konzentriert, um dem Einfluss grösserer statistischer Fluktuationen in anderen Gebieten auszuweichen. Damit setzt man sich aber auch der Gefahr aus, relevante Aktivität in anderen Gebieten zu verpassen.

- **Das Problem der Aussagekraft der Bilder:** Die durch fMRI gewonnen Bilder haben eine starke suggestive Wirkung. Hierbei muss aber berücksichtigt werden, dass unterschiedliche Resultate der statistischen Analyse kommuniziert werden können: Die Position des am stärkste aktivierten Voxels, die Position des Schwerpunkts des signifikant aktivierten Clusters, der Rand dieser Cluster oder das gesamte Aktivitätsmuster. Je nach Darstellungsweise erzeugt man unterschiedliche Bilder: So sind bei einer Studie den Versuchspersonen unterschiedliche Objekte gezeigt worden und die Auswertung des Resultats hat ergeben, dass bei jedem Objekt das maximal aktivierte Voxel an einem anderen Ort ist, was zu drei unterschiedlichen Bildern führt. Werden hingegen die jeweils statistisch signifikant aktivierten Regionen gezeigt, so sind die Bilder viel ähnlicher, weil bei allen Präsentationen alle drei Regionen statistisch relevant aktiv waren (Beispiel aus [146]). Dieses Problem verlangt offenbar, dass man jeweils möglichst das gesamte Aktivitätsmuster aufzeigen sollte. Hier zeigt sich im Übrigen ein oft unpräziser Sprachgebrauch, indem man diese Strukturen als "aktiviertes Netzwerk" definiert, ohne dass man die Architektur dieses Netzwerks aufzeigt [117]:1235. Theoretische Studien wecken zudem Zweifel, ob durch simples Beobachten solcher Netzwerke die Hierarchie der Netzwerkverbindung ermittelt werden kann [93]. Zu nennen ist in diesem Kontext auch das Problem der Falschfarben, welche für die Darstellung statistischer Signifikanz verwendet werden. Diese werden so gewählt, um beim Betrachter einen psychologischen Effekt dergestalt auszulösen, dass die Schlussfolgerungen der Autoren gestützt werden. Es macht vermutlich einen Unterschied beim Betrachter, wenn solche Bilder grau-skaliert gezeigt werden oder mit einer Farbskala, so dass rot eine höhere Intensität bedeutet. Hier wäre eine Theorie der Wahrnehmung solcher Bilder durchaus nützlich. Im weiteren ist noch nicht klar, welcher Zusammenhang zwischen dem fMRI-Signal und der Information, welche diese Aktivität repräsentieren soll, besteht. Eine informationstheoretische Modellstudie zu diesem Thema kommt zum Schluss, dass dieser Zusammenhang hochgradig nichtlinear ist [126]. Dies bedeutet insbesondere, dass das Voxel mit der höchsten Aktivität nicht notwendigerweise jenem Voxel entspricht, dass die meiste sensorische Information kodiert. Um diesen Effekt zu korrigieren, müsste in der Auswertung der Daten Modelle über die bekannten *tuning properties* von Neuronen berücksichtigt werden. Schliesslich muss noch gesagt sein, dass eine durch fMRI festgestellte erhöhte Aktivität bestimmter Regionen auch auf die Aktivität *inhibitorischer* Neuronen zurückgeführt werden könnte. Demnach würde diese festgestellte erhöhte Aktivität in Tat und Wahrheit einer reduzierten Aktivität des betreffenden Gebietes im Netzwerk entsprechen. Inhibitorische Neuronen sind im Regelfall mit lokalen Neuronen verknüpft und würden demnach beispielsweise die auf andere kortikale Felder projizierenden Neuronen blockieren. All diese Aspekte zeigen Grenzen in der Aussagekraft von fMRI Bildern auf. Hier entstehen Probleme, wenn fMRI Bilder durch Laien beurteilt werden – beispielsweise im Rahmen eines Geschworenenprozesses. Die damit verbundenen neuroethischen Fragen werden im Abschnitt 4.2 weiter behandelt.

- **Das Problem des Umgangs mit fMRI-Daten:** Die Speicherung und Verteilung der durch fMRI gewonnenen Daten ist mit weiteren Problemen belastet. Nach der Jahrtausendwende sind im Rahmen des *BrainMap*-Projektes erhebliche Anstrengungen für den Aufbau von *Imaging*-Datenbanken unternommen worden.[5] Die dabei entstehenden Probleme sind aber weit schwieriger zu lösen als bei den Genomik und Proteomik-Datenbanken – ebenfalls ein Grossprojekt hinsichtlich der Speicherung wissenschaftlicher Daten. Wie oben deutlich geworden ist, spielt bei fMRI-Experimenten jedes Detail der experimentellen Anordnung eine Rolle. Wie erfasst man diese Vielfalt und wie kann man damit Aussagen erreichen, die über die bereits Mitte des 20. Jahrhunderts bekannte Lokalisierung hinausgehen? Noch 2001 wurde beklagt, dass es keine akzeptierten Standards für die Struktur und den Inhalt solcher Datenbanken gibt (Datenformate variieren, wie klassifiziert man das Verhalten der Versuchspersonen, etc.) [129]. Auch Fragen der Datensicherheit stellen sich (siehe Abschnitt 4.4). Beim Aufbau von *Imaging*-Datenbanken sind im Übrigen nicht nur eine Reihe technologischer Probleme zu lösen, sondern es stellen sich auch wissenschaftssoziologische Fragen hinsichtlich der Etablierung eines Standards (das Talairach-System), der Kooperation verschiedener wissenschaftlicher Disziplinen (medizinisch orientiertes *Imaging* vs. Datenbankexperten) und zur ganzen Frage der Theoriebeladenheit von Beobachtungen. Diese Fragen können hier nicht weiter diskutiert werden, es sei auf die Arbeiten von Anne Beaulieu verwiesen [12, 13, 14].

Die Auflistung dieser Vielzahl von Problemen soll nicht dahingehend interpretiert werden, dass *Imaging* ein unbrauchbarer Ansatz für die Untersuchung der Struktur und Funktion neuronaler Systeme ist. Ganz abgesehen vom unbestrittenen Nutzen dieser Technologien in der Medizin, können Verbesserungen im Bereich fMRI und Kombinationen verschiedener Technologien (beispielsweise fMRI mit EEG) durchaus einen tieferen Einblick in die raumzeitliche Dynamik der neuronalen Aktivität geben. *Imaging* allein wird aber das Problem, wie diese Aktivierungsmuster mit dem damit einhergehenden Verhalten verknüpft ist, nicht lösen können.

2.2.4 Methoden der experimentellen Ökonomie

Die experimentelle Ökonomie untersucht, wie sich Menschen in definierten Randbedingungen (in so genannten Spielen) in ökonomische Interaktion (z.B. Kauf- und Verkaufshandlungen, Investitionen) verhalten. Diese Spiele wurzeln in Konzepten, die im Rahmen der Spieltheorie ab der Mitte des 20. Jahrhunderts (unter anderem von John von Neumann und Oskar Morgenstern) entwickelt wurden. Die Spieltheorie (*game theory*) ist ein Teilgebiet der Mathematik, des Operations Research und der Wirtschaftswissenschaften und beschäftigt sich mit der Analyse von Handlungsstrategien in Systemen mit vorgegebenen Regeln. Dazu untersucht die Spieltheorie vorhergesagtes und tatsächliches Verhalten von Akteuren in Spielen und leitet optimale Strategien her. Für letzteres wurden im Rahmen der Ökonomie experimentellen Spiele angewandt, mit welchen man mittels ökonomischen Anreizen das Verhalten von Spielern untersucht, um Konzepte wie Vertrauen und Kooperation genauer zu analysieren [62]. Wie Urs Fischbacher vom Institut für Empirische

[5] Siehe `http://ric.uthscsa.edu/projects/brainmap.html`

Wirtschaftsforschung der Universität Zürich im Gespräch ausführte, wird mit solchen Untersuchungen ein psychologisch angereichertes Modell des *homo oeconomicus* angestrebt, zumal mehrere experimentellen Situationen bekannt sind die zeigen, dass sich Versuchspersonen in solchen Situationen nicht gemäss den Voraussagen des Modells des *homo oeconomicus* verhalten. Ziel dieser Spiele ist es, quantitative Aussagen über die Motive der Versuchspersonen zu gewinnen, beispielsweise Vertrauen (siehe Abschnitt 2.4.6) oder Grosszügigkeit, also durchaus moralnahe Konzepte. Gewiss lassen sich solche Motive nicht immer eindeutig feststellen, obgleich Kontrollexperimente die Zahl möglicher Alternativen durchaus einzuschränken vermögen. Dennoch sind solche Spiele ein interessanter Ansatz, um quantitative Aussagen (beispielsweise messbar durch die Geldbeträge, welche in einem Spiel ausgetauscht werden) über solche Motive zu gewinnen. Die Struktur der Spiele erlaubt es zudem, den Effekt von Institutionen einzubeziehen – beispielsweise einer strafenden Partei. In den vergangenen Jahren wurde deshalb eine Vielzahl solcher experimenteller Spiele entwickelt und angewandt. Solche Experimente wurden auch in unterschiedlichen Kulturen durchgeführt, um den Einfluss kultureller Komponenten abschätzen zu können [92]. In jüngerer Zeit wurden solche Spiele auch mit Methoden der Bildgebung kombiniert, um mehr über die Vorgänge auf der neuronalen Ebene zu untersuchen, wenn die Versuchspersonen solche Spiele spielen. An dieser Stelle soll deshalb eine kurze Übersicht über einige dieser Spiele gegeben werden (Quellen: [28, 58, 62, 114]):

- **Diktator-Spiel:** In diesem einfachsten Zwei-Personen-Spiel treten Spieler A und B aufeinander. A verfügt über einen (normierten) Betrag 1 und entscheidet, welchen Betrag $0 \leq x \leq 1$ Spieler B erhalten soll. Dieses Spiel ermöglicht eine Quantifizierung des Motivs Grosszügigkeit.

- **Ultimatum-Spiel:** In diesem zweistufigen Zwei-Personen-Spiel treten Spieler A und B aufeinander. A verfügt über einen (normierten) Betrag 1 und entscheidet, welchen Betrag $0 \leq x \leq 1$ Spieler B erhalten soll. Spieler B kann nun aber x akzeptieren oder zurückweisen. Akzeptiert B, so erhält B den Betrag x und A den Betrag $1 - x$. Lehnt B das Angebot ab, so erhalten beide den Betrag 0. Dieses Spiel ermöglicht eine Quantifizierung des Motivs Fairness.

- **Gefangenen-Dilemma:** Zwei Spieler A und B sind mit dem Problem konfrontiert, eine binäre Entscheidung zwischen den Varianten "kooperieren" (*cooperate*: C) oder "den anderen verraten" (defektieren, *defect*: D) zu treffen. Die für jeden Spieler vorgesehenen Belohnungen (*payoffs*) werden wie folgt bestimmt: $DC > CC > DD > CD$ und $CC > (CD + DC)/2$. Jeder Spieler würde also jeweils dann am meisten erhalten, wenn man selbst den anderen verrät, der andere aber kooperieren würde. Defektieren ist also die dominierende Strategie, aber eine pareto-Verbesserung ist möglich. Das Gefangenen-Dilemma lässt sich parallel (beide Spieler entscheiden unabhängig voneinander, ob sie C oder D spielen) oder sequenziell (zuerst entscheidet A über die Wahl von C oder D, dann entscheidet B basierend auf dem Wissen über die Entscheidung von A) spielen.

- **Vertrauens-Spiel:** In diesem zweistufigen Zwei-Personen-Spiel, das man als eine Variante des Gefangenen-Dilemmas auffassen kann, treten Spieler A und B aufeinander. Spieler A entscheidet in einem ersten Schritt über (in der Regel zwei) mögliche

Varianten, wie Geldbeträge aufgeteilt werden. In der ersten Variante wird ein geringer Geldbetrag fair (d.h. *fifty-fifty*) zwischen beiden Spielern verteilt. In der zweiten Variante erhält Spieler B die Kompetenz, einen weit grösseren Betrag zwischen beiden Spielern zu verteilen. A geht mit dieser zweiten Wahl aber das Risiko ein, dass B die Wahl derart trifft, dass A weniger erhält als in der ersten Variante. In diesem Spiel wird das Vertrauen eines Spielers in die Reziprozität des anderen Spielers gemessen (im Gegensatz etwa zum Vertrauen in die Fähigkeiten eines anderen).

- ***Public-good*-Spiele:** Von diesem (eher schwierig zu interpretierenden) n-Personen-Spiel ($n \geq 2$) gibt es verschiedene Varianten. Die Grundstruktur sieht wie folgt aus: Jeder Spieler $A_1, \ldots A_n$ besitzt am Anfang den gleichen Betrag 1 und entscheidet, welchen Betrag $0 \leq x_i \leq 1$ er an ein Gemeinschaftsgut $X = F(\sum_{i=1}^{n} x_i)$ (*public good*) abgeben will. Das Gemeinschaftsgut X wird nachher wieder unter den Spielern fair verteilt (d.h. jeder erhält den Anteil X/n). X ist eine Funktion aller Einzahlungen dergestalt, das die *payoffs* die Struktur des Gefangenen-Dilemmas aufweisen: Alle würden profitieren, wenn alle Spieler einen hohen Beitrag an X leisten. Der einzelne Spieler A_i profitiert jedoch dann am meisten, wenn alle andern viel geben, A_i aber nichts (Trittbrettfahrer). Geben alle nur wenig, verlieren alle.

- **Wettbewerbs-Spiele:** Hier handelt es sich um n-Personen-Spiele, die der Grundstruktur des Ultimatum-Spiels folgen. Man unterscheidet Wettbewerbs-Spiele mit Angebots-Wettbewerb und Wettbewerbs-Spiele mit Nachfrage-Wettbewerb (*responder*-Wettbewerb):

  - Im Wettbewerbs-Spiel mit Angebots-Wettbewerb treten $A_1 \ldots A_{n-1}$ Anbieter und ein Nachfrager B auf. Jeder Anbieter macht B ein Angebot im Sinn des Ultimatum-Spiels. B kann nun das höchste aller Angebote entweder ablehnen (dann erhält niemand etwas) oder annehmen (dann erhalten B und der höchste Anbieter A_i einen *payoff* gemäss dem Angebot).
  - Im Wettbewerbs-Spiel mit Nachfrage-Wettbewerb tritt ein Anbieter A und $B_1, \ldots B_{n-1}$ Nachfrager auf. Der Anbieter macht ein Angebot im Sinn des Ultimatum-Spiels. Alle Nachfrager entscheiden gleichzeitig, ob sie das Angebot annehmen oder ablehnen wollen. Lehnen alle ab, gehen alle leer aus. Nehmen einige Nachfrager an, so wird unter dieser Population ein Nachfrager B_i zufällig ausgewählt und der *payoff* wird zwischen A und B_i entsprechend dem Angebot verteilt.

- **Drittperson-Bestrafungs-Spiele:** Eine Reihe der genannten Spiele (z.B. Gefangenen-Dilemma) lassen sich so abwandeln, dass eine dritte Person C das Verhalten der beiden anderen Spieler beobachten und danach bestrafen kann, indem den anderen Spielern ein Geldbetrag abgezogen wird. Dieses *third-party-punishment* kann so gespielt werden, dass der Akt der Bestrafung für C gratis oder *costly* ist, d.h. im zweiten Fall muss C für den Akt der Bestrafung selbst etwas bezahlen. Ein *public good game* wiederum kann so gespielt werden, dass die einzelnen Spieler erfahren, wie jeder gespielt hat und sie können danach Trittbrettfahrer bestrafen, indem sie deren *payoff* reduzieren.

All diese Experimente können derart ausgestaltet werden, dass die Spieler nur ein einziges Mal aufeinander treffen (*one-shot games*) oder aber mehrfach aufeinander treffen. Zudem können die Auswahlvarianten begrenzt werden.

2.3 Skizze des wissenschaftlichen Umfeldes

Die bibliometrische Untersuchung zeigt das in jüngster Zeit gestiegene Interesse der Neurowissenschaft an einer Untersuchung der neuronalen Grundlagen komplexer menschlicher Verhaltensweisen und psychischer Konzepte. Ein Beispiel ist das zunehmende Interesse an der Suche nach neuronalen Korrelaten des Bewusstseins. Dieses weit gefächerte Forschungsgebiet steht sicherlich in einem Bezug zu Fragen der moralischen Kognition, zumal die Definition zentraler Begriffe der Ethik wie "Handlung", "Autonomie" und "Entscheidung" davon ausgeht, dass die bei der Wahrnehmung dieser Fähigkeiten auftretenden kognitiven Vorgänge bewusste Vorgänge sind [128]. Eine umfassende Einführung in die Bewusstseinsforschung kann hier nicht erfolgen. Ein summarischer Blick zeigt jedoch, dass spezifisch ethische Fragestellungen in der engeren Bewusstseinsforschung vermutlich eher selten auftauchen. Daneben gibt es eine Reihe von Forschungsgebieten, welche in der Literatur über die neurobiologischen Grundlagen der Moral oft auftauchen und unterstützende Konzepte liefern. Diese Gebiete erforschen die neurobiologischen Grundlagen des Sozialverhaltens (*social cognitive neuroscience*), der Emotionen und die so genannten Spiegelneuronen. Selbstverständlich gehört eine Reihe von weiteren, in der Psychologie wurzelnde Gebiete in das wissenschaftliche Umfeld der Erforschung der Grundlagen der Moral (Entwicklungspsychologie, Moralpsychologie), welche nicht weiter vorgestellt werden, da diese in den Bereich der zweiten Pilotstudie fallen.

2.3.1 Neuronale Grundlagen des Sozialverhaltens

Das erst in jüngerer Zeit entstandene Teilgebiet der *social cognitive neuroscience* hat die neuronalen Grundlagen des Sozialverhaltens sowohl von Tieren wie Menschen als Forschungsgegenstand [3, 21] und die Untersuchung der Neurobiologie der Moral lässt sich demnach als Teilgebiet dieses Forschungsbereichs auffassen. Der Boom in der *social cognitive neuroscience* geht einher mit einem wachsenden Interesse an der Erforschung von Emotionen (siehe Abschnitt 2.3.2) und hat damit auch verwandte Fragestellungen. Was genau Gegenstand der *social cognitive neuroscience* sein soll, wird unterschiedlich ausgeführt. Adolphs [3] nennt drei Grundfragen: In welchem Verhältnis stehen Kognition und Emotion? In welchem Verhältnis stehen Wahrnehmung und Handlung? Worin besteht der Unterschied in der Wahrnehmung der eigenen Person gegenüber der Wahrnehmung anderer Personen? Blakemore et al. [21]:216 fokussieren die Frage: Lässt sich soziales Verhalten unter Rückgriff auf bestehende Erkenntnisse über allgemeine kognitive Fähigkeiten wie Wahrnehmung, Sprache, Gedächtnis und Aufmerksamkeit erklären, oder treten bei sozialen Interaktionen spezifische, neue kognitive Prozesse auf? Insel und Russell [99] schliesslich identifizieren vier Grundfragen: Wie werden soziale Signale wahrgenommen? Wie bilden sich Gedächtnisinhalte über soziale Aspekte? Was ist die Motivation für das Eingehen sozialer Bindungen (z.B. Eltern-Verhalten)? Was sind die neuronalen Konsequenzen von sozialem Verhalten? Gewiss sind viele dieser Fragen nicht wirklich neu. Auch stellt sich generell die Frage, wie genau das "Soziale" in den einzelnen Untersuchungen der *social cognitive neuroscience* identifiziert und von nicht-sozialen Aspekten angegrenzt werden kann. Die damit verbundenen Probleme sind den Forschern in diesem Bereich nicht verborgen geblieben. So wird festgestellt [99], dass die experimentellen Paradigmen für Sozialverhalten meist viel einfacher sind als in der realen Interaktion. In Tierversuchen gehe zudem oft vergessen, dass Labortiere ein an-

deres Verhaltensspektrum mitbringen als deren Artgenossen in natürlichen Lebensräumen. Schliesslich würden bereits bestehende Kenntnisse der Verhaltenswissenschaften oft nicht genügend berücksichtigt. Es ist demnach offenbar unklar, bis zu welchem Ausmass sich die *social cognitive neuroscience* hinsichtlich ihres Forschungsgegenstandes mit anderen Gebieten wie der Sprachforschung überschneiden. Diese allgemeinen Probleme sollen hier aber nicht weiter ausgeführt werden. Vielmehr werden einige exemplarische Problemstellungen der *social cognitive neuroscience* etwas genauer vorgestellt, insofern sie einen Bezug zu jenen Studien haben, die explizit Moral thematisieren.

- Zum ersten erschöpft sich die Mehrzahl der untersuchten Studien in der Lokalisation von Hirnregionen, die bei sozialer Kognition eine Rolle spielen sollen. Demnach soll es im Temporallappen mehrere Regionen geben, welche sozial relevante Stimuli verarbeiten. Die Amygdala, der rechte somatosensorische Kortex, der orbotofrontale Kortex und das Cingulum sollen solche Wahrnehmungen mit Aspekten wie Motivation, Emotion und Kognition verknüpfen [5, 4]. In diesem allgemeinen Sinn helfen jedoch solche Charakterisierungen für das Verständnis der sozialen Kognition nicht weiter. Auffällig sind vielmehr die doch sehr allgemein gehaltenen Problemstellungen. So fragt beispielsweise Adolphs: "Are the information-processing demands made by social cognition different from those made by non-social cognition?" [4]:231. Einmal vorausgesetzt, man könne klar zwischen *social cognition* und *non-social cognition* unterscheiden: Was genau besagt der Ausdruck "information-processing demands"? Ist damit lediglich das Ausmass (im Sinn von genauer Lokalisierung und Anzahl) der übermässig aktivierten Gebiete im Gehirn gemeint? Dazu muss grundsätzlich folgendes festgehalten werden: Niemand bestreitet, dass Gedanken und Handlungen von Menschen von Aktivierungen bestimmter Hirnregionen begleitet sind. Vergleicht man unterschiedliche Gedanken und Handlungen, wird man gewissermassen *per definitionem* unterschiedliche Aktivierungen sehen – sofern die raumzeitliche Auflösung der Bildgebung für die Ermittlung der Aktivierung geeignet gewählt ist. Insofern beantwortet sich die Frage von Adolphs trivialerweise mit Ja, was letztlich zeigt, dass die Frage in dieser Form kein relevantes Problem anspricht.
- Ein zentraler Aspekt ist *decision making* in einem sozialen Kontext. "Entscheiden" wird dabei definiert als der Prozess, eine Wahl zu treffen oder zu einer Konklusion zu kommen [131]. Dieser Prozess wird unterteilt in die Bildung von Präferenzen bezüglich möglicher Optionen, die Wahl und Ausführung einer Handlung und die Erfahrung und Evaluation des Resultats der Handlung. Es wird dabei aber anerkannt, dass in den Studien der *social cognitive neuroscience* oft ein eher rudimentärer Begriff von "entscheiden" verwendet wird. Dies führt zur Forderung nach einer Taxonomie von Entscheidungssituationen [131]. Konkret werden verschiedene Facetten eines Entscheidungsprozesses untersucht:
  - Eine Frage ist, wie genau Wahrnehmungen eine Entscheidung formen und beeinflussen [3]. Gibt es beispielsweise Wahrnehmungen, welche quasi automatisch zu bestimmten Entscheidungen führen? Hier stellt sich natürlich die Anschlussfrage, ob dies überhaupt noch "Entscheidung" genannt werden soll. Interessanter ist hier die Frage, wie genau gewisse "Färbungen" von Wahrnehmungen einen Entscheidungsprozess unbewusst beeinflussen können. So sind beispielsweise Gesichter ein

für Menschen wichtiger Stimulus und gewisse Gesichtsausdrücke vermögen unbewusst Entscheidung vorzuspuren. Fraglich ist nur, ob der derzeit zur Verfügung stehende methodische Apparat der *social cognitive neuroscience* (in erster Linie *Imaging*) ausreicht, um hier Erkenntnisse zu gewinnen, die über das bekannte Wissen der Wahrnehmungspsychologie hinausgeht. In dem sich langsam entwickelnden Modell einer neuronalen Grundlage von Moral wird dem Aspekt des Automatismus bzw. der "Färbung" von Wahrnehmungen übrigens eine wichtige Rolle zugeordnet. So wird postuliert, dass viele so genannt moralische Entscheidungen einen automatischen Charakter haben und dass Rechtfertigungen dafür erst *post fakto* generiert werden (mehr dazu unter Abschnitt 3.5).

– Eine zweite Frage ist, wie Handlungen Anderer wahrgenommen werden. Auch diese Frage wird derzeit primär im Sinn einer Lokalisation abgehandelt, indem beispielsweise untersucht wird, welche Hirnregionen besonders aktiv werden, wenn man die Handlungsabsichten anderer erkennen will (dies soll der superiore temporale Sulcus sein) [132]. Dies ist eine wichtige Komponente sozialer Kognition, weil die Erkennung solcher Absichten die eigenen Handlungen beeinflussen.

– Eine dritte Forschungsrichtung sucht nach einer Verbindung zwischen dem mittels spieltheoretischer Methoden untersuchten Konzept der *rational choice* und dem empirisch festgestellten Entscheidungsverhalten (siehe dazu auch Abschnitt 2.4.7). So wird seit längerem postuliert, dass das Konzept der *rational choice* das tatsächliche menschliche Verhalten nicht erklärt (siehe dazu die Diskussion in [41]). Berns schlägt diesbezüglich vor, mittels *Imaging* jene Hirnregionen zu identifizieren, die den verschiedenen möglichen Gleichgewichten der spieltheoretischen Erwägungen entsprechen: "The interaction of different pools of neurons in the brain may result in phenotypic behavior that appears to be irrational, but it is possible that the rational agents are the neurons, not the person." [41]:156. Derartige Vorschläge haben zur Idee geführt, die Vorgänge im Gehirn mit den Begriffen der Ökonomie zu untersuchen. So fragen Montague und Berns beispielsweise, ob es eine Art "Währung" im Gehirn gebe, mit welcher man den im Gehirn ablaufenden Prozess einer Entscheidung universell messen oder bewerten könne [123]: "Without internal currencies in the nervous system, a creature would be unable to assess the relative value of different events like drinking water, smelling food, scanning for predators, sitting quietly in the sun, and so forth. To decide on an appropriate behavior, the nervous system must estimate the value of each of these potential actions, convert it to a common scale, and use this scale to determine the course of action." Sie schlagen vor, dass die Aktivität im *orbitofrontal-striatal circuit* die gewünschte "Währung" sein könne, wobei auf eine *Imaging*-Studie verwiesen wird, welche die erhöhte Aktivierung dieses Bereichs in Korrelation zu *monetary reward* gezeigt habe. Auch Glimcher benutzt ein ökonomisches Vokabular für die Vorgänge im Gehirn [73]: Das Gehirn sei dazu da, effiziente Entscheide (im Sinn der Ökonomie) zu treffen. Glimcher präsentiert diese These als Ausdruck einer sich veränderten Sichtweise auf das Gehirn. Früher seien Entscheidungen in einem senso-motorischen Kontext gesehen worden – man suchte nach *stimulus-response* Schemata im Sinn des Behaviorismus. Die neuronalen Systeme, welche Bewertungen, Wahrscheinlichkeiten

und Nützlichkeit von Handlungen umfassen, seien damals jedoch nicht untersucht worden, weil die Methoden für deren Untersuchung fehlten. Das *Imaging* habe dies nun aber verändert und ermögliche einen Einblick in die *black box* der Behavioristen (das Gehirn). Diese Entwicklungen lassen sich als den Versuch einer Naturalisierung des Entscheidungsbegriffes auffassen, was längerfristig zweifellos auch für die Ethik Auswirkungen haben wird. Zum jetzigen Zeitpunkt lassen sich aber eine Reihe von Problemen identifizieren. So ist zum ersten zweifelhaft, ob *Imaging* hinsichtlich der räumlichen wie zeitlichen Auflösung wirklich die gewünschte Genauigkeit aufweist, um die bei Entscheidungsprozessen auftretenden neuronalen Vorgänge zu erfassen. Begriffe wie "Währung" werden zudem auf einen ganz einfachen Kern (einheitliche Repräsentation unterschiedlicher zu bewertender Sachverhalte) zurückgeführt. Doch was sollen andere Facetten des Währungsbegriffs (z.B. Umtausch, Inflation) in einem neuronalen Kontext bedeuten?

- Ein weiteres Forschungsthema ist der so genannte "soziale Schmerz" (*social pain*). Gemeint sind damit jene kognitiven Zustände, welche aus einer Ausgrenzung aus bestimmten sozialen Kontexten (Beziehungen, Gruppen) resultieren. Solcher sozialer Schmerz ist sicher ein Bereich, der für die Ermittlung von neuronalen Grundlagen von Moral wichtig werden könnte. Eisenberger und Lieberman [56] haben postuliert, dass die neuronalen Systeme für die Erzeugung von physischem Schmerz und sozialem Schmerz bis zu einem gewissen Grad überlappen. Dies stütze sich aus Läsionsstudien bei Tieren und *Imaging*-Experimenten bei Menschen – wobei nachgeprüft werden müsste, wie in den dort zitierten Arbeiten sozialer Schmerz bei Tieren bestimmt worden ist. Diese Überlappung zeige sich dadurch, dass bei beiden Formen von Schmerz der *anterior cingulate cortex* involviert sein soll. Derartige Studien können beispielsweise auf die Möglichkeit hinweisen, dass künftige pharmakologisch-therapeutische Interventionen beide Formen von Schmerz kurieren könnten.

Insgesamt gesehen zeigen sich bei den untersuchten Übersichtsarbeiten zu *social cognitive neuroscience* eine Reihe von Schwierigkeiten, die später auch bei den Studien über die neuronalen Grundlagen der Moral auftauchen werden. So ist es insbesondere schwierig, die komplexen Phänomene der realen Welt derart zu vereinfachen, dass sie für experimentelle Studien zugänglich sind und dennoch als "sozial" gelten können. Gerade die Verwendung der *Imaging*-Technik ist mit dem Problem verbunden, dass die Versuchspersonen vereinzelt in einem Scanner liegen und nur beispielsweise mit einem Film oder mit Bildern in einen sozialen Interaktionskontext gestellt werden können. Aus diesem Grund ist vorgeschlagen worden, künftig simultane Hirnscans durchzuführen, indem Personen (immer noch via Videokamera vermittelt) miteinander interagieren und gleichzeitig einem *Imaging* unterworfen werden [21]:221.

2.3.2 Emotionen als Thema der Neurowissenschaft

Seit den 1990er Jahren sind Emotionen wieder ein bevorzugtes Thema der Neurowissenschaft geworden [30]:415-416. Interessanterweise waren Emotionen bereits in der ersten Hälfte des 20. Jahrhunderts ein Schwerpunkt der Hirnforschung gewesen, wurden dann aber gewissermassen ein Opfer der "kognitiven Revolution" Mitte des 20. Jahrhunderts, als neuronale

Prozesse vermehrt unter dem Blickwinkel von Informationsverarbeitung untersucht wurden [107]. Zudem glaubte man, mit dem Konzept des limbischen Systems ein brauchbares Modell für die Lokalisation und Erklärung des Wirkmechanismus von Emotionen gefunden zu haben. In der jüngeren Zeit wuchsen jedoch die Zweifel an der Brauchbarkeit dieses Konzepts für die Erklärung von Emotionen [107]:157. In den 1980er Jahren führte unter anderem die Entdeckung der Neuroanatomie der Angstkonditionierung, in welcher die Amygdala eine wichtige Rolle spielt, mit zur genannten Renaissance der Emotionen in der Neurowissenschaft. Popularisiert wurde diese unter anderem durch die Bücher von Antonio Damasio [42, 45, 46].

Ein umfassender Überblick zur Emotionsforschung, welche auch die Psychologie mit einschliesst, kann an dieser Stelle nicht geleistet werden (eine kompakte Übersicht bietet [160]). Fokussiert werden lediglich jene Aspekte, welche im Rahmen der Erforschung der Grundlagen der Moral oft genannt werden. Generell stellt sich auch bei der Emotionsforschung das Problem, dass es keine generell akzeptierte Definition des Begriffs "Emotion" gibt und deshalb in experimentellen neurowissenschaftlichen Studien oft ein unpräziser Begriff von Emotion verwendet wird [107]. Daran schliesst sich die Forderung an, dass man sich in solchen Studien auf einen psychologisch gut definierten Aspekt von Emotionen beschränken soll und ein experimentelles Paradigma wählt, dass diesen Aspekt genügend vereinfacht und auf nicht präzise definierbare Begriffe verzichtet.

Es wird heute kaum mehr bestritten, dass die Erklärung komplexer kognitiver Fähigkeiten wie auch des Verhaltens von Tieren und Menschen eine Theorie der Emotionen verlangt.[6] So ist beispielsweise gezeigt worden, dass die emotionale Färbung von Erfahrungen, die sich in einer erhöhten Aktivierung der Amygdala ausdrückt, für die Bildung einer stabilen Erinnerung dieser Erfahrung zentral ist [116]. Auch besteht eine recht gute Übersicht über die funktionelle Neuroanatomie hinsichtlich der Verarbeitung emotionaler Stimuli durch das Gehirn. In einer Meta-Analyse verschiedener (meist *Imaging*-) Studien durch Phan et al. werden folgende Regionen identifiziert [134]: 1) Generell geht die Verarbeitung emotionaler Stimuli mit einer erhöhten Aktivität im medialen präfrontalem Kortex einher. 2) Bei der Verarbeitung von angstauslösenden Stimuli (bzw. bei dem dadurch entstehenden Gefühl der Angst) ist die Amygdala involviert. 3) Das Gefühl der Traurigkeit ist mit einer erhöhten Aktivität im *subcallosal cinculate* assoziiert. 4) Visuell induzierte Emotionalität ist mit Aktivität im occipitalen Kortex und der Amygdala verknüpft. 5) Emotionalität induziert durch Rückerinnerung an vergangene Ereignisse sowie die Umsetzung kognitiver Aufgaben mit einer emotionalen Komponente rekrutieren den *anterior cingulate* und die Insula. Die Autoren erinnern bei dieser Übersicht an die Probleme solcher Meta-Analysen. So sind die untersuchten Studien jeweils sehr unterschiedlich und die bekannten methodischen Probleme von *Imaging*-Studien (*Baseline*-Problem, präzise Definition des Stimulus etc., siehe Abschnitt 2.2.3) erschwert die Vergleichbarkeit erheblich. Zudem bestehe bei vielen Studien die Gefahr einer zu engen Fokussierung auf bestimmte Hirnregionen, so dass erhöhte Aktivierungen anderer Regionen übersehen werden. Die Autoren schliessen mit der Forderung nach einer Standardisierung der Methoden, der experimentellen Paradigma und der Datenformate im Bereich der Emotionsforschung.

Diese Probleme sind Ausdruck der Tatsache, dass es heute weder eine allgemein akzep-

[6] Dies geht heute so weit, dass man auch für künstliche Systeme wie Roboter die Integration einer Form von Emotionalität postuliert, damit diese dereinst höhere kognitive Aufgaben, wie beispielsweise "Entscheidungen fällen", lösen können [8]. Inwiefern solche rein funktionale Emotionen in künstlichen Systemen, die (wahrscheinlich) kein inneres Erleben dieser Emotionen haben, realisiert werden sollen, bleibt aber offen.

tierte Theorie von Emotionen gibt, noch eine allgemein anerkannte Klassifikation der verschiedenen Arten von Emotionen. Weithin akzeptiert ist lediglich die Unterscheidung zwischen primären und sekundären Emotionen [160]:388: Das primäre Emotionssystem basiert auf angeborene Emotionen (beispielsweise die Fähigkeit, Angst zu empfinden) und benötigt keine kognitive Verarbeitung. Das primäre Emotionssystem kann auch bei Wahrnehmungsvorgängen aktiv werden, ohne dass dies bewusst bemerkt wird. Dies zeigt sich beispielsweise in einer Studie von Whalen et al. [163]. In dieser wurden den Versuchspersonen über eine kurze Zeitspanne (33 ms) ängstliche oder glückliche Gesichter gezeigt, unmittelbar gefolgt von einer längeren Präsentation neutraler Gesichter. Erstere Gesichter können von den Versuchspersonen nicht bewusst wahrgenommen werden und werden durch die neutralen Gesichter gewissermassen maskiert. Dennoch zeigte sich im Scanner eine stärkere Aktivierung der Amygdala bei maskierten ängstlichen Gesichtern im Vergleich zu maskierten glücklichen Gesichtern. Dies ist ein Beispiel einer unbewussten Färbung von Wahrnehmungen durch primäre Emotionen. Das kognitiv-affektive Emotionssystem wiederum beruht auf der Verknüpfung primärer Emotionen mit gelernten Informationen. Emotionale Zustände sind an bekannte physiologische Marker (z.B. erniedrigter elektrischer Hautwiderstand [29]:1168, erhöhte Herzschlagfrequenz) gekoppelt, wobei diese aber keine zuverlässigen Hinweis für eine bestimmte Emotion geben.

Im Kontext der Erforschung der neurobiologischen Grundlagen der Moral ist die These aufgekommen, dass es so genannte "moralische Emotionen" gebe (Adolphs nennt diese "soziale Emotionen" [3]:166). Für Moll et al. gehören in Anlehnung an Haidt (siehe unten) Emotionen, welche einen Bezug zur Wohlfahrt einer Gruppe oder eines anderen Individuums als der Betreffende selbst haben, in diese Klasse [120]:299. Der Konflikt zweier sich widersprechender moralischer Emotionen ist dann gemäss Moll das Wesensmerkmal eines moralischen Dilemmas. Eine ausführliche Beschreibung moralischer Emotionen stammt von Haidt. Er definiert diese als "those emotions that are linked to the interest of welfare either of society as a whole or at least of persons other than the judge or agent" [84]:853. Moralische Emotionen beziehen sich also explizit auf soziale Interaktionen, während nichtmoralische Emotionen sich primär auf den emotionalen Zustand des Agenten als Folge irgendwelcher äusserer Einflüsse beziehen (z.B. Angst oder Glück). Eine scharfe Abgrenzung lässt sich gemäss Haidt aber nicht ziehen. Er gliedert moralische Emotionen in folgende vier Familien:

- Die andere verurteilenden Emotionen (*other-condemning emotions*): Gemäss Haidt fallen Zorn/Wut (*anger*), Ekel/Abscheu (*disgust*) und Verachtung/Geringschätzung (*contempt*) in diese Klasse. Zorn/Wut ist nach ihm die wohl am meisten unterschätze moralische Emotion – auch deshalb weil sie oft als "unmoralische" Emotion aufgefasst werde. Zorn/Wut sei aber primär als eine Reaktion auf eine ungerechtfertigte Beurteilung durch Andere aufzufassen und motiviert Straf- und Racheverhalten, d.h. künftige Aktionen gegen jene, welche Zorn/Wut bei der betreffenden Person ausgelöst haben. Der evolutionäre Ursprung von Ekel/Abscheu wiederum ist wahrscheinlich die Fähigkeit, Erinnerung an schlechte Nahrungsmittel aufbauen zu können. In einem sozialen Kontext unterstützen diese Emotionen Abgrenzungen gegenüber anderen Gruppen – als Beispiel nennt Haidt das Kastenwesen in Indien. Ekel/Abscheu führt dazu, den Kontakt mit solchen ausgegrenzten Gruppen zu vermeiden und löst im Falle eines unbeabsichtigten Kontakts rituelle Reinigungshandlungen aus. Verach-

tung/Geringschätzung schliesslich ist eine Emotion zur Stützung von Hierarchien und Prestige in einer Sozialgemeinschaft – vor allem als Abgrenzung gegenüber hierarchisch tieferen Gruppen. Die Emotionen dieser Familie spielen gemäss HAIDT für den Aufbau der (negativen) Reputation einzelner Individuen in einer sozialen Gemeinschaft eine wichtige Rolle.

- Die bewusst-machenden Emotionen (*self-conscious emotions*): Diese Emotionen tauchen bei einem Individuum auf, wenn dieses gegen gewisse soziale Normen verstossen hat. Sie spielen eine wichtige Rolle für den Aufbau der moralischen Persönlichkeit eines Individuums innerhalb einer Gruppe. Mit "moralischer Persönlichkeit" ist die positive Reputation in der Gruppe gemeint, d.h. das Zeigen von Respekt gegenüber den in der Gruppe geltenden moralischen Regeln. Gemäss HAIDT fallen Scham (*shame*), Verlegenheit/Peinlichkeit (*embarrassment*) und Schuld (*guilt*) in diese Kategorie. Scham wie Verlegenheit/Peinlichkeit sind Emotionen, die sich als Regelverletzungen in einem hierarchisch geprägten Umfeld ergeben, d.h. als Folge der Beurteilung des Fehlverhaltens durch eine als sozial höher gestellten Person. Bei Scham und Verlegenheit/Peinlichkeit zeigen sich aber interessante kulturelle Unterschiede: In einem westlichen Kontext werden diese beiden Emotionen klar unterschieden: Scham ist eine Folge der Verletzung moralischer Regeln, während Verlegenheit/Peinlichkeit eine Folge der Verletzung sozialer Konventionen ist. In nichtwestlichen Kulturen hingegen werden die beiden Emotionen als weit ähnlicher betrachtet. Schuld schliesslich ist im Gegensatz zu Scham eine Emotion, welche sich auf schlechte Handlungen beziehen und nicht auf eine generelle Beurteilung des eigenen Selbst. Schuld motiviert Verhaltensweisen wie Gestehen oder Entschuldigen, welche die Wiederherstellung sozialer Beziehungen anstreben. Insgesamt gesehen dienen die Emotionen dieser Familie zur Aufrechterhaltung bestimmter sozialer Regelsysteme und Ordnungen.

- Die Mitleids-Emotionen (*other-suffering emotions*): In diese Familie fallen "klassische" moralische Emotionen wie Mitgefühl/Mitleid (*compassion*), Sympathie (*sympathy*) und Empathie (*empathy*) – wobei gemäss HAIDT Empathie aber nicht als Emotion aufzufassen sei, sondern als Fähigkeit zu fühlen, was eine andere Person in einer gewissen Situation fühlt. Für ihn ist Mitgefühl/Mitleid die zentrale Emotion dieser Familie, welche Verhaltensweisen wie Helfen und Unterstützen motiviert.

- Die andere preisenden Emotionen (*other-praising emotions*): Die Emotionen dieser Familie sind "positiv" im Sinn dass sie nicht bei der Verletzung von Regeln auftreten, sondern bei deren positiven Erfüllung. In diese Kategorie fallen die Emotionen Dankbarkeit (*gratitude*), Ehrfurcht (*awe*) und Erhabenheit (*elevation*). Dankbarkeit ist eine wichtige Emotion zur Unterstützung von *reciprocal altruism*. Ehrfurcht und Erhabenheit wiederum treten auf, wenn ein Individuum mit einer Person konfrontiert wird, die innerhalb einer Gruppe eine ausserordentlich hohe moralische Reputation hat. Diese Emotionen dienen als generelle Motivatoren, die eigene moralische Reputation zu verbessern.

Diese Liste ist nicht als abschliessend zu betrachten, zumal HAIDT selbst feststellt, dass umstritten ist, wie viele verschiedenen Emotionen es überhaupt gibt. Auch sind gewisse Emotionen schwer diesem Schema zuordbar – HAIDT nennt diesbezüglich insbesondere die

Emotion Liebe (*love*). Auch dürfte der kulturelle Kontext bei der genauen Einordnung der einzelnen Emotionen eine wichtige Rolle spielen.

Zum Schluss soll an dieser Stelle noch die *somatic-marker*-Hypothese von Antonio Damasio etwas genauer vorgestellt werden, da diese Theorie bei der Untersuchung der neurobiologischen Grundlagen der Moral oft auftaucht. Damasio klassifiziert Emotionen in drei Grundtypen: Primäre Emotionen (Angst, Wut, Glück, Traurigkeit, Ekel, Überraschung), sozialen Emotionen (z.B. Peinlichkeit, Stolz, Eifersucht, Schuld) und Hintergrundemotionen (*background emotions*, z.B. Ruhe, Spannung, Wohlergehen – andere bezeichnen solche Emotionen als Stimmungen (*moods*)) [44]. Physiologisch gesehen sind Emotionen im Sinn von Damasio Veränderungen des Körperzustandes. Falls das Nervensystem eine koordinierte Repräsentation dieser Zustandsänderung erreichen kann, wird diese Emotion auch gefühlt, d.h. Gefühle (*feelings*) sind Wahrnehmungen von Körperzustandsveränderungen [108]:273. Gefühle sind dabei nicht notwendigerweise an die Existenz eines Bewusstseins geknüpft – ein Bewusstsein ermöglicht es lediglich zu wissen, dass man eine Emotion gefühlt hat. Emotionen sind gemäss Damasio evolutionär späte und Gefühle noch spätere Formen der Steuerung von Lebensvorgängen. Weiter unterscheidet Damasio zwischen einer echten emotionalen Reaktion und einer emotionalen als-ob-Reaktion. Bei ersteren löst ein bestimmter Hirnzustand (hervorgerufen beispielsweise durch eine bestimmte Wahrnehmung) eine Veränderung des Körperzustands aus (höherer Puls etc.), der dann wiederum in den somatosensorischen Rindenfeldern wahrgenommen wird. Die als-ob-Reaktion umgeht diese Körperschleife, indem der die emotionale Reaktion auslösende Hirnzustand direkt auf die somatosensorischen Rindenfelder wirkt. Die *somatic marker hypothese* postuliert, aufbauend auf diese Theorie der Emotionen, ein Modell des Zustandekommens von Entscheidungen (*decision-making*). Folgende Voraussetzungen für diese Hypothese werden genannt [15]:295: Denken und Entscheiden beruhen auf einer Vielzahl neuronaler Prozesse, wobei nur einige davon mit bewussten Wahrnehmungen und Denkvorgängen (kognitive Prozesse) verknüpft sind. Wichtig sind vielmehr auch unterstützende Prozesse wie Aufmerksamkeit, Gedächtnis und Emotionen. Vor allem letztere im Sinn von Körperzustandsänderungen sind demnach bei (kognitiven) Entscheidungsprozessen ständig präsent. Sie markieren eine (oft unbewusste) Bewertung der kognitiven Prozesse, die zu einer bestimmten Entscheidung führen und beeinflussen diese damit wesentlich. Einmal etabliert, kann die Überlagerung von kognitiven Inhalten und somatischen Markern auch ohne die oben genannte Körperschleife erfolgen. Diese Simulation der körperlichen Rückmeldung erlaubt dann auch eine emotionale Bewertung fiktiver Situationen, was für Planungen von grosser Hilfe ist. Die *somatic marker* Hypothese postuliert, dass Defekte in Hirnregionen, die in die Verarbeitung emotionaler Zustände involviert sind, zu einer verminderten Entscheidungsfähigkeit führen, obgleich die betreffenden Personen über ausreichende kognitive Fähigkeiten verfügen.[7] Dieses Modell wird bei der Erklärung moralischer Pathologien herangezogen (siehe Abschnitt 3.4.1).

[7]Natürlich können sich Emotionen bei geeigneter Versuchsanordnung auch unvorteilhaft auf Entscheidungen auswirken. Shiv et al. zeigen in einem *investment*-Experiment, dass Patienten mit Läsionen in ventromedialen Regionen des präfrontalen Kortex besser abschneiden, weil sie weniger risikoavers sind [148] – was natürlich eine Folge der Tatsache ist, dass sich im Experiment Risiko auszahlt.

2.3.3 Spiegelneuronen

Die Entdeckung und Erforschung von so genannten Spiegelneuronen erlaubt die Postulierung eines Wirkmechanismus von moralnahen Fähigkeiten wie Empathie. Aus diesem Grund finden sich in der Literatur über die neurobiologischen Grundlagen der Moral oft Verweise auf die Erforschung der Spiegelneuronen. Diese Neuronen gehören zu den so genannten *visuomotor neurons* und wurden ursprünglich im Areal F5 des prämotorischen Kortex von Affen entdeckt [142]. Der Begriff "visuomotor" sagt aus, dass diese Neuronen sowohl dann aktiv sind, wenn der Affe bestimmte Bewegungen vollzieht, wie auch, wenn der Affe dieselbe Bewegung bei anderen Affen beobachtet. Diese Eigenschaft ist offenbar unabhängig von der Art des bewegten Objekts, so dass diese Neuronen also für bestimmte Klassen von Bewegungen Aktivität zeigen und nicht für Klassen von bewegten Objekten. Innerhalb der Spiegelneuronen lassen sich zudem weitere Unterarten differenzieren: So genannte *ingestive mirror neurons* zeigen Aktivität bei nahrungsbezogenen Handlungen (z.B. das Bewegen von Nahrung zum Mund) und *communicative mirror neurons* werden bei Gesten mit einer mutmasslich kommunikativen Funktion aktiv.

Hinsichtlich der funktionalen Rolle von Spiegelneuronen werden mehrere Hypothesen diskutiert [142]: So dürften sie eine wichtige Rolle für die Imitation von gewissen Handlungen spielen – und demnach auch beim Erlernen dieser Handlungen. Weiter wird spekuliert, dass Spiegelneuronen auch notwendig sind, die Handlungen anderer zu verstehen – sie wären demnach in kognitive Prozesse eingebunden. Manche Forscher vermuten zudem, dass Spiegelneuronen bei der Evolution von Sprache eine Rolle spielen. Es ist jedoch schwierig, die Existenz von Spiegelneuronen bei Menschen schlüssig nachzuweisen, weil die in den Tierexperimenten verwendeten Methoden der elektrischen Ableitung von Nervenzellen nicht verwendet werden können. Man muss auf *Imaging*, sowie EEG und MEG zurückgreifen, welche (wie in Abschnitt 2.2.2 erläutert) eine weit schlechtere räumliche Auflösung haben. Es gibt jedoch eine Reihe von *Imaging*-Experimenten, welche mit der Existenz von Spiegelneuronen beim Menschen kompatibel sind. So wurde beispielsweise gezeigt, dass Aktivierungsmuster im inferioren Parietallappen, dem *pars opercularis*, des inferioren frontalen Gyrus und dem angrenzenden *precentral gyrus* unabhängig davon sind, ob der Versuchsperson das Beissen eines Menschen, eines Affen oder eines Hundes gezeigt wird, was die Eigenschaft der Bildung einer Klasse von Bewegungen zeigt. Ein schlüssiger Beweis für die Existenz von Spiegelneuronen bei Menschen steht aber offenbar noch aus.

2.4 Moralnahe Verhaltensweisen und Fähigkeiten

Der Begriff "moralisches Verhalten" umfasst eine Vielzahl möglicher Verhaltensweisen und Voraussetzungen für solches Verhalten, die ihrerseits wieder Gegenstand von Studien werden können, ohne dass damit eine umfassende Theorie über die Grundlagen von moralischem Verhalten angestrebt wird. In diesem Abschnitt soll eine Übersicht über Forschungen über Verhaltensweisen gegeben werden, die entweder als Grundlagen für moralisches Verhalten dienen können oder ein Ausdruck von moralischem Verhalten sein können. In erstere Kategorie gehört die Erforschung von Religiosität, Empathie, Intuition, Vertrauen, Bedauern und Enttäuschung sowie Lügen. In letztere Kategorie gehören Kooperation und Altruismus. Nachfolgend soll ein Überblick über neuere Arbeiten in diesen Bereichen gegeben werden.

2.4.1 Religiosität

Von der Vielzahl komplexer menschlicher Verhaltensweisen, welche in der jüngeren Zeit Gegenstand der Neurowissenschaft geworden ist, hat auch religiöses Verhalten die Aufmerksamkeit der Neurowissenschaft gewonnen. Ein Ausgangspunkt dieser Studien ist die Beobachtung, dass religiöse Vorstellungen und Praktiken in der menschlichen Kultur und Geschichte weit verbreitet sind. Deshalb stellt sich die Frage, ob sich eine Art (neuro-)biologisches Fundament für Religiosität identifizieren lässt. Ein Wesensmerkmal der untersuchten religiösen Phänomene ist der Glaube an die Existenz abstrakter *agents* wie Götter oder Geister von Vorfahren, die einen personalen Charakter haben und metaphysischen Eigenschaften aufweisen [10]. Die Menschen, welche an die Existenz dieser abtsrakten *agents* glauben, treten mit diesen in ein (inneres) Gespräch und behandeln diese oft so, wie sie auch andere Personen behandeln. Für die Forscher stellen sich dazu folgende drei Fragen [10]:29: Wie repräsentieren Personen Vorstellungen und Begriffe übernatürlicher Wesen? Wie erwerben Leute solche Konzepte? Wie reagieren sie auf solche Konzepte mittels spezifischer Handlungen wie religiöse Rituale? Die Forscher bemerken dazu, dass diese übernatürlichen Wesen, welche Gegenstand der Untersuchung sind, hinsichtlich der Zahl der Eigenschaften wie auch ihrer Einbindung in einen grösseren theoretisch-theologischen Zusammenhang wesentlich einfacher sind als theologische Gottesbegriffe. Zudem sollen diese *agents* eine eher geringe Zahl unplausibler (z.B. im Widerspruch zu den Eigenschaften normaler physikalischer Objekte stehende) Eigenschaften aufweisen, was die Aufrechterhaltung des Glaubens an diese abstrakten *agents* sicher erleichtert. Gewiss existiert heute keine ausgefeilte Theorie über die neurobiologischen Grundlagen der Religion. Interessant ist jedoch, das die Forscher in diesem Bereich durchaus Bezüge zu moralischen Fragen herstellen, weil die untersuchten abstrakten *agents* eine wichtige Rolle beim Aufbau von Überzeugungssystemen spielen können, welche ihrerseits moralische Handlungen von Einzelpersonen leiten. So sind diese *agents* in der Lage, den moralischen Gehalt einer Situation einzuschätzen und geben Ratschläge, wie man sich in dieser verhalten soll. Sie könnten also einer moralischen Intuition ein Gesicht geben und sind ein möglicher Ausdruck von *moral feelings* [26]:122. Die derzeit vorliegenden Erkenntnisse zum Problemkreis der Naturalisierung von Religion haben auch heute erst einen allgemeinen Charakter. So soll die *mental machinery*, welche für den Erwerb und die Repräsentation religiöser Konzepte relevant ist, zum grossen Teil nicht bewusst zugänglich sein. Dies wird aus dem Sachverhalt geschlossen, dass die tatsächlichen religiösen Konzepte von Personen sich von jenen unterscheiden, welche die Personen glauben zu haben. Aus diesem Grund könnten, so eine Schlussfolgerung, offizielle religiöse Institutionen, Theologien und Dogmen für die Beschreibung der Inhalte und der Gründe von persönlichen Glaubensüberzeugungen inadäquat sein. Weiter soll Religiosität nicht das Resultat eines spezifischen kognitiven Prozesses sein, sondern ist “a whole collection of distinct mental systems”. Religiöse Kognition schliesslich unterscheidet sich nicht dramatisch von üblicher Kognition. Vergleichbare allgemeine Einschätzungen finden sich, wie später deutlich werden wird, auch bei der Charakterisierung moralischer Kognition.

2.4.2 Empathie

Empathie ist ein Konzept, das primär in der Psychologie untersucht wird und das sicherlich in einem engen Bezug zu moralischem Verhalten steht. Bestimmte normative Ethiken wie

die *care ethics* sehen in der Empathiefähigkeit gar eine zentrale Voraussetzung für moralisches Verhalten [53], während rationalistische Konzepte die Bedeutung von Empathie zwar nicht verleugnen, dennoch aber nicht ins Zentrum setzen. Zu Empathie gibt es in der wissenschaftlichen Literatur zahlreiche Studien (eine Suche nach dem Stichwort "*empathy*" ergibt in *MedLine* gegen 8000 Treffer). Da hier eine Überschneidung mit der zweiten Pilotstudie zu erwarten ist, erfolgte keine systematische Untersuchung der neueren (primär psychologischen) Literatur zu Empathie, sondern es werden jene Studien fokussiert, welche neuronale Korrelate für Empathie suchen.[8] Bereits in der für diesen Abschnitt verwendeten Literatur ist aber deutlich geworden, dass es keine allgemein akzeptierte und präzise Definition von Empathie gibt. Vreeke und Vandermark definieren Empathie primär durch drei Fähigkeiten: zu wissen, was eine andere Person empfindet (*role taking*), tatsächlich zu fühlen, was die andere Person fühlt (*emotional congruence*) und schliesslich gegenüber der anderen Person unter Berücksichtigung der beiden ersten Fähigkeiten zu handeln (*sympathetic concerns*) [161]:178. Sie halten aber auch fest, dass weitere Ausdifferenzierungen von Empathie möglich seien. Der Einbezug der Möglichkeit, dass Empathie auch bei Tieren (insbesondere Primaten) vorkommen könnte[9], erweitert das Definitionsfeld von Empathie weiter, wie die Diskussionsbeiträge zur Arbeit von Preston und de Waal [137] zeigen. Charakteristische Aspekte von Empathie sind demnach: Sympathie, generell prosoziales Verhalten (wie Hilfsverhalten), oder die Beeinflussung des eigenen Verhaltens durch emotionale Zustände Anderer. Es wird aber auch vermutet, dass der Begriff Empathie gar kein kohärentes Phänomen beschreibt, sondern eher ein *umbrella term* für eine Reihe verschiedener Phänomene ist (siehe den Beitrag von Davis in [137]:32-33).

In der allgemeinsten Formulierung wird Empathie als ein *shared-state*-Phänomen aufgefasst [138]:287, was natürlich zur bereits aufgeworfenen Frage führt, welcher Art dieser "geteilte Zustand" denn sein soll. Grundsätzlich bietet sich *Imaging* als Ansatz für das Erkennen solcher *shared states* im Sinne gleichartiger neuronaler Aktivierungsmuster an. Der zweite oben genannte definitorische Aspekt von Empathie – *emotional congruence* – könnte eine solche Ähnlichkeit von Aktivitätsmustern ebenfalls postulieren. Demnach könnte zwischen dem Wahrnehmen von Emotionen Anderer und dem Erzeugen eigener Emotionen eine Art *shared code* bestehen (siehe den Beitrag von Iacoboni/Lenzi in [137]:39-40). Nahe liegend ist in in diesem Zusammenhang die (derzeit noch spekulative) Idee, dass Spiegelneuronen (siehe Abschnitt 2.3.3) einen Beitrag zum neuronalen Mechanismus von Empathie leisten können [137]:11. Vreeke und Vandermark hingegen binden Empathie eng an die Fähigkeit zu einer reichen Kommunikation zwischen den beteiligten Partnern [161]. So

[8] Es wurde ebenfalls nicht systematisch geprüft, welche psychologische Studien einen expliziten Zusammenhang zwischen Empathie und moralischem Verhalten untersuchten. Die Zahl dieser Studien scheint aber nicht allzu gross zu sein. Generell scheinen diese Studien auf eine gewisse Scheinheiligkeit hinzuweisen im Sinn dass Versuchspersonen zwar gegen Aussen hin moralisch korrekt erscheinen wollen, die Kosten des moralischen Verhaltens aber zu vermeiden versuchen (vgl. die Diskussion in [11]:1190-1191). Wenig überraschend ist, dass die Versuchsanordnung entscheidet, welchen Einfluss Empathie bei moralischen Handlungen hat: Wenn Versuchspersonen sich und einer zweiten Person zwei Aufgaben zuteilen müssen, von welcher eine deutlich unangenehmer ist, so teilen sich die angesprochenen Personen in der Regel den angenehmeren Job zu – unabhängig davon, ob sie sich zuvor in die jeweils andere Person einfühlen mussten oder nicht. Besteht hingegen ein Auswahlszenario, wonach entweder die erste Form von Wahl getroffen werden muss, oder aber alternativ beiden Personen eine moderat angenehme Aufgabe zugeteilt werden soll, beeinflussen empathische Vorübungen den Entscheid [11].

[9] Bereits in den 1950er und 1960er Jahren soll es eine Vielzahl von Studien gegeben haben, welche versucht haben, den Nachweis von Empathie bei Tieren zu führen [138]:293.

definiert, dürfte Empathie ein für *Imaging*-Experimente schwierig zu erfassendes Phänomen sein, weil ein reicher kommunikativer Austausch zwischen Versuchspersonen, die sich in Scannern befinden, schwer möglich ist – vor allem hinsichtlich jener nichtsprachlicher Kommunikationslemenente (Gesichtsausdruck, Körpersprache), welche für das Erfassen des emotionalen Zustandes des Gegenüber zentral sind.[10] Aufgrund der oben abgesprochenen Komplexität des Empathiebegriffs ist kaum zu vermuten, dass eine eigentliche Lokalisation der mit Empathie verbundenen neuronalen Repräsentationen möglich ist (vgl. den Beitrag von Blair/Perschardt in [137]:27-28.).

Ein Beispiel für die Schwierigkeit, Empathie (und andere komplexe psychische Phänomene) durch *Imaging*-Studien untersuchen zu wollen, bildet die Studie von Tom Farrow et al., welche Unterschiede hinsichtlich der neuronalen Korrelate von *empathy* und *forgiveness* mittels fMRI untersucht hat [61]. Im experimentellen Paradigma wird Vergebung dadurch geprüft, dass den Versuchspersonen zuerst ein Szenario schriftlich vorgestellt wird und danach in zwei kurzen Sätzen zwei Varianten von Handlungsweisen vorgestellt werden (z.B. "stealing milk from a doorstep" und "stealing milk from a cornershop"). Die Versuchspersonen mussten per Knopfdruck entscheiden, welche der beiden Varianten eher vergeben werden könne. Empathie wurde auf vergleichbare Weise gemessen. Als *baseline* wird ein so genanntes *social reasoning judgement* verwendet, z.B. mussten die Versuchspersonen den möglichen Grund eines Verkehrsstaus anhand auf der gleichen Weise wie zuvor präsentierter Alternativen bestimmen. Unterschieden wurde zwischen einer 16 Sekunden dauernden Lesephase und einer 35 Sekunden dauernden Entscheidungsphase, wobei nur die in der Entscheidungsphase gemessenen Daten verwendet wurden. Die Studie kommt zum Schluss, das die folgende Hirnregionen bei beiden Konzepten aktiviert sind: der linke superiore frontale Gyrus, der orbitofrontale Gyrus und der untere Pränukleus. Empathie hingegen soll zudem den *left anterior middle temporal gyrus* und den *left inferior frontal gyrus* überdurchschnittlich aktiveren, während Vergebung den *posterior cingulate gyrus* aktivieren soll. Daraus schliessen sie "that forgiving others is neuro-physiologically distinct from empathising with them" [61]:2438, und "these results give a strong indication that activations of very high level cognitive processes are recordable, and that abnormal mental states could be reasonably predicted to show differing patterns of activation, which may be amenable to 'normalization' through cognitive interventions." Diese Studie zeigt die Schwierigkeiten von *Imaging* Experimenten über komplexe psychologische Phänomene deutlich auf: So werden komplexe Phänomene wie Empathie und Vergebung auf die Bewertung einfacher Sätze reduziert, die aber ihrerseits durchaus Anlass zu kritischen Diskussionen geben können (z.B. beinhaltet die Beurteilung des Aktes von Stehlen nicht auch empathische Überlegungen etwa im Sinn, wie sich die Versuchsperson fühlen würde, wenn sie den Akt beginge?). Am Schluss wird dann behauptet, mit solchen Methoden liessen sich dann *abnormal states* (etwa die Unfähigkeit zu vergeben?) identifizieren. Methodisch wird nicht deutlich gemacht, warum die sehr unterschiedlichen *social reasoning judgments* als angemessene *baseline* zu gelten haben, da diese eine ganze Bandbreite von Überlegungen in den Versuchspersonen auslösen könnten, die sich schwer als einheitliches Phänomen charakterisieren lassen. Auch wird nicht klar, warum die (sich über 16 Sekunden erstreckende!) Lesephase nicht untersucht wird, zumal hier je nach Szenario unterschiedliche Vorverar-

[10] In [138]:301 wird diesbezüglich auf das Problem hingewiesen, dass die zunehmend anonymen Kommunikationsmittel wie e-Mail es erschweren, den emotionalen Zustand des Gesprächspartners zu ermitteln, was die Wahrnehmung von Empathie beeinträchtigen könnte.

beitungen des Problems passieren könnten. Schliesslich ist die generelle Aussagekraft der grundlegenden Schlussfolgerung aus den bereits erwähnten Gründen zu bezweifeln: Unterschiedliche psychologische Phänomene sind bei ausreichender Messauflösung praktisch *per definitionem* mit unterschiedlichen neuronalen Aktivierungsmustern verbunden.

2.4.3 Intuition

Der Begriff "Intuition" taucht in der neurowissenschaftlichen Literatur erstaunlich selten auf.[11] Der Begriff tritt auch bei jenen Arbeiten selten auf, welche die neuronalen Grundlagen moralischen Verhaltens untersuchen. GREENE beispielsweise spricht von Intuitionen (*intuitions*, *gut-feelings*), ohne aber den Begriff weiter auszuführen [78]). In der Psychologie findet sich der Begriff bei HAIDT beispielsweise hinsichtlich der Problemstellung, wie eine intuitive gegenüber einer rationalen Entscheidungsstrategie abschneidet [83]. In der Neurowissenschaft wiederum wird der Begriff zuweilen auch im Rahmen der Emotionsforschung untersucht, indem Intuitionen letztlich emotionsgesteuerte, von rationalen kognitiven Prozessen abtrennbare Prozesse sein sollen [158]. Intuitionen können demnach als Entscheidungshilfe-Gefühle aufgefasst werden, die verhältnismässig wenig kognitive Ressourcen beanspruchen sollen. Eine explizite Definition moralischer Intuitionen, welche mit dieser Auffassung vereinbar ist, liefert HAIDT: "Moral intuition appears to be the automatic output of an underlying, largely unconscious set of interlinked moral concepts [83]:825. Er verwendet den Intuitionsbegriff auch in seinem Modell über moralische Entscheidungen, das in Abschnitt 3.5 näher vorgestellt wird.

Experimentell lassen sich Intuitionen beispielsweise durch Mustererkennungs-Experimente erfassen, da sich Ähnlichkeiten zwischen Objekten oft rasch feststellen, sprachlich aber nur schwer in Worte fassen bzw. erklären lassen. In einem Beispiel [22] werden Versuchspersonen drei Worte gezeigt. Diese mussten abschätzen, ob die drei Wörter semantisch kohärent sind und ob ein viertes Wort existiert, welche diese Kohärenz ausdrückt. Ein weiterer Ansatz für die Untersuchung von Intuition ist implizites Lernen. LIEBERMAN schlägt vor [109], Intuition generell als das Ergebnis von Prozessen impliziten Lernens zu betrachten. Intuition wird dabei definiert als "the subjective experience of a mostly nonconscious process that, dependent on exposure to the domain or problem space, is capable of accurately extracting probabilistic contingencies" [109]:111. Damit wären – im Gegensatz etwa zu einem Instinkt – bestimmte Aspekte von Intuition wie die Fähigkeit, eine Intuition zu erfassen und zu nutzen, durch Lernen verbesserbar. Lieberman verweist auf *Imaging*-Studien, welche implizites Lernen untersuchen und eine erhöhte neuronale Aktivität im *Caudatum*, dem Putamen und den Basalganglien finden. Dass diese Regionen aber auch bei Intition eine Rolle spielen sollen, ergibt sich natürlich aus der definitorisch vorgegebenen Verknüpfung von implizitem Lernen und Intuition.

[11] Eine Suche nach dem Stichwort "*intuition*" ergibt in *Medline* insgesamt 1145 Einträge. Wird die Suche aber auf unsere Referenzmenge (siehe Abschnitt 2.1) eingeschränkt, so verbleiben lediglich 71 Einträge. Eine genauere Untersuchung zeigt, dass der Begriff *intuition* viel öfter in Arbeiten über Pflege und Medizin auftaucht, wo beispielsweise das Verhältnis zwischen Ärzten/Pflegenden und Patienten diskutiert wird.

2.4.4 Bedauern und Enttäuschung

In einer konsequentialistische Ethik spielen Aspekte wie Bedauern und Enttäuschung eine wichtige Rolle, da sich diese bei der Bewertung der Folgen von Handlungen manifestieren können. Es wurde aber nicht systematisch nach Studien gesucht, welche solche Gefühle thematisieren. An dieser Stelle soll lediglich an einem Beispiel gezeigt werden, wie eine solche Untersuchung stattfinden kann. In einer Arbeit von Camille et al. [29] lässt sich dabei erneut die Problematik der Analyse komplexer psychischer Konzepte deutlich machen. In dieser Studie wurde anhand eines einfachen Glücksspiels (eine Art Roulette) das Verhalten von normalen Versuchspersonen mit solchen verglichen, die einer Schädigung des obitofrontalen Kortex aufweisen. Die Personen konnten zwischen zwei Spielen auswählen: Beim ersten Spiel waren eine geringe Belohnung und eine geringe (finanzielle) Bestrafung gleich wahrscheinlich. Beim zweiten Spiel war der finanzielle Verlust gleich hoch wie beim ersten, aber wahrscheinlicher, während der Gewinn weit höher, aber unwahrscheinlicher war. Das Experiment verlief dann in zwei Varianten: In der ersten Variante wählten die Personen die gewünschte Spielweise aus, das Ergebnis des Roulette erfolgte und die Versuchspersonen bewerteten schliesslich ihren emotionalen Zustand auf einer Skala zwischen den Extremen *extremely sad* und *extremely happy*. Dies ergibt ein Mass für Enttäuschung (*disappointment*). In der zweiten Variante sahen die Versuchspersonen nach ihrer Wahl der Spielweise sowohl was passiert, wie auch, was passiert wäre, wenn sie die andere Spielweise gewählt hätten. Danach mussten die Personen erneut ihren emotionalen Zustand einschätzen. Dies bestimmt die Variable Bedauern (*regret*). Parallel wird in beiden Varianten die Hautleitfähigkeit gemessen. Danach kam es zu einer neuen Runde. Das Resultat zeigte, dass Personen ohne Schäden im obitofrontalen Kortex aus ihren emotionalen Erfahrungen derart lernten, dass das Gefühl des Bedauerns schliesslich weniger oft auftritt. Dies war bei Personen mit solchen Schädigungen nicht der Fall. Zudem war die emotionale Reaktion bei *regret* höher als bei *disappointment* bei normalen Versuchspersonen, nicht aber bei solchen mit Schäden im orbitofrontalen Kortex. Dies wird als Hinweis darauf gewertet, dass *regret* und *disappointment* durch unterschiedliche neuronale Prozesse verursacht werden. Zudem wird aus dem Ergebnis geschlossen, dass Menschen mit Schäden im orbitofrontalen Kortex das Gefühl des Bedauerns nicht haben könnten. Diese Schlussfolgerungen haben die bekannten Schwächen: Solche Ergebnisse beweisen letztlich nichts anderes, dass mentale Aspekte durch neuronale Aktivitäten verursacht werden und dass, wenn die Aspekte unterschiedlich sind, auch die Prozesse unterschiedlich sein müssen. Zudem wird der Fehler begangen, von einem operationalisierten Begriff von Bedauern zu einem sehr viel komplexeren Begriff von Bedauern zu schliessen.

2.4.5 Lügen

Lügen ist ein klassisches Beispiel eines unmoralischen Verhaltens. Da Lügen natürlich auch in rechtlicher Hinsicht bedeutsam ist und im Zug des *war against terror* zusätzlich auf Interesse gestossen ist (Identifikation von Attentätern auf Flughäfen), existieren seit langem Bestrebungen, Lügen mittels so genannter Lügendetektoren zu erkennen. Klassische Lügendetektoren messen Änderungen des Hautwiderstands, des Pulsschlags und der Atmungsfrequenz als Folge von Lügen. Diese physiologischen Änderungen werden nicht als eine direkte Folge des Lügens angesehen. Die Hypothese ist vielmehr, dass der Lügner ja weiss,

dass er lügt und dass dieses Wissen im Kontext einer Befragung durch den Lügendetektor Angstzustände (nämlich die Angst, dass das Lügen entdeckt wird) hervorruft, welche die genannten physiologischen Veränderungen hervorrufen. Die Sicherheit des Nachweises wird bestritten, weil vermutet wird, dass skrupellose Lügner solche Angstzustände vermindert aufweisen, während es wiederum Personen geben kann, die bei solchen Befragungen übermässige Angstzustände haben können, so dass der Detektor ausschlägt, obwohl der Betreffende die Wahrheit sagt.

Im Zeitalter des *Imaging* wird nun versucht, mehr über den Vorgang des Lügens zu erfahren – nicht zuletzt auch um alternative Lügendetektoren entwickeln zu können. Zum einen kann man ausgehend von obiger Hypothese hinsichtlich der Funktionsweise klassischer Lügendetektoren annehmen, dass die durch das Lügen hervorgerufenen Angstzustände sich auch in entsprechenden Aktivitätsmustern im Gehirn niederschlagen. Phan et al. untersuchten dies in einer fMRI-Studie und kamen zum Schluss, dass Lügen mit erhöhter Aktivität im ventrolateralen präfrontalen Kortex, dem dorsolateralen präfrontalen Kortex, dem dorsalen medialen präfrontalen Kortex und dem superioren temporalen Suclus einhergehen soll [134]. Offen ist aber, ob ein auf *Imaging* beruhender Lügendetektor mit den oben genannten Problemen besser fertig wird als klassische Detektoren. Zum anderen könnte aber auch geprüft werden, ob die mit dem Akt des Lügens einhergehende Bildung fiktiver Szenarien mit charakteristischen Aktivitätsmustern einhergehen. Ein derartiger Lügendetektor könnte die oben genannten Probleme umgehen und ist offenbar in Entwicklung [164].

2.4.6 Vertrauen

Vertrauen ist eine Verhaltensdisposition die in jüngerer Zeit unter Benutzung der Methoden der experimentellen Spieltheorie genauer untersucht worden ist. Diese Methoden ermöglichen eine Operationalisierung von Vertrauens-Verhalten, indem dieses durch Investitions-Experimente erfasst und durch den Betrag des investierten Geldes quantifiziert wird. Natürlich findet sich auch psychologische Forschung über Vertrauen, die an dieser Stelle aber nicht weiter vorgestellt werden soll. Vielmehr soll anhand zweier Studien aufgezeigt werden, wie nach den neurobiologischen Grundlagen für Vertrauen gesucht wird.

Ein Thema ist die Frage nach neuronalen Korrelaten von Vertrauen. In einer Studie von King-Casas et al. wird diese Frage unter Mithilfe von Vertrauensspielen angegangen [102] (zu den Erklärung der Spiele siehe Abschnitt 2.2.4). Die beteiligten Partner spielen dabei mehrfach miteinander, so dass sich im Verlauf der Interaktion drei Arten von Verhalten des Investors (A) ausgebildet haben: Benevolentes Verhalten (A investiert mehr, obwohl B zuvor weniger Geld als erwartet zurückgegeben hat), neutrales Verhalten (das Verhalten des Investors ändert sich nicht im Vergleich zum vorangegangenen Spiel) und malevolentes Verhalten (A investiert weniger, obwohl B zuvor mehr zurückgegeben hat). Die Hirnaktivität beider Personen wurde während der Interaktion (via Video) mittels fMRI erfasst. Als wesentliche Hirnregion für die Untersuchung des neuronalen Korrelats von Vertrauen wurde das dorsale Striatum identifiziert, wobei Stärke wie Zeitpunkt der maximalen Aktivierung in Person B im Zentrum standen. Die Studie kommt zum Schluss, dass sich aus diesen Daten Hinweise über die Absicht, im nächsten Spiel zu kooperieren, gewinnen lassen. Demnach soll die Stärke der Aktivierung des dorsalen Striatum die Absicht, in einem kommenden Spiel zu kooperieren, reflektieren (je stärker die Aktivierung, desto grösser die Wahrscheinlichkeit, dass B im kommenden Spiel kooperiert). Zudem verschiebt sich dieses

intention to trust Signal (also der Zeitpunkt der maximalen Aktivierung) in B: Zu Beginn erfolgt dieses, nachdem A seinen Entscheid über den Investitionsbetrag gefällt hat. Später erfolgte die maximale Aktivierung, bevor dieser Entscheid bekannt gegeben wurde. Die Autoren der Studie interpretieren dies so, dass diese Verschiebung die Reputation von A in B zu Ausdruck bringen soll. Die Autoren schliessen zusammenfassend: "Taken together, these results suggest that the head of the caudate nucleus receives or computes information about (i) the fairness of a social partner's decision and (ii) the intention to repay that decision with trust." Schliesslich vermöge diese Studie auch Hinweise über *social processing deficits* zu geben. Auch bei dieser Studie sind eine Reihe kritischer Fragen zu stellen: So erstaunt zum Ersten, dass das Signal in B als neuronales Korrelat von Vertrauen aufgefasst wird, zumal ja A vertrauen muss, dass sein Investitionsentscheid von B honoriert wird. Zweitens stellen sich eine Reihe methodischer Fragen: Der genaue Versuchsablauf (z.B. wann genau wird gemessen?) ist unpräzise beschrieben, die Unterschiede im gemessenen Signal sind nur sehr klein und es fehlen auch die Fehlerbalken bei den entsprechenden Graphen. Drittens schliesslich scheint der Schluss, dass sich aufgrund solcher Experimente Hinweise auf *social processing deficits* gewinnen liessen, doch etwas gewagt.

Ein zweites Thema ist, inwieweit chemische Substanzen Vertrauen zu beeinflussen vermögen. Aufsehen hat diesbezüglich eine Studie von Kosfeld et al. erregt, welche in einer Variante des *trust game* den Einfluss des Neuropeptids Oxytocin auf der Veralten der Versuchspersonen untersucht hat [104]. In der Studie wurde festgestellt, dass die Gabe von Oxytocin bei der Versuchsperson A diese dazu verleitet, im Schnitt mehr Geld zu investieren, was als eine Zunahme von Vertrauen von A in B interpretiert wurde. Ist der Partner von A aber ein Computer (A weiss das), der seine Entscheide zufällig trifft, so ändert die Gabe von Oxytocin das Investitionsverhalten von A nicht. Dies wurde so interpretiert, dass Oxytocin seine Wirkung nur in einem sozialen Kontext entfaltet.[12] Bei B wiederum hat die Gabe von Oxytocin keinen Einfluss auf dessen Entscheide gehabt. Interessant bei solchen Studien ist, dass sie eine Zusammenhang zwischen einem komplexen sozialen Verhalten und der Applikation einer einzigen Substanz schaffen.

2.4.7 Kooperation

Kooperation ist Gegenstand einer Vielzahl spieltheoretischer wie verhaltensbiologischer Untersuchungen, so dass an dieser Stelle kein umfassender Überblick gegeben werden kann. Fokussiert werden im Folgenden neuere Studien, welche Kooperation mittels experimenteller Spiele untersuchen. Die wichtigsten experimentellen Spiele, welche für diese Studien verwendet werden, sind in Abschnitt 2.2.4 erläutert. Im Kontext solcher Studien wird Kooperation als ein Signal aufgefasst, mit welchem sich ein Mitglied einer Gruppe gegenüber anderen Mitgliedern als zuverlässiger Partner präsentiert. Da Kooperation meist einen gewissen Aufwand der kooperierenden Partner erfordert, wird Kooperation oft unter dem Stichwort des *costly signalling* untersucht [71]. Ein möglicher Aspekt der Kosten für Kooperation

[12] Dieser Unterschied kann natürlich auch so interpretiert werden, dass man emotional anders involviert ist, wenn der Gegenüber ein Mensch oder ein Computer ist. So prüften McCabe et al. in einem Vertrauensspiel die Hirnaktivierung von Spieler A, wenn B ein Mensch bzw. ein Computer ist. Wütend wurden die Spieler A nur dann, wenn ein Mensch Vertrauen missbrauchte, nicht aber der Computer – und dies zeigte sich auch in einer entsprechenden Aktivierung von Regionen, die mit emotionalen Reaktionen in Verbindung gebracht werden [114].

beinhaltet den Aufwand, allfällig nicht-kooperierende Gruppenmitglieder zu bestrafen. Diese Überlegungen führten zum Konzept der starken Reziprozität (*strong reciprocity*) [70]. Grundidee dieses Konzeptes ist, dass ein Gruppenmitglied, das gemäss starker Reziprozität handelt, sich grundsätzlich kooperativ verhält, bei allfällig defektem Verhalten des Partners nicht mehr kooperiert und Nichtkooperierende bestraft, soweit dies einen gewissen Aufwand nicht überschreitet. Dieses Grundverhalten ist unabhängig davon, ob Verwandschaftsbeziehungen zwischen den Gruppenmitgliedern bestehen oder ob die Wahrscheinlichkeit eines künftigen Zusammentreffens klein ist. Schliesslich trifft in spieltheoretischen Überlegungen oft auch der Begriff "Norm" auf, der aber unterschiedlich verstanden werden kann. Im Kontext der evolutionären Spieltheorie ist vorgeschlagen worden, Normen als stabile Strategien aufzufassen [18]. Basierend auf Modellstudien (analytische Untersuchungen [70, 71] wie auch *agent-based modelling* [25]) wurde gezeigt, dass das Modell der starken Reziprozität zu stabilen Strategien führen soll. Dies motiviert die Idee, dass Normen als Verhaltensregeln aufzufassen sind, die durch Sanktionen gestützt sind. Die Art der Sanktion kann aber zweierlei sein: Internen Normen werden durch Gefühle wie Scham, Schuld und Verlust von Selbstwert sanktioniert. Sie entstehen durch einen Prozess der Internalisierung. Externe Normen werden durch Belohnung und Bestrafung durch Gruppenmitgliedern gestützt [72]. Externe Normen können weiter differenziert werden zwischen Normen, die durch den direkten Interaktionspartner wahrgenommen werden (d.h. eine allfällige Strafe erfolgt durch jene Person, mit der nicht kooperiert wurde) und solchen, bei denen die Sanktion durch eine dritte Partei erfolgt. Letztere werden soziale Normen genannt [18].

Das Konzept der starken Reziprozität und die Bedeutung der Strafe bei Nicht-Kooperation ist in jüngerer Zeit Gegenstand der experimentellen Ökonomie geworden. Dies, weil Strafen – vor allem ein mit Aufwand verbundenes Strafen – im Widerspruch zum klassischen Modell des *homo oeconomicus* steht, denn *costly punishment* vermindert den maximalen *payoff*. Verschiedene experimentelle Studien zeigten aber, dass sich viele Menschen in einer Reihe experimenteller Spiele nicht gemäss dem Standardmodell des *homo oeconomicus* verhalten. So geben beispielsweise in Diktator-Spielen 20 bis 40 Prozent der Versuchspersonen A der Person B die Hälfte des zur Verfügung stehenden Betrags, selbst wenn die beiden Personen sich in nachfolgenden Spielen nicht mehr treffen [62]:847-848. In anderen Spielen wiederum verhalten sich die Versuchspersonen durchaus egoistisch im Sinn des *homo oeconomicus* Modells. Zur Erklärung dieses Unterschieds verwenden Fehr und Schmitt das Konzept der *Fairness* als Aversion gegen Ungleichheit (*self-centered inequity aversion*) [62]:819. Sie postulierten, dass Versuchspersonen in experimentellen Spielen bereit sind, einen Teil des eigenen *payoffs* abzugeben, um das System in Richtung einer grösseren Gleichheit zu bringen. Weiter wird postuliert, dass sich lediglich eine Subpopulation von interagierenden Versuchspersonen von solcher Fairness leiten lassen muss, damit Kooperation ein stabiles Phänomen wird. In weiteren Spielen wurde der Einfluss des Strafens untersucht. Das Zusammenspiel von Strafen und Entstehung von Kooperation wurde Gegenstand einer Vielzahl von experimentellen Studien. In einem *one shot public good* Spiel (d.h. zwei Personen einer Gruppe von interagierender Versuchspersonen trafen jeweils nur einmal aufeinander) haben Fehr und Gächter gezeigt, dass Kooperation sich dann entwickelte, wenn *costly punishment* möglich war [63]. Nachträgliche Befragungen der Versuchspersonen haben gezeigt, dass Trittbrettfahrer zu stark negativen emotionalen Reaktionen bei den "Opfern" geführt haben. Trittbrettfahrer wiederum erwarteten ihrerseits, dass die anderen solche Reaktionen haben werden. Natürlich hängt das Auftreten von Kooperation von der genauen Ausgestal-

tung des Spiels aus: FALK et al. haben dazu Versuchspersonen ein Ultimatum-Spiel mit unterschiedlichen Varianten spielen lassen [57]. Dabei verändert sich das Verhältnis der angenommenen zu den abgelehnten Angeboten je nach zur Verfügung stehender Alternative, aber unabhängig vom eigentlichen Angebot (A wählt aus den Varianten x_1 und x_2 aus, wobei x_1 immer gleich ist, x_2 aber ändert. B kennt die Wahlmöglichkeiten von A, wenn er entscheidet, ob er x_1 annimmt oder ablehnt). Dies wird als Hinweis gewertet, dass der Signalcharakter des Angebots ändert, wenn die Alternative fairer bzw. unfairer ist – was wohl eine plausible Interpretation ist. Ein weiteres, intuitiv einleuchtende Ergebnis zeigt eine Studie von FALK und FISCHBACHER [58]: Nicht nur das Angebot in einem Ultimatum-Spiel bestimmt, wie entschieden wird, sondern auch die Absicht, die B dem Spieler A unterschiebt.

In einem weiteren Schritt wurde untersucht, mit welchen emotionalen Aspekten der Akt der Bestrafung einher geht. So postulierten FEHR und FISCHBACHER, dass Sanktionen sehr wichtig sind, um Normen durchzusetzen und damit Kooperation zu ermöglichen, und dass der Akt des Sanktionierens für den Ausführenden befriedigenden ist [64]. Die Frage ist nur, ob die damit einhergehenden Emotionen kausal mit dem Akt des Strafens verbunden sind, oder ob die Emotionalität lediglich ein Epiphänomen darstellt. Unter Hinweis auf Studien im Bereich der Neurobiologie von Emotionen wurde postuliert, dass ersteres der Fall ist [65].

Diese Überlegungen führten zu einer Verbindung von experimentellen Spielen mit Imaging-Experimenten, da man prüfen wollte, ob die in diesen Spielen auftretende emotionale Komponente sich auch in entsprechenden Hirnaktivierungen zeigt. HASELHUHN und MELLERS untersuchen diesen Zusammenhang mit modifizierten Ultimatum- und Diktatorspielen [86]. Die Spiele wurde derart verändert, dass die Versuchspersonen verschiedene Angebots-Szenarien hinsichtlich ihrer Präferenz und des damit erwarteten Lustgewinns (*pleasure*) beurteilten. Es zeigten sich zwei Gruppen von Spielern: Die erste Gruppe gewinnt *pleasure* aus grösserem *payoff*. Diese Spieler kooperieren im Ultimatumspiel, nicht aber im Diktatorspiel. Die zweite Gruppe gewinnt *pleasure* aus fairen Handlungen und die Spieler kooperieren demnach in beiden Spielen. Umgekehrt konnte aus den dadurch ermittelten neuronale Aktivierungsmustern vorausgesagt werden, welches Verhalten die Versuchspersonen in den Spielen tatsächlich zeigen werden. Dies sagt aber nicht mehr aus als die bekannte Tatsache, dass unterschiedliche Verhaltensweisen mit unterschiedlichen neuronalen Aktivierungen einhergehen.

DEQUERVAIN et al. untersuchten mit Hilfe eines Vertrauens-Spiels, in welchen Hirnregionen der Akt des Strafens eine erhöhte Aktivität auslöst [48]. Das Spiel wurde so modifiziert, dass, falls Spieler B nicht kooperiert, Spieler A diesen bestrafen kann. Dazu standen vier Varianten zur Verfügung: *intentional/costly, intentional/free, intentional/symbolic* und *nonintentional/costly*. Der Begriff *intentional* bedeutet, dass B bewusst nicht kooperiert hat, während B im Fall von *nonintentional* aufgrund einer äusseren Anweisung nicht kooperierte (was A wusste). Der Begriff *costly* bedeutet, dass der Akt des Strafens A etwas kostet, im Fall von *free* ist dies nicht der Fall und im Fall von *symbolic* findet keine eigentliche Bestrafung statt (d.h. kein Geld wird abgezogen). A hatte eine Minute Zeit, zu diesem Entscheid zu kommen und die neuronale Aktivität wurde in dieser Zeitspanne mittels PET erfasst. Die Autoren vermuteten, dass lediglich die beiden ersten Varianten für A befriedigend seien. Diese Vermutung wurde bestätigt. So zeigte sich zum ersten, dass der Akt der Bestrafung mit einer erhöhten Aktivität im dorsalen Striatum einher geht. Zum zweiten zeigte sich, dass diese Aktivität umso höher ist, desto mehr man für den Akt des Strafens investierte.

RILLING et al. schliesslich suchten nach neuronalen Korrelaten von kooperativem Ver-

halten [141], wobei sie Personen scannten, die das parallele Gefangenen-Dilemma mehrfach hintereinander spielten. In einem ersten Schritt wurden drei Szenarien unterschieden: 1) Zwei Personen spielten ohne äussere Einflussnahme. 2) Zwei Personen spielen, wobei aber die zweite Person einem gewissen Spielschema folgte (ohne dass die gescannte Person das wusste). 3) Eine Person spielte gegen einen Computer, der im ersten Spiel defektierte und danach einer *tit-for-tat*-Strategie folgte. Als Kontroll-Experiment mussten die Versuchspersonen im Scanner eine Wahl zwischen Geldbeträgen treffen (*baseline*-Bedingung). In einem zweiten Schritt wurde die Mensch-Computer-Interaktion genauer untersucht. Ohne auf das weitere Prozedere genauer einzugehen, haben die Autoren folgende Hirnregionen identifiziert, welche bei gegenseitiger Kooperation überdurchschnittlich aktiviert sein sollen: den *nucleus accumbens*, den *caudate nucleus*, den ventromedial-frontalen Kortex, den orbitofrontalen Kortex und den rostralen anterioren *cingulate cortex*. All diese Experimente zeigen, dass die Lokalisierung von Hirnregionen, welche bei den verschiedenen *tasks* überdurchschnittlich aktiv sein sollen, bislang im Zentrum jener Forschungsbemühungen standen, welche *Imaging* als Methode einsetzten.

2.4.8 Altruismus

Während Kooperation vor allem im Kontext der experimentellen Ökonomie ein Thema geworden ist, wird Altruismus eher in der Verhaltensforschung untersucht. Auch im Kontext der evolutionären Ethik spielt Altruismus bzw. die Frage nach seiner Entstehung eine wichtige Rolle. Nachfolgend soll keine umfassende Einführung in dieses Thema gegeben werden, weil dies den Rahmen dieser Pilotstudie sprengen würde. Generell lässt sich aber auch hier vermuten, dass es offenbar keine einheitliche Definition des Begriffs Altruismus gibt. Rachlin beispielsweise schlägt folgende Definition vor: "A truly altruistic act is always part of a pattern of acts (highly valued by both the actor and the community) particular components of which are dispreferred by the actor to their immediate alternatives [139]:244. Offen ist die Frage, ob überhaupt (siehe dazu Danielsson in [139]:257) und wie Altruismus von Kooperation unterschieden werden soll. Ein möglicher Unterschied ist, dass eine Person durch altruistisches Verhalten einen unmittelbaren *payoff* in Form eines positiven Gefühls erhält, während man bei kooperativem Verhalten erst später einen *reward* erwartet [139]:258. Auffallend ist dennoch, dass in einigen Studien über Kooperation eine recht enge begriffliche Verbindung zwischen Altruismus und Kooperation eingegangen wird. So wird beispielsweise der Akt des Strafens als altruistisch aufgefasst (*altruistic punishment* [48]), weil eben dieser Akt dem Strafenden eine unmittelbare Befriedigung verschafft. In dieser Studie wird auch zwischen psychologischem und biologischem Altruismus unterschieden. So sei im ersten Fall das Handeln durch ein altruistisches Motiv motiviert, während im zweiten Fall das Handeln deshalb altruistisch sei, weil der Strafer einen Aufwand für die Strafe leisten müsse (*costly punishment*) und gleichzeitig dafür sorgt, dass der Bestrafte in Zukunft mehr kooperiert (demnach wäre also das Modell der *strong reciprocity* ein Modell für biologischen Altruismus).

Chapter 3

Neuronale Grundlagen der Moral

3.1 Einführung

In jüngerer Zeit findet sich eine zunehmende Anzahl neurowissenschaftlicher Studien, welche explizit Moral bzw. moralisches Verhalten zum Gegenstand haben. Diese Studien werden in diesem Kapitel anhand der Einteilung von Abschnitt 1.3.1 vorgestellt. Aufgrund der Literaturrecherche wurden insbesondere folgende Forschergruppen identifiziert, welche sich in den vergangenen Jahren mit dieser Fragestellung befasst haben:

- **William Casebeer**: Arbeitet vorab in den Bereichen *moral cognition* und *moral decision making*. Adresse: Department of Philosophy, United States Air Force Academy, 2354 Fairchild Drive USAF Academy, Colorado 80840, William.casebeer@usafa.af.mil.

- **Frans de Waal**: Arbeitet im Bereich Verhaltensforschung von Primaten und über Vorformen von Moral in Primaten. Adresse: Department of Psychology, Emory University, 532 Kilgo Circle Atlanta, GA 30322, dewaal@emory.edu.

- **Joshua Greene**: Arbeitet vorab im Bereich *moral decision making* und dem Wechselspiel zwischen emotionalen und kognitiven Aspekten bei *moral cognition*. Derzeit noch in Princeton, ab Juli 2006: Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge MA 02138, jdgreene@princeton.edu.

- **Hauke Heekeren**: Arbeitet im Bereich *(moral) decision making*, dem Wechselspiel zwischen emotionalen und kognitiven Aspekten bei *moral cognition* und dem Einfluss individueller Aspekte wie Alter. Adresse: Junior Research Group Neurocognition of Decision-Making. Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, heekeren@mpib-berlin.mpg.de.

- **Jorge Moll**: Arbeitet im Bereich *moral cognition*, Verarbeitung moralischer Stimuli. Adresse: Cognitive Neuroscience Section, National Institute of Neurological Disorders

and Stroke, Building 10, room 5C205, 10 Center Drive, MSC 1440 Bethesda MG, mollj@ninds.nih.gov.

- **Tania Singer**: Arbeitet in Bereich neuronale Grundlagen des menschlichen Sozialverhaltens, insbesondere Empathie, Fairness, Vertrauen und Reziprozität. Adresse: Wellcome Department of Imaging Neuroscience & Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, singer@fil.ion.ucl.ac.uk.

Casebeer und Churchland haben in einem der ersten Übersichtsartikel zu diesem Thema festgehalten [33], dass diese Studien vor allem moralische Emotionen, moralische soziale Kognition und abstraktes moralisches Denken zum Gegenstand haben. All diese Studien zeichneten sich durch sehr vereinfachende Annahmen über *moral reasoning* aus, welche auch in entsprechend einfachen experimentelle Paradigmen untersucht wurden. Bereits durch diese Studien sei klar geworden, dass es mit Sicherheit kein abgrenzbares *moral module* oder "Moralzentrum" im Gehirn gebe, was im Übrigen auch unplausibel sei. Drei Motivationen für solche Studien werden angegeben: Erstens sollen solche Studien Hilfestellung zur Identifizierung moralischer Pathologien, die eine biologische Ursache haben, liefern. Zweitens sollen Hinweise für eine Verbesserung der Moralerziehung gewonnen werden. Drittens schliesslich sollen solche Studien auch Beiträge für die Lösung normativ-ethischer und metaethischer Fragen liefern. Casebeer beispielsweise argumentiert, dass die jetzigen Erkenntnisse der Neurowissenschaft die aristotelische Tugendethik gegenüber deontologischen und utilitaristischen Konzeptionen auszeichnen würden [31, 32]. Er argumentiert zusammengefasst wie folgt: Die utilitaristische Moralpsychologie erfordere einen Lernmechanismus derart, dass erkannt werden könne, welche Aktionen oder Regeln Glück produzieren würden: "In terms of raw computations, a utilitarian moral psychology would require some mechanisms for learning what actions or rules would eventually produce happiness" [31]:841. Die Kant'sche Moralpsychologie hingegen braucht die Fähigkeit einer affektfreien rationalen Prüfung der Universalitätsfähigkeit von Maximen: "It would require at least the ability to check universalized maxims for logical consistency in a manner that is separable from the taint of affect and emotion" [31]:842. Die aristotelische Tugendethik schliesslich sei eine "whole-psychology, whole-brain affair" [31]:842. In einer wenig einsichtigen Argumentation wird nun behauptet, dass die Vielzahl an Hirnregionen, die bei *moral cognition* involviert sein sollen, offenbar am besten mit der Tugendethik verträglich sei – oder in den Worten von Casebeer: "There is clear consilience between contemporary neuroethics and Aristotelian moral psychology" [31]:845.

Greene, ein weiterer Exponent dieser Forschung, geht nicht so weit. Er glaubt aber dass metaethische Fragen besser untersucht werden können, indem Neurowissenschaft das *common sense* Verständnis von Moral untersuchen könne. Zudem könne diese Forschung die Frage nach der Gültigkeit eines ethischen Naturalismus entscheiden helfen: "Understanding how we make moral judgments might help us to determine whether our judgments are perceptions of external truths or projections of internal attitudes." [78]:849 Er macht aber nicht klar, wie man mittels *Imaging*-Studien zwischen diesen beiden Möglichkeiten unterscheiden könne.

Moll schliesslich ergänzt die Grundzielsetzung dieser Forschung mit dem Ziel "[to] help to shape environmental, psychological and medical interventions aimed at promoting prosocial behaviours and social welfare" [119]. Im Rahmen seiner umfassenden Übersichtsarbeit

nennt er drei grundlegende Beschränkungen für die neurobiologische Erforschung der Moral: Die Kontextabhängigkeit moralischen Verhaltens; die Schwierigkeit, den Effekt von Gehirnläsionen und Abnormitäten auf das Verhalten abschätzen zu können; und den kulturellen Relativismus. Angesichts dieser doch sehr grundlegenden Einschränkungen erstaunt der optimistische Grundton in seinen Ausführungen. Die Voraussagen des von ihm vorgeschlagenen Modells schliesslich sind entweder trivial oder möglicherweise gar keine Voraussagen: So ist die "general prediction [is] that different neural subdivisions store distinct knowledge or motivational states" [119] – was wohl niemand bestreitet. Es folgen dann Voraussagen hinsichtlich der Auswirkungen bestimmter Läsionen auf das Sozialverhalten. Doch wie stehen diese Voraussagen zur zuvor gemachten grundlegenden Einschränkung und inwiefern sind diese Voraussagen nicht bereits bekannte Phänomene in der Neurologie? Diese kurze Übersicht weckt Zweifel an der zuweilen geäusserten Behauptung, das Forschung im Bereich "neurobiologische Grundlagen der Moral" einen wesentlichen Einfluss auf die normative Ethik und die Metaethik hat.

3.2 Moralische Stimuli

Die experimentelle Untersuchung von moralischem Verhalten verlangt nach der präzisen Definition eines moralischen Stimulus, der dieses Verhalten auslöst. Dies ist die erste Komponente des in dieser Pilotstudie skizzierten Modell eines *moral agent* (Abschnitt 1.3.1). In den *Imaging*-Experimenten gehen die Forscher davon aus, dass man solche moralischen Stimuli definieren und gegenüber anderen Stimuli auszeichnen könne.[1] Man muss also – insbesondere bei fMRI-Experimenten – nicht nur den eigentlichen moralischen Stimulus genügend präzise charakterisieren können, sondern auch andere Stimuli mit einer vergleichbaren Wahrnehmungskomplexität aber ohne moralischen Gehalt finden. In praktischer Hinsicht werden unterschiedliche, immer aber visuell präsentierte Stimuli verwendet: Fotografien von Situationen oder Gesichtern, einfache Sätze oder ganze, mit Wort und Bild unterlegte Beschreibungen von Szenarien und ethischen Dilemmas. Die genaue Ausgestaltung dieser Stimuli soll im Folgenden ausgeführt werden. Deren Verwendung in Experimenten zwecks Suche nach neuronalen Korrelaten von moralischen Entscheidungen wird in Abschnitt 3.3.2 erläutert.

- **Bilder:** Fotografien von (in meisten Fällen) Menschen, die einen moralischen Gehalt zum Ausdruck bringen sollen, sind ein oft verwendeter Stimulus. Als Beispiel soll eine Studie von Moll et al. dienen. Die Autoren charakterisieren moralische Bilder als "portraying emotionally charged, unpleasant social scenes, representing moral violations" [121]. Beispiele sind Bilder, die physische Bedrohungen zum Ausdruck bringen, Kriegsszene, oder Bilder verlassener Kinder. Als Kontrollstimulis werden unangenehme Bilder ohne (behauptete) moralische Konnotation verwendet wie beispielsweise verletzte Körper, gefährliche Tiere oder Kot. Zuweilen werden auch weitere Bildklassen verwendet: "angenehme" Bilder (Personen, Landschaften), "interessante" Bilder (surreale Bilder, Sportszenen), "neutrale" Bilder und verrauschte Bilder

[1] Gemäss Moll et al. soll dies bereits in psychologischen Studien der 1980er und 1990er Jahre gezeigt worden sein (siehe [122]:697 und die dortigen Referenzen, siehe auch der Beitrag von Eslinger et al. in [137]:34-35).

ohne Bildinhalt. Weiter können auch Bilder verwendet werden, die nicht einen direkt zugänglichen, sondern einen gelernten moralischen Gehalt aufweisen. SINGER et al. verwendet in einer Studie Gesichter von Personen, die zuvor in einem Spiel (Gefangenendilemma) als kooperierende oder als defektierende Personen aufgetreten sein sollen [150]. Hier figurieren die Bilder als Repräsentationen von moralisch wünschenswertem oder anzulehnendem Verhalten. Die Verwendung von Bildstimuli sind mit einer Reihe von Problemen behaftet: Generell dürfte es schwierig sein, den moralischen Gehalt von Bildern präzise zu charakterisieren – vor allem im Vergleich zu den Referenzstimuli. So weisen HEEKEREN et al. darauf hin, dass zwischen der emotionalen Komponente des Stimulus selbst (z.B. hervorgerufen durch eine Gewaltabbildung) und den durch das *moral decision making* aufkommenden Emotionen unterschieden werden müsse [91]. Dieser Zusammenhang wurde auch experimentell geprüft, indem Bilder verwendet wurden, die Körperverletzungen zeigten [90]. Dabei hat sich gezeigt, dass Reaktionszeiten wie auch Aktivierungen gewisser Hirnregionen (*temporal pole*) sich ändern, wenn derartige Bilder verwendet werden – unabhängig davon, ob von der Versuchsperson eine moralische oder eine semantische Entscheidung verlangt wurde. Unerwartete Effekte kann beispielsweise auch die Rasse der abgebildeten Personen haben. So hat sich beispielsweise in einem *Imaging*-Experiment gezeigt, dass Gesichter der gleichen Rasse wie die Versuchsperson nicht nur schneller erkannt werden, sondern auch zu unterschiedlichen Aktivierungen führen [135]. Diese Beispiele zeigen, dass unerkannte Korrelationen in den verwendeten Bildern die Resultate verfälschen können, da eben diese Korrelationen zu den in den *Imaging*-Studien beobachteten neuronalen Aktivierungen führen können.

- **Einfache Sätze:** Eine Möglichkeit, dieses Problem zu umgehen, bietet die Verwendung einfacher Sätze, weil deren Inhalte im Vergleich zu Bildern explizit zugänglich sind. So werden kurze Sätze verwendet, die moralisch einfach zu bewertende Sachverhalte zum Ausdruck bringen (z.B. "Töten eines Unschuldigen"). Diese werden dann von den Versuchspersonen hinsichtlich der Alternativen "moralisch richtig" und "moralisch falsch" bewertet. Als Vergleichs-Stimulus bieten sich Sätze an, die semantisch zu bewertende Sachverhalte zum Ausdruck bringen (z.B. "Das Viereck ist rund") und dann hinsichtlich "korrekt" oder "inkorrekt" bewertet werden. Ein solcher Ansatz verwendete HEEKEREN et al. [91]. MOLL et al. benutzten in einer ähnlichen Studie drei Arten von Sätzen: Nichtmoralisch-neutrale Sätze (z.B. "He never uses the seat belt."), nichtmoralische Sätze mit einer unangenehmen emotionalen Konnotation (z.B. "He licked the dirty toilet.") und Sätze mit moralischem Inhalt (z.B. "He shot the victim to death"). Diese Sätze mussten dann von den Versuchspersonen als "richtig" oder "falsch" klassifiziert werden. Die Auflistung der Sätze wie auch die Bewertungsmethode lassen aber Fragen offen. Ist beispielsweise das erste Beispiel wirklich frei von einer moralischen Konnotation? Ist die Einordnung dieses Satzes als "richtig" oder "falsch" nicht eine moralische Aussage? Die Zuteilung der Sätze zu den drei Kategorien soll zwar mittels einer Begleituntersuchung mit anderen Versuchspersonen validiert worden sein, dennoch erscheint es seltsam, warum immer noch derart mehrdeutige Beispiele in der Studie genannt werden.
- **Dilemmas:** Die bisher komplexesten Stimuli bestehen aus Beschreibungen moralischer Dilemmas, was üblicherweise mit längeren Texten, teilweise kombiniert mit Bildin-

halten passieren kann. Diese Dilemmas bringen miteinander konfligierende moralische Werte in einen Zusammenhang und verlangen von der Versuchsperson eine Entscheidung, welche dieser beiden Werte verletzt werden soll. Für fMRI-Studien muss zudem ein alternatives Dilemma gefunden werden, das eine vergleichbare Komplexität aufweist, sich aber im moralischen Gehalt unterscheidet. GREENE, der am meisten mit solchen Stimuli arbeitet [80, 78, 77], unterscheidet dazu zwei Arten von Dilemmas: In *personal dilemmas* ist die Versuchsperson aufgefordert sich vorzustellen, direkt (d.h. durch Einsatz seines Körpers) in das Szenario einzugreifen. Drei Kriterien definieren ein *personal dilemma* [77]: Erstens, eine der beiden im Dilemma involvierten moralischen Werte ist von solcher Art, dass seine Verletzung zur Schädigung eines menschlichen Körpers führt. Zweitens, diese Schädigung betrifft eine konkrete Person oder Gruppe von Personen. Drittens, diese Schädigung ist keine Reaktion im Sinn eines Abwehrverhaltens von Gefahren. Dies lasse sich im Satz "me hurt you" zusammenfassen. Ein unpersönliches Dilemma erfüllt diese Kriterien nicht. GREENE verwendet folgende persönlichen Dilemmas in seinen Studien: Im *footbridge dilemma* ist die Versuchsperson mit dem Problem konfrontiert, dass ein unkontrolliert gewordener Strassenbahnwagen auf eine Gruppe von Personen zurast. Dieser könne nur dadurch aufgehalten werden, indem die Versuchsperson eine beleibte Person von einer Brücke auf das Geleise stösst, so dass der Wagen aufgehalten, die Person aber getötet werde. Die Versuchsperson muss entscheiden, ob die Person von der Brücke gestossen werden solle oder nicht. Im *crying baby dilemma* befindet sich eine Gruppe von Flüchtlingen im Keller eines Hauses, das von feindlichen Truppen durchsucht wird. Würde die Gruppe entdeckt, würden alle Personen erschossen. Die Versuchsperson muss sich mit einer Mutter identifizieren, die ihr Baby im Arm trägt. Das Baby beginnt zu weinen und gefährdet damit die ganze Gruppe. Die Versuchsperson muss sich dafür entscheiden, entweder die Entdeckung der Gruppe zuzulassen, oder dem Baby im Arm den Mund zuzuhalten mit der Folge, dass das Baby erstickt. In einem weiteren Dilemma schliesslich versetzt sich die Versuchsperson in einen Autofahrer, der mit seinem neuen Auto mit teuren Ledersesseln unterwegs ist. Er trifft am Strassenrand auf einen stark blutenden Fremden und ist mit der Frage konfrontiert, ob er diese Person mitnehmen soll (und damit seine teuren Ledersessel ruiniert) oder nicht.

Folgende unpersönliche Dilemmas werden als Kontrolldilemmas verwendet: Im *trolley dilemma* ist die Ausgangssituation vergleichbar mit dem *footbridge dilemma*. Nur befindet sich die Einzelperson auf einem Nebengeleise und die Versuchsperson muss entscheiden, ob sie entweder nichts tut (und der Strassenbahnwagen rammt eine Gruppe von Menschen) oder eine Weiche umlegt, so dass nur die Einzelperson getötet wird. Im Infantizid-Dilemma muss die Versuchsperson bewerten, ob eine Teenagerin nach der Geburt ihr unerwünschtes Kind töten darf. Im Spenden-Dilemma schliesslich muss die Versuchsperson entscheiden, ob sie eine gewisse Summe entweder für den eigenen Konsum, oder für eine anonyme Spende an ein Hunger-Hilfswerk verwendet. GREENE verweist auf den psychologisch gut untersuchten Unterschied, dass persönliche Dilemmas intuitiv anders entschieden werden als unpersönliche Dilemmas, obgleich eine utilitaristisch begründete Entscheidung jeweils zum gleichen Resultat führen würde (z.B. würde sowohl im *footbridge dilemma* wie auch im *trolley dilemma* eine Person geopfert, um eine ganze Gruppe von Personen zu retten). Das Ziel von GREENE ist es, diesen

Unterschied gewissermassen neurobiologisch festzunageln, indem unterschiedliche Aktivierungsmuster mittels *Imaging* ermittelt werden.

Es lassen sich eine Reihe von Problemen bei der Verwendung solcher Stimuli diagnostizieren. So ist zum ersten unklar, wann genau gemessen werden soll, weil das Erfassen des Dilemmas wie auch das Fällen des Entscheids sich über viele Sekunden erstreckt. Weiter ist die Unterscheidung zwischen persönlichen und unpersönlichen Dilemmas nicht unbedingt klar, da sich die Versuchsperson beim Entscheidungsprozess durchaus in beiden Fällen in die involvieren Personen hinein versetzen und damit vergleichbare emotionale Empfindungen haben kann. Festgestellte Unterschiede im *Imaging* würden dann auf andere Unterschiede in den Dilemmas hinweisen. Da es sich dabei um komplexe Situationen handelt, sind solche, vom Versuchsleiter unerkannte Unterschiede, durchaus möglich.

Zur Charakterisierung moralischer Stimuli ist weiter vorgeschlagen worden, zwischen einfachen und komplexen moralischen Stimuli zu unterscheiden [91]. Einfache Stimuli zeichneten sich durch einen eindeutigen, nicht-dilemmatischen Charakter aus, während komplexe Stimuli einen starken emotionalen Charakter haben. Es ist unklar, ob diese Unterscheidung für alle Arten von Stimuli gleich einfach zu bewerkstelligen ist. Wie wollen Bilder einen moralischen Sachverhalt zum Ausdruck bringen sollen, gleichzeitig aber keine emotionale Reaktion auslösen? Bei Sätzen wiederum scheint diese Unterscheidung eher möglich, was wohl auch der Grund ist, dass Heekeren et al. für ihre Untersuchung Sätze verwendet haben. Hinsichtlich der Charakterisierung moralischer Stimuli müssten zudem die bekannten Einwände des ethischen Relativismus berücksichtigt werden. So kann die Einordnung eines Stimulus als "moralisch" kulturrelativ sein kann. Dies wird deutlich in einer japanischen Studie, wo die Sätze in die Kategorien "neutral" ("I used a cellular phone in the park"), "Schuld ausdrückend" ("I used a cellular phone in the hospital") oder "peinlich" ("I was not dressed properly for the occasion") klassifiziert wurden [154]. Sätze der zweiten und dritten Kategorie sollen dabei moralische Emotionen auslösen, weil sie als Folge von Verletzungen moralischer Konventionen auftreten. In einem europäischen Kontext würden solche Sätze wohl aber eher als Verletzungen von Benimmregeln aufgefasst, die nicht notwendigerweise eine moralische Komponente beinhalten müssen. Weiter ist es auch möglich, dass sich bei der Klassifizierung von Stimuli als "moralisch" *gender*-Unterschiede zeigen. Moll et al. sprechen dieses Problem an und verweisen darauf, dass dies offenbar noch nicht untersucht worden sei [121]. In ihren Daten liessen sich aber bisher keine deutlichen Hinweise auf solche Unterschiede nachweisen.

3.3 Moralische Kognition

3.3.1 Was ist moralische Kognition?

Der Prozess der Aufnahme und Verarbeitung moralischer Stimuli unter Einbezug möglicher Prädispositionen und Färbungen dieses Verarbeitungsprozesses bildet der nächste Gegenstand des skizzierten Modells von Moral. Von verschiedener Seite wurden Defitionen einer moralischen Kognition (*moral cognition / moral reasoning*) vorgeschlagen, welche diesen Prozess genauer umschreiben sollen. So definieren Casebeer und Churchland moralische

Kognition wie folgt: "Moral reasoning deals with cognitive acts and judgments associated with norms, or with facts as they relate to norms" [33]:171. CASEBEER nennt im weiteren sechs charakteristische Eigenschaften moralischer Kognition: Sie involviert Emotionen (ist "*hot*"), bezieht sich auf einen sozialen Aspekt, ist abhängig vom Kontext, orientiert sich auf ein Ziel hin, ist *distributed* und *genuine* [31]:845-846. Diese Eigenschaften sind hingegen entweder wenig überraschend oder schlecht definiert (was meint "genuin" in diesem Zusammenhang?). Hinsichtlich der involvierten neuronalen Prozesse liessen sich aber nach Ansicht von CASEBEER und CHURCHLAND keine genuinen Unterschiede zwischen *practical reasoning* und *moral reasoning* machen – was in dieser Form natürlich auch die Aussagekraft von *Imaging*-Experimenten in diesem Bereich weiter abschwächt.

Soll moralische Kognition mittels fMRI untersucht werden, stellt sich die Frage nach dem Vergleichsszenario. Welche Form von Kognition hat eine vergleichbare Komplexität, ist aber nicht moralisch? Eine Möglichkeit wäre, die Unterscheidung zwischen moralischen Regeln und konventionellen Regeln zu verwenden (siehe dazu [20]). Doch auch hier wird hingewiesen, dass der Einbezug von Emotionen bei der Beurteilung konventioneller Regeln berücksichtigt werden muss [127]. Die Verletzung von Konventionen, welche eine affektiven Charakter haben (z.B. Ekel erzeugen wie auf den Tisch spucken) dürften anders beurteilt werden als Konventionen, welche keine solche affektive Komponente haben.

3.3.2 Neuronale Korrelate moralischer Kognition

Imaging-Methoden werden bei der Untersuchung moralischer Kognition mit dem Ziel eingesetzt, neuronale Korrelate (d.h. die bei solchen Prozessen involvierten Hirnregionen) dieser Prozesse zu finden. Dazu werden die Versuchspersonen mit den bereits beschriebenen moralischen Stimuli konfrontiert und zu einem *moral decision making* aufgefordert. Ein Grundproblem solcher Studien ist, ob es überhaupt eine einheitliche Klasse moralischer Entscheidungen gibt, oder ob es vielmehr sehr unterschiedliche Arten von moralischen Entscheidungen gibt – etwa eine Äusserung von Missfallen, ein Handeln nach akzeptierten Gruppennormen oder ein Entscheid nach Überdenken eines Dilemmas. Es ist demnach nicht klar, ob die unter Verwendung verschiedener experimenteller Paradigma gewonnenen Erkenntnisse von überdurchschnittlicher Aktivierung einzelner Hirnregionen überhaupt zu einem einheitlichen Bild verbunden werden können.

Eine Leithypothese vieler Imaging-Studien über moralische Kognition ist, das eine Reihe von Teilprozessen, die zum *moral decision making* gehören, unbewusst ablaufen und diesen Prozess entscheidend prägen (siehe Abschnitt 3.5) [79]. Kognition wie Emotion spielen zwar gleichermassen eine Rolle, doch unbewusst ablaufende emotionale Prozesse sollen eine entscheidende Rolle spielen. Diese quasi automatisierten Prozesse führen bei Personen, die mit moralischen Dilemmas konfrontiert sind, zu einer unmittelbaren Einschätzung der Situation hinsichtlich "moralisch gut" oder "moralisch schlecht". Diese Hypothese wird im direkten Gegensatz zum KOHLBERGschen Modell der moralischen Entwicklung gesetzt [103], welches bei einem ausgereiften moralischen Bewusstsein rationalen Prozessen der moralischen Entscheidungsfindung eine zentrale Rolle einräumt. Im Folgenden soll nun eine Übersicht über eine Reihe von Studien gegeben werden, die nach neuronalen Korrelaten für moralische Kognition suchen:

- In einer ersten Studie von GREENE et al. werden Versuchspersonen im Scanner

(fMRI) mit dem persönlichen *footbridge* Dilemma und dem unpersönlichen *trolley* Dilemma sowie so genannten nicht-moralischen Kontrolldilemmas (z.B. die Wahl eines geeigneten Verkehrsmittels) konfrontiert [80]. Geprüft wird die Frage, inwiefern Dilemmas mit starkem emotionalen Gehalt (persönliche Dilemmas) andere Hirnregionen aktivieren als unpersönliche Dilemmas. Diese Vermutung wurde bestätigt, d.h. Hirnregionen wie der medale frontale Gyrus, der posteriore *cingulate gyrus*, und der *angular gyrus*, die mit *emotional processing* zu tun haben sollen, sind bei *personal dilemmas* stärker involviert als bei *impersonal dilemmas* und den Kontrolldilemmas. Allerdings ist die prozentuale Abweichung im Bereich von lediglich 0.2-0.3% und es fehlen die Fehlerbalken bei den Grafiken. Im weiteren zeigte sich, dass die Reaktionszeit (d.h. die Zeitdauer, in der das Dilemma beurteilt wurde bis eine Antwort gegeben wurde) jener (wenigen) Versuchspersonen, die sich auch beim persönlichen Dilemma für das Töten der Einzelperson zwecks Retten der anderen Personen entschieden haben, länger ist als bei jenen, die sich anders entschieden haben. Dies werten die Autoren als eine Art "moralischen Stroop-Effekt"[2]. Problematisch könnte bei dieser Studie könnte zudem die Wahl des Referenzstimulus sein: Als *Baseline* wird nämlich die Messungen im Zeitabschnitt zwischen den Dilemmas verwendet.

- In einer weiteren Studie von Greene et al. werden Versuchspersonen im Scanner (fMRI) mit mit einer Reihe verschiedener Dilemmas konfrontiert [77]. In einem ersten Schritt wurden die Versuchspersonen mit *personal* vs. *impersonal* Dilemmas konfrontiert mit dem Ziel, die Ergebnisse der oben erläuterten Studie zu reproduzieren. In einem zweiten Schritt wurden sie mit so genannt "schwierigen" Dilemmas (wie das *cry baby* Dilemma) und "leichten" Dilemmas (wie das Infantizid-Dilemma) konfrontiert (diese, wie auch die anderen Unterscheidung wurde durch eine vorgängige Kategorisierung der verwendeten Dilemmas durch andere Testpersonen validiert). Ein schwieriges Dilemma ist dadurch charakterisiert, dass eine emotional problematische Komponente (z.B. Töten eines Babies beim *cry baby* Dilemma) gegen eine rational-utilitaristische Überlegung (Retten von Personen) in Konflikt geraten. In einem dritten Schritt wurde untersucht, welche Unterschiede sich bei der Beurteilung schwieriger Dilemmas hinsichtlich der Aktivierung von Hirnregionen ergeben, je nachdem ob eine utilitaristische oder eine nicht-utilitaristische Lösung gewählt wurde. Untersucht wird mit diesem Szenario insbesondere das in der ersten Studie aufgetretene Phänomen eines moralischen Stroop-Effekts. So soll erstens geprüft werden, ob die mit dem Stroop-Effekt einhergehende höhere Aktivierung des *anterior cingulate cortex* und des dorsolateralen präfrontalen Kortex auch beim moralischen Stroop-Effekt auftaucht. Zweitens soll die Hypothese geprüft werden, ob die mit der oben genannten höheren Aktivierung einhergehenden Kontrollprozesse "work against the social-emotional responses described above and in favor of utilitarian judgments" [77]:390. Diese Hy-

[2]Der Stroop-Effekt ist ein bekanntes Phänomen der Wahrnehmungspsychologie, wo automatisierte Verarbeitsungsprozesse mit kontrollierten Verarbeitungsprozessen in Konflikt geraten. Der klassische Test verläuft wie folgt: Zuerst werden einer Versuchsperson Farbnahmen vorgelegt, die so rasch als möglich abgelesen werden sollen. Dann werden der Person Farbstreifen vorgelegt, die so rasch als möglich korrekt benannt werden müssen. Schliesslich werden farbig gedruckte Farbnamen vorgelegt und die Versuchspersonen müssen den Namen der Druckfarbe so rasch als möglich nennen – wobei aber die Druckfarbe nicht immer gleich der Farbbezeichnung ist (z.B. ein blau gedrucktes Wort "grün"). Diese Aufgabe dauert länger als die beiden vorangegangenen, was als Stroop-Effekt bezeichnet wird.

pothese erscheint seltsam, da man annehmen kann, dass sie automatisch erfüllt ist, ist doch der moralische Stroop-Effekt ja der Ausdruck des Konflikts einer utilitaristischen Überlegung in einem *personal dilemma* – d.h. eine positive Bestätigung der ersten Hypothese führt automatisch zu einer positiven Bestätigung der zweiten Hypothese. Die Studie bestätigte die Ergebnisse der vorangehenden Studie (Schritt 1) sowie die beiden genannten Hypothesen. Die Autoren kommen zum Schluss, dass die genannten Hirnregionen, welche mit abstraktem Denken und kognitiver Kontrolle verbunden sind, verstärkt aktiv sein sollen, wenn utilitaristische Überlegungen gegen emotionale Widerstände obsiegen. Nebst dem bereits gesagten lassen sich zu dieser Studie folgende kritische Anmerkungen machen. Erstens ist beim *cry baby* Dilemma die Unterscheidung zwischen einer emotional geprägten Entscheidung (gegen das Töten des Babies) und einer rationalen Entscheidung (für das Töten des Babies, um die Gruppe zu retten) unklar. Auch letztere Entscheidung beinhaltet emotionale Komponenten – etwa hervorgerufen durch die Vorstellung, was passieren werde, wenn die Gruppe entdeckt würde (auch in diesem Fall stirbt ja das Baby). Zweitens stellen sich kritische Fragen hinsichtlich der Methodik: so sind die Unterschiede in der Aktivierung sehr klein und in den entsprechenden Kurven fehlen die Fehlerbalken. Drittens behaupten die Autoren, die Studie könne bekannte Spannungen zwischen deontologischen und utilitaristischen Ethiken erhellen – was aber in keiner Weise einsichtig gemacht wird.

- In einer fMRI-Studie von Moll et al. werden Versuchspersonen mit verschiedenen Bildstimuli (moralische Bilder vs. Varianten unangenehmer Bilder, siehe oben) konfrontiert [121]. Ziel ist zu prüfen, ob sich bei den vorgängig als moralisch taxierten Bildern ausgezeichnete neuronale Aktivitätsmuster finden lassen, ohne dass die Versuchspersonen eine Handlung im Scanner vollbringen müssen. Erst danach werden die Bilder von den Versuchspersonen hinsichtlich emotionalem und moralischem Gehalt bewertet Die Autoren kommen zum Schluss, dass das Betrachten moralischer Bilder insbesondere mit einer erhöhten Aktivierung des orbitalen und medialen präfrontalen Kortex und des superioren temporalen Sulcus einher geht. Es stellen sich bei dieser Studie die bereits oben ausgeführten Probleme hinsichtlich der Kategorisierung der Bilder und möglicher unerkannter Korrelationen zwischen den Bildern.

- In einer weiteren fMRI-Studie von *Moll* et al. wird zwischen emotionalen Stimuli in Form von Sätzen mit und ohne moralischem Gehalt unterschieden [122] – und zwar unter Verwendung der Kategorien "nicht-moralisch neutral", "nicht-moralisch unangenehm" und "moralisch" (siehe oben). Die Autoren kommen zum Schluss, dass bei der Bewertung moralischer Sätze insbesondere der mediale orbitofrontale Kortex, der temporal pole und der superiore temporale Sulcus der linken Hemisphäre aktiv sein sollen. Bei der Bewertungen von nicht-moralisch unangenehmen Sätzen hingegen waren die linke Amygdala, die lingual Gyri und der laterale orbitale Gyrus aktiv.

- In einer fMRI-Studie von Heekeren et al. wurden Versuchspersonen mit kurzen Sätzen (einfachen moralischen Stimuli, siehe oben) konfrontiert, die entweder einen moralisch oder einen semantisch richtig oder falschen Sachverhalt beschreiben [91]. Die Versuchspersonen mussten bei jedem Satz einen entsprechende Bewertung vornehmen. Die Autoren gingen von früheren Ergebnissen (u.a. erzielt von Greene und Moll) mit komplexen moralischen Stimuli aus, die besagten, dass die Verarbeitung solcher Stimuli

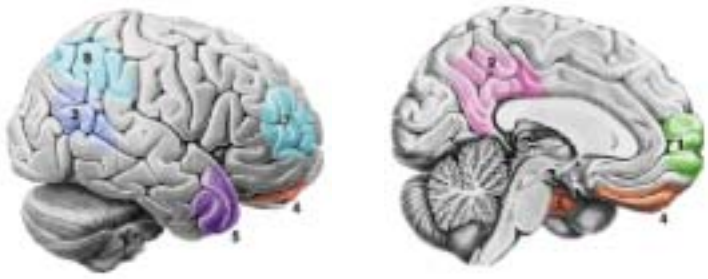

Figure 3.1: Das "moralische Gehirn" I: Hirnregionen, die bei *moral decision making* besonders aktiv sein sollen (Abbildung aus [79], die Nummern verweisen auf Tabelle 3.1).

mit einer vermehrten Aktivierung des ventromedialen präfrontalen Kortex (vmPFC), dem linken posterior superior temporal sulcus (pSTS) und dem posterior cingulate cortex einher gehen soll. Geprüft wurde, welche Aktivierungen sich bei einfachen moralischen Stimuli zeigen. Die Autoren kommen zum Schluss, dass die Verarbeitung einfacher moralischer Stimuli mit einer vermehrten Aktivierung des linken pSTS, dem mittleren temporalen Gyrus, den bilateralen temporal poles, dem linken lateralen PFC und dem bilateralen vmPFC einhergeht. Basierend auf diesem Resultat schliessen die Autoren, dass die Aktivierung des pSTS und des vmPFC ein gemeinsames neuronales Korrelat einer moralischen Entscheidung sei. Problematisch bei diesem Vergleich ist jedoch, dass die jeweiligen Aktivierungen unter Verwendung unterschiedlicher Stimuli und unterschiedlicher Baseline-Bedingungen erzielt wurden. Es ist demnach fraglich, ob einfach die Schnittmenge der aktivierten Hirnregionen das gewünschte neuronale Korrelat liefern soll.

- In einer Studie von Singer et al. wurden Gesichter von Personen als moralischer Stimulus verwendet [150]. Das experimentelle Verfahren erfolgte in drei Schritten. Zuerst wurde den Versuchspersonen Fotografien gezeigt. Ihnen wurde gesagt, dass die betreffenden Personen zuvor in einem experimentellen Spiel (Gefangenendilemma) als kooperierende, als nicht-kooperierende oder als diesbezüglich neutrale Personen aufgetreten sein sollen – die Fotografien dienten also als Platzhalter für Personen, welche sich zuvor sozial wünschenswert bzw. inkorrekt verhalten haben. Danach haben die Versuchspersonen in einem zweiten Schritt das mehrfach hintereinander das sequenzielle Gefangenendilemma mit den (fiktiven) Personen gespielt (die Versuchsperson war immer der *first mover*). Die fiktiven Personen haben sich danach wiederholt entweder wie Kooperierende verhalten (d.h. im *second move* Kooperation von A mit Kooperation beantwortet), wie Nicht-Kooperierende (auf Kooperation mit Nicht-Kooperation geantwortet) oder mit Null-Spielen (d.h. neutral). Damit konnten sich die Versuchspersonen gewissermassen über den moralischen Charakter der betreffenden Personen überzeugen (moralisches Lernen). Um den Effekt des moralischen Lernens von durch die Belohnung (*payoff* im Spiel) ausgelöste Effekte zu unterscheiden, wurde den Versuchspersonen in den Spielen zuvor jeweils gesagt, ob Spieler B

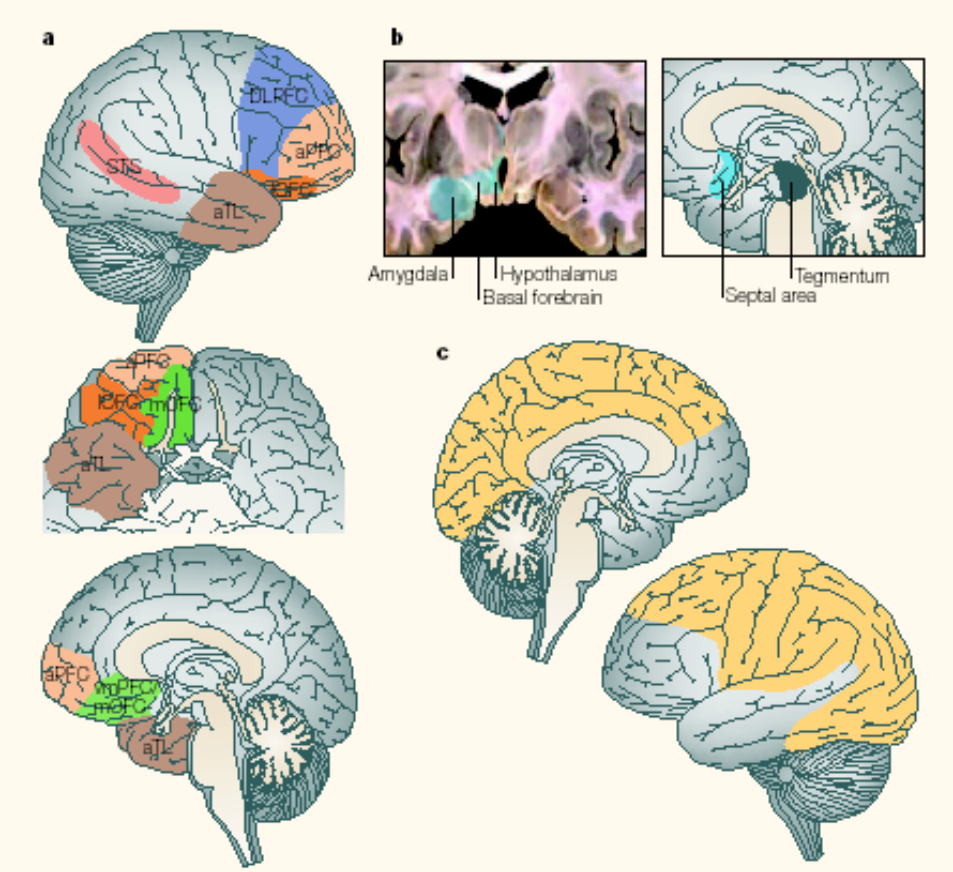

Figure 3.2: Das "moralische Gehirn" II : Hirnregionen, die bei *moral decision making* besonders aktiv sein sollen. a) Kortikale Regionen: antreriorer, medialer, lateraler, dorsolateraler und ventromedialer präfrontaler Kortex (aPFC, mOFC, lOFC, DLPFC, mvPFC), anteriorer *temporal pole* (aTP) und superiorer *temporal sulcus* (STS). b) Subkortikale Strukturen. c) Für folgende Hirnregionen besteht noch Unsicherheit hinsichtlich ihrer Bedeutung in *moral cognition*: Parietal- und Occipitallappen, weitere Bereiche des Frontal- und Temporallappen, Hirnstamm, Basalganglien und weitere subkortikale Strukturen (Abbildung aus [119]).

intentional oder nach einem vorgegebenen Schema handelt. In einem dritten Schritt wurde dann den Personen in einem Scanner die Bilder der verschiedenen Personen, die als Spieler *B* aufgetreten sind, gezeigt und es wurde gemessen, in welchen Regionen eine erhöhte Aktivität aufgetreten ist. Es zeigte sich, dass die stärksten Aktivierungen bei der Präsentation der Bilder der intentional Kooperierenden auftraten (und nicht bei den Nicht-Kooperierenden). Hier zeigte sich insbesondere eine verstärkte Aktivierung in der linken Amygdala, der bilateralen Insula, dem fusiformen Gyrus,

dem superioren temporalen Sulcus und weiteren, *reward-related* Gebieten. Daraus schliessen die Autoren, dass Menschen deshalb für Kooperation empfänglich sind, weil die damit verbundenen (moralischen) Gefühle könnten stärker sind als jene, die bei Nicht-Kooperierenden auftreten.

Brain Region	**Associated moral task**	**Other associated task**	**Social pathology from damage**	**Likely functions**
1. Medial frontal gyrus	Personal moral judgments, Impersonal moral judgments (relative to non-moral), simple moral judgments*, viewing moral pictures, forgivability judgments* (*also lateral frontopolar)	Attributing intentionality to moving shapes and cartoon characters*, theory of mind (ToM) stories and cartoons*, representing a historical figure's mental states*, viewing angry/sad faces, pleasant pictures, negative pictures (with emotional report), reward, viewing and/or recall of happy, sad, and disgusting films, emotional autobiographical recall, emotional planning, 'Rest' *(focus in the paracingulate sulcus)	Poor practical judgment, reactive aggression and (primarily in developmental cases) diminished empathy and social knowledge	Integration of emotion into decision-making and planning, esp. for conscious processes, ToM
2. Posterior cingulate, precuneus, retrosplenial cortex (BA 31/7)	Personal moral judgments, Impersonal moral judgments (relative to non-moral), simple moral judgments, forgivability judgments, moral pictures	Hearing affective autobiographical episodes, threat words, reading coherent stories, esp. ToM stories, viewing ToM cartoons, familiar faces, disgusted faces, sad faces, snake video, previously experienced robbery video, combat pictures (and imagery), sad autobiographical recall (men), recognizing neutral words from negative context, emotional planning, recall of happy personal life episodes, imaginable word pairs, 'Rest'	Impaired recognition memory for faces, capgras delusion?	Integration of emotion, imagery (esp. precuneus), and memory, esp. for coherent social narratives
3. Superior temporal sulcus, inferior parietal lobe (BA 39)	Personal moral judgments, simple moral judgments, moral pictures	Viewing biological motion (hands, faces, eyes, body), sad faces, happy, sad, and disgusting films, ToM cartoons, reading coherent stories with self-perspective and with characters , esp ToM attributing intentionality to moving shapes, representing a historical figure's mental states, recognizing neutral words from negative context, recall of imaginable word pairs, judgment of indoor/outdoor vs. subjective response to (un)pleasant pictures, emotional film viewing vs. recall, 'Rest'	Impaired judgment from eye gaze (monkeys), capgras delusion?	Supporting representations of socially significant movements, and possibly complex representations of 'personhood', ToM
4. Orbitofrontal/ ventromedial frontal cortex (BA 10/11)	Simple moral judgments, moral pictures	Reward/punishment, sad autobiographical recall, recognizing words from positive context, viewing angry faces, 'Rest' (Note: absent in many PET studies of emotion)	Poor practical judgment, reactive aggression and (primarily in developmental cases) diminished empathy and social knowledge, difficulty with advanced ToM tasks	Representation of reward/punishment value, control of inappropriate/disadvantageous behavior, 'hot' ToM
5. Temporal pole (BA 38)	Simple moral judgments	Reading coherent stories (with characters), ToM stories attributing intentionality to moving shapes and cartoon characters, Representing a historical figure's mental states Recall of familiar faces and scenes, hearing affective autobiographical episodes, recognition of emotional pictures, viewing emotional pictures (with subjective report), angry/sad faces, viewing and recall of happy, sad (viewing only), and disgusting film, emotional autobiographical recall	Impaired autobiographical memory	Imparting affective tone to experience and memory, ToM
6. Amygdala	Moral pictures	Recognition of emotional pictures [59] Viewing emotional film, sad faces, viewing racial outgroup faces	Poor social judgment from faces and movement	Rapid assessment of reward/punishment value, esp. visual and negative
7. Dorsolateral prefrontal cortex (BA 9/10/46); 8. Parietal lobe (BA 7/40)	Impersonal moral judgment	Working memory and other 'cognitive' tasks		Working memory and other 'cognitive' functions

Table 3.1: *The moral brain*: Übersicht gemäss Referenz [79].

Basierend auf diesen (und vergleichbaren) Untersuchungen haben Greene und Haidt [79], sowie Moll et al. [119, 120] in Übersichtspapieren eine erste Skizze des "moralischen Gehirns" geliefert – also eine Übersicht über jene Hirnregionen, die in moralischer Kognition involviert sein sollen. Das moralische Gehirn soll demnach insbesondere folgende Hirnregionen umfassen (vgl. mit Tabelle 3.1 und Bilder 3.1, 3.2 – die nachfolgende Aufzählung beinhaltet Regionen, die in beiden Arbeiten genannt werden): a) den medialen frontalen Gyrus. b) Das posteriore Cingulum, den Precuneus und den retrosplenialen Kortex. c) Den superioren temporalen Sulcus und inferioren Parietallappen. d) Den orbitofrontalen/ventromedialen frontalen Kortex. e) Das frontpolare und anteriore Cingulum. f) Den anterioren temporalen Kortex. g) Den temporalen Pol. h) Die Amygdala. i) Den dorsolateralen präfrontalen Kortex und den Parietallappen. j) Die Insula und den Precuneus. k) Den Thalamus. l) Das Mittelhirn. m) Das basale Vorderhirn (*basale forebrain*). Anhand dieser Vielfalt an Hirnregionen, die bei moralischer Kognition beteiligt sein soll, wird deutlich, dass eine eigentliche Lokalisation der moralischen Kognition wie erwartet unsinnig ist. Vielmehr bestehe eine moralische Entscheidung aus einer Vielzahl unterschiedlicher Prozesse affektiver wie kognitiver Art. Obgleich emotionale Aspekte eine wichtige Rolle spielen würden, dürfe die rationale Kognition nicht unterschätzt werden – insbesondere bei unpersönlichen moralischen Werturteilen und bei solchen, wo emotionale und rationale Komponenten in einen Konflikt geraten. Die Frage ist nun, ob die bisherigen Experimente überhaupt eine deutliche Abtrennung zwischen rationalen und emotionalen Komponenten ermöglicht haben.

3.4 Moralisches Handeln

3.4.1 Moralische Pathologien

Eine Möglichkeit, Moral besser zu verstehen, besteht in der Untersuchung von Personen mit abnormen moralischem Verhalten (moralische Pathologien) [79, 119]. Diese Personen zeigen Defizite bei der Wahrnehmung moralrelevanter Fähigkeiten (z.B. Empathie), der moralischen Kognition (Entscheidungsfindung in sozialen Kontexten) und handeln moralisch unangemessen (z.B. unmotivierte Gewaltausbrüche). Gewiss stellt sich bei solchen Forschungen die Frage nach den Abgrenzungskriterien, welche ein gewisses Verhalten als "abnorm" definieren. Andererseits gibt es durchaus klare Fälle von moralischen Pathologien, bei welchen sich Hirnschädigungen und moralische Defizite in Verbindung bringen lassen. Derartige Forschungen sind in jüngerer Zeit insbesondere durch die Arbeiten von Damasio einem breiteren Publikum (neu) bekannt geworden. Diese Forschung fokussiert die Entscheidungsfähigkeit der Betroffenen in moralischen und/oder sozialen Kontexten.

Die Untersuchung moralischer Pathologien bilden einen wichtigen historischen Baustein der Neuropsychologie. Vorab in der US-amerikanischen Literatur wird der Fall des Eisenbahnarbeiters Phineas Gage, bei welchem 1845 durch einen Sprengunfall eine Eisenstange durch den vorderen Schädelbereich getrieben wurde [101]:515-516, als historischer Referenzfall genannt. Die durch die Verletzung des präfrontalen Kortex hervorgerufene Verhaltensänderungen führten zu einem zunehmenden Interesse der Neurologie an moralischen Pathologien und bereits im 19. Jahrhunderts hatten (vorab deutsche) Neurologen breite Kenntnisse über solche Fälle, wie Peter Brugger und Marianne Regard im Gespräch

ausführten. Auch in der modernen Neuropsychologie bildet die Erforschung moralischer bzw. sozialer Pathologien (z.B. Psychopathie) ein wichtiges Thema und es kann dazu kein umfassender Überblick gegeben werden, zumal auch eine Überschneidung mit der zweiten Pilotstudie zu erwarten ist. Nachfolgend sollen in erster Linie ein Forschungsthema genauer untersucht werden: die Unterscheidung der Fähigkeit einer abstrakten moralischen Kognition von einer emotional unterlegten moralischen Kognition. Moralische Pathologien würden sich demnach dadurch auszeichnen, dass nur erstere, nicht aber letztere Fähigkeit genutzt werden könne.

Eine der ersten Studien die Hinweise auf eine solche Unterscheidung lieferte, wurde von Blair durchgeführt [20]. Er suchte nach einem Modell für die Erklärung von Psychopathologie. Dieses basiert auf der in Verhaltensstudien entwickelten Hypothese, dass bei Tiere mit ausgeprägten Sozialstrukturen ein Hemmmechanismus für aggressives Verhalten aktiviert wird, wenn einer von beiden Konflikpartnern Unterwerfungssignale aussendet. Blair postuliert die Existenz eines vergleichbaren Mechanismus (*violence inhibition mechanism*) bei Menschen, welcher bei Psychopathen aber nicht mehr funktionieren soll. Das Modell (das nicht weiter vorgestellt werden soll) sagt voraus, dass Psychopathen die *moral/conventional* Unterscheidung – als das in Menschen offenbar konsistent beobachtete Verhalten, dass moralische Regelverletzungen im Vergleich zu Konventionsverletzungen als gravierender beurteilt werden[3] – nicht zu treffen vermögen. Blair definiert moralische Regeln als solche, welche die Wohlfahrt einer Gemeinschaft betreffen. Konventionen wiederum sind Verhaltenskonstanten, welche zur Strukturierung von sozialen Interaktionen beitragen. Die Studie wurde mit insgesamt zehn Personen durchgeführt, welche allesamt in Gefängnissen einsassen und der Kategorie der '*Psychopatic Disorders*' zugeordnet wurden. Diese Personen wurden einem zweiten psychologischen Test für Psychopathie (Befragung) unterworfen. Es ergab sich hinsichtlich der Testresultate eine klare Zweiteilung in eine nachfolgend "Psychopathen" genannten Gruppe und einer zweiten Gruppe von "Nicht-Psychopathen". In Einzelgesprächen wurden mit den einzelnen Personen verschiedene Szenarien über Gewaltvorkommnisse an Schulen diskutiert. Die Versuchspersonen mussten diese Akte beurteilen und bewerten. Die Auswertung der Antworten ergab drei Schlussfolgerungen. So konnte erstens die Voraussage, dass Psychopathen die *moral/conventional* Unterscheidung nicht zu treffen vermögen, bestätigt werden. Zweitens zeigte sich dieses mangelnde Unterscheidungsvermögen aber nicht derart, dass moralische Regelverletzungen als vergleichbar mit Konventionsverletzungen angesehen wurden, sondern vielmehr umgekehrt. Drittens schliesslich wurden derartige Regelverletzungen von den Psychopathen nicht mit Bezügen auf die konkrete Wohlfahrt des Opfers (z.B. hinsichtlich der Schmerzen von Opfern) beurteilt, sondern mit Hinweis auf abstrakte Normen (Beispielsweise: "It's wrong, it's not socially acceptable"). Diese Studie gibt einen Hinweis darauf, dass moralische Pathologien nicht notwendigerweise mit einer Unfähigkeit der Wahrnehmung einer abstrakten moralischen Kognition einher geht, sondern vielmehr mit einer Unfähigkeit zur Setzung emotionaler Bezüge. Letzteres könnte der Grund sein, warum Psychopathen zwar ihre Handlungen durchaus als moralisch falsch taxieren können, dieses Wissen aber nicht für ihr eigenes Handeln verwenden.

Eine zweite Studie, die Hinweise in diese Richtung ergab, untersuchte das Phänomen der antropomorphen Interpretation natürlicher Vorgänge in der Welt. Ein klassisches Experi-

[3] Andere für diese Pilotstudie untersuchten Arbeiten lassen aber die Vermutung zu, dass diese Unterscheidung nicht in allen Kulturen gleichermassen klar getroffen werden kann [84, 154]

ment dazu wurde 1940 von den deutschen Psychologen HEIDER und SIMMEL durchgeführt. Diese drehten einen Film mit animierten geometrischen Figuren. Die Ereignisse im Film werden vom Betrachter in der Regel als Interaktion dreier verschiedener menschlicher Charaktere angesehen. Mit diesem Film lässt sich demnach die Tendenz erfassen, Vorgänge in der Welt mittels sozialer Analogien zu erfassen. HEBERLEIN und ADOLPHS benutzten dieses Paradigma von HEIDER und SIMMEL für Experimente mit Versuchspersonen mit einer bilateralen Schädigung der Amygdala [88]. Diese Personen beschrieben den Film mit einer Sprache, die sich rein auf die abstrakten geometrischen Vorgänge bezog, ohne auf das Mittel der Anthropomorphisierung zurückzugreifen. Kontrolltests zeigten, dass dies nicht an einer Schädigung der visuellen Wahrnehmung oder an einem generellen Unvermögen liegt, soziale Stimuli zu beschreiben. Gewiss ist die Datenbasis – die Experimente wurden an nur zwei Versuchspersonen durchgeführt – klein. Dies ist eine grundlegende Schwierigkeit bei solchen Untersuchungen, da es in der Regel nur wenige Patienten gibt, welche die gewünschte Läsion aufweisen. Diese Studie gehört aber zu einer ganzen Reihe von Untersuchungen, welche die Beteiligung der Amygdala in einer Vielzahl sozialer Verhaltensweisen nachweisen. Die Autoren vermuten aufgrund ihrer und anderer Ergebnisse, dass die Amygdala an der Verarbeitung grundlegender Emotionen wie auch an komplexeren sozialen Entscheidungen und Werturteilsbildungen beteiligt ist. Diese Verarbeitung erfolgt hingegen auf einer unbewussten Ebene. Fehlt diese durch die Amygdala vermittelte "emotionale Färbung" der Wahrnehmung, so können die Phänomene zwar durchaus abstrakt erfasst und kommentiert werden. Es fehlt offenbar aber ein inneres emotionales Erleben der Wahrnehmung derart, dass dies sich sowohl auf die Art der Beschreibung der Wahrnehmung auswirkt, wie auch auf deren Nutzung für eigenes (moralisches bzw. soziales) Handeln (siehe auch [120]).[4]

Aufsehen hat die Studie von ANDERSON et al. erregt, in welcher über zwei Fälle frühkindlich erworbener Schäden im präfrontalen Kortex berichtet wird [7]. Beide Fälle stammen aus unauffälligen Mittelklasse-Familien, deren Familiengeschichte nicht auf das Vorhandensein psychischer oder neurologischer Probleme hinweist. Beide Patienten bestanden Tests für die Prüfung intellektueller Fähigkeiten problemlos, scheiterten aber in Tests, welche die Beurteilung sozialer Dilemmas und angemessene Reaktionen auf soziale Verhaltensweisen prüften. Im Gegensatz zu Patienten mit später erworbenen präfrontalen Schäden, konnten diese auch in den Test für die Einordnung der Personen in das Entwicklungsschema von KOHLBERG nicht bestehen und wurden der *preconventional phase* zugeordnet. Auch die Anamnese zeigt, dass bei den betreffenden Personen das Ausmass des asozialen Verhaltens grösser ist und verschiedene Lernprogramme zur Erlangung von korrektem sozialen Verhalten scheiterten. Die Autoren interpretierten diese Resultate so, dass ein Zeitfenster für die Entwicklung moralischer Kognition vorhanden ist. Auch andere Autoren stützen diese Vermutung. DOLAN vermutet, dass frühkindlich erworbene Hirnschäden es den Betroffenen verunmöglicht, für Verhalten nutzbares Wissen über akzeptierte moralische Standards zu erwerben [50]. Erwachsene mit Schäden im präfrontalen Kortex zeigen durchaus auch, in unterschiedlichem Ausmass, Anzeichen von abnormem sozialen bzw. moralischen Verhalten [43]. Dennoch vermögen diese Personen im KOHLBERG-Paradigma (z.B. unter Verwendung des HEINZ-Dilemmas[5]) zu bestehen. Dies wird so interpretiert, dass Menschen mit vorher

[4] Im Kontext der experimentellen Spieltheorie wird dieser Sachverhalt so beschrieben, dass die Betroffenen die Fähigkeit verlieren, Normen zu internalisieren [72]:2

[5] Im HEINZ-Dilemma wird der Fall beschrieben, wonach der Ehemann einer an Krebs erkrankten Frau vor der Frage steht, einem Apotheker ein überteuertes Medikament zu stehlen, weil der Mann dieses nicht

normalen Gehirnen zwar neue Strategien für die Bewältigung sozialer und moralischer Probleme entwickeln können, diese aber nicht in Echtzeit (also im Alltag) anwenden können.

Diese Studien scheinen darauf hinzuweisen, dass relativ klar abgrenzbare Hirnregionen für die Fähigkeit zur moralischen Kognition entscheidend sind. In den Gespräche mit Fachpersonen der Universität Zürich wurde aber zu einer vorsichtigen Interpretation dieser Resultate geraten. Anton Valavanis erklärte im Gespräch aufgrund seiner klinischen Erfahrungen, dass Personen nach Hirntumor-Operationen und der dabei notwendig gewordenen Schädigung bestimmter Hirnregionen zunächst durchaus die mit der Schädigung dieser Regionen assoziierten Abnormitäten im Sozialverhalten zeigen. Später jedoch verschwinden diese Abnormitäten wieder fast vollständig, was Ausdruck der hohen Plastizität des Gehirns ist. Peter Brugger und Marianne Regard von der neuropsychologischen Abteilung des Universitätsspitals Zürich betonen im Gespräch das Problem, wie moralisches Verhalten operationalisiert werden kann, um moralische Pathologien angemessen beschreiben zu können. Sie betrachten solche Pathologien als eine Form von "Mini-Autismus" (was hier nicht weiter erläutert werden soll).

3.4.2 Vorformen von moralischem Verhalten

Die Untersuchung von Vorformen von Moral in sozial organisierten Tierpopulationen fällt in den Bereich der Verhaltensforschung und Primatologie. Solche Vorformen können letztlich nur anhand der beobachtbaren Interaktionen der Tiere ermittelt und untersucht werden. Damit stellt sich in diesem Bereich das bekannte Problem der Verhaltensforschung: die Gefahr einer Anthropomorphisierung von Beobachtungen. Die Frage nach Vorformen von Moral kann unter mehreren Perspektiven untersucht werden: So kann gefragt werden, ob sich im Verhaltensreperoire gewisser Tiere bestimmte Verhaltensweisen finden lassen, die als "moralisch" bzw. als eine Vorform von moralischem Verhalten charakterisiert werden können. Von einer evolutionären Perspektive her kann gefragt werden, wie sich diese Verhaltensweise auf einer evolutionären Zeitskala entwickelt haben könnten. Dazu können auch Modelle (evolutionäre Spieltheorie, Agenten-basierte Modellierung) verwendet werden. In diesem Zusammenhang wird auch spekuliert, dass die sich in der Evolution entwickelnden Regeln eine Art "moralische Charakteristik der Natur" widerspiegeln sollen [96]:323. Dies soll nicht als Begründung für die Gültigkeit dieser Regeln gelten; das Argument zielt eher in Richtung der evolutionären Erkenntnistheorie, wonach die Tatsache der Evolution moralisch geprägter Regeln etwas über die Natur der Realität aussagen. Diese Gedanken weisen auf die Tendenz einer Naturalisierung der Moral hin, die nicht weiter besprochen werden soll. Im nachfolgenden wird lediglich auf den ersten Aspekt eingegangen.

Generell besteht im Bereich der Primatologie durchaus die Ansicht, dass Tiere Fähigkeiten besitzen, die auf das Vorhandensein gewisser Vorformen von Moral hinweisen könnten. Frans de Waal identifiziert eine Reihe von Verhaltensweisen, welche dies dokumentieren sollen [49, 69]: Das Teilen von Futter, Versöhnungsverhalten nach Konflikten, Interventionen Dritter bei Konflikten, Schlichtungsverhalten und Strafverhalten (eine ausführliche Beschreibung solcher Verhaltensweisen findet sich im 1996 erschienenen Buch *Good Natures. The Origins of Right and Wrong in Humans and Other Animals*). Basierend auf diesen Untersuchungen haben Flack und de Waal vier Aspekte von Moralität (*ingredients of morality*,

bezahlen kann. Die Art der Antworten gibt Hinweise auf den Grad der Moralentwicklung gemäss dem Stufenmodell von Kohlberg [103].

[69]:3) bei Tieren (und Menschen) identifiziert: Erstens, die Fähigkeit zur sympathetischem Verhalten, ausgedrückt in sozialen Bindungen zwischen Tieren einer Population, einer Bereitschaft, verletzten oder behinderten Mitgliedern der Gruppe zu helfen und empathisches Verhalten. Zweitens, die Fähigkeit zur regelbezogenem Verhalten. Dies drückt sich aus im Vorhandensein gewisser sozialer Regeln, insbesondere von Strafverhalten, und gewissen Erwartungen, wie andere mit einem umgehen. Drittens, eine gewisse Reziprozität, also einem System von Austauschverhältnissen und gewissen Formen von "Rache". Viertens schliesslich ein Komplex von Verhaltensweisen, welche der Stabilität einer Gruppe dienen. Dazu gehören insbesondere die verschiedenen Verhalten für die Aufrechterhaltung von friedlichem Verhalten und für das Lösen von Konflikten. Gewiss lassen sich bei all diesen Aspekten eine Reihe kritischer Fragen stellen (siehe dazu die Kommentare in [69]:31-65). Diese zielen in den meisten Fällen in die Richtung einer fehlenden Intentionalität dieser Verhaltensweisen, d.h. die beschriebenen Verhaltensweisen treten bei den Primaten-Gruppen zwar auf, doch die Mitglieder würden über diese nicht verfügen. Im Gespräch wies Carel van Schaik zudem darauf hin, dass die Hypothesen von de Waal in den meisten Fällen lediglich auf Beobachtungen basieren, nicht aber auf Verhaltensexperimenten.

Die Erforschung von Kooperation in Tieren ist ebenfalls ein Gebiet mit einer langen Tradition. Bereits Charles Darwin hat sich mit diesem Problem beschäftigt und es ist heute unbestritten, dass viele soziale Tierarten (Primaten, Fledermäusen, Ameisen etc.) Formen von Kooperation (z.B. gemeinsames Jagen, Teilen von Nahrungsmitteln) ausüben. Für die Erklärung der Entstehen von Kooperation existieren im wesentlichen vier theoretische Ansätze [55]: Kooperation aufgrund Verwandschaftsbeziehungen (*kin selection*), aufgrund der Etablierung von Austauschverhältnissen (*reciprocity*), aufgrund so genanntem *byproduct mutualism* und aufgrund Gruppenselektion. Zu all diesen Konzepten existiert eine Vielzahl von Literatur, die im Detail nicht vorgestellt werden kann. Hier soll lediglich eine kurze Darstellung dieser vier Konzepte erfolgen (basierend auf [54]). Das bekannteste Modell für die Erklärung von Kooperation ist *kin selection*. Kooperation erklärt sich hier aus der Tatsache enger Verwandschaftsbeziehungen zwischen den beteiligten Partnern. Die kognitive Voraussetzung für *kin selection* ist die Fähigkeit, Verwandte als solche erkennen zu können. Im Konzept der *reciprocity* wird der Fokus auf Gegenseitigkeit gelegt: Wenn Tier A mit Tier B kooperiert, so könne A erwarten, dass B künftig auch mit ihm kooperiert. Dieses Konzept verlangt gewisse kognitive Voraussetzungen. Insbesondere müssen die Tiere ein Gedächtnis für den Akt der Kooperation derart haben, dass Kooperation von beiden beteiligten Partnern als positives und deshalb künftig zu wiederholendes Verhalten charakterisiert wird. Da dieses System betrugsanfällig ist, müssen die beteiligten Partner auch die Möglichkeit haben, allfällige Nicht-Kooperierende zu bestrafen. Diese Bedeutung des Strafens für das Modell der *reciprocity* ist sowohl für Tierstudien [55]:472 wie auch für experimentelle Verhaltensstudien mit Menschen (siehe Abschnitt 2.4.7) betont worden. Im Konzept des *by-product mutualism* wird Kooperation als eine Art Epiphänomen aufgefasst. So kann die Aktivität von einem Tier (z.B. das Graben einer Höhle als Ruhestätte für die Nacht) von einem anderen Tier (z.B. als Ruhestätte für den Tag, wenn der Ersteller der Höhle diese nicht bewohnt) genutzt werden. Hier wird bestritten, ob man dies überhaupt Kooperation nennen will. Es wird vermutet, dass sich solche Formen von Kooperation insbesondere in einer lebensfeindlichen Umwelt ausbildet, in der sich die Aktivitäten einzelner Tiere derart verzahnen, dass sie für das gegenseitige Überleben nützlich sind. Dieses theoretische Konzept braucht keine kognitive Voraussetzungen wie Gedächtnis früherer Interaktionen oder das Erkennen von Ver-

wandschaftsbeziehungen. Das Konzept der Gruppenselektion schliesslich wird heutzutage sehr kritisch beurteilt und soll nicht weiter vorgestellt werden. Studien über Kooperation in Tieren sind, vorab unter Feldbedingungen, mit einer Reihe von Schwierigkeiten konfrontiert. Die gemessenen Grössen sind in der Regel der Fortpflanzungserfolg oder die Menge (bzw. der Energiewert) der aufgenommenen Nahrung. Es ist aber nicht einfach, eine kausale Beziehung zwischen kooperativen Akten und diesen Grössen herzustellen. Zudem ist unklar, welch andere Formen eines direkten *reward* kooperierende Tiere haben könnten [17].

Die Charakterisierung von tierischem Veralten als "moralisch" stösst auf eine Reihe von Schwierigkeiten. So finden sich zwar immer wieder Medienberichte, wonach sich Tiere quasi "moralisch" (auch gegenüber Menschen) verhalten haben sollen und diese Berichte stossen meist auf ein grosses Interesse. In populären Berichten wie wissenschaftlichen Studien werden diese Verhaltensweise mit Begriffen wie *empathetic, sympathetic, compassionate, trustworthy, shameful, vengeful, conciliatory, friendship, loyalty, justice* und *fairness* umschrieben. Auch Forscher, welche neurobiologische Grundlagen der Moral in Menschen untersuchen, verweisen auf Studien, welche die Existenz von Vorformen von Moral in Tieren behaupten (siehe z.B. [120]:299). Wissenschaftler im Bereich der Verhaltensforschung wiederum verweisen darauf, dass die chemischen Systeme für Emotionen bei Menschen wie Tiere ähnlich sein sollen und es demnach plausibel sei, dass zumindest Vorformen von moralischem Verhalten auch bei gewissen Tieren gefunden werden könnten [17].

Die Frage ist, ob die Charakterisierung dieses Verhaltens als "moralisch" gerechtfertigt ist, wie Gruen festhält [82]. Dazu müsse der Begriff der *moral agency* genauer umschrieben werden. Gemäss Gruen umfasse dieser mindestens die Fähigkeiten, moralische Empfindungen zu haben, sich in moralische Bewertungen einlassen zu können, moralische Entscheidungen zu treffen und moralische Handlungen vollziehen zu können. Abgesehen vom Problem, dass diese Charakterisierung das Definitionsproblem auf den Begriff "moralisch" abschiebt, sei gemäss Gruen unbestritten, dass gewisse Eigenschaften des Personseins notwendig sind, um als *moral agent* zu gelten. Sie nennt dazu eine Reihe von Fähigkeiten: Erstens die Fähigkeit, Zustände von *beliefs* und Bedürfnisse zu haben und sich ihrer bewusst zu sein. Zweitens müssen diese Zustände durch äussere Informationen geändert werden können. Drittens muss der *moral agent* in der Lage sein, Konflikte hinsichtlich Wissen und Wünschen zu erkennen, nach Lösungen zu suchen und entsprechend zu handeln. Viertens muss der *moral agent* auch gewisse Fähigkeiten hinsichtlich Voraussage und Planung von Handlungen haben. Im Gespräch mit dem Anthropologen Carel van Schaik von der Universität Zürich ergab sich dazu, dass diese Kriterien zumindest bei einigen Tieren tatsächlich erfüllt sein könnten, wobei sich aber das Problem stellt, dass deren Vorhandensein möglicherweise empirisch gar nicht nachgeprüft werden kann (z.B. wie kann man herausfinden, welche Bedürfnisse gewissen Tieren bewusst sind). Auch nannte er *ingroup/outgroup*-Verhalten als eine weitere mögliche Voraussetzung von moralischem Verhalten. Eine Population von Tieren muss demnach in der Lage sein, alle Mitglieder in einem gewissen Sinn als "gleich" anzusehen. In seinem Forschungsgebiet – der Erforschung von Orang Utangs – ist ein solches *ingroup/outgroup*-Verhalten nicht festzustellen. Orang Utangs zeigen auch keine Form von Kooperation, was auf einen gewissen Zusammenhang zwischen *ingroup/outgrup*-Verhalten und dem Auftreten von Kooperation hinweist.

Gruen bezweifelt, dass der Begriff des *moral agent* bereits genügend ausdifferenziert ist, um Vorformen von Moral in Tieren ausreichend präzise definieren zu können. Auch Bekoff hält in diesem Zusammenhang fest, dass es schlicht an genügend Daten mangelt, um eine

Taxonomie der verschiedenen notwendigen kognitiven und emotionalen Fähigkeiten eines *moral agent* zu erstellen [17]:511. Dabei stellt sich auch die Frage, welche Verhaltensweisen nebst Kooperation sich für eine Untersuchung möglicher Vorformen von Moral bei Tieren eignen könnten. Bekoff nennt das Spielverhalten von Tieren als möglichen Kandidaten [17]. Denkbar ist künftig auch, dass nebst Verhaltensstudien auch direkt Messungen in Gehirnen von sozial interagierenden Tieren gemacht werden (z.B. mit chronisch implantierten Elektroden). Hier stellt sich ein klassisches Problem der Tierethik aber verstärkt: Je komplexer das Phänomen ist, das man untersuchen will – und die Untersuchung von Vorformen von Moral gehört zweifellos in diese Kategorie –, desto ethisch fragwürdiger wird der Eingriff, da man solchen Tieren ja einen gewissen Personenstatus zubilligen müsste.

3.5 Zur Rolle von Begründungen

In der philosophischen Ethik zeichnen sich moralische Handlungen dadurch aus, dass sie begründet werden können. Diese Begründungen fundieren auf Theorien der normativen Ethik. Es stellt sich demnach die Frage, inwieweit dieses Verständnis moralischer Handlungen mit den tatsächlichen, als "moralisch" taxierten Handlungen übereinstimmt. Dieses Problem hat mehrere Facetten. Zum einen ist natürlich seit langem bekannt, dass eine Kluft zwischen einer idealen, d.h. auf guten und konsistenten Gründen beruhenden, moralischen Handlungen und tatsächlichen moralischen Handlungen besteht. Letztere können auf falsche oder inkonsistente Begründungen beruhen. Eine Theorie über die biologischen Grundlagen moralischer Handlungen kann dazu beitragen, diese Diskrepanz zu verstehen, ohne dass damit das philosophische Ideal einer moralischen Handlung aufgegeben werden muss. Gerade für Entscheidungsfindungsprozesse auf der Ebene von Institutionen könnte solches Wissen bedeutsam sein. Zum anderen kann aber die Bedeutsamkeit von Begründungen generell angegriffen werden, indem man diesen beispielsweise den Charakter nachträglicher Rechtfertigungen verleiht. Die Rolle von Begründungen moralischer Handlungen sind dabei in zweierlei Hinsicht Gegenstand empirischer Untersuchungen: Zum einen muss geprüft werden, wie Begründungen mit der Hypothese einer (zumindest teilweise) automatisiert verlaufenden moralischen Kognition vereinbar sind. Zum anderen ist die These zu prüfen, inwieweit Begründungen tatsächlich als *post fakto* Ereignisse zu werten sind, die moralische Handlungen nicht kausal verursachen, sondern erst nachträglich für deren Rechtfertigung konstruiert werden.

Wie Bargh und Chartrand feststellen, beinhaltet der Automatisierungsprozess in (moralischen) Entscheidungen mindestens drei Komponenten [9]: Einen unbewussten Einfluss von Wahrnehmungen auf Handlungen, ein automatisiertes Streben nach gewissen Zielen und eine kontinuierliche Evaluation der eigenen Erfahrungen. Sie sehen eine wichtige Rolle des *Imaging* darin, nach neuronalen Aktivierungen zu suchen, welche diese automatisch ablaufenden Prozesse repräsentieren könnten. Die oben geschilderte Suche nach neuronalen Korrelaten für bestimmte Aspekte moralischer Handlungen fällt genau in dieses Programm. Die Frage ist nur, wie man gemessene Aktivierungen als verursacht durch automatische bzw. bewusste Prozesse unterscheiden kann? Der Zustand des Bewusstseins ist ja selbst für dessen Träger nie genau definierbar und begrifflich mitteilbar, so dass in einem Experiment die gewünschte Zuordnung vollzogen werden kann. Hier besteht die Gefahr, dass sich das Argument gewissermassen in den Schwanz beisst, bzw. dass man von vornherein davon

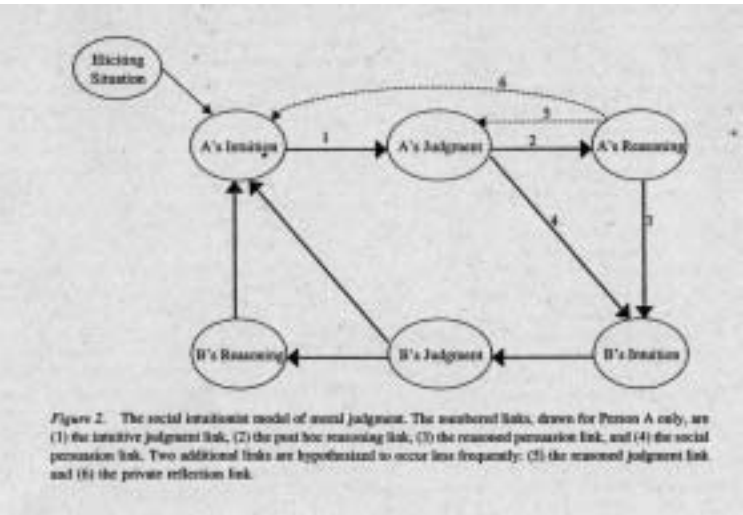

Figure 2. The social intuitionist model of moral judgment. The numbered links, drawn for Person A only, are (1) the intuitive judgment link, (2) the post hoc reasoning link, (3) the reasoned persuasion link, and (4) the social persuasion link. Two additional links are hypothesized to occur less frequently: (5) the reasoned judgment link and (6) the private reflection link.

Figure 3.3: Das *social intuitionist* Modell von HAIDT (Abbildung aus [83]

ausgeht, dass eine erhöhte Aktivierung in einem bestimmten Gebiet als bewusster oder eben automatischer Prozess zu gelten habe. Ein Beispiel für dieses Problem bildet die Studie von WINSTON et al., in welcher das Verhältnis zwischen automatisierten und intentionalen Aspekten von Wahrnehmungen untersucht wird [165]. Versuchspersonen müssen hier im Scanner die Vertrauenswürdigkeit von Gesichtern beurteilen. Die Autoren postulieren, dass dieser Prozess in eine automatische Komponente, basierend auf einer erhöhten Aktivität der Amygdala, und eine intentionalen Komponente, basierend auf einer erhöhten Aktivität des superioren temporalen Sulcus (STS) unterteilt ist. Tatsächlich finden die Autoren dann, dass bei den Versuchspersonen beide Regionen aktiv sind. Doch lässt sich aus dieser Tatsache die postulierte funktionale Dissotiation wirklich nachweisen?

Im folgenden soll nun noch ein Modell des Moralpsychologen HAIDT vorgestellt werden, das die Bedeutung automatisierter Komponenten bei moralischen Entscheidung betont und Begründungen als *post fakto* Ereignisse charakterisiert [83]. Dieses Modell wird in der Literatur über die neurobiologische Grundlagen von Moral oft zitiert (vgl. mit Abbildung 3.3). HAIDT greift in seinem Modell kognitivistische Varianten des *moral decision making*, basierend etwa auf KOHLBERG [103], an. Demnach seien moralische Entscheidungen keine Folge von *moral reasoning*, vielmehr ist letzteres ein nachträgliches Konstrukt, um die getroffene Entscheidung zu rechtfertigen. *Moral reasoning* definiert er dabei wie folgt: "Moral reasoning can now be defined as conscious mental activity that consits of transforming given information about people in order to reach a moral judgment [83]:818. Als Gegenmodell zur kognitivistischen Variante präsentiert er sein *social intuitionist model*, welches der *moral cognition* einen geringeren Stellenwert einräumt, dafür aber kulturelle, soziale und emotionale Komponenten betont. Moralische Entscheidungen sind seiner Ansicht nach die Folge schneller, weitgehend automatisierter Evaluationen (auch Intuitionen genannt). Er definiert moralische Entscheidungen wie folgt: "Moral judgments are (...) defined as eval-

uations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture" [83]:817. In diesen moralischen Entscheidungen spielen Intuitionen eine bedeutende Rolle. Definiert werden diese wie folgt: "Moral intuition can be defined as the sudden appearance in consciousness of a moral judgement, including an affective valence (good–bad, like–dislike) without any conscious awareness of having gone through steps of searching, weighting evidence, or inferring a conclusion" [83]:818. HAIDT nennt eine Art *protomorality* in Primaten als möglichen Ursprung solcher Intuitionen, welche aber in menschlichen Gesellschaften durchaus eine kulturelle Prägung aufweisen. Damit jedoch eine solche Prägung möglich wird, müssen sie externalisiert werden. HAIDT spekuliert, dass diese Externalisierung ein wesentlicher Aspekt der sozialen Entwicklung eines Kindes sein. Im Prozess dieser Externalisierung können einzelne solche Intuitionen auch verloren gehen, was individuelle verschiedene Moralsysteme erklären könnte.

Das Modell von HAIDT zweifelt insbesondere die kausale Rolle der moralischen Vernunft – also die Bedeutung von rationalen Begründungen – beim alltäglichen moralischen Handeln an. Dafür führt er mehrere Gründe an: Erstens hätten die Sozialpsychologie wie die Kognitionswissenschaften gezeigt, dass automatisierte Evaluationen in vielen Entscheidungsprozessen eine Rolle spielen würden. Es sei demnach plausibel anzunehmen, dass dies auch bei moralischen Entscheiden der Fall sei. Zweitens agiere die moralische Vernunft – so das Bild von HAIDT – mehr wie ein Anwalt, der seinen Kunden verteidigt, als ein Wissenschaftler, der die Wahrheit suche. Dies zeige sich anhand der Motive und Kriterien, an denen sich der *reasoning process* orientiere: Man orientiert sich in seinen Urteilen oft an seinem Umfeld und man verteidige Entscheide, die im Einklang mit dem Selbstbild stünden. Würden zudem Begründungen für moralische Handlungen mit psychologischen Methoden geprüft, so zeigten sich regelmässig Schwächen. So betonen die Befragten unwichtige Aspekte und vergessen wichtige Überlegungen. Schliesslich, so HAIDT unter Verweis auf die Arbeiten von DAMASIO, würden moralische Handlungen stärker mit moralischen Emotionen als mit *moral reasoning* co-variieren. All diese Punkte liessen darauf schliessen, dass in alltäglichen moralischen Handlungen von Einzelpersonen nicht Begründungen, sondern primär von moralischen Emotionen gefärbte Intuitionen diese Handlungen kausal verursachen.

Zu den Überlegungen von HAIDT stellen sich eine Reihe von Fragen. So ist zum ersten nicht ganz klar, was mit "alltäglichen" moralischen Handlungen gemeint ist. So lassen sich in sozialen Interaktionen sicher viele Handlungen als moralisch charakterisieren, die in der Tat nicht Resultat eines ausgefeilten *moral reasoning* sind und in diesen Kontexten auch nicht begründet werden – ausser man fragt explizit nach einer Begründung nachdem die Handlung vollzogen wurde. Dass diese Begründungen dann konstruiert erscheinen, überrascht dann aber nicht. Gewiss enthält das Modell durchaus plausible und wichtige Elemente. Dennoch stellt sich die Frage, ob es eben nicht doch wichtige moralische Entscheide gibt, bei welchen durchaus Begründungen den Ausschlag für bestimmte Handlungen geben. In diesen Fragen dürfte das Zusammenspiel automatisierter und intentionaler Prozesse weit komplexer sein, als das Modell suggeriert. Schliesslich findet das Modell lediglich Anwendung auf die moralischen Handlungen auf der Stufe einzelner moralischer Agenten – nicht aber auf der Stufe von Entscheidungsfindungen in einem institutionellen Rahmen. Die interessante Frage ist nun, wie – unter der Voraussetzung der Korrektheit des Modells von HAIDT – Entscheidungsprozesse auf dieser Stufe stattfinden.

Chapter 4

Neuroethik

4.1 Zum Begriff der Neuroethik

Angesichts des anhaltenden Booms der Neurowissenschaft stellte sich in jüngerer Zeit vermehrt die Frage, welche ethischen Probleme aus ihren Anwendungen erwachsen könnten. Diese Probleme lassen sich in drei Gruppen ordnen [23]: Zum ersten fallen einige in den Bereich der klassischen Bio- und Medizinethik und dürften damit mit den dort entwickelten Methoden und normativen Theorien untersucht werden. Die Bandbreite der in diesen Bereich fallenden Probleme ist breit [52, 60, 144]: Es stellen sich praktische Fragen der Forschung wie Studiendesign und Datenschutz, medizinethische Fragen hinsichtlich des *informed consent* von hirngeschädigten Personen, der Transplantation von Hirngewebe und des Hirntod-Konzeptes, bioethische Fragen im Hinblick auf neuronale Stammzellen und schliesslich Fragen der Auswirkung von Erkenntnissen der Neurowissenschaft auf Recht (z.B. hinsichtlich des Begriffs der rechtlichen Verantwortung) und Moral (z.B. hinsichtlich tierethischer Fragen auf Folge über zunehmendes Wissen von Bewusstsein bzw. Schmerz bei Primaten [143]).

Zweitens dürften aber auch eine Reihe neuartiger Probleme auftauchen, die teilweise einen Bezug zu anderen philosophischen Fragen (wie etwa Willensfreiheit) haben. Diese Probleme gruppieren sich um eine Art Hirnzentrismus, wonach Interventionen auf pädagogischer, sozialer und rechtlicher Ebene zu Interventionen auf der Ebene des Gehirns umgedeutet werden könnten. Hier stellt sich die Frage, was ein "normales" Gehirn sein soll und wie dieses in struktureller und funktioneller Hinsicht identifiziert werden soll. Derartige Diskussionen dürften praktische Auswirkungen auf versicherungstechnische Fragen haben, vergleichbar mit den Diskussionen über die Ergebnisse von Gentests. Denkbar wäre beispielsweise, dass Arbeitgeber für gewisse Aufgaben (etwa für den Transport gefährlicher Güter) einen Hirnscan verlangen um sicher zu stellen, das die betreffende Person kein abnormes Gehirn hat. Fragen stellen sich hier aber auch hinsichtlich von Interventionen mit der Absicht einer Verbesserung der Hirnleistung (*neuro enhancement*), was vorab mit chemischen Substanzen angestrebt werden dürfte. Im Rahmen dieses Hirnzentrismus könnten schliesslich auch Konzepte wie Autonomie und Biografie eine Neubeurteilung erfahren.

Drittens schliesslich könnten aus neurowissenschaftlichen Erkenntnissen auch neue technologische Anwendungen resultieren, die mit den Mitteln der Technikfolgenabschätzung un-

tersucht werden müssten. Dies betrifft zum einen die Entwicklung von autonomen technischen Systemen, die biologische Eigenschaften wie Selbstreparatur und Lernen aufweisen könnten. Hier stellen sich Fragen der Haftpflicht bis hin zu einer möglichen Zuschreibung von Verantwortung, welche den praktischen Einsatz solcher Systeme mit sich bringen kann [39]. Zum anderen betrifft dies technische Möglichkeiten, die sich an der Cyborg-Idee orientieren – also Formen einer "Verschmelzung" von biologischen und technischen Systemen, wie sie beispielsweise in der Neuroprothetik und der Entwicklung so genannter *brain-machine-interfaces* angestrebt werden [40]. Auch hier können unerwartete ethische Probleme auftreten, wie das Beispiel der Cochlea Implantate und die daraus folgenden Diskussion über die Beeinträchtigung der Gehörlosen-Kultur zeigt [38]. Denkbar ist auch, dass dereinst ein technologisches *neuro enhancement* angestrebt wird, beispielsweise für militärische Anwendungen. Solche Ideen sind nicht neu. So hat beispielsweise die deutsche Luftwaffe während des zweiten Weltkriegs die Anwendung von EEG für die verbesserte Steuerung von Flugzeugen untersucht [24]. Auch heute investiert das amerikanische DARPA Programm beträchtliche Summen in Neurotechnologie [94].

Diese Vielfalt an Problemen hat die Aufmerksamkeit der praktischen Ethik gewonnen. Bereits 1997 finanzierte die EU Kommission ein Projekt "Ethical, legal, and social aspects of brain research". Doch erst ab dem Jahr 2002 fanden eine Reihe von Konferenzen zu Fragen der Neuroethik statt, welche dieses Thema breiter bekannt machten. Im Januar 2002 finanzierte die *American Association for the Advancement of Science* und die Zeitschrift *Neuron* gemeinsam ein Treffen zum Thema "Die neuralen Grundlagen komplexer Verhaltensweisen verstehen: Implikationen für Wissenschaft und Gesellschaft". Im Februar 2002 führten das *University of Pennsylvania Center for Bioethics* und das *Penn's Center für Neuroscience* eine gemeinsame Konferenz unter dem Titel "Bioethik und die Revolution der kognitiven Neurowissenschaften" durch. Ihr folgte im März 2002 ein Meeting an der *Royal Institution* in London über die Zukunft der Neurowissenschaften und damit zusammenhängende ethische Fragen. Im Mai 2002 schliesslich finanzierte die *Dana Foundation* eine von der Universität Stanford und der Universität von Kalifornien (San Francisco) organisierte Konferenz mit dem Titel: *Neuroethics: mapping the field*. Diese Konferenzen fanden auch Widerhall in Publikumsmedien (siehe z.B. die Beiträge in der Zeitschrift *The Economist* [156]). Im folgenden sollen die wichtigsten derzeit diskutierten Probleme der Neuroethik kurz dargestellt werden.

4.2 Interpretation und Schutz von *Imaging*-Daten

Generell wird heute das Imaging als Hauptproblemfeld der Neuroethik angesehen [98], was sich auch in entsprechenden Berichten in der Wissenschaftspresse niedergeschlagen hat [130, 35]. Bereits in diesem Bereich lassen sich eine Vielzahl von Problemen identifizieren (die forschungsethischen Fragen werden in Abschnitt 4.4 thematisiert):

- Generell stützt der anhaltende Imaging-Boom die Idee eines *brainotyping* (analog zum *genotyping*) – also die Suche nach neuronalen Korrelaten von Verhaltens- und Charaktereigenschaften [60]. Angesichts der Vielzahl methodischer Probleme des Imaging (siehe Abschnitt 2.2.3) besteht die Gefahr einer grossen Diskrepanz zwischen der propagierten und der tatsächlichen Aussagekraft solcher Resultate. Dazu kommt

die Problematik, dass man geneigt ist, diese Bilder zu manipulieren, um ihre "Aussagekraft" weiter zu erhöhen. Gerade bei stark kulturell geprägten Verhaltensweisen wie Sexualität bzw. sexuelle Ausrichtung besteht die Gefahr einer Simplifizierung, zumal, wie WOLPE festhält: "History has shown us again and again that society tends to use science to reinforce the moral assumptions and biases of the cultural moment'" [166]:1032

- Diskutiert wird in diesem Zusammenhang die Idee eines *forensic neuroimaging* (beispielsweise bei Kindern) mit dem Ziel, Gewaltverhalten und Psychopathie frühzeitig zu erkennen [30]:420. Je nach Resultat könnten daraus unterschiedliche therapeutische Ansätze resultieren (so könnte gewalttätiges Verhalten aufgrund eines Amygdala-Schaden anders therapiert werden als solches Verhalten aufgrund einer Abnormität im frontalen Kortex). Fraglich ist hier, mit welcher Sicherheit diese Technologie derartige Abnormitäten im Verhalten nachweisen kann und inwiefern eine naive Anwendung von *Imaging* zu einem übertriebenen und unethischen Gebrauch dieser Technik führen kann.

- Die Gefahr einer naiven Interpretation von Imaging-Daten besteht insbesondere dort, wo Geschworene (Laien) im Rahmen eines Gerichtsprozesses solche Bilder beurteilen müssen. Dieses Problem ist durchaus relevant, da vermutet werden muss, dass Resultate von *Imaging* vermehrt vor Gericht verwendet werden, um die Unfähigkeit der Verhaltenskontrolle zu belegen, weil dies einfacher und statistisch härter ist als ein Verhaltenstest [30].

Auch die im Rahmen dieser Studie befragten Experten teilen die Ansicht, dass die Fehlinterpretation von *Imaging*-Daten ein Hauptproblem der Neuroethik darstellt (siehe Abschnitt 2.2.3). Der Neuropsychologie LUTZ JÄNCKE vom Psychologisches Institut der Universität Zürich äussert im Gespräch grosse Bedenken hinsichtlich des Aufkommens von Bindestrich-Wissenschaften mit dem Präfix "Neuro" wie beispielsweise Neuro-Marketing, welche unzulässige Schlussfolgerungen aus *Imaging*-Daten ziehen würden. Diese methodischen Probleme erhalten Bedeutung in der Diskussion um die gesellschaftliche Relevanz der Resultate die durch *Imaging* gewonnen werden. Eine Untersuchung in der englischsprachigen Presse über die Art und Weise,wie über fMRI berichtet wird (Zeitraum: 1991-2004), zeigt interessanterweise, dass diese Medien gegenüber der Aussagekraft und ethischen Relevanz der Resultate im Allgemeinen kritischer eingestellt sind als die Fachpresse [140]. Insbesondere fehlt in der Fachpresse oft ein Eingehen auf die ethischen Probleme, welche durch diese Resultate aufgeworfen werden.

4.3 Eingriffe in das Gehirn

Ein zweiter Themenkomplex betrifft die Möglichkeit, durch Eingriffe in das Gehirn dessen Funktionsweise gezielt zu verändern. Dies kann therapeutische Ziele haben oder auch mit dem Ziel einer Verbesserung der Hirnleistung verbunden sein. Folgende Bereiche sind hier von Interesse [59]: Neuropharmakologie, Neurochirurgie und der Bau technischer Systeme für das Erfassen von Hirnprozessen bzw. für die Stimulation von Hirngewebe:

- *Neuro enhancement* mittels chemischer Substanzen ist nebst *Imaging* das am meisten diskutierte neuroethische Problem [27, 59, 85]. Die Idee, geistige Fähigkeiten des Menschen zu verbessern, ist natürlich nicht neu – man denke etwa an das in den 1960ern aufgekommene Konzept der Bewusstseinserweiterung mittels geeigneter Drogen. Heute gehen die Vorstellungen, Zeitgeist bedingt, eher in Richtung einer Verbesserung der Leistung kognitiver Funktionen wie Lernfähigkeit und Gedächtnis. Hier ergeben sich gleitende Übergänge zwischen therapeutischen Interventionen und *neuro enhancement*, wie das Beispiel von Ritalin deutlich macht. Diese für die Behandlung von Aufmerksamkeitsstörungen eingesetzte Medikament findet immer mehr auch Anwendung für therapiefremde Ziele wie Stressabbau oder Leistungsförderung bei "normalen" Personen. Diese Vermischung zwischen therapeutischen Zielen und einem eigentlichen *enhancement* dürfte auch das Arzt-Patient-Verhältnis ändern, da psychische Aspekte nicht mehr im Jargon der Psychoanalyse, sondern mit pharmakologischen Begriffen diskutiert werden. Die Bandbreite der praktischen Anwendungsmöglichkeiten wie ethischen Probleme des *neuro enhancement* ist gross: Ein Anwendungsbereich betrifft die Förderung der Lernfähigkeit und des Gedächtnisses: Wie stabil ist das unter dem Einfluss solcher Substanzen gewonnene Wissen? Wie ändert unsere Vorstellung von Lernen, verstehen wir doch derzeit Lernen als einen aufwendigen und mühevollen Prozess? Lassen sich die durch diese Forschung gewonnenen Erkenntnisse auch so nutzen, dass man unerwünschte Gedächtnisinhalte dereinst "löschen" kann? Derartige Anwendungen dürften bedeutende Auswirkungen auf unseren Begriff einer persönlichen Biografie haben. Zudem sind Anwendungen im Rahmen des Strafrechts denkbar, die aber noch einen sehr spekulativen Charakter haben. Ein weiterer Anwendungsbereich betrifft die Förderung von *executive functions* – also von Funktionen wie Entscheiden unter Stress. Denkbar ist, dass dereinst gewisse Berufsgruppen verpflichtet sind, solche Möglichkeiten zu nutzen. Auch die militärische Anwendung solcher Substanzen ist zweifellos ein Thema. Schliesslich stellen sich Fragen wie jene nach den Kosten solcher Anwendungen und auch solche nach der Transformation grundlegender Begriffe wie unser Menschenbild, unsere Vorstellung des Personseins und der Gesundheit wie auch den Wert des nicht Perfekten. Derartige Fragen haben beispielsweise den Philosophen Metzinger zur Forderung nach einer Neuroanthropologie inspiriert [118].

- Die zunehmenden Möglichkeiten für chirurgische Eingriffe in das Gehirn, die zu einer Renaissance der Psychochirurgie führen könnten, bilden einen weiteren Schwerpunkt der neuroethischen Diskussion. Auch dieser Problemkreis kann auf eine gewisse Geschichte zurückblicken. Erinnert sei beispielsweise an den Zeitraum der 1930ern bis 1970er, wo Psychochirurgie (insbesondere Lobotomie) praktiziert wurde, was zu einer breiten Debatte geführt hatte. Bereits damals ging es um die Erklärungskraft biologischer Modelle hinsichtlich des menschlichen Bewusstseins, menschlicher Emotionen und menschlicher Freiheit [67, 124]. Solche Eingriffe wurden später zu Gunsten pharmakologischer Interventionen aufgegeben. Heute könnte sich im Zug einer *Imaging*-gesteuerten Neurochirurgie aber eine neue Phase der Psychochirurgie etablieren. Anton Valavanis erklärte dazu im Gespräch, dass das Gehirn langsam den Status eines "heiligen Organs" verlieren würde und Neurochirurgen in zunehmendem Masse Eingriffe wie Biopsien etc. vornehmen würden. Fraglich ist hier natürlich, wie solche

Eingriffe mit der bekannten Plastizität des Gehirns vereinbar sind. Nebst der ethischen Frage stellt sich also auch die praktische Frage, ob chirurgische Eingriffe für die Verhaltenskontrolle überhaupt dauerhaft den gewünschten Effekt haben können.

- Technologien für das Erfassen oder das Beeinflussen von Gehirnprozessen können schliesslich ebenfalls auf eine Geschichte zurückblicken – erinnert sei beispielsweise an das Buch von Delgado über die *Physical control of the mind* [47]. Die moderne Forschung strebt diesbezüglich den Bau von *brain-machine interfaces* für die Steuerung von Prothesen, die Kommunikation mit *locked-in* Patienten oder die gezielte Stimulation für therapeutische Zwecke (*deep brain stimulator*) an [1]. Nebst diesen therapeutisch orientierten Forschungen existieren aber auch solche, welche durch gezielte Eingriffe Erkenntnisse über die Verhaltenssteuerung gewinnen wollen. Bekanntere Beispiele sind die Entwicklung einer Art primitiven Fernsteuerung für das Navigieren von Ratten [155] oder die genetische Manipulation von Neuronen in *Drosophila* derart, dass deren Feuern durch Bestrahlen mit Laserlicht ausgelöst werden kann [111]. Derartige Forschungen sind natürlich für Verhaltensstudien interessant, da man Verhalten auf eine neue Art und Weise steuern kann. Dies weckt aber auch Spekulationen, ob dereinst auch komplexes Verhalten einer gewissen Steuerung zugänglich ist.

4.4 Forschungsethische Fragen

Im Rahmen der modernen neurowissenschaftlichen Forschung stellen sich auch eine Reihe von forschungsethischen Fragen. Aktuell diskutiert werden hier insbesondere Fragen bezüglich der Anwendung von *Imaging* bei neurowissenschaftlichen Experimenten [105]:

- Sollen Versuchspersonen informiert werden, wenn unerwartete Abnormitäten bei einem *Imaging*-Experiment bemerkt werden (man schätzt, das sich bei zwei bis acht Prozent aller gescannten Versuchspersonen klinisch relevante Abnormitäten findet lassen [36])? Wie steht es in diesem Zusammenhang mit dem "Recht auf Nichtwissen"?
- Welche Studien fallen in welche bisher bestehende rechtliche Regelungen (in den USA beispielsweise fallen nicht alle solche Studien in den Bereich der *human subject research*, was mit unterschiedlichen, den Forschern offenbar aber meist unklaren, rechtlichen Anforderungen verbunden ist)?
- Wie steht es um den Schutz von *Imaging*-Daten? So können Datenpunkte, die Aufschluss über das Gesicht einer Person geben, zwar entfernt werden, was aber die Auswertung bestimmter Scans erschwert. Müssen Versuchspersonen informiert werden, wenn ihre Scans in Datenbanken gespeichert und durch Drittpersonen ausgewertet werden? Wie steht es um dem Schutz von Forschern, wenn Dritte mit Hilfe derer Daten schneller neue Erkenntnisse finden?

Im Rahmen des *neuro enhancement* stellen sich ebenfalls eine Reihe von Fragen. So ist grundsätzlich offen, ob Versuche mit solchen Wirkstoffen überhaupt als therapeutische Forschung zu gelten hat oder nicht vielmehr ähnlich wie kosmetische Forschung zu beurteilen sind. Eine entsprechende Einordnung hat natürlich erhebliche Auswirkungen darauf, welche Arten von Experimenten überhaupt unternommen werden dürfen. Deshalb ist anzunehmen,

dass Versuche in diese Richtung immer in einen therapeutischen Mantel gekleidet werden dürften. Zu bemerken ist hier auch die Beobachtung, dass gerade im Bereich der Psychopharmakologie eine enorme Zahl von Studien nicht von den Forschern selbst, sondern von *ghostwritern* verfasst wird. Entsprechende Schätzungen erreichen einen Anteil von bis zu fünfzig Prozent [87], was das ökonomische Interesse in diesem Bereich deutlich macht. Schliesslich stellen sich eine Reihe spezifischer Fragen bei der Erforschung komplexer psychischer Phänomene. Wird beispielsweise Intelligenz untersucht, so müssten aus statistischen Gründen auch Menschen mit schwacher Intelligenz Gegenstand der Untersuchung sein. Die Frage ist dann aber, ob diese Personen überhaupt einen *informed consent* zu solchen Studien geben können [75].

4.5 Auswirkungen auf Philosophie und Recht

Die Auswirkungen neurowissenschaftlicher Erkenntnisse auf philosophische Probleme wie auch auf das Recht bilden einen weiteren Themenbereich der Neuroethik. Ein breit diskutiertes Thema betrifft die Frage, inwiefern Konzepte wie Autonomie, Willensfreiheit und Verantwortung durch neurowissenschaftliche Erkenntnisse tangiert werden. So behauptet beispielsweise Wolf Singer, dass der Unterschied zwischen freien und unfreien Handlungen neurobiologisch gar nicht besteht, weil der Unterschied nur in verschiedenen Graden der Bewusstheit der Motive bestehe [149]:150. Es gibt eine Reihe guter Gründe, welche die Tragweite solcher Behauptungen anzweifeln (vgl. beispielsweise [16, 128, 162]). Diese Debatte soll angesichts der Vielzahl von Literatur zu diesem Thema an dieser Stelle nicht weiter ausgeführt werden.

Ein Problemkreis, der jedoch noch etwas genauer vorgestellt werden soll, betrifft den Einfluss der Neurowissenschaft auf das Recht. Dieser steht natürlich in einem Zusammenhang mit der generellen Diskussion um eine mögliche Abschwächung des Verantwortungsbegriffs, zumal das Rechtssystem, vorab das Strafrecht, einen Teil seiner Legitimation über die in der Gesellschaft vorherrschenden Ansichten hinsichtlich Schuld und Strafe zieht [115]. Eine veränderte Auffassung über Freiheit und Verantwortung mag damit durchaus einen indirekten Einfluss auf die künftige Ausgestaltung des Strafrechts haben. Eine radikale Änderung des Strafrechts ist aber nicht zu erwarten. Schliesslich geht das Recht schon seit langem davon aus, dass Menschen zwar die Befähigung haben, rationale Entscheidungen zu treffen. Das Recht berücksichtigt aber durchaus auch Grenzen dieser Fähigkeit. Es ist nicht anzunehmen, dass Neurowissenschaft die minimalen Anforderungen an die Fähigkeit solcher rationaler Entscheidungen untergräbt [76]. So sind einige der Aussagen in der aktuellen Diskussion über die Auswirkungen der Neurowissenschaft auf das Recht kaum spektakulär. So soll beispielsweise *legal reasoning* mehr Hirnprozesse rekrutieren als *intuitive moral judgment* [74] – was ja stimmen mag aber im weiteren nicht interessant erscheint. Chorvat und McCabe meinen, dass die kognitiven Neurowissenschaften insbesondere für die Untersuchung der Genese von Gesetzen und des möglichen Einflusses von Gesetzen auf Verhalten wertvolle Hinweise liefern könnten [37]. Sie nennen fünf aktuelle Forschungsgebiete, welche dafür von besonderem Interesse sein sollen: Erstens sei die Untersuchungen im Sinne von Greene über moralische Dilemmas von Interesse (vgl. dazu mit Abschnitt 3.3). Das zweite relevante Gebiet sei die Anwendung der experimentellen Spieltheorie, weil dadurch die Kontextabhängigkeit von Entscheidungsprozessen genauer untersucht werden

könne. Das dritte Thema sei die Untersuchung der Bildung von Vertrauen in sozialen Situationen. Das vierte Thema sei die genauere Analyse des Faktums, dass Zurückweisung als eine Form von Schmerz empfunden werde – ein Aspekt, der das Gesetz nutzen solle (und es sicher bereits schon tut!). Fünftens schliesslich solle der automatische Charakter gewisser Verhaltensweisen genauer untersucht werden, um damit ein genaueres Verständnis der Verantwortungsbegriffs zu gewinnen. Ihr Fazit erscheint ausserordentlich optimistisch: "By understanding the neural mechanism, which we use to solve problems, we can hope to create laws and other rules that will help to foster socially optimal behavior." Das Grundproblem bei solchen Hoffnungen ist, dass die Neurowissenschaft nicht zu definieren vermag, welche Verhaltensweisen sozial optimal sind.

Abgesehen von derart weit gehenden Ansprüchen lassen sich eine Reihe von praktischen Problemen in der Rechtssprechung identifizieren, zu deren Verständnis die Neurowissenschaften durchaus beitragen können: Von besonderer Bedeutung ist die Gedächtnisforschung. So weiss man, dass sich Gedächtnisinhalte in der Zeit ändern und teilweise gar neu geschaffen werden können und sich auf Ereignisse beziehen können, die nie stattgefunden haben [112]. Diese Problematik von Zeugenaussagen und suggestiven Befragungsmethoden waren natürlich schon vorher bekannt, die Neurowissenschaft könnte aber mehr über den Mechanismus der Gedächtnis(fehl)bildung herausfinden, so dass Zeugenaussagen besser eingeschätzt werden können. Im weiteren könnte die Auswirkungen emotional belastender Beweismaterialien (Fotografien von Gewaltopfern etc.) auf die Geschworenen untersucht werden. Auch bei gewissen praktischen Problemen, beispielsweise bei der in der Praxis festgestellten Differenz der Beurteilung von Eigentumsdelikten im Vergleich zu Verletzungen des geistigen Eigentums (z.B. Software-Raubkopien), könnten interessante Erkenntnisse gewonnen werden. So wird die Hypothese aufgestellt, letztere könnten kein subjektives Gefühl der moralischen Falschheit solchen Verhaltens erzeugen, was die Durchsetzung entsprechender Gesetze erschwere [74].

Chapter 5

Weiterführende Fragestellungen

5.1 Welches Problem soll gelöst werden?

Die bisherigen Ausführungen haben deutlich gemacht, dass die moderne Neurowissenschaft und Verhaltensforschung noch weit davon entfernt sind, die "biologischen Grundlagen" von Moral bzw. moralischem Verhalten ausreichend beschreiben zu können. Dies hängt damit zusammen, dass das zu lösende Problem offenbar noch nicht ausreichend detailliert beschrieben worden ist. Nachfolgend soll versucht werden, die bisherige Formulierung des Problems zusammenfassend zu beschreiben, auf die damit verbundenen Schwierigkeiten aufmerksam zu machen und eine strukturierte Beschreibung des zu lösenden Problemkomplexes zu liefern.

Das **Grundproblem**, das sich die Forscher im Bereich "neurobiologische Grundlagen der Moral" offenbar stellen, kann wie folgt umrissen werden (vgl. dazu mit Abbildung 5.1): Man geht von der (in diesem Sinn wohl unbestrittenen) Annahme aus, das so genannt moralisches Verhalten aus bestimmten Prozessen im Gehirn resultiert. Demnach gibt es für *moral agents* bestimmte Anlässe (moralische Stimuli), die Anlass zu einem moralischen Verhalten geben, was wiederum Ausdruck eines *moral decision making* ist. Im Prozess dieses *decision making* ist ein bestimmtes Netzwerk von Hirnregionen involviert – gewissermassen die neurobiologische Infrastruktur der Moralfähigkeit. Das Erkennen der neurobiologischen Grundlage von Moral besteht dann offenbar darin, dieses Netzwerk zu bestimmen. Unter "Bestimmen" ist dabei nicht nur eine Lokalisation der Knoten dieses Netzwerkes gemeint, sondern auch deren funktionellen Interaktion, sowie die phylogenetischen wie ontogenetischen Aspekte, welche die Ausprägung des Netzwerkes determinieren. Phylogenetische Aspekte sind beispielsweise die Evolutionsgeschichte des Netzwerkes bzw. genetische Aspekte, welche dessen Struktur mitbestimmen. Die ontogenetischen Aspekte betreffen den einzelnen *moral agent*, dessen neurobiologische Infrastruktur der Moralfähigkeit durch dessen sozialen Interaktionen oder auch durch mögliche Schäden am Netzwerk beeinflusst wird. Im ersten Fall kann die Ontogenese des Netzwerkes aufgrund "falscher" Interaktionen (beispielsweise schlimme Familienverhältnisse) dahingehend beeinträchtigt werden, dass der betreffende *moral agent* in moralischer Hinsicht als "anormal" eingestuft wird. In zweiten Fall können Läsionen im Netzwerk (beispielsweise aufgrund einer Hirnblutung im frontalen Kortex) diese neurobiol-

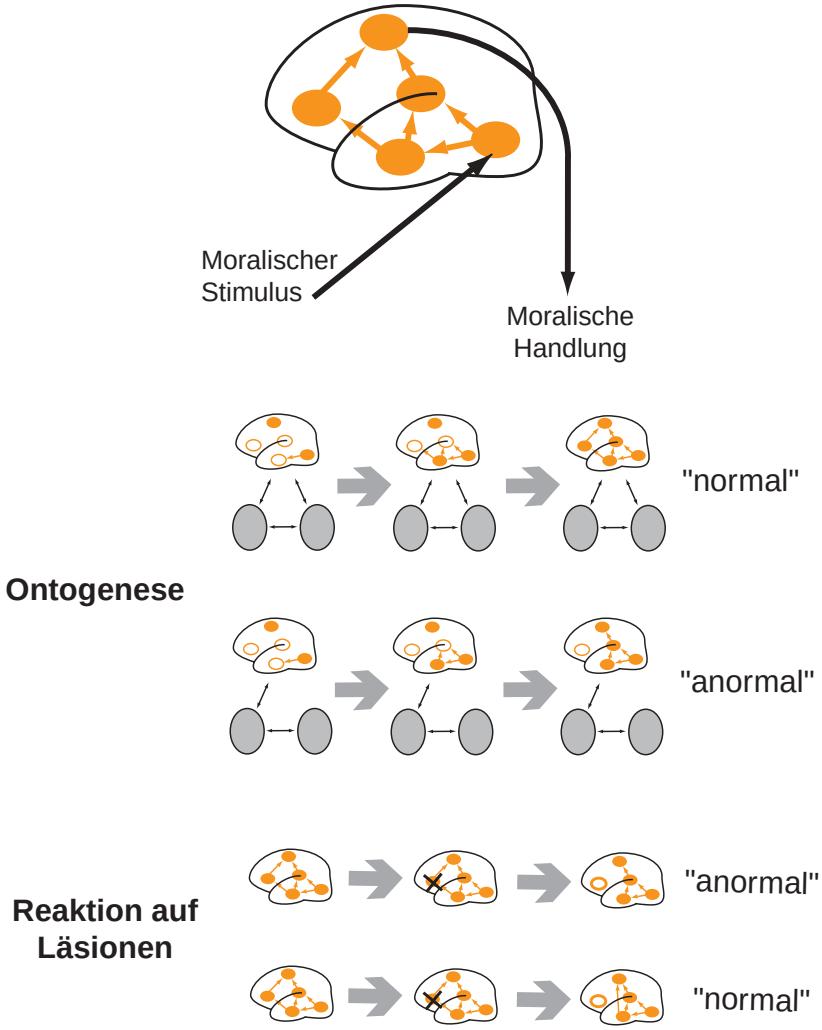

Figure 5.1: Schematische Darstellung des Grundproblems, das sich die Forscher im Gebiet "neurobiologische Grundlagen der Moral" stellen.

ogische Infrastruktur ebenfalls derart beeinträchtigen, dass ein anormales moralisches Verhalten resultiert. In bestimmten Fällen hingegen scheint die Plastizität des Gehirns eine Art "Reparatur" des Netzwerkes dahingehend zu ermöglichen, dass der *moral agent* nach einer Übergangszeit wieder als "normal" eingestuft werden kann. Dieses Phänomen könnte man "moralische Plastizität" nennen. In methodischer Hinsicht bieten sich demnach das Studium von Läsionspatienten, die Anamnese sozialer Pathologien und *Imaging*-Studien an, um die neurobiologische Infrastruktur der Moralfähigkeit im oben ausgeführten Sinne zu ermitteln.

Hinsichtlich des **Motivs** dieser Studien lässt sich zusammengefasst folgendes sagen (vgl.

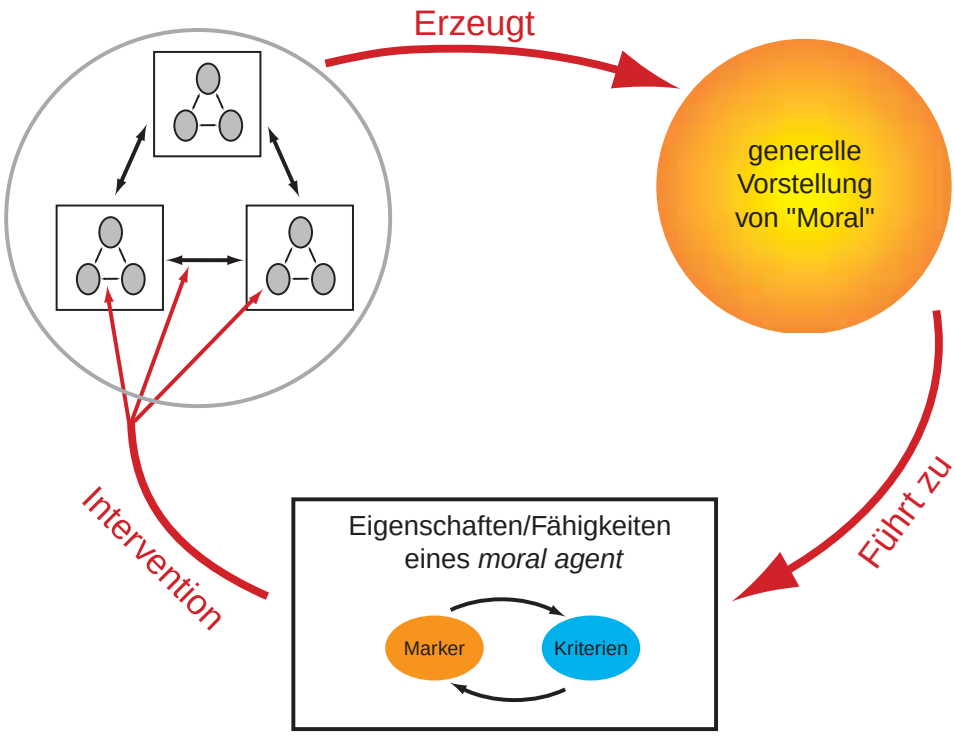

Figure 5.2: Schematische Darstellung des Motivs für die Erforschung der "neurobiologischen Grundlagen der Moral".

dazu mit Abbildung 5.2): Man geht davon aus, dass die Interaktion von *moral agents* sowie die Interaktion von Institutionen von *moral agents* Moral im Sinne der deskriptiven Ethik generieren – d.h. eine Gesamtheit von Wertvorstellungen, Normen, etc., welche ausmachen, was in einem gewissen Kulturkreis als moralisch richtig bzw. falsch gilt. Diese allgemeine Vorstellung von Moral dient den Forschern dann als Ausgangslage, um die Eigenschaften bzw. Fähigkeiten eines *moral agent* zu bestimmen. Das "Bestimmen" dieser Eigenschaften geschieht in einem Wechselspiel des Festsetzens von Kriterien – was der intentionalen Ebene zugeordnet wird – und dem Finden von Markern – d.h. messbaren Entitäten auf der Ebene des biologischen Systems, deren Vorhandensein die Erfüllung des gewünschten Kriteriums anzeigen. Was letztlich als Marker gelten kann, hängt natürlich von den zur Verfügung stehenden Methoden ab – ein Beispiel wäre eine mittels *Imaging* feststellbare Läsion in einem für die neurobiologische Infrastruktur der Moralfähigkeit notwendigen Hirnregion. Das Finden der Marker in Kombination mit ausreichenden Erkenntnissen über die genannte neurobiologische Infrastruktur (also beispielsweise über den Zusammenhang zwischen den Beziehungen, in die ein *moral agent* eingebunden ist und der Ontogenese dieses *moral agent*) dient dann für die Bestimmung von Interventionen auf der Ebene des einzelnen *moral agent*, auf der Ebene der Beziehungen zwischen *moral agents* oder auf der Ebene der Interaktio-

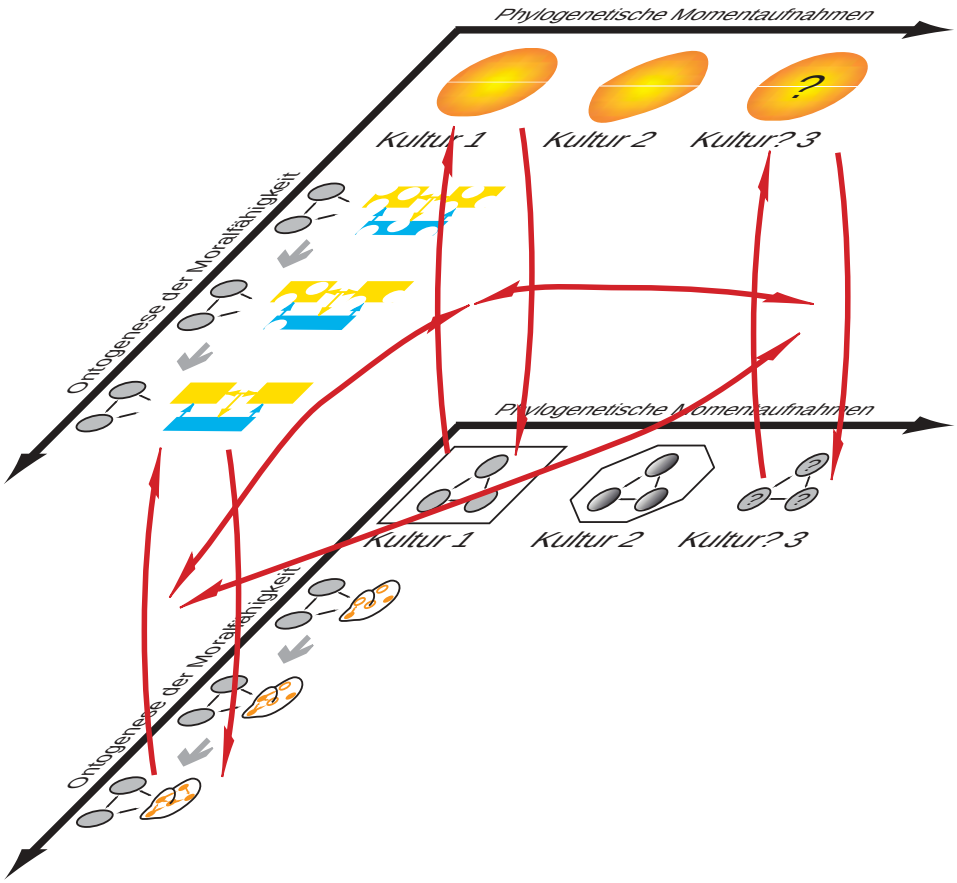

Figure 5.3: Schematische Darstellung des "Problems der zwei Ebenen".

nen zwischen Institutionen von *moral agents*. Beispiele wären der Einsatz eines bestimmten Psychopharmakas, eine Art neurobiologisch gestützte Familientherapie oder eine Anpassung des Rechtssystems in Form einer Festlegung, auf welche *moral agents* ein rechtlicher Verantwortungsbegriff anwendbar wäre. Soweit die "schöne neue Welt" mit dem Ziel "[to] help to shape environmental, psychological and medical interventions aimed at promoting prosocial behaviours and social welfare" [119].

Wo sind nun die Probleme in diesem Projekt? Auf die vielfältigen Einwände hinsichtlich Studiendesign, methodischen Schwierigkeiten etc. soll hier nicht eingegangen werden. Vielmehr soll das grundsätzliche Problem angesprochen werden. Dieses kann als das "Problem der zwei Ebenen" bezeichnet werden (vgl. dazu mit Abbildung 5.3). Dieses ist letztlich Ausdruck der zentralen Problemstellung der Philosophie des Geistes, das klassisch als das Leib-Seele-Problem bzw. modern als das Problem der Interaktion des Physischen und des Psychischen bezeichnet wird. Es ist grundsätzlich nicht erstaunlich, dass dieser Problemtyp im Bereich der Moral auftaucht, zumal schon das Alltagsverständnis von Moral viel

mit Vorstellungen des sittlich Guten bzw. Falschen, mit Begründungen, etc. zu tun hat. So interessiert beispielsweise bei der Ontogenese eines *moral agent* nicht nur dessen neurobiologische Infrastruktur der Moralfähigkeit, sondern genauso auch sein Selbstbild und der Raum der Gründe, innerhalb dessen der *agent* sich bewegt, wenn er moralische Entscheide treffen will. Die Geschichte dieses Selbstbildes wie auch der Aufbau dieses Raums der Gründe sind demnach gleichermassen interessant. Auf der phylogenetischen Achse wiederum sind Untersuchungen in verschiedenen Kulturen wie auch in Tiergemeinschaften (bei welchen zweifelhaft ist, ob überhaupt *moral agents* interagieren) als Momentaufnahme in einem Kontinuum von evolutionären Entstehungsmöglichkeiten zu verstehen, die Einblick in die Spanne möglicher Moralsysteme geben. Hier ist der kulturwissenschaftliche Blick auf die Geschichte dieser Moralsysteme relevant. Was in Abbildung 5.2 den Anschein eines Regelkreises macht, ist in Wirklichkeit ein Umherhüpfen zwischen diesen beiden Ebenen. Doch gerade bei diesen Sprüngen zwischen den Ebenen verbergen sich die entscheidenden Fragen: Ist überhaupt eine Phänomenologie des *moral agent* ausreichend definiert, um wissen zu können, welche neuronalen Module man der neurobiologischen Infrastruktur der Moralfähigkeit zuordnen will? Inwiefern prägen unsere historisch gewachsenen Vorstellungen eines bestimmten (d.h. unseres) Moralsystems dessen Untersuchung beispielsweise mit Methoden der experimentellen Ökonomie? Inwiefern beeinflussen unsere Vorstellungen von einer evolutionär entstandenen Moral die Beobachtung von Tierverhalten, welchen wir einen moralischen Charakter zugestehen wollen? Dies sind nur einige der wohl zahlreichen Fragen, welche das Problem der zwei Ebenen ausdifferenzieren. Nachfolgend soll nun eine Skizze eines Forschungsprojektes geliefert werden, welches diesem Problem Rechnung tragen könnte.

5.2 Skizze eines Forschungsprogramms

Die bisherigen Ausführungen haben deutlich gemacht, dass unterschiedliche Forschungsgebiete bei der Untersuchung der Grundlagen moralischer Orientierung beteiligt sind bzw. künftig beteiligt sein sollten. Diese Gebiete und die damit verbundenen Fragestellungen können wie folgt den zwei Ebenen zuordnet werden (vgl. dazu mit Abbildung 5.4)

- Die Neurobiologie des *moral agent* ist lediglich ein Aspekt des ganzen Programms. Gewiss dürfte eine Bestimmung des funktionalen Netzes, welches die genannte Infrastruktur der Moralfähigkeit ausmacht, kein unbedeutender Aspekt sein – auch wenn bereits die jetzigen Hinweise darauf schliessen lassen, dass sehr viele Hirnbereiche eine kritische Rolle beim *moral decision making* spielen. Die derzeit durchgeführten *Imaging*-Studien, bei welchen Versuchspersonen mit diversen moralischen Stimuli konfrontiert werden, erscheinen deshalb als wenig erfolgsversprechender Ansatz. Viel bedeutsamer dürfte eine genau Spezifizierung der oben genannte moralischen Plastizität sein – also der offenbar vorhanden Fähigkeit des Gehirns, das Netzwerk der neurobiologische Infrastruktur der Moralfähigkeit derart umorganisieren zu können, dass der betreffende *moral agent* immer noch ein "normales" moralisches Verhalten zeigen kann. Die genaue Bestimmung des Ausmasses dieser Plastizität dürfte nicht einfach sein. So dürfte beispielsweise die jetzige *Imaging*-Technologie noch keine ausreichende Auflösung haben, um das genaue Ausmass des Funktionsausfalls in einem bestimmten geschädigten Hirnteil zu erfassen. Auf der Ebene der medizinische Kasuistik wiederum

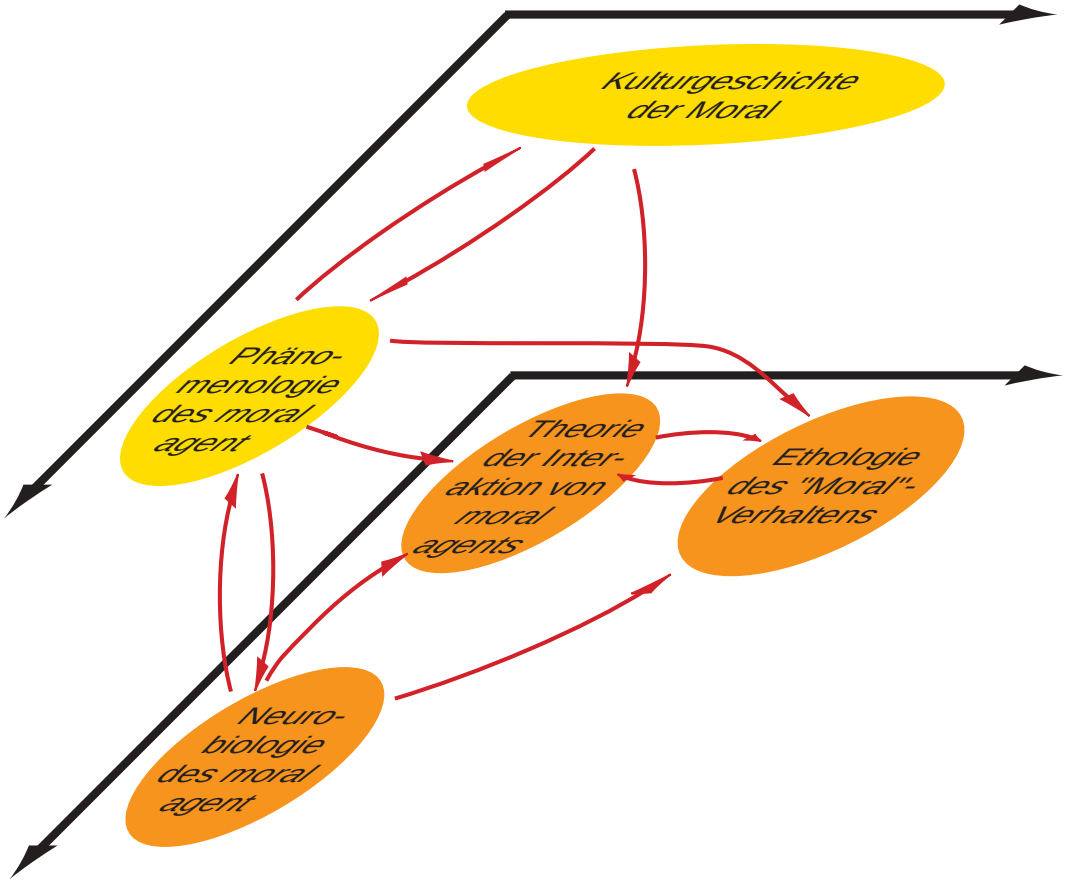

Figure 5.4: Forschunsgebiete, welche in der Untersuchung der Grundlagen moralischer Orientierung involviert sind.

könnte das Problem auftauchen, dass in der Medizin die einzelnen Fälle nicht standardisiert beschrieben worden sind, so dass man mit einer grossen Vielfalt an Varianten von moralischer Abnormität konfrontiert ist, welche natürlich in historischer Hinsicht zusätzlich durch die wandelnden Vorstellungen des moralisch Abnormen kontaminiert sind. Die Erfassung der moralischen Plastizität ist demnach ein schwieriges Unterfangen.

- Es erscheint absolut notwendig, dass die Suche nach der Neurobiologie des *moral agent* mit der Bestimmung einer Phänomenologie des *moral agent* einher gehen muss. Gewiss sind dazu viele Erkenntnisse, etwa im Bereich der Moralpsychologie und deskriptiven Ethik, bereits erarbeitet worden. Diese müssten nun aber zu einem generellen Modell eines *moral agent* zusammengestellt und ergänzt werden. Ob die in dieser Pilotstudie erstellte erste Skizze als Ausgangslage dienen kann, ist offen. Begriffe wie

"Selbstbild" oder "Raum der Gründe" müssten aber integriert werden. Ein weiterer zu klärender Kernbegriff in diesem Kontext ist "Autonomie" – zumal diese im Bereich der philosophischen Ethik untrennbar zum Begriff der Moral gehört. Im Rahmen dieser Phänomenologie des *moral agent* müsste insbesondere auch die Konzepte eines moralischen Stimulus einer moralischen Entscheidung und einer moralischen Handlung ausdifferenziert werden.

- Im weiteren wird eine Theorie der Interaktion von *moral agents* benötigt. Hier können die bereits entwickelten Methoden der experimentellen Ökonomie Anwendung finden, zumal eine Reihe moralnaher Konzepte mit dieser Methode recht gut untersucht werden können. Es wäre lohnen zu prüfen, inwiefern spieltheoretische Modelle für Konzepte wie "Lügen", "Betrügen", "moralischer Ruf" etc. entwickelt werden können. Basierend auf ersten rudimentären Modellen eines *moral agent* könnte auch die Agentenbasierte Modellierung Anwendung finden – beispielsweise um den Effekt gewisser Vorstellungen von *moral agents* auf die Bildung von Institutionen zu untersuchen. Schliesslich gehört auch die experimentelle (Moral-)Psychologie in dieses Feld – also Untersuchungen im Sinne von Piaget oder Kohlberg.

- Wichtige Erkenntnisse könnte eine Ethologie des Moralverhaltens liefern. Diese Forschungen hingegen brauchen ein Modell eines *moral agent*, dessen Eigenschaften empirisch prüfbar sind. Nur so kann untersucht werden, ob Tiere Teilaspekte bzw. Vorformen von Moralität realisieren. Solche Forschungen sollten sich nicht lediglich auf Beobachtungen abstützen, sondern benötigt auch Verhaltensexperimente. Es wäre viel versprechend, wenn beispielsweise Spiele der experimentellen Ökonomie derart abgewandelt werden könnten, dass sie auch von Primaten gespielt werden können.

- Dieses hier skizzierte Forschungsvorhaben müsste schliesslich auch eine Kulturgeschichte der Moral mit einschliessen – gewiss ein Projekt, das bereits von einigen Wissenschaftlern in Angriff genommen wurde. Eine historische Perspektive dürfte in so manchem Feld vor zu raschen Schlussfolgerungen schützen. Sie könnte beispielsweise untersuchen, bis zu welchem Grad das oft gehörte Argument des ethischen Relativismus überhaupt eine historische Abstützung findet.

Nicht ausser acht gelassen werden dürfen bei dieser Aufzählung die diversen Schnittstellen-Aspekte, wie bereits weiter oben angedeutet wurde: So müsste eine Phänomenologie des *moral agent* Moral in Zusammenarbeit mit der Untersuchung der Neurobiologie der Moralfähigkeit entwickelt werden. Das daraus erwachsende Modell eines *moral agent* wird dann sowohl für die Theorie der Interaktion dieser *agents* wie auch für die Ethologie des Moralverhaltens wichtig werden. Die Kulturgeschichte einer Moral wiederum könnte durchaus auch wichtige Impulse für das Verständnis von Theorien und Modellen für die Interaktion von *moral agents* liefern. Sicherlich ist das hier beschriebene Programm noch sehr rudimentär und bei weitem nicht ausformuliert. Immerhin können auf dem Platz Zürich eine Reihe von Personen identifiziert werden, welche grundsätzliches Interesse an diesen Fragen haben:

- Der Problemkreis einer Phänomenologie des *moral agent* fällt in den Bereich der philosophischen Ethik (also des Ethikzentrums der Universität Zürich, dem Auftraggeber dieser Pilotstudie) und in den Bereich der Moralpsychologie (Interessent: Lutz Jäncke, l.jaencke@psychologie.unizh.ch).

- Der Problemkreis einer Neurobiologie des *moral agent* könnte in Zürich vorab in den Kontext der klinischen Forschung gestellt werden (Interessierte: Peter Brugger, peter.brugger@usz.ch, Marianne Regard, mregard@npsy.unizh.ch und Anton Valavanis, nra.dir@dmr.usz.ch). Eine (im Rahmen der Pilotstudie nicht angefragte) weitere Person wäre eventuell Victor Candia vom *Collegium Helveticum* (candia@collegium.ethz.ch).

- Für den Problemkreis der Theorie der Interaktion von *moral agents* könnten im Umfeld der Fehr-Gruppe interessierte Personen gewonnen werden (Urs Fischbacher, fiba@iew.unizh.ch und evt. Daria Knoch, dknoch@iew.unizh.ch, nicht angefragt).

- Für den Problemkreis Ethologie der Moral könnten Forscher am Anthropologischen Institut der Universität Zürich gewonnen werden (Grundsätzlich interessiert: Carel van Schaik, vschaik@aim.unizh.ch).

- Für den Problemkreis der Kulturgeschichte der Moral schliesslich könnten Personen aus dem Umfeld des ETH-Zentrums "Geschichte des Wissens" gewonnen werden (Michael Hagner, hagner@wiss.gess.ethz.ch).

Im Rahmen dieser Skizze sollten schliesslich neuroethische Fragen nicht ganz vergessen werden. Gemeint sind hier weniger die bereits von verschiedener Seite aufgeworfenen Probleme des *Imaging* und des *neural enhancement*. Bedeutsamer ist vielmehr, dass eine Erforschung des *moral agent* auch mit einer Veränderung des Verständnisses von Moral einher gehen kann. Es macht beispielsweise einen Unterschied, ob man gewisse ethische Regeln als Ergebnis eines rationalen Entscheidungsprozesses auffasst, oder vielmehr als eine Art emergentes Phänomen der Interaktion von *moral agents*. Die damit verbundenen Fragen verdienen eine nähere Aufmerksamkeit. Hier ist zu bemerken, dass von Seiten des Zentrums für Neurowissenschaften Zürich durchaus Interesse an einer verstärkten Berücksichtigung neuroethischer Fragen besteht (Wolfgang Knecht, wknecht@neuroscience.unizh.ch).

Bibliography

[1] Abbott A (2005): Deep in thought. Nature 436: 18-19.

[2] Adelman G, and BH Smith (eds.) (2004): *Encyclopedia of neuroscience*, third edition. Elsevier, Amsterdam.

[3] Adolphs R (2003): Cognitive neuroscience of human social behaviour. Nature Reviews: Neuroscience 4: 165-178.

[4] Adolphs R (2001): The neurobiology of social cognition. Current Opinion in Neurobiology 11: 231-239.

[5] Adolphs R (1999): Social cognition and the human brain. Trends in Cognitive Sciences 3(12): 469-479.

[6] Andersen RA, JW Burdick, S Musallam, B Pesaran, and JG Cham (2004): Cognitive neural prosthetics. Trends in Cognitive Sciences 8(11): 486-493.

[7] Anderson SW, A Bechara, H Damasio, D Tranel, and AR Damasio (1999): Impairment of social and moral behavior related to early damage in human prefrontal cortex. Nature Neuroscience 2(11): 1032-1037.

[8] Arbib MA, J-M Fellous (2004): Emotions: from brain to robot. Trends in Cognitive Sciences 8(12): 554-561.

[9] Bargh JA, and TL Chartrand (1999): The unbearable automaticity of being. American Psychologist 54(7): 462-479.

[10] Barrett JL (2000): Exploring the natural foundations of religion. Trends in Cognitive Sciences 4(1): 29-34.

[11] Batson CD, DA Lisher, A Carpenter, L Dulin, S Harjusola-Webb, EL Stocks, S Gale, O Hassan, and B Sampat (2003): "... as you would have them do unto you": does imaging yourself in the other's place stimulate moral action? Personality and Social Psychology Bulletin 29(9): 1190-1201.

[12] Beaulieu A (2001): Voxels in the brain: Neuroscience, informatics and changing notions of objectivity. Social Studies of Science 31(5): 635-680.

[13] Beaulieu A (2002): A space for measuring mind and brain: interdisciplinarity and digital tools in the development of brain mapping and functional imaging, 1980-1990. Brain and Cognition 49: 13-33.

[14] Beaulieu A (2004): From brainbank to database: the informational turn in the study of the brain. Studies of the History and Philosophy of the Biological & Biomedical Sciences 35: 367-390.

[15] Bechara A, H Damasio, and AR Damasio (2000): Emotion, decision making and the orbitofrontal cortex. Cerebral Cortex 10: 295-307.

[16] Beckermann A (2005): Biologie und Freiheit. Zeigen die neueren Ergebnisse der Neurobiologie, dass wir keinen freien Willen haben? In: H Schmidinger, und C Sedmak (Hrsg.): Der Mensch – ein freies Wesen? Autonomie – Personalität – Verantwortung. Wissenschaftliche Buchgesellschaft, Darmstadt: 111-124.

[17] Bekoff M (2004): Wild justice and fair play: cooperation, forgiveness, and morality in animals. Biology and Philosophy 19: 489-520.

[18] Bendor J, and P Swistak (2001): The evolution of norms. American Journal of Sociology 106(6): 1493-1545.

[19] Bennett MR, and PMS Hacker (2003): *Philosophical foundations in neuroscience*. Blackwell Publishing, Malden MA.

[20] Blair RJR (1995): A cognitive developmental approach to morality: investigating the psychopath. Cognition 57: 1-29.

[21] Blakemore S-J, JS Winston, and U Frith (2004): Social cognitive neuroscience: where are we heading? Trends in Cognitive Sciences 8(5): 216-222.

[22] Bolte A, T Goschke, and J Kuhl (2003): Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. Psychological Science 14(5): 416-421.

[23] Bondolfi A, and M Christen (2001): Folgenreiche Fortschritte in den Neurowissenschaften. Neue Zürcher Zeitung, 02.05.2001.

[24] Borck C (2005): *Hirnströme. Eine Kulturgeschichte der Elektroenzephalographie*. Wallstein Verlag, Göttingen.

[25] Bowles S, and H Gintis (2004): The evolution of strong reciprocity: cooperation in heterogeneous populations. Theoretical Population Biology 65: 17-28.

[26] Boyer P (2003): Religious thought and behaviour as by-product of brain function. Trends in Cognitive Sciences 7(3):

[27] Breithaupt H, K Weigmann (2004): Manipulating your mind. EMBO Reports 5(3):230-232.

[28] Camerer CF, and E Fehr (2002): Measuring social norms and preferences using experimental games: a guide for social scientists. IEW Working Paper 97, Institute for Empirical Research in Economics, Universität Zürich.

[29] Camille N, G Coricelli, J Sallet, P Pradat-Diehl, J-R Duhamel, and A Sirigu (2004): The involvment of the orbitofrontal cortex in the experience of regret. Science 304: 1167-1170.

[30] Canli T, and Z Amin (2002): Neuroimaging of emotion and personality: Scientific evidence and ethical considerations. Brain and Cognition 50: 414-431.

[31] Casebeer WD (2003): Moral cognition and its neural constituents. Nature Reviews: Neuroscience 4: 841-847.

[32] Casebeer WD (2003): Natural ethical facts. Evolution, connectionism, and moral cognition. MIT Press, Cambridge.

[33] Casebeer WD, and PS Churchland (2003): The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. Biology and Philosophy 18: 169-194.

[34] Chalmers DJ (2000): What is a neural correlate of consciousness? In: T Metzinger (ed.): Neural correlates of consciousness: empirical and conceptual questions. MIT Press, Cambridge Mass.

[35] Check E (2005): Ethicists urge caution over emotive power of brain scans. Nature 435: 254-255.

[36] Check E (2005): Brain-scan ethics come under the spotlight. Nature 433: 185.

[37] Chorvat T, and K McCabe (2004): The brain and the law. Philosophical Transactions of the Royal Society of London B 359: 1727-1736.

[38] Christen M (2005): Der Einbau von Technik in das Gehirn. Das Wechselspiel von Informationsbegriffen und Technologieentwicklung am Beispiel des Hörens. In: B Orland (ed): *Interferenzen. Studien zur Kulturgeschichte der Technik* Volume 9: 197-218.

[39] Christen M (2004): Schuldige Maschinen? Systemautonomie als Herausforderung für das Konzept der Verantwortung. *Jahrbuch für Wissenschaft und Ethik* 9: 163-191.

[40] Christen M (2003): Die Ontologie künstlicher Körper. *Studia Philosophica* 63: 65-82.

[41] Colman AM (2003): Cooperation, psychological game theory, and limitations of rationality in social interactions. Behavioral and Brain Sciences 26: 139-198.

[42] Damasio AR (2003): Looking for Spinoza. Joy, sorrow, and the feeling brain. Harvest Book, Orlando.

[43] Damasio H (2002): Impairment of interpersonal social behavior caused by acquired brain damage. In: SG Post, LG Underwood, JP Schloss, and WB Hurlbut: Altruism & altruistic love. Oxford University Press, Oxford: 272-283

[44] Damasio AR (2002): A note on the neurobiology of emotions. In: SG Post, LG Underwood, JP Schloss, and WB Hurlbut (eds.): Altruism & altruistic love. Oxford University Press, Oxford: 264-271.

[45] Damasio AR (1999): The feeling of what happens – Body and emotion in the making of consciousness. Harcourt Brace, New York.

[46] Damasio AR (1994): Descartes' error – emotion, Reason, and the human brain. Penguin Putnam, New York.

[47] Delgado JMR (1969): Physical control of the mind. Harper & Row Publishers, New York.

[48] De Quervain DJ-F, U Fischbacher, V Treyer, M Schellhammer, U Schnyder, A Buck, and E Fehr (2004): The neural basis of altruistic punishment. Science 305: 1254-1258.

[49] De Waal F (1996): Good natured. The origins of right and wrong in humans and other animals. Harvard University Press, Cambridge.

[50] Dolan RJ (1999): On the neurology of morals. Nature Neuroscience 2(11): 927-929.

[51] Donaldson DI (2004): Pasring brain activity with fMRI and mixed designs: what kind of a state is neuroimaging in? Trends in Neurosciences 27(8): 442-444.

[52] Dudai Y (2004): The neurosciences: the danger that we will think that we have understood it all. In: D Rees, and S Rose: The new brain sciences. Perils and prospects. Cambridge University Press, Cambridge: 167-180.

[53] Düwell M, C Hübenthal, MH Werner (Hrsg.) (2002): Handbuch Ethik. Verlag J.B. Metzler, Stuttgart, Weimar.

[54] Dugatkin LA, and MS Alfieri (2002): A cognitive approach to the study of animal cooperation. In: M Bekoff, C Allen, and GM Burghardt (eds.): The cognitive animal. The MIT Press, Cambridge: 413-419.

[55] Dugatkin LA (2002): Cooperation in animals: an evolutionary overview. Biology and Philosophy 17: 459-476.

[56] Eisenberger NI, and MD Lieberman (2004): Why rejection hurts: a common neural alarm system for physical and social pain. Trends in Cognitive Sciences 8(7): 294-300.

[57] Falk A, E Fehr, and U Fischbacher (1999): On the nature of fair behavior. Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 17.

[58] Falk A, and U Fischbacher (2000): A theory of reciprocity. Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6.

[59] Farah MJ, J Illes, R Cook-Deegan, H Gardner, ER Kandel, P King, E Parens, B Sahakian, and PR Wolpe (2004): Neurocognitive enhancement: what can we do and what should we do? Nature Reviews: Neuroscience 5: 421-425.

[60] Farah MJ (2005): Neuroethics: the practical and the philosophical. Trends in Cognitive Sciences 9(1): 34-40.

[61] Farrow TFD, Y Zheng, ID Wilkinson, SA Spence, JFW Deakin, N Tarrier, PD Griffiths, and PWR Woodruff (2001): Investigating the functional anatomy of empathy and forgiveness. NeuroReport 12(11): 2433-2438.

[62] Fehr E, and KM Schmidt (1999): A theory of fairness, competition and cooperation. The Quarterly Journal of Economics, August: 817-868.

[63] Fehr E and S Gächter (2002): Altruistic punishment in humans. Nature 415: 137-140.

[64] Fehr E, U Fischbacher (2004): Social norms and human cooperation. Trends in Cognitive Sciences 8(4): 185-190.

[65] Fehr E, and B. Rockenbach (2004): Human altruism: economic, neural, and evolutionary perspectives. Current Opinion in Neurobiology 14: 784-790.

[66] Fehr E (2005): Mit Neuroökonomik das menschliche Wesen ergründen. Neue Zürcher Zeitung 25./26. Juni 2005: 29.

[67] Fins JJ (2004): Neuromodulation, free will and determinism: lessons from the psychosurgery debate. Clinical Neuroscience Research 4: 113-118.

[68] Fischer J (2005): Antrag für ein Teilprojekt "Grundlagen der Moral" im Rahmen des universitären Forschungsschwerpunktes Ethik. Zürich.

[69] Flack, JC, and FBM de Waal (2000): Any animal whatever. Darwinian building blocks of morality in monkeys and apes & Response to commentary discussion. Journal of Consciousness Studies 7(1-2): 1-29, 67-77.

[70] Gintis H (2000): Strong reciprocity and human sociality. Journal of Theoretical Biology 206: 169-179.

[71] Gintis H., H Smith, and S Bowles (2001): Costly signalling and cooperation. Journal of Theoretical Biology 213: 103-119.

[72] Gintis H (2003): The hitchhiker's guide to altruism: Gene-culture coevolution and the internalization of norms. The Journal of Theoretical Biology 220(4): 407-418.

[73] Glimcher PW (2002): Decisions, decisions, decisions: Choosing a biological science of choice. Neuron 36: 323-332.

[74] Goodenough OR, K Prehn (2004): A neuroscientific approach to normative judgment in law and justice. Philosophical Transactions of the Royal Society of London B 359: 1709-1726.

[75] Gray JR.; and PM Thompson (2004): Neurobiology of intelligence: science and ethics. Nature Reviews: Neuroscience 5: 471-482.

[76] Greene J, and JD Cohen (2004): For the law, neuroscience changes nothing and everything. Philosophical Transactions of the Royal Society of London B 359: 1775-1785.

[77] Greene J, LE Nystrom, AD Engell, JM Darley, and JD Cohen (2004): The neural bases of cognitive conflict and control in moral judgment. Neuron 44: 389-400.

[78] Greene J (2003): From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology? Nature Reviews: Neuroscience 4: 847-851.

[79] Greene J, and J Haidt (2002): How (and where) does moral judgment work? Trends in Cognitive Sciences 6(12): 517-523.

[80] Greene J, RB Sommerville, LE Nystrom, JM Darley, and JD Cohen (2001): An fMRI investigation of emotional engagement in moral judgment. Science 293: 2105-2108.

[81] Greenstein B, and A Greenstein (2000): Color atlas of neuroscience. Thieme, Stuttgart, New York.

[82] Gruen L (2002): The morals in animal minds. In: M Bekoff, C Allen, and GM Burghardt (eds.): The cognitive animal. The MIT Press, Cambridge: 437-442.

[83] Haidt J (2001): The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review 108(4): 814-834.

[84] Haidt J (2003): The moral emotions. In: RJ Davidson, KR Scherer, HH Goldsmith (eds.): Handbook of affective sciences. Oxford University Press, Oxford: 852-870.

[85] Hall W (2004): Feeling 'better than well'. EMBO Reports 5(12):1105-1109

[86] Haselhuhn MP and BA Mellers (2005): Emotions and cooperation in economic games. Cognitive Brain Research 23: 24-33.

[87] Healy D (2004): Psychopharmacology at the interface between the market and the new biology. In: D Rees, and S Rose: The new brain sciences. Perils and prospects. Cambridge University Press, Cambridge: 232-248.

[88] Heberlein AS, and R Adolphs (2004): Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. Proceedings of the National Academy of Science USA 101(19): 787-7491.

[89] Heeger DJ, AC Huk, WS Geisler, and DG Albrecht (2000): Spikes versus BOLD: what does neuroimaging tell us about neuronal activity. Nature Neuroscience 3(7): 631-633.

[90] Heekeren HR, I Wartenburger, H Schmidt, K Prehn, H-P Schwintowski, and A Villringer (2005): Influence of bodily harm on neural correlates of semantic and moral decision-making. NeuroImage 24: 887-897.

[91] Heekeren HR, I Wartenburger, H Schmidt, H-P Schwintowski, and A Villringer (2003): An fMRI study of simple ethical decision-making. NeuroReport 14(9): 1215-1219.

[92] Henrich J, R Boyd, S Bowles, C Camerer, E Fehr, H Gintis, and R McElreath (2001): In search of Homo Economicus: Experiments in 15 small-scale societies. American Economic Review 91(2): 73-79.

[93] Hilgetag C-C, MA O'Neill, MP Young (1996): Indeterminate organization of the visual system. Science 271: 776.

[94] Hoah H (2003): Remote control. *Nature* 423: 796-799.

[95] Hoyningen-Huene P (1994): Zu Emergenz, Mikro- und Makrodetermination. In: W Lübbe (ed.): Kausalität und Zurechnung. Über Verantwortung in komplexen kulturellen Prozessen. Walter de Gruyter, Berlin, New York.

[96] Hurlbut WB (2002): Empathy, evolution, and altruism. In: SG Post, LG Underwood, JP Schloss, and WB Hurlbut: Altruism & altruistic love. Oxford University Press, Oxford: 309-327.

[97] Illes J, MP Kirschen, and JDE Gabrieli (2003): From neuroimaging to neuroethics. Nature Neurosci. 6(3): 205.

[98] Illes J, and TA Raffin (2002): Neuroethics: a emerging new discipline in the study of brain and cognition. Brain and Cognition 50: 341-344.

[99] Insel TR, and RD Fernald (2004): How the brain processes social information: searching for the social brain. Annual Review of Neuroscience 27: 697-722.

[100] Kandel ER, JH Schwartz, and TM Jessell (2000): Principles of neural science. McGraw-Hill, New York.

[101] Karnath HO, P Thier (Hrsg.) (2003): Neuropsychologie. Springer Verlag, Berlin, Heidelberg, New York.

[102] King-Casas B, D Tomlin, C Anen, CF Camerer, ST Quartz, and PR Montague (2005): Getting to know you: reputation and trust in a two-person economic exchange. Science 308: 78-83.

[103] Kohlberg L (1995): Die Psychologie der Moralentwicklung. Suhrkamp Verlag, Frankfurt a.M.

[104] Kosfeld M, M Heinrichs, PJ Zak, U Fischbacher, and E Fehr (2005): Oxytocin increases trust in humans. Nature 435: 673-676.

[105] Kulynych J (2002): Legal and ethical issues in neuroimaging research: human subjects protection, medical privacy, and the public communication of research results. Brain and Cognition 50: 345-357.

[106] Landrigan C (2001): Preventable deaths and injuries during magnetic resonance imaging. New England Journal of Medicine 345(13): 1000-1001.

[107] LeDoux JE (2000): Emotion circuits in the brain. Annual Review of Neuroscience 23: 155-184.

[108] Lenzen W (2004): Damasios Theorie der Emotionen Facta Philosophica 6: 269-309.

[109] Lieberman MD (2000): Intuition: a social cognitive neuroscience approach. Psychological Bulletin 126(1): 109-137.

[110] Lillehammer H (2003): Debunking morality: evolutionary naturalism and moral error theory. Biology and Philosophy 18: 567-581.

[111] Lima SQ, and G Miesenböck (2005): Remote control of behavior through genetically targeted photostimulation of neurons. Cell 121: 141-152.

[112] Loftus E (2003): Our changeable memories: legal and practical implications. Nature Reviews: Neuroscience 4: 231-234.

[113] Logothetis NK, J Pauls, M Augath, T Trinath, and A Oeltermann (2001): Neurophysiological investigation of the basis of the fMRI signal. Nature 412: 150-157.

[114] McCabe K, D Houser, L Ryan, V Smith, and T Trouard (2001): A functional imaging study of cooperation in two-person reciprocal exchange. Proceedings of the National Academy of Science USA 98(20): 11832-11835.

[115] McCall Smith A (2004): Human action, neuroscience and the law. In: D Rees, and S Rose: The new brain sciences. Perils and prospects. Cambridge University Press, Cambridge: 103-122.

[116] McGaugh JL (2004): The amygdala modulates the consolidation of memories of emotionally arousing experiences. Annual Review of Neuroscience 27: 1-28.

[117] McIntosh AR, SM Fitzpatrick, and KJ Friston (2001): On the marriage of cognition and neuroscience. Neuroimage 14: 1231-1237.

[118] Metzinger T (2005): Neuroethik. Unterwegs zu einem neuen Menschenbild. Gehirn & Geist 11/2005: 50-54.

[119] Moll J, R Zahn, R de Oliveira-Souza, F Krueger, and J Grafman (2005): The neural basis of human moral cognition. Nature Reviews Neuroscience 6: 799-809.

[120] Moll J, R de Oliveira-Souza, and PJ Eslinger (2003): Morals and the human brain: a working model. NeuroReport 14(3): 299-305.

[121] Moll J, R de Oliveira-Souza, PJ Eslinger, IE Bramati, J Mourão-Miranda, PA Andreiuolo, and L Pessoa (2002): The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. The Journal of Neuroscience 22(7): 2730-2736.

[122] Moll J, R de Oliveira-Souza, IE Bramati, and J Grafman (2002): Functional networks in emotional moral and nonmoral social judgments. NeuroImage 16: 696-703.

[123] Montague PR, and GS Berns (2002): Neural economics and the biological substrates of valuation. Neuron 36: 265-284.

[124] Moreno JD (2003): Neuroethics: an agenda for neuroscience and society. Nature Reviews: Neuroscience 4: 149-153.

[125] Mukamel R, H Gelbard, A Arieli, U Hasson, I Fried, and R Malach (2005): Coupling between neuronal firing, field potentials, and fMRI in human auditory cortex. Science 309: 951-954.

[126] Nevado A, MP Young, and S Panzeri (2004): Functional imaging and neural information coding. NeuroImage 21: 1083-1095.

[127] Nichols S (2002): Norms with feelings: towards a psychosocial account of moral judgment. Cognition 84: 221-236.

[128] Nida-Rümelin J (2005): Über menschliche Freiheit. Reclam, Stuttgart.

[129] Organization of Human Brain Mapping, the Governing Council (2001): Neuroimaging databases. Science 292: 1673-1676.

[130] Olson S (2005): Brain scans raise privacy concerns. Science 307: 1548-1550.

[131] Paulus MP (2005): Neurobiology of decision-making: quo vadis? Cognitive Brain Research 23: 2-10.

[132] Pelphrey KA, JP Morris, and G McCarthy (2004): Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. Journal of Cognitive Neuroscience 16(10): 1706-1716.

[133] Phan KL, A Magalhaes, TJ Zielmevicz, DA Fitzgerald, C Green, and W Smith (2005): Neural correlates of telling lies. Academic Radiology 12: 164-172.

[134] Phan KL, T Wager, SF Tayler, and I Liberzon (2002): Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage 16: 331-348.

[135] Phelps EA (2001): Faces and races in the brain. Nature Neuroscience 4(8): 775-776.

[136] Pigliucci M (2003): On the relationship between science and ethics. Zygon 38(4): 871-894.

[137] Preston SD, and FBM de Waal (2002): Empathy: its ultimate and proximate bases. Behavioral and Brain Sciences 25: 1-72.

[138] Preston SD, and FBM De Waal (2002): The communication of emotions and the possibility of empathy in animals. In: SG Post, LG Underwood, JP Schloss, and WB Hurlbut: Altruism & altruistic love. Oxford University Press, Oxford: 284-308.

[139] Rachlin H (2002): Altruism and selfishness. Behavioral and Brain Sciences 25: 239-296.

[140] Racine E, O Bar-Ilan, and J Illes (2005): fMRI in the public eye. Nature Reviews: Neuroscience 6: 159-164.

[141] Rilling JK, DA Gutman, TR Zeh, G Pagnoni, GS Berns, and CD Kilts (2002): A neural basis for social cooperation. Neuron 35: 395-405.

[142] Rizzolatti G, and L Craighero (2004): The mirror-neuron system. Annual Review of Neuroscience 27: 169-192.

[143] Robinson WS (1997): Some nonhuman anilams can have pains in a morally relevant sense. Biology and Philosophy 12: 51-71.

[144] Roskies A (2002): Neuroethics for the new millenium. Neuron 35: 21-23.

[145] Rottschaefer WA (1997): Evolutionary ethics: an irrestisible Temptation: some reflections on Paul Farber's The Temptation of Evolutionary Ethics. Biology and Philosophy 12: 369-384.

[146] Savoy RL (2001): History and future directions of human brain mapping and functional neuroimaging. Acta Psychologica 107: 9-42.

[147] Sellars W (1956): Empiricism and the Philosophy of Mind. Harward University Press, Cambridge, Mass.

[148] Shiv B, G Loewenstein, and A Bechara (2005): The dark side of emotion in decision-making: when individuals with decreased emotional reactions make more advantageous decisions. Cognitive Brain Research 23: 85-92.

[149] Singer W (2005): Selbsterfahrung und neurologische Fremdbeschreibung. Zwei konfliktträchtige Erkenntnisquellen. In: H Schmidinger, und C Sedmak: Der Mensch – ein freies Wesen? Autonomie – Personalität – Verantwortung. Wissenschaftliche Buchgesellschaft, Darmstadt: 135-160.

[150] Singer T, SJ Kiebel, JS Winston, RJ Dolan, and CD Frith (2004): Brain responses to the acquired moral status of faces. Neuron 41: 653-662.

[151] Spektrum Akademischer Verlag (2001): Lexikon der Neurowissenschaft in vier Bänden. Heidelberg, Berlin.

[152] Stark CEL, LR Squire (2001): When zero is not zero: The problem of ambiguous baseline conditions in fMRI. Proceedings of the National Academy of Sciences USA 98(22):12760-12766.

[153] Stoop R, and N Stoop (2004): Computation by natural systems defined. *Proceedings of ISCAS* 5: 664-667.

[154] Takahashi H, N Yahata, M Koeda, T Matsuda, K Asai, and Y Okubo (2004): Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. NeuroImage 23: 967-974.

[155] Talwar SK, S Xu, ES Hawley, SA Weiss, KA Moxon, and JK Chapin (2002): Behavioural neuroscience: rat navigation guided by remote control. Nature 417: 37-38.

[156] The Economist (2002): The future of mind control / Open your mind. Issue of May 23th.

[157] Thompson JK, MR Peterson, and RD Freeman (2003): Single-neuron activity and tissue oxygenation in the cerebral cortex. Science 299: 1070-1072.

[158] Turnbull OH, CEY Evans, A Brunce, B Carzolio, and J O'Connor (2005): Emotion-based learning and central executive resources: An investigation of intuition and the Iowa gambling task. Brain and Cognition 57: 244-247.

[159] Uttal WR (2001): The new phrenology. The limits of localizing cognitive processes in the brain. MIT Press, Cambridge.

[160] Vaas R (2000): Emotionen. In: Spektrum Akademischer Verlag: Lexikon der Neurowissenschaft in vier Bänden. Heidelberg, Berlin.

[161] Vreeke GJ, and IL van der Mark (2003): Empathy, an integrative model. New Ideas in Psychology 21: 177-207.

[162] Walter H (1999): Neurophilosophie der Willensfreiheit. Mentis, Paderborn.

[163] Whalen PJ, SL Rauch, NL Etcoff, SC McInerney, MB Lee, and MA Jenike (1998): Masked presentations of emotional face expressions modulate amygdala activity without explicit knowledge. The Journal of Neuroscience 18(1): 411-418.

[164] Wild J (2005): Brain imaging ready to detect terrorists, say neuroscientists. Nature 437: 457.

[165] Winston JS, BA Strange, J O'Doherty, and RJ Dolan (2002): Automatic and intentional brain response during evaluation of trustworthiness of faces. Nature Neuroscience 5(3): 277-283.

[166] Wolpe PR (2004): Ethics and social policy in research on the neuroscience of human sexuality. Nature Neuroscience 7(10): 1031-1033.

Index