M A N U S K R I P T - Fassung vom 09. April 2009
189’000 Zeichen

Naturalisierung von Moral?

Einschätzung des Beitrags der Neurowissenschaft zum Verständnis moralischer Orientierung[1]

1. Einführung

In der Einführung werden Zielsetzung und Methoden dieser Studie vorgestellt. Zudem werden begriffliche Grundlagen eingeführt, die bei der Analyse der Arbeiten aus dem Bereich der so genannten *neuroscience of ethics* (Roskies 2002) eine Rolle spielen. Thematisiert werden insbesondere der Begriff der *moral agency* und die Struktur ihrer experimentellen Untersuchung.

1.1 Zielsetzung

Welche Rolle spielt die „Intuition“ und das „Gefühl“ bei der Bildung und dem Gebrauch der moralischen Urteilsfähigkeit beim Menschen? Diese Frage verweist auf ein altes Grundproblem der philosophischen Ethik, die von David Hume in seiner Untersuchung der Grundlagen der Moral (Hume 1751) eingängig formuliert wurde, nämlich „ob Moral aus Vernunft abgeleitet sei oder vom Gefühl“. Vorab der Bezug zum Gefühl eröffnet ein Einfallstor für eine empirische Untersuchung der Moralfähigkeit des Menschen und berührt aktuelle Forschungsgegenstände der Neurowissenschaft. Diese Arbeit soll aufzeigen, auf welche Weise die „biologischen Grundlagen“ der Moral im Rahmen der so genannten *neuroscience of ethics* (Roskies 2002, später begann man auch von *moral neuroscience* zu sprechen) zum Gegenstand empirischer Untersuchungen geworden sind, welche Ergebnisse bislang erzielt worden sind und wie diese einzuschätzen sind.

Ausgangspunkt bildet eine umfassende Literaturrecherche, die den aktuellen Stand der Neurowissenschaft bezüglich des Verständnisses moralischer Orientierung wiedergibt. Basierend darauf werden Kernpunkte und Motive dieses Programms, sowie mögliche Probleme des empirischen Ansatzes aus der Perspektive der philosophischen Ethik herausgearbeitet. Eine Einführung in die Konzepten und Methoden, die in diesen Forschungsgebieten verwendet werden, soll den Einstieg in die Thematik erleichtern. Die Studie liefert damit eine aktuelle Übersicht zu einem Forschungsgebiet, das vermehrt auf das Interesse der Philosophie stößt.

[1] Dieser Text ist eine aktualisierte Fassung der Pilotstudie „Abklärung des möglichen Beitrags der Neurowissenschaft und der Verhaltensforschung zum Verständnis moralischer Orientierung (Universität Zürich, 2006).

1.2 Methoden

Die Studie stützt sich auf eine Literaturrecherche[2] und eine Expertenbefragung[3]. Aus historischer Perspektive ist anzumerken, dass empirische Aspekte der Moralfähigkeit des Menschen natürlich schon früher in verschiedenen Gebieten untersucht worden sind (beispielsweise in der so genannten „evolutionären Ethik"). Die Untersuchung neuronaler Korrelate moralischen Verhaltens und die darauf aufbauende Theoriebildung im Rahmen einer *neuroscience of ethics* ist jedoch vergleichsweise jung und hängt mit der Renaissance der Emotionen in der Hirnforschung ab den 1990er Jahren sowie den neuen Möglichkeiten der bildgebenden Verfahren zusammen. Die große Menge zu beachtender Literatur wurde wie folgt eingeschränkt:

- Die Untersuchung der neuronalen Korrelate von moralischem Verhalten mittels bildgebender Verfahren (*Imaging*) ist ein neues und überschaubares Forschungsgebiet, das weitgehend abgedeckt wird. Da die Beurteilung der erzielten Resultate auch von methodischen Aspekten abhängt, werden Verfahren und Grundprobleme des *Imaging* und weiterer verwendeter Methoden kurz aufgezeigt.

- Die Untersuchung „moralischer Pathologien" ist mindestens seit dem 19. Jahrhundert ein Thema. Gemeint ist damit die In-Bezug-Setzung bzw. Erklärung von moralisch abnormem Verhalten mit bestimmten Hirnschäden, was heute insbesondere in der Neuropsychologie ein Thema ist. Dieses große Gebiet kann nicht umfassend abgehandelt werden. 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* und weiterer Aspekte des menschlichen Sozialverhaltens hat in den vergangenen Jahren im Rahmen der *social neuroscience* einen enormen Aufschwung erfahren. Dieses Forschungsgebiet kann nur in seinen Grundzügen skizziert werden, indem Arbeiten (u.a. mit dem Einbezug der experimentellen Ökonomie) untersucht wurden, in denen sich explizit ein Bezug zu *moral decision making* finden ließ. Arbeiten zum Themenkomplex Neurowissenschaft und Willensfreiheit und die daraus sich ergebenden möglichen Konsequenzen für das Recht werden hier nicht behandelt.

---

[2] Die Literaturrecherche verlief in zwei Phasen: Im ersten Schritt (Pilotstudie) wurden folgende Zeitschriften im Zeitraum 2000 bis 2005 durchgängig auf relevante Arbeiten durchsucht: Annual Review of Neuroscience, Behavioral and Brain Sciences, Behavioral Neuroscience, Biology and Philosophy, Current Opinion in Neurobiology, Human Brain Mapping, International Journal for Psychophysiology, Journal of Cognitive Neuroscience, Journal of the History of Behavioral Sciences, Nature, Nature Neuroscience, Nature Reviews: Neuroscience, NeuroImage, Neuron, Proceedings of the National Academy of Sciences USA, Progress in Neurobiology, Science, The Journal of Neuroscience, Trends in Cognitive Sciences, Trends in Neurosciences. Ergänzt wurde diese Suche mit einer Stichwortsuche in PubMed. In einem zweiten Schritt wurden im Zeitraum 2006-2008 jene Zeitschriften obiger Liste umfassend untersucht, in welchen Arbeiten aus dem Umfeld der *neuroscience of ethics* erschienen sind, wiederum ergänzt mit einer Stichwortsuche in PubMed sowie in der Datenbank SCI expanded (vgl. dazu Fussnote 21) sowie einer Analyse der erst jüngst gegründeten Zeitschriften im Bereich *social neuroscience*.

[3] Im Rahmen der Pilotstudie wurden folgende Personen befragt: Peter Brugger (Neuropsychologie, Universitätsspital Zürich), Urs Fischbacher (experimentelle Ökonomie, Universität Zürich), Alumit Ishai (Neuroimaging, Universität Zürich), Lutz Jänke (Neuropsychologie, Universität Zürich), Daniel Kiper (Sinnesphysiologie, Universität Zürich), Eric Kubli (Zoologie, Universität Zürich), Kevan Martin (Neurophysiologie, Universität Zürich), Marianne Regard (Neuropsychologie, UniversitätsSpital Zürich), Anton Valavanis (Neuroradiologie, Universitätsspital Zürich) und Carel van Schaik (Anthropologie, Universität Zürich). Einem Teil dieser Personen (Brugger, Fischbacher, Jänke, Regard, Valavanis, van Schaik) wurde 2005 der Entwurf der Pilotstudie zwecks Prüfung und weiterer Kommentierung zugestellt. Die Aktualisierungen der 2006 abgeschlossenen Studie erfolgten im Rahmen des SNF-Projektes „Die neurobiologische Untersuchung des *moral agent*: Eine Spezifizierung aus philosophischer und kulturwissenschaftlicher Perspektive".

- Die Emotionsforschung hat sich ebenfalls zu einem schwer überschaubaren Gebiet entwickelt, das sich mit mehreren klassischen Disziplinen überlappt (Psychologie, Neurowissenschaft, Philosophie). In der vorliegenden Studie wurden vorab ausgewählte Arbeiten zu so genannten „moralischen Emotionen" untersucht und solche, die einen prägenden Einfluss vorbewusstlicher, emotions-gesteuerter Prozesse auf (vermeintlich rationale) moralische Entscheidungen postulieren.

- Die in den Verhaltenswissenschaften (und anderswo) diskutierte Frage nach den evolutionären Ursprüngen von Moral sowie nach „Vorformen" von Moral in Tiergemeinschaften – vorab Primaten – wird hier nicht behandelt, obgleich diese Themen in Arbeiten zu neurobiologischen Grundlagen der Moral regelmäßig zitiert und in die Theoriebildung eingebracht werden (siehe dazu den Beitrag von Carel van Schaik und Claudia Rudolf von Rohr in diesem Band).

1.3 Begriffliche Grundlagen

1.3.1 Moral und Ethik

Die Literaturrecherche hat ergeben, dass die Unschärfe hinsichtlich der Verwendung von Begriffen wie „Moral/moralisch", „Intuition", „Gefühl/Emotion", etc. teilweise erheblich ist. Insbesondere werden in vielen der untersuchten naturwissenschaftlichen Arbeiten „Ethik" (*ethics*) und „Moral" (*moral*) synonym verwendet, was mit einem unterschiedlichen Gebrauch dieser Begriffe im englischen Sprachraum zusammenhängen mag oder auch mit einer fehlenden Differenzierung dieser geisteswissenschaftlichen Begriffe in naturwissenschaftlichen Arbeiten. Dieser Abschnitt soll deshalb eine Übersicht zu zentralen Begriffen der philosophischen Ethik geben, die durch die untersuchten Arbeiten angesprochen werden.[4]

In dieser Studie (wie auch üblicherweise in der deutschsprachigen ethischen Literatur) werden „Moral" und „Ethik" wie folgt unterschieden: ersterer bezeichnet eine bestimmte Gesamtheit normativ geprägter Sachverhalte oder Verhaltensmuster, wobei aus der Binnenperspektive eines jeweiligen Moralsystems der Begriff „moralisch" gleichbedeutend mit „gut" oder „gefordert" verwendet wird. Letzterer bezeichnet das wissenschaftliche Gebiet, das Moral zum Gegenstand hat, wobei beispielsweise die Moralität einer gewissen Norm eines bestimmten Moralsystems angezweifelt werden kann. Ethik wird gemeinhin in die Bereiche deskriptive Ethik, normative Ethik und Metaethik wie folgt unterteilt:

- Aufgabe der **deskriptiven Ethik** ist es, einen strukturierten Überblick über ein Moralsystem[5] 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 (z.B. implizite und explizite Normen, Werte) und Verhaltensmuster für ein bestimmtes System miteinander interagierender moralischer Agenten (*agent*). Der Begriff „moralischer Agent" (*moral agent*, im Deutschen 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.

[4] Den nachfolgenden Erläuterungen liegen die Beiträge des „Handbuch Ethik" (Düwell et al. 2002) zugrunde.
[5] Der Begriff „Moralsystem" wird hier in einem weiten Sinn verstanden und umfasst die in einer (abgrenzbaren) Gesellschaft oder Gruppe geltenden Normen, Regeln und Verhaltensmuster mit moralischem Charakter. Dabei sind zwei Abgrenzungsleistungen nötig: einerseits zur Umgrenzung der Gruppe, andererseits zur Unterscheidung moralisch vs. nicht moralisch.

Welche Eigenschaften dies sind, ist Gegenstand von Debatten, die durchaus vergleichbar sind mit der Suche nach Kriterien für die Bestimmung des Begriffs „Person“. Deskriptive Ethik kann einhergehen mit (empirischer) Theoriebildung über die Funktion von Moral – so etwa die Idee, ein Moralsystem resultiere als „evolutionär gewachsene Lösung“ der sozialen Probleme, welche die Interaktion von *moral agents* mit sich bringen. 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, wobei man ein konkretes Begründungssystem für ein gegebenes Moralsystem ebenfalls als eine *bestimmte* normative Ethik bezeichnet. Innerhalb des Felds der normativen Ethik lassen sich zwei unterschiedliche Forschungsbereiche unterscheiden. Zum einen geht es darum, Begründungsstrategien für Moralsysteme zu entwickeln und zu verteidigen. Bekannte Strategien sind teleologische und deontologische Ethiken, sowie sogenannt schwach normative bzw. kontextualistische Ansätze. Zum anderen werden – im Sinn einer Angewandten Ethik und basierend auf solchen Begründungsstrategien – Moralsysteme für konkrete soziale und politische Probleme ausgearbeitet oder kritisch untersucht mit dem Ziel, Grundlagen für (meist politische) Entscheidungen zu liefern. Bekannte Anwendungsgebiete dieser so genannten Bereichsethiken sind die Bioethik, die Medizinethik, die Umweltethik, die Technikethik und die Wissenschaftsethik.

- Die **Metaethik** untersucht Fragen, die sich bei der Konstruktion und Begründung normativer Ethiken ergeben. Diese Fragen lassen sich vier Bereichen zuordnen: Erstens können sprachphilosophische Aspekte, beispielsweise hinsichtlich der Bedeutung normativer Aussagen, untersucht werden. Zweitens 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 können epistemologische Fragen hinsichtlich der Rechtfertigung und Begründbarkeit moralischer Urteile untersucht werden. Eine scharfe Grenzziehung zwischen normativer Ethik und Metaethik ist nicht immer möglich, da etwa das Finden von Begründungsstrategien für eine bestimmte normative Ethik auch die Beantwortung metaethischer Fragen mit einschließen kann.

Aus dieser Dreiteilung ergeben sich Anforderungen an empirische Untersuchungen über neurobiologische Grundlagen der Moral. Die erste (empirische) Anforderung zielt darauf, dass solche Untersuchungen eine präzise Vorstellung davon vermitteln müssen, *welchen* moralischen Sachverhalt sie untersuchen wollen und in welchen größeren Kontext dieser zu stellen ist. Dies bedeutet insbesondere die Spezifikation des moralischen Stimulus für bestimmte Experimente oder die Operationalisierung einer moralischen Handlung, so dass diese quantitativ erfassbar wird.

Die zweite (normative) Anforderung betrifft die Frage, in welchem Sinn derartige Untersuchungen Argumente für oder gegen *bestimmte* normative Ethiken liefern können. Unbestritten ist, dass empirische Daten (beispielsweise ermittelt durch soziologische Studien) über die praktische Umsetzbarkeit bestimmter normativer Theorien gewonnen werden können. Fraglich ist aber, inwieweit damit eine bestimmte normative Ethik gegenüber anderen ausgezeichnet werden kann (siehe dazu Abschnitt 3.1). Dieser Argumentationsschritt lässt sich nicht rein empirisch absichern.

Die dritte (metaethische) Anforderung verlangt, dass solche Untersuchungen Stellung dazu beziehen sollten, inwieweit sie sich als Teil des Projektes der Naturalisierung von Moral verstehen. Gemeint ist mit „Naturalisierung“ die Auffassung, dass moralische Tatsachen natürliche Tatsachen seien bzw. in einem zu bestimmenden Sinn durch natürliche Tatsachen konstituiert seien. Die Diskussion der Frage, ob ein solches Projekt erfolgversprechend ist, hat in der Ethik eine lange Tradition – erinnert sei an die Debatte um den naturalistischen Fehlschluss bzw. um das Moore'sche Argument der offenen Frage (siehe z.B. Düwell et al. 2002: Kapitel 3). Einige der in dieser Studie untersuchten Arbeiten nennen explizit das Ziel einer Naturalisierung von Moral, doch die meisten Autoren vertreten die (weniger kontroverse) Ansicht, 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.

1.3.2 Moral agency

Fokus der weitaus meisten Arbeiten über die neurobiologischen Grundlagen der Moral ist das (moralisch) handelnde Subjekt. Sie leisten demnach einen Beitrag zur theoretischen Erfassung von *moral agency*. Diese „*moral agency*“ wird hier so verstanden, dass ein umfassendes Verständnis des folgenden Satzes erreicht werden soll:

> *„X ist aufgrund der Fähigkeiten {Y_n} im Kontext K ein moral agent.“*

Dieser Satz zeigt die Komplexität des Unterfangens, eine Theorie von *moral agency* zu entwickeln, indem jeder unterstrichene Satzteil auf miteinander verknüpfte Folgefragen verweist:

- X: Welche Entität X soll (potenziell) als *moral agent* klassifiziert werden können? Diese Frage hat eine ontogenetische (ist z.B. ein Säugling auch schon ein *moral agent*?), phylogenetische (sind gewisse Primaten *moral agents*?), pathologische (sind z.B. Demenzkranke noch *moral agents*?) und quantitative Dimension (sind Kollektive *moral agents*?).

- Fähigkeiten: Welche Fähigkeiten {Y_n} sind als notwendig bzw. hinreichend zu bezeichnen? Inwieweit sind dies Fähigkeiten, über die X bewusst verfügt bzw. die er (oder andere, z.B. Moralerziehung) entwickeln kann? Inwieweit sind gewisse Fähigkeiten durch andere kompensierbar? Was ist der biologische Marker dieser Fähigkeiten?

- Kontext: Wie sehen die strukturellen und zeitlichen Komponenten des Kontextes aus, in denen der *moral agent* agiert? Zu thematisieren sind beispielsweise der Zwangscharakter bestimmter Handlungssituationen oder die Genese eines bestimmten Handlungskontextes.

- *moral*: Wodurch ist die Moralität der jeweiligen Handlungssituation charakterisiert (z.B. welche Normen sind involviert?)? Inwieweit muss X über diese Charakteristika verfügen können? Welche Begründungsstärke verlangt die jeweilige Handlung von X? Inwieweit ist „normative Kreativität“ erlaubt – also das Einbringen neuer moralischer Charakteristika (z.B. Normen) durch X?

- *agency*: Wann wird ein Verhalten zu einer Handlung? Welchen Stellenwert haben hier Konzepte wie Willensfreiheit?

Zweifellos kann die Vielzahl dieser Fragen nie gemeinsam angegangen werden. Gerade empirische Untersuchungen fixieren einige dieser „Parameter", um andere zu bestimmen. In der Regel werden *X* (gesunde Menschen oder Personen mit spezifischen Hirnschäden als Versuchspersonen) und *moral* (z.B. durch die Festlegung des moralischen Stimulus) festgelegt, während $\{Y_n\}$ und *agent* (also z.B. das konkrete Verhalten) quasi die „Messvariablen" sind. Die Kontexte werden je nach Studie definiert oder thematisiert.

1.3.3 Struktur moralischer Experimente

Die Struktur der meisten Experimente der empirischen Moralforschung (abgesehen von reinen Beobachtungsstudien) folgen dem *Stimulus-Response*-Schema, das sich im Fall der Untersuchung von *moral agency* wie folgt unterteilen lässt: 1) in eine handlungsauslösende Komponente; 2) in eine Komponente des *decision making*; 3) in die dadurch verursachten Handlungen; 4) in die Kriterien für die Sicherung bzw. Stützung der moralischen Gültigkeit der Handlung. Diese vier Bereiche lassen sich wie folgt charakterisieren:

- **Stimulus:** Die erste Komponente ist ein für den betreffenden *moral agent* sinnlich erfahrbares raumzeitliches Ereignis – im Kontext eines Experiments beispielsweise ein moralischer Stimulus.

- ***Decision making:*** Die zweite Komponente lässt sich in einen intentionalen und einen automatisch ablaufenden Prozess eines *decision making* trennen (es wird der englische Ausdruck verwendet, weil das deutsche „entscheiden" nur den intentionalen Aspekt umfasst). Die genaue Charakterisierung und Unterscheidung dieser Prozesse ist Gegenstand von Untersuchungen. Die Prozesse werden durch Prädispositionen phylogenetischen oder ontogenetischen Ursprungs, die dem *moral agent* in unterschiedlichem Masse zugänglich sind, beeinflusst und schränken die Bandbreite der möglichen Handlungen als Folge der sinnlichen Erfassung des moralischen Stimulus ein. Zusammenfassend werden diese einschränkenden Faktoren hier „handlungsleitende Faktoren" genannt. Auf der Wahrnehmungsebene umfassen diese beispielsweise Prozesse wie „Filterung" oder „Färbung" von Stimuli. Im Rahmen einer naturwissenschaftlichen Erfassung des Phänomens geht man davon aus, dass diese Prozesse zumindest prinzipiell messbar sind – etwa im Sinn eines charakteristischen Erregungsmusters im Gehirn des betreffenden *agent*, das sich nach einem moralischen Stimulus aufbaut.

- **Handlung:** Die dritte Komponente umfasst das Wirken des Agenten in der Raumzeit als Folge der handlungsauslösenden Komponenten und des *decision making*. Die zentrale Messvariable der empirischen Moralforschung ist in der Regel eine Handlung[6] – sei dies ein tatsächlich körperliches Agieren oder ein Sprachhandeln, etwa im Sinn einer beurteilenden Aussage. Dies ist nicht weiter erstaunlich, da die reine, nicht gegen außen kommunizierte Introspektion über ein als moralisch erkanntes Problem durch ein bestimmtes Individuum nicht Gegenstand einer empirischen Untersuchung sein kann. Auch die Bildgebung misst nicht die Introspektion, sondern die biologischen Prozesse des *decision making*. Hier muss angefügt werden, dass das Kundtun einer

[6] Empirische Studien verwenden oft auch den Begriff „*moral behavior*" (moralisches Verhalten), womit die in der Ethik vertretene und insbesondere für moralische Kontexte wichtige Trennung zwischen Handlung (als intentionales, von Gründen geleitetes Tun) und Verhalten (als ein Reagieren aus äussere Einflüsse) quasi unterminiert wird. Dies ist Ausdruck der unterschiedlichen Perspektiven auf das Problem: während die Ethik nur Handlungen als „moralisch" auszeichnen will, interessiert sich die empirische Forschung für jene Faktoren, die den „Verhaltensaspekt" moralischer Handlungen ausmachen, und sieht das Tun des *moral agent* demnach nicht als moralische Handlung im strengen Sinn an.

Entscheidung in einem experimentellen Kontext (sei dies verbal oder mittels Knopfdruck), ebenfalls als Handlung zu verstehen ist.[7] 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.

- **Begründung:** Die vierte Komponente betrifft den Komplex von Begründungen, welche der betreffende *moral agent* für die Rechtfertigung seiner Handlung angibt (bzw. angeben würde oder könnte). In der philosophischen Ethik basiert ein moralisches *decision making* auf dem selbstständigen Verhalten im logischen Raum der Gründe – ein von Sellars geprägter Begriff (Sellars 1956). Hier würde man also die vierte Komponente mit einem intentionalen *decision making* identifizieren. In diesem Fall vertritt man eine rationalistische Auffassung von moralischem Handeln, gemäß der es die einem *moral agent* bewusst zugänglichen Gründe sind, die zu einer moralisch (guten) Handlung führen. Eine Gegenposition ist, dass vorab unbewusste, automatisierte Vorgänge eine moralische Handlung leiten und Rechtfertigungen erst *post facto* (falls die Situation dies verlangt) durch den betreffenden Agenten erzeugt werden (in Form einer Sequenz von Aussagen für die Rechtfertigung der Handlung gegenüber Anderen). Empirisch wird diese vierte Komponente mit Fragebögen, Interviews etc. erfasst.

Eine empirische Untersuchung der neurobiologischen Grundlagen von Moral müsste grundsätzlich in der Lage sein, alle vier Komponenten genau zu spezifizieren und deutlich zu machen, wie man diese quantifizieren will. In den untersuchten Studien geschieht dies in dieser Umfassendheit oft nicht, wie nachfolgend deutlich wird. Bei der Messbarkeit einiger der oben erwähnten Komponenten stellt sich zudem vom Standpunkt der philosophischen Ethik ein grundlegendes Problem – vor allem bei der vierten Komponente. Auch wenn man davon ausgeht, dass der Prozess des Suchens, Bewertens und Abwägens von Gründen auf bestimmten neuronalen Prozessen beruht, ist es unklar, aufgrund welcher Kriterien man derart komplexe psychische Entitäten mit den messbaren physiologischen Entitäten in Beziehung setzen will (siehe dazu den folgenden Abschnitt 2.2.3). Dies mag ein Grund dafür sein, warum in der empirischen Moralforschung das Modell entwickelt wurde, wonach 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.

1.3.4 Raum- und Zeitskalen von moral agency

Moral agency entfaltet sich in Raum und Zeit auf unterschiedlichen, mehr oder weniger gut unterscheidbaren Skalen. In räumlicher Hinsicht ist folgende Unterscheidung gut etabliert (vgl. auch mit Abbildung 1):

[7] Die Unterscheidung zwischen den Begriffen „Entscheidung“ und „Handlung“ – beispielsweise im Sinn, dass nur jene Akte als Handlungen aufgefasst werden sollen, denen eine Entscheidung vorangeht (Nida Ruemelin 2005) – soll damit nicht untergraben werden. Diese Beobachtung soll lediglich darauf hinweisen, dass „moralische Experimente“ im Regelfall keine komplexe Handungsstruktur (Handlungssequenzen etc.) kennen, sondern beispielsweise das Kundtun einer Beurteilung eines moralischen Dilemmas oder der Spielzug in einem Ultimatumspiel als „Handung“ gilt. Handlungstheoretische Erwägungen bieten hier sicher Ansatzpunkte für Kritik, was in dieser Studie aber nicht weiter verfolgt wurde.

- **Individuum:** Erstens kann ein einzelner *moral agent* Gegenstand der Untersuchung sein. Dies ist das übliche Vorgehen bei *Imaging* Experimenten, in welchen eine Versuchsperson beispielsweise mit moralischen Dilemmas konfrontiert wird. Stimuli werden in solchen Experimenten praktisch immer visuell präsentiert. Handlungsleitende Faktoren in diesem Kontext nennt man auf der intentionalen Ebene Überzeugungen oder Präferenzen, die in einem größeren Begründungskontext stehen.

- **Gruppe:** Zweitens kann eine Gruppe direkt interagierender *moral agents* Gegenstand der Untersuchung sein. „Direkt interagieren" bedeutet dabei, dass sich die beteiligten Partner über längere Zeiträume regelmäßig begegnen, so dass Beziehungen entstehen, die einzelnen *agents* einen Ruf erwerben und Vorstellungen über die jeweils anderen *agents* entwickeln können – Abgrenzungskriterium gegenüber der dritten räumlichen Skala ist demnach die Art der Beziehung der *agents* untereinander. Ein Grundproblem solcher Kleingruppen ist beispielsweise die Verteilung eines öffentlichen Guts, wobei keine zentralisierten Strukturen bestehen und die einzelnen *agents* keinen durch soziale Institutionen definierten Status haben (Bowles & Gintis 2004). Handlungsleitenden Faktoren auf dieser Ebene sind beispielsweise Gruppennormen, die durch Belohnungen und/oder Bestrafungen gestützt werden.

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

ABBILDUNG 1:

Abbildung 1: Räumliche Skalen von *moral agency:* a) Interaktion des Individuums mit sich selbst. b) (Wiederholte) Interaktion in der Gruppe. c) Interaktion via Institutionen.

Dies ist eine idealtypische Unterscheidung insofern, als die den einzelnen Ebenen zugeordneten, handlungsleitenden Faktoren auch auf anderen Ebenen wirken. Individuelle Überzeugungen ändern im Zug der Interaktion in Kleingruppen, Gruppennormen können institutionelle Regeln unterwandern. All dies geschieht in der Zeit, wobei hier üblicherweise ebenfalls drei Zeitskalen unterschieden werden:

- **Zeitskala der unmittelbaren Handlung:** Sind *agents* mit einem (moralischen) Problem konfrontiert, kann dies eine vergleichsweise rasche Handlung oder Stellungnahme erfordern (Sekunden bis Minuten, evt. Stunden). Die meisten Experimente der empirischen Moralforschung sind auf dieser Skala angesiedelt.

- **Ontogenetische Zeitskala:** *Moral agents* entwickeln Vorstellungen von sich und Reputationen gegenüber anderen *agents*, mit denen sie interagieren. Die damit verbundenen Prozesse (darunter auch Lernen) verlaufen auf einer Zeitskala, die von der Größenordnung Wochen bis Jahre umfasst. Die Pädagogik interessiert sich beispielsweise für die Genese relevanter Eigenschaften von *moral agency* auf einer Zeitskala von 10 bis 20 Jahren (ein Beispiel ist das Stufenmodell von Kohlberg, 1995).

- **Phylogenetische Zeitskala:** Gewisse Entitäten von Moralsystemen (z.B. Werte) überdauern die Existenz einzelner *moral agents*. Entsprechend wird von einer phylogeneti-

schen Zeitskala gesprochen, die – je nach Forschungsgebiet – noch feiner unterteilt werden kann. Die Politikwissenschaften beispielsweise sind an der Frage interessiert, wie sich handlungsleitende Faktoren auf der Ebene der Interaktion von Institutionen auf der Zeitskala von mehreren Generationen verändern. Die evolutionsbiologische Perspektive interessiert sich für die Frage nach der Entstehung von Normensystemen über viele Jahrtausende (das Themengebiet der evolutionären Ethik).

Auch hier sind Überschneidungen vorhanden. So können beispielsweise Einzelereignisse Reputationen dauerhaft verändern. Die neurobiologisch inspirierte empirische Moralforschung agiert meistens auf der ersten Zeitskala, wenngleich natürlich Erkenntnisse anderer Bereiche in die Theoriebildung Eingang finden.

1.3.5 Ein Arbeitsmodell des moral agent

Obige Differenzierungen zeigen deutlich die Komplexität, der sich eine empirische Untersuchung von Moral stellen muss. Dass eine *neuroscience of ethics* hier Vereinfachungen machen muss, kann ihr nicht vorgeworfen werden. Bedeutsam wird nachfolgend sein, welche Folgen diese Vereinfachungen auf das Verständnis von *moral agency* haben werden.

Die wichtigsten Themen, die nachfolgend untersucht werden sollen, sind in Abbildung 2 zusammengefasst. Dieses Arbeitsmodell eines *moral agent* zeigt die zentralen Punkte, denen sich die *neuroscience of ethics* bedient, zu welchen sie Beiträge liefern will und die als Probleme identifiziert werden können. Die Stichworte „moralischer Stimulus“ und „moralische Handlung“ beziehen sich auf das experimentelle Setting, denen ein *moral agent* unterworfen wird. Das Stichwort „*decision making*“ bezeichnet das zentrale wissenschaftliche Interesse dieses Gebietes, wobei insbesondere die Beziehung zwischen bewussten bzw. kognitiven und unbewussten Prozessen und deren biologischen Träger fokussiert werden. Die Stichworte „Selbstbild“ und „Raum der Gründe“ bezeichnen jene Aspekte von *moral agency*, auf die insbesondere die philosophische Kritik zielt, wenn sie die empirische Untersuchung von *moral agency* untersucht. Unter dem Selbstbild soll dabei das moralische Bild des *moral agent* von sich selbst verstanden werden, dem dieser sich auch verpflichtet fühlt – also nicht nur die Reputation des *agent* gegenüber anderen *agents* (hier dürften klassische moralische Begriffe wie „Gewissen“ und „Ehre“ lokalisiert werden). Unter dem „Raum der Gründe“ wird das moralische Begründungssystem verstanden, das der *moral agent* im Laufe seines Lebens erwirbt und dessen er sich bedient, wenn er beispielsweise zu einer Stellungnahme aufgefordert wird. Zu beiden Begriffen lässt sich zweifellos mehr sagen, was hier nicht geschehen kann (siehe auch die anderen Beiträge in diesem Band).

ABBILDUNG 2:

Abbildung 2: Ein Arbeitsmodell eines *moral agent*, das die Struktur moralischer Experimente mit den Komponenten moralischer Stimulus, *decision making* (dunkelgrau die unbewussten Prozesse, hellgrau die bewussten Prozesse, mit unscharfer Trennlinie), moralische Handlung und Begründung (Selbstbild und Raum der Gründe, siehe Text) aufgreift.

2. Methoden und wissenschaftliche Einbettung der *neuroscience of ethics*

Dieser Abschnitt liefert Hintergrundinformationen, um die Resultate der untersuchten Arbeiten aus der *neuroscience of ethics* besser beurteilen zu können. Zum einen werden die wichtigsten Definitionen aus der Neuroanatomie sowie zentrale neurowissenschaftliche Methoden (insbesondere die funktionelle Magnet-Resonanz-Tomographie) vorgestellt. Zum anderen wird mittels bibliometrischen Methoden aufgezeigt, welchen Stellenwert die Untersuchung von moralischem Verhalten in der empirischen Forschung generell hat, und wie das Projekt der *neuroscience of ethics* in vergleichbare naturwissenschaftliche Projekte eingeordnet werden kann.

2.1 Anatomische und methodische Grundlagen des neurowissenschaftlichen Zugangs zu *moral agency*

2.2.1 Hirnanatomie

Viele der untersuchten Arbeiten streben (auch) eine Lokalisation der neuronalen Prozesse an, die bei moralischen Entscheidungen bzw. Handlungen aktiv sind. Deshalb soll hier eine kurze Einführung in die Anatomie des menschlichen Gehirns gegeben werden (vgl. dazu Greenstein & Greenstein 2000, Kandel et al. 2000).

Für die 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 ersten steht, werden Orte nahe dem Nullpunkt als medial bezeichnet, Orte fern des Nullpunkts als lateral. Auf das menschliche Gehirn bezogen (man stelle sich das Gehirn in der Position eines vor einem sitzenden Menschen vor), liegen die vorderen Hirnregionen (Stirnbereich) rostral, die hinteren Regionen (Hinterkopf) caudal. Der obere Hirnbereich (oberes Kopfende) liegt dorsal und die unteren Regionen (gegen den Rachen hin) liegen ventral. Die mittlere Hirnregion schließlich liegt medial, während die Seitenbereiche (links oder rechts) lateral liegen. Für die Schnittebenen (z.B. beim *Imaging*) werden folgende Bezeichnungen verwendet. 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: 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, das 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 umfassen unter anderem die so genannten Basalganglien (u.a. den *nucleus caudatus*) und die Amygdala. Beim 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 und Glia-

zellen und weist eine Schichtenstruktur auf (sechs Schichten, mit unterschiedlicher Ausprägung je nach Ort), letztere besteht primär aus myelinisierten[8] Axonen von Nervenzellen, welche die verschiedenen Bereiche des Kortex miteinander verbinden. Eine weitere grundlegende Differenzierung ist 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: Den Frontallappen (rostral), den Occipitallappen (caudal), den Temporallappen (lateral) und den Parietallappen.

Die kortikale Anatomie lässt sich 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) heißen. Die wichtigsten Sulci trennen die einzelnen Lappen des Kortex: Der zentrale Sulcus trennt den Frontallappen vom Parietallappen, der laterale Sulcus den Temporallappen vom Parietal- und Frontallappen und 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 insulare Kortex (*Insula*) und das Cingulum. Im Verlauf einer sich über viele Jahrzehnte erstreckenden Lokalisationsforschung hat man einzelnen Regionen des Kortex (Rindenfelder) bestimmte Funktionen zuordnen können, wobei insbesondere die Regionen, welche Informationen der Sinnesorgane aufnehmen (sensorische Rindenfelder) oder Signale an das Bewegungssystem abgeben (motorische Rindenfelder) gut erforscht sind. Schwieriger ist die Lokalisation „höherer" Leistungen des zentralen Nervensystems.

Im Rahmen der Erforschung der „Neuroanatomie der Moral" sind in den für diese Studie untersuchten Arbeiten zahlreiche Regionen genauer untersucht worden (siehe dazu Abbildung 3). Dabei muss auf das Problem hingewiesen werden, dass dieselben Hirnregionen in unterschiedlichen neurowissenschaftlichen Gebieten (Neuroanatomie, Sinnesphysiologie, *Imaging*) unterschiedlich bezeichnet werden, was die Vergleichbarkeit von Studien erschwert. Ein allgemeiner, von allen Neurowissenschaftlern akzeptierter Hirnatlas existiert offenbar nicht (Kevan Martin: persönliche Mitteilung).

ABBILDUNG 3:

Abbildung 3: Die wichtigsten Hirnregionen (kortikale Bereiche in Grautönen, außer insulärer Kortex) und spezifische Sulci/Gyri, die in der *neuroscience of ethics* als (potentiell) bedeutsam identifiziert wurden (keine vollständige Auflistung). Insbesondere die subkortikalen Strukturen sind in dieser Abbildung nur ungefähr lokalisiert (bzw. unter der Oberfläche verborgen). Die Abbildung wurde adaptiert und ergänzt aus Moll et al. (2005).

2.2.2 Methoden der Neurowissenschaft[9]

Die Neurowissenschaft kann auf zahlreiche Methoden für die Messung und Beeinflussung neuronaler Systeme zurückgreifen. Die große Mehrzahl der für diese Arbeit relevanten Studien verwendet eine heute stark verbreitete Form der Bildgebung (*Imaging*): die funktionelle Magnet-Resonanz-Tomographie (fMRT oder fMRI). Dieses Verfahren und die damit verbundenen methodischen Schwierigkeiten werden im nächsten Abschnitt vorgestellt. Hier soll ein kurzer Überblick über weitere Methoden erfolgen, ohne Anspruch auf Vollständigkeit. Gegliedert wird diese Übersicht in Methoden, die Aufschluss über strukturelle bzw. funktionelle

[8] 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.

[9] Der Autor dankt Deborah Ann Vitacco von der Abteilung für Neuropsychologie des Universitätsspitals Zürich für eine kritische Durchsicht dieses Abschnittes.

Beziehungen im Gehirn geben und solche, mit denen in das neuronale System eingegriffen werden kann (die Ausführungen beruhen u.a. auf Jäncke 2005, Spektrum 2001, Uttal 2001).

Strukturelle Untersuchungen geben Aufschluss über die Anatomie des Gehirns, der Verknüpfung verschiedener Gehirnbereiche und der Mikrostruktur der neuronalen Verknüpfung (z.B. mittels Färbetechniken). Sie bilden seit vielen Jahrzehnten einen zentralen Bestandteil der Hirnforschung. Folgende Verfahren der strukturellen Bildgebung werden zuweilen in Arbeiten verwendet, die für diese Studie untersucht wurden:

- Röntgenstrahlen bilden insbesondere in der Neurologie ein wichtiges diagnostisches Mittel. Das physikalische Prinzip dieser Methode beruht auf unterschiedlichen Durchdringungseigenschaften verschiedener biologischer Gewebe für Röntgenstrahlen, die erfasst und unter Nutzung von Computern zu einem zwei oder dreidimensionalen Bild verrechnet werden können (**Computer-Tomographie**, CT). Durch Zugabe bestimmter Kontrastmittel kann auch das Blutgefässsystem des Gehirns dargestellt werden. Mittels der Computertomografie können beispielsweise Läsionen oder andere Schädigungen des Gehirns in Regionen erkannt werden, denen eine Rolle im moralischen Verhalten zugesprochen wird. Bei Anwendung der CT ist zu beachten, dass die Versuchsperson mit Röntgenstrahlen belastet wird.

- Die Einführung der **Magnet-Resonanz-Tomographie** (*magnetic resonance imaging*, MRI) in der Medizin (und der Neurowissenschaft) gilt als eine der wichtigsten wissenschaftlichen Innovationen der vergangenen Jahrzehnte. Da MRI auch Grundlage für die funktionelle MRI ist, wird das physikalische Prinzip dieser Methode etwas detaillierter vorgestellt: Bestimmte Atome haben einen (quantenmechanisch erklärbaren) Spin. In biologischen Geweben ist das Wasserstoffatom (gebunden u.a. in Wasser) die dominierende Atomart mit dieser Eigenschaft. Dieser Spin erzeugt einen magnetischen Dipol. Die Richtungen dieser Dipole sind im biologischen Material zufällig verteilt. Durch Anlegen eines starken Magnetfeldes werden diese Dipole in eine Richtung ausgerichtet (z-Richtung) und dann durch einen weiteren magnetischen Puls, in eine (zunächst phasengleiche) Präzession gebracht – der Dipolvektor enthält also Komponenten in Richtung der x-y-Ebene. Die Frequenz der Präzession (Lamorfrequenz) ist proportional zur Stärke des Magnetfeldes, das die Spins ausrichtet. Dieses Magnetfeld hat einen Gradienten, so dass die Lamorfrequenz an unterschiedlichen Orten verschieden ist, was zur Lokalisierung der folgenden zwei Komponenten des messbaren Signals dient. Erstens richten sich die Dipole nach dem magnetischen Puls wieder in z-Richtung aus. Dieser Vorgang ist mit einer messbaren Energieabgabe verbunden. Die Zeitkonstante dieses Vorgangs heißt T1-Zeit, sie hängt von der Stärke des Magnetfeldes wie auch von gewebetypischen Aspekten ab. Die T1-Zeit bestimmt, wie schnell sich die Dipole von der Anregung durch den magnetischen Puls erholen und wieder anregbar werden. Wird die Repetitionszeit kurz gewählt, so bestimmt primär 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 magnetischen Puls 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 Bild der verschiedenen Gewebetypen im Messobjekt. Heutige MRI-Scanner erreichen eine Bildauflösung von etwa einem Kubikmillimeter. Das Volumenelement, welches diese Auflösung definiert,

nennt man Voxel (analog zum Pixel – der Auflösungsgrenze eines digitalen Bildes). Höhere Magnetfeldstärken erlauben eine bessere Auflösung, doch stellen sich dann zusätzliche Fragen, z.B. hinsichtlich der Sicherheit der Versuchspersonen (mehr dazu unter Abschnitt 2.2.3).

- Bei der **Diffusions-Tensor-Tomographie** handelt es sich um eine Variante der MRI, die zunehmend an Bedeutung gewinnt, um die Feinstruktur von Faserverbindungen im Gehirn *in vivo* zu erfassen. Kurz gefasst werden hierbei die Diffusions-Bewegungen von Wassermolekülen erfasst. Da diese Bewegung durch die Faserstruktur der *white matter* des Gehirns gerichtet wird, kann dadurch indirekt auf die Richtung der Fasern geschlossen werden. Es ist zu erwarten, dass künftig vermehrt Arbeiten erscheinen werden, die Anomalien in solchen Feinstrukturen mit auffälligem moralischem Verhalten zu korrelieren versuchen.

Funktionelle Untersuchungen fokussieren direkt oder indirekt auf Prozesse im Gehirn. In der ersten Phase der modernen Hirnforschung konnten funktionelle Beziehungen lediglich durch Läsionsstudien ermittelt werden – entweder indem im Tierversuch bestimmte Hirnbereiche zerstört wurden und die daraus folgenden Verhaltensdefizite untersucht wurden, oder indem bei Menschen mit Hirnschädigungen (verursacht z.B. durch Verletzungen oder einem Hirnschlag) Verhaltensstörungen festgestellt und dann bei *post-mortem*-Untersuchungen mit festgestellten Läsionen im Gehirn korreliert werden konnten. Später sind dann auch nichtinvasive Verfahren entwickelt worden, wobei für den Kontext dieser Studie – nebst der im nächsten Abschnitt behandelten fMRI – noch folgende Methoden zu nennen sind:[10]

- Eine bekannte, nichtinvasive Methode ist die in den 1920er Jahren entwickelte **Elektro-Enzephalo-Graphie** (EEG), das später auch durch evozierte Potentiale (EEG-Signale die mit bestimmten Stimuli in Bezug gesetzt werden können) und weiteren Methoden ergänzt wurde (Borck 2005). 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). In den für diese Studie relevanten Arbeiten wird EEG eher selten eingesetzt.

- Da transiente elektrische Ströme von Magnetfeldern begleitet sind, finden auch magnetische Messverfahren Anwendung. Mit der so genannten **Magnet-Enzephalo-Graphie** (MEG) werden hirnelektrische Ströme (vorab des Kortex) gemessen, wobei – im Unterschied zur EEG – Leitfähigkeitsunterschiede der verschiedenen Gewebetypen des Kopfes weitgehend vernachlässigt werden können. Die zu messenden Magnetfelder sind extrem schwach und können nur durch hochempfindliche Detektoren erfasst werden, was messtechnisch hohe Anforderungen stellt.

- Bedeutsamer als die beiden erstgenannten Verfahren ist für den Kontext dieser Studie die **Positron-Emissions Tomographie** (PET). Dabei werden in den Körper des Patienten 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 Radionuklide (^{11}C, ^{18}F, ^{13}N, ^{15}O) setzen bei ihrem Zerfall

[10] Methoden zur Messung der elektrischen Aktivität von einzelnen Neuronen oder ganzen Neuronengruppen mittels in das Gewebe eingeführter Elektroden sind zwar für die Neurowissenschaft generell ebenfalls von grosser Bedeutung – nicht aber für die in dieser Studie untersuchten Arbeiten. Sie werden hier deshalb nicht vorgestellt.

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 errechnet aus dem gleichzeitigen Eintreffen solcher Gammaquanten, die aus demselben Zerfallsereignis herrühren, den Ort des Zerfalls in einer Auflösung von 3-5 mm. Mit diesem Verfahren können, je nach verwendetem Radionuklid, verschiedene Informationen gewonnen werden – beispielsweise kann durch die Verwendung von radioaktiver Glucose Information über erhöhte Stoffwechselaktivität gewonnen werden, da an diesen Orten vermehrt Glucose konsumiert wird. Kein anderes Verfahren erlaubt einen derart spezifischen Einblick in bestimmte metabolische Prozesse im Gehirn. Diese Methode hat aber auch Nachteile: Unter anderem unterwirft sich die Versuchsperson einer gewissen Strahlenbelastung, was die mehrfache Wiederholung einer PET-Messung verbietet. Zudem ist die zeitliche Auflösung vergleichsweise ungenau und erlaubt kaum Aussagen über den Verlauf der untersuchten Prozesse.

Werden strukturelle oder funktionelle Messungen mit Verhaltensuntersuchungen verknüpft, erreicht man lediglich eine Korrelation zwischen Verhaltensmustern und bestimmten Struktur- bzw. Funktionsmessungen. Um kausale Aussagen zu gewinnen, müssen zusätzliche Informationen in die Theoriebildung einfließen. Solche können durch geeignete Stimulation des Gehirns bzw. bestimmter Hirnregionen gewonnen werden. Bedeutsam sind folgende Verfahren:[11]

- Die nichtinvasive **transkraniale magnetische Stimulation** (TMS) ist in den frühen 1980er Jahren entwickelt worden und hat in den letzten Jahren zunehmend an Bedeutung gewonnen. Hierbei wird ein transientes magnetisches Feld aufgebaut, das im Kortex elektrische Ströme induziert, die im betroffenen Areal erregend oder hemmend wirken können (was unter anderem durch die Frequenz der Stimulation bestimmt ist). Die genaue physiologische Wirkungsweise von TMS ist weiter Gegenstand von Forschungen (Bestmann 2008). Die Feldstärken sind deutlich niedriger als bei fMRI und die Anwendung von TMS gilt – unter Beachtung entsprechender Richtlinien – als sicher. Langzeiteffekte sind bislang nicht bekannt (Jäncke 2005). TMS wurde zunächst vorab zur klinischen Diagnostik verwendet. Heute gewinnt das Verfahren auch in der (kognitiven) Neurowissenschaft an Interesse, weil man damit beispielsweise eine Funktionshemmung in bestimmten Arealen des Kortex induzieren und dadurch die kausale Rolle dieses Areals für bestimmte Verhaltensaufgaben untersuchen kann. Diskutiert werden auch therapeutische Anwendungen von TMS (Ridding & Rothwell 2007). Als Alternative zur TMS in Verhaltensexperimenten wird die Anwendung schwacher elektrischer Ströme untersucht (*transcranial direct current stimulation*, siehe Knoch et al. 2008). Hierbei wird durch am Kopf befestigte Elektroden ein schwaches elektrisches Feld erzeugt, das ebenfalls (je nach Polarität des Feldes) im darunter liegenden neuronalen Gewebe erregend oder hemmend wirken kann. Auch diese Methode gilt als ungefährlich und ist zudem vergleichsweise leicht anzuwenden.

- Die Gabe von **neuroaktiven Substanzen** (Hormone, Psychopharmaka etc.) ist ebenfalls ein Verfahren, um kausal auf das Gehirn einzuwirken und wird beispielsweise in

[11] Die direkte elektrische Stimulation von Nervenzellen bzw. Hirngewebe mittels ins Gewebe eingeführter Elektroden ist in der Hirnforschung ebenfalls seit langem ein wichtiges Verfahren (vorab für die Grundlagenforschung, zunehmend auch für therapeutische Zwecke mittels der so genannten Tiefen Hirnstimulation), wird aber in Arbeiten, die hier untersucht wurden, nicht angewendet.

der Psychopharmakologie breit untersucht. Um die Spezifität dieser Kausalität zu ermitteln, müssen zusätzliche Informationen bekannt sein (z.B. Kenntnisse über die Dichte von Rezeptoren für die betreffenden Wirkstoffe in unterschiedlichen Hirnregionen). Im Kontext der hier interessierenden Arbeiten ist insbesondere Oxytocin, das sowohl als Neuropeptid als auch als Hormon wirksam ist, von zahlreichen Studien untersucht worden, weil die Gabe von Oxytocin offenbar das Vertrauen interagierender Versuchspersonen zu steigern vermag (vgl. Abschnitt 3.4.2).

2.2.3 Funktionelle MRI

Eine wichtige Erweiterung erfuhr MRI zu Beginn der 1990er Jahre durch die funktionelle Magnet-Resonanz-Tomographie (*funktional MRI*, fMRI). Hierbei wird ausgenutzt, dass im MRI-Signal auch Informationen enthalten sind, die über funktionelle Beziehungen Aufschluss geben. Die diesbezüglich häufigste Variante ist der so genannte BOLD-Kontrast (BOLD steht für *blood oxygenation level dependent*). Dieses Verfahren beruht auf der Entdeckung, dass der Blutfarbstoff Hämoglobin unterschiedliche magnetische Eigenschaften hat, je nachdem ob sich dieser in einem oxygenisierten Zustand (also wenn Sauerstoff zur Zelle transportiert wird) bzw. desoxygenisierten Zustand (d.h. Sauerstoff ist an die Zelle abgegeben worden) befindet.[12] Wird an einer bestimmten Stelle im Gehirn viel Sauerstoff verbraucht, so verändert sich das Verhältnis zwischen oxygenisiertem und desoxygenisiertem Hämoglobin (die Hämodynamik). Das BOLD-Signal misst die Veränderung dieses Verhältnisses, wobei im Regelfall eine räumliche Auflösung von der Größenordnung von einigen Kubikmillimetern und eine zeitliche Auflösung von der Größenordnung von mehreren hundert Millisekunden erreicht werden. Es sind auch bessere Werte erreichbar (u.a. abhängig von der Stärke des verwendeten Magnetfeldes), wobei die räumliche und zeitliche Auflösung gegenläufig optimiert werden. Zunehmend wird es sogar möglich, gewissermaßen „im Moment des Geschehens" die Hirnaktivität mittels fMRI zu erfassen (*real-time* fMRI, siehe deCharms 2008), was neue mögliche Anwendungen dieser Technologie eröffnet.

Das BOLD-Signal ist Resultat eines komplexen Geschehens, das durch drei Prozesse bestimmt ist: dem cerebralen Blutvolumen, dem Blutfluss und der eigentliche Hämodynamik. Die dem BOLD-Signal zugrunde liegenden physiologischen Mechanismen sind weiterhin Gegenstand der Forschung (für eine aktuelle Übersicht siehe Logothetis 2008). Untersuchungen weisen darauf hin, dass das BOLD Signal tatsächlich mit der Aktivität der Neuronen korreliert – aber nicht, wie ursprünglich vermutet, mit der Feuerrate, sondern mit dendritischen Prozessen. Auch dürfte die Aktivität von Astrozyten das BOLD-Signal mit beeinflussen (Takano et al. 2006). Diese Unklarheiten hinsichtlich der physiologischen Basis des fMRI-Signals (und andere Gründe, siehe unten) nähren auch innerhalb der Neurowissenschaft gewisse Zweifel an der Reichweite der mittels fMRI erzielten Resultaten, insbesondere, wenn höhere kognitive Funktionen untersucht werden (siehe z.B. Poldrack 2008).[13]

ABBILDUNG 4:

[12] Dieser Unterschied kann auch mit optischen Methoden (z.B. mithilfe von Laserlicht) festgestellt werden, weil das Hämoglobin eine andere Farbe hat, je nachdem ob es sich im oxygenisierten bzw. desoxygenisierten Zustand befindet. Bislang werden optische Methoden meist in invasiven Experimenten eingesetzt und werden demnach in den für diese Studie untersuchten Arbeiten nicht verwendet. Zunehmend werden aber auch nichtinvasive Einsatzformen entwickelt, so dass solche optischen Methoden künftig vermehrt Anwendung finden können.

[13] So ist beispielsweise nicht gesichert, welcher Zusammenhang zwischen dem fMRI-Signal und der Information, die diese Aktivität repräsentieren soll, besteht. Eine informationstheoretische Modellstudie zu diesem Thema kommt zum Schluss, dass dieser Zusammenhang hochgradig nichtlinear ist (Nevado et al. 2004). Dies bedeutet insbesondere, dass das Voxel mit der höchsten Aktivität nicht notwendigerweise jenem Voxel entspricht, das die meiste Information kodiert.

Abbildung 4: Die vier Stufen eines Bildgebungs-Experiments (exemplarisch dargestellt an der funktionellen MRI).

Um die Nutzung von fMRI und die damit verbundenen methodischen Probleme besser zu verstehen, kann der Prozess in vier Phasen unterteilt werden (Dumit 2004, vgl. mit Abbildung 4): Design des Experiments, Messung der Hirnaktivität, Auswertung der Daten, Präsentation der Daten. Jede Phase ist mit eigenen Fragestellungen und Problemen verbunden, die nachfolgend (ohne Anspruch auf Vollständigkeit) skizziert werden (hauptsächlich verwendete Quellen: Amaro & Barker 2006, Bergman 2006, Canli & Amin 2002, Jäncke 2005, Poldrack 2006, Savoy 2001, Uttal 2001).[14]

In der Phase des *Designs des Experiments* stellen sich grundlegende methodische Fragen, die teilweise auch wissenschaftstheoretischen Charakter haben:

- **Lokalisationen:** Ein fMRI-Experiment kann dazu dienen, bestimmte psychologische Funktionen in bestimmten Hirnregionen zu lokalisieren. Daraus kann geschlossen werden, ob unterschiedliche psychologische Funktionen gleiche oder unterschiedliche Hirnregionen (bzw. gleiche Regionen in unterschiedlichem Masse) aktivieren, was dann Schlüsse über die physiologische Realisierung dieser Prozesse erlaubt. Eine Frage ist hier, ob man das zu untersuchende psychischen Phänomene hinreichend genau definieren kann, um es mit den gemessen Aktivitätsmustern in Beziehung setzen zu können? Möglich ist, dass diese Phänomene vielleicht gar keine gar keine psychobiologischen Entitäten sind, sondern Manifestationen der verwendeten experimentellen Methoden und Theorien (Uttal 2001). Dieses Argument behauptet implizit, dass es nicht möglich sei, eine Ontologie klar abgrenzbarer psychischer Phänomene (eine „kognitive Ontologie“) zu schaffen.

- ***Reverse Inference:*** Zunehmend werden (insbesondere in den Arbeiten, die für diese Studie untersucht wurden) aus gemessenen Aktivierungsmustern Rückschlüsse auf die in einem *task* involvierten psychischen Prozesse gezogen – man spricht von *reverse inference* (Poldrack 2006). Logisch gesehen ist ein solcher Schluss nur dann korrekt, wenn die gemessene Region *nur* bei diesem psychischen Prozess übermäßig aktiv ist – eine Bedingung, die (gegeben die jetzige raumzeitliche Auflösung von fMRI) klarerweise nicht erfüllt ist. Insofern stehen solche Studien unter dem Verdacht, den syllogistischen Fehlschluss der „Bejahung der Konsequenz“ zu begehen. Um diesem Problem auszuweichen, sollten, wie von Poldrack (2006) vorgeschlagen, vermehrt Methoden der Bayes’schen Analyse Anwendung finden. Auch hier müssen aber die involvierten psychischen Phänomene im Sinn einer kognitiven Ontologie erfassbar und geordnet sein, wobei die kognitiven Ontologien in Bildgebungs-Experimenten meist deutlich rudimentärer sind als in der kognitiven Psychologie.

- **Wahl der Referenz:** Da während einer BOLD-MRI Messung das gesamte Gehirn Sauerstoff braucht, muss man das BOLD-Signal gegenüber einer Referenz bestimmen, um eine Aussage darüber zu gewinnen, ob eine spezifische Aktivität den Sauerstoffverbrauch an einem spezifischen Ort erhöht. Dies kann auf unterschiedliche Weise geschehen (Amaro & Barker 2006): Mittels einer Subtraktion einer *active*-Bedingung gegen eine *control*-Bedingung, mittels einem Faktor-Design (hierbei werden die As-

[14] Probleme von fMRI, die der Neuroethik zugeordnet werden (z.B. der Umgang mit nichtintendierten Befunden bei Versuchspersonen) oder den Umgang mit fMRI-Daten betreffen (Aufbau von Datenbanken, Austausch von Daten), werden hier nicht besprochen.

pekte des zu untersuchenden kognitiven Phänomens während des *tasks* in den einzelnen Blöcken unterschiedlich zusammengesetzt), mittels einem parametrischen Verfahren (indem Aspekte des zu untersuchenden kognitiven Phänomens in unterschiedlicher Stärke präsentiert werden, falls das möglich ist) oder mittels einer Konjunktions-Analyse (hierbei wird untersucht, welche Aspekte des zu untersuchenden kognitiven Phänomens ein gleiches BOLD-Signal verursacht). Zunehmend kommen komplexe Mustererkennungs-Algorithmen in Gebrauch, um bestimmte kognitive Zustände (die von der Versuchsperson berichtet werden) mit bestimmten Klassen von Aktivierungsmustern zu korrelieren (*multi-voxel pattern analysis*, siehe Norman et al. 2006). Diese Varianten sind mit unterschiedlichen methodischen Schwierigkeiten verbunden, die hier im Einzelnen nicht besprochen werden können, aber grundsätzlich beherrschbar sind (abgesehen vom oben angesprochenen Problem, inwieweit die psychologischen Phänomene überhaupt als biophysiologische Entitäten zu werten sind).

Die eigentliche *Durchführung des Experiments* stellt dann unter anderem die folgenden methodischen Schwierigkeiten:

- **Beeinträchtigungen der Versuchspersonen:** Der nichtinvasive Charakter von fMRI soll nicht darüber hinwegtäuschen, dass Versuchspersonen durchaus gewisse Beeinträchtigungen erfahren, die auf das Ergebnis bestimmter Experimente rückwirken können. So ist der Raum, in dem sich die Personen befinden, eng und laut (bis 120 Dezibel), was entsprechende psychische Folgen haben kann und beispielsweise gewisse Patientenstudien erschwert, wenn nicht verunmöglicht (etwa bei Angststörungen). Im Weiteren gilt festzuhalten, dass die verwendeten magnetischen Felder zwar nichtionisierend sind, hingegen Wärme erzeugen können. Sehr starke Felder können auch Ströme induzieren, deren möglichen Auswirkungen noch nicht hinreichend untersucht worden sind. Schließlich gilt anzumerken, dass die Anwesenheit magnetischer Gegenstände in der Nähe der Scanner gravierende Auswirkungen haben können, da diese von den Magneten angezogen werden und unter Umständen in den von der Versuchsperson besetzten Messbereich kommen können. In Einzelfällen ist es zum Tod von Versuchspersonen gekommen (Landrigan 2001). Diese Risiken sind aber in der Regel gut kontrollierbar.[15]

- **Fehlerquellen im Messprozess:** Bei der Messung muss 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 neuroscience* interessanten Regionen (präfrontaler Kortex) erschwert. Zahnimplantate, Haarspangen etc. können zu (leicht erkennbaren) Bildartefakten führen. Auch diese Probleme sind beherrschbar und werden z.T. bei der Vorbearbeitung von fMRI-Daten (z.B. Bewegungskorrektur) angegangen.

- **Das Probleme der Wiederholbarkeit:** Um statistische Aussagekraft zu erreichen, müssen Versuchspersonen mehrfach denselben Stimuli (bzw. derselben Klasse von Stimuli) ausgesetzt werden. Dies wird dann zu einem Problem, wenn ein Stimulus verwendet wird, der eine emotionale Reaktion bei der Versuchsperson auslösen soll. Es

[15] Bis Ende 2005 wurden weltweit gegen 200 Millionen MRI und fMRI Untersuchungen durchgeführt, wobei 14 Todesfälle (vorab durch Inaktivierung von Herzschrittmachern) und gegen 100 Verletzte (z.B. wegen unentdeckter Metallsplitter in den Augen) registriert wurden (Quelle: Blockkurs „Bildgebende Verfahren“, Wintersemester 2005/06, Universität Zürich).

ist bekannt, dass ein Gewöhnungseffekt auftritt (bzw. der Stimulus erreicht bei mehrfacher Präsentation eine immer geringere emotionale Reaktion). Zudem hat sich gezeigt, dass unterschiedliche Kontrollbedingungen unterschiedliche Aktivierungsmuster erzeugen können (Stark & Squire 2001). Um dieses Problem anzugehen, werden unterschiedliche Strategien bei der Präsentation der Stimuli verwendet (*block-design*, *event-related-design*, *mixed-design*; siehe Amaro & Barker 2006). Jede Variante ist mit unterschiedlichen methodischen Schwierigkeiten verbunden, die hier im Einzelnen nicht besprochen werden können

Die Erzeugung eines fMRI-Bilds ist mit einem erheblichen statistischen Aufwand verbunden, z.B. weil sich die gemessene Aktivität im Testfall im Vergleich zum Referenzfall meist nur im Promillebereich unterscheidet. Die *statistische Auswertung* von fMRI-Daten stellt demnach hohe Anforderungen:

- **Schwellenwert-Bestimmung:** 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 (Savoy 2001). Auch 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 der Gefahr aus, relevante Aktivität in anderen Gebieten zu verpassen.

- **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 *inter-individual* 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 ist entsprechend schwierig und könnte zu einer nur vermeintlichen Lokalisierung führen. Zudem ist es möglich, dass die Anatomie verschiedener Hirnregionen eine unterschiedliche Varianz aufweist, was die Mittelwertbildung weiter erschwert.

- **Auswahl der zu korrelierenden Regionen:** Um die Aktivität bestimmter Hirnregionen mit bestimmten Verhaltensmustern zu korrelieren, muss im Auswertungsprozess ein Kriterium definiert werden, um aus der großen Zahl gemessener Voxel jene zu identifizieren, die für die Berechnung der Korrelation in Frage kommen. Beim Selektionsprozess sollten dabei nicht bereits die Korrelation selbst als Auswahlkriterium genommen werden, da aufgrund der enorm hohen Zahl an Voxel man zufällig immer eine gewisse Menge an Voxel finden wird, die signifikant mit dem interessierenden psychischen Phänomen korreliert. Eine im Dezember 2008 vorab veröffentliche Untersuchung (Vul et al. 2008) weckte den Verdacht, dass bei über der Hälfte von (in bedeutenden Zeitschriften erschienen) Arbeiten in den Sozialen Neurowissenschaften dieser Fehler begangen worden sei, so dass die dort berichteten Korrelationen als nicht stichhaltig bezeichnet werden könnten. Auf diese Kritik wurde inzwischen geantwortet, indem den Autoren ihrerseits statistische Fehler vorgeworfen wurden. Die Kontroverse über diese Studie war zum Zeitpunkt der Fertigstellung dieser Studie noch im Gang; sie zeigt auf, wie anspruchsvoll der statistische Umgang mit fMRI-Daten ist.

Nach der Analyse der fMRI-Daten stellen sich schließlich auch Fragen hinsichtlich der *Präsentation der Ergebnisse*, was nicht zuletzt wissenschaftssoziologische Aspekte berührt:

- **Auswahl des Darzustellenden:** Es können unterschiedliche Resultate der statistischen Analyse kommuniziert werden: 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 verschiedene 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 Savoy 2001). Diese (nicht in jedem Fall gegebene) Wahlmöglichkeit kann einen gewissen Spielraum eröffnen, um Resultate von Studien dem Studienzweck entsprechend zu präsentieren.

- **Fehlerhafte Präsentation der Resultate:** Im Zug der Pilotstudie haben Gespräche mit Fachleuten ergeben (u.a. mit Anton Valavanis), dass bei einigen der publizierten Bildgebungs-Studien Resultate fehlerhaft dargestellt würden. Eine Auswertung aller im Zeitraum 2001-2004 publizierten *Imaging*-Studien zeigte auf, dass in rund 60 Prozent aller Abbildungen von fMRI-Messungen die anatomischen Bezeichnungen inkorrekt sind (Valavanis: persönliche Mitteilung). Nach Meinung von Valavanis würde 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) ungenügend berücksichtigt, was die Aussagekraft der Resultate vermindere. Zudem erhöhe die digitale, computergestützte Visualisierung das Risiko einer bis hin zur Manipulation und Fälschung reichende Veränderung wissenschaftlicher Daten. 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.

- **Wirkkraft der Bilder:** Die durch fMRI gewonnen Bilder können eine starke suggestive Wirkung haben. Zu nennen sind in diesem Kontext die Falschfarben, die für die Darstellung statistischer Signifikanz verwendet werden. Diese könnten so gewählt werden, 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 (beispielsweise) rot eine höhere Intensität bedeutet. Zumindest bei gewissen Hirnbildern lässt sich ein derart (intendierter) Einsatz solcher Bilder zeigen (Dumit 2004). Hier wäre eine Theorie der Wahrnehmung solcher Bilder durchaus nützlich.

Die Auflistung dieser Probleme soll nicht dahingehend interpretiert werden, dass *Imaging* (und insbesondere fMRI) ein ungeeignetes Mittel für die Untersuchung der neurowissenschaftlichen Prozesse sind, die mit psychischen Phänomenen einher gehen. Bei fMRI geht es insbesondere nicht nur um ein Abbilden des Gehirns, sondern um eine Art *parsing*, d.h. das Aktivierungsmuster wird als eine Sequenz von Aktivierungen unterschiedlicher Hirnregionen verstanden (Donaldson 2004). Es ist demnach falsch, fMRI lediglich als eine Methode anzusehen, die eine neue Form von Lokalisation von Funktionen erlaubt, denn man will das mit einer psychischen Funktion verbundene zeitliche Aktivitätsmuster im Gehirn ermitteln, nicht

nur die aktiven Regionen. Damit lassen sich zweifellos wichtige Informationen über die Genese (und Störung) psychischer Phänomene finden.

Zudem muss bemerkt werden, dass bei genügender (derzeit möglicherweise noch nicht in allen Fällen realisierbarer) raumzeitlicher Auflösung für die Ermittlung von Aktivierungsmustern man *per definitionem* unterschiedliche Muster bei unterschiedlichen kognitiven oder emotionalen *tasks* finden wird. Dies folgt aus der (üblicherweise nicht bestrittenen) metaphysischen Voraussetzung, dass Gedanken und Handlungen von Menschen von Aktivierungen bestimmter Hirnregionen begleitet sind – ohne damit mehr über die Art dieser „Begleitung" (Korrelation, Kausalität, Emergenzbeziehung) sagen zu müssen. Die Konstatierung des Unterschieds allein ist demnach kein wissenschaftlich überraschendes Ergebnis.

2.2. Methoden für die Erfassung von *moral agency*

2.2.1 Zum Verhältnis von Methode und inhaltlicher Festlegung

Die empirische Erfassung von *moral agency* ist angesichts der beschriebenen Komplexität des Phänomens keineswegs trivial. Sie ist insbesondere mit dem Erfordernis verbunden, dass die Erfassung des Phänomens mit inhaltlichen Festlegungen im Hinblick auf das Moralsystem verbunden ist, anhand deren die Handlungen des *moral agent* bemessen werden soll. Dies ist insofern schwierig, weil sowohl definiert werden muss, welche Inhalte (beispielsweise: welche Normen) denn nun die Kernelemente des Moralsystems bilden sollen, als auch sich *moral agency* nicht notwendigerweise darin erschöpft, ob und bis zu welchem Grad der *agent* diesen Kernelementen folgt. So könnte gerade die begründete Ablehnung gewisser Normen (beispielsweise aufgrund gewandelter Kontexte) *moral agency* auszeichnen, so dass der *Umgang* mit diesen inhaltlichen Festlegungen ebenfalls Gegenstand des empirischen Interesses sein kann.[16] Gemäß der in Abschnitt 1.3.1 eingeführten Unterscheidung kann demnach sowohl die „Moralität" des *moral agent* (inwiefern erfüllt der *agent* die Erfordernisse eines – mittels deskriptiver Ethik definierten – Moralsystem einer bestimmten Gesellschaft?) als auch seine „Ethizität" (wie stellt sich der *agent* zu diesem Moralsystem?) Gegenstand einer „Moral-Messung" sein. Eine umfassende Untersuchung müsste beide Aspekte erfassen, stößt bei empirischen Studien aber sofort auf die Schwierigkeit, dass die Versuchsanordnungen schnell schwer beherrschbar wird und zu viele freie Parameter enthält, so dass das Experiment keine klare Aussage mehr erlaubt (zweifellos ein generelles Problem bei der Erforschung derartiger Phänomene).

ABBILDUNG 5:

Abbildung 5: Festlegungen hinsichtlich „Moralität" und „Ethizität" bei der empirischen Erfassung von *moral agency*.

Methoden zur empirischen Bestimmung von *moral agency* müssten demnach Festlegungen sowohl hinsichtlich der Inhalte des Moralsystems als auch hinsichtlich des Umgangs mit diesem Moralsystem treffen (vgl. mit Abbildung 5). Eine gängige Strategie der Vereinfachung besteht darin, moralische Inhalte zu wählen, die möglichst unstrittig sind und in Kontexten präsentiert werden, die eine klare Interpretation zulassen – also beispielsweise ein Szenario, in dem Unschuldige bewusst und mit bösartiger Absicht verletzt oder getötet werden und dem-

[16] Um dies an einem (notorischen) Beispiel zu zeigen: *Moral agency* daran zu bemessen, wie oft die Norm „du sollst nicht lügen" im Alltag des *agent* erfüllt wird oder nicht, stösst an das Problem, dass es Kontexte gibt (die Geheimpolizei eines totalitären Staates klopft an die Türe der eigenen Wohnung, in der man Widerstandskämpfer versteckt), die Moralität der Erfüllung dieser Norm wiederum zweifelhaft werden lässt.

nach das Prinzip „Nichtschaden“ (Beauchamp & Childress 2001) unzweideutig verletzt wird. Hier lässt sich aber fragen, inwieweit derart gewonnene Erkenntnisse über *moral agency* Auskunft darüber geben, wie *moral agents* sich gegenüber ethischen Fragen verhalten, die in aktualen Gesellschaften diskutiert werden – also auf Fragen, die oft keine eindeutige Antworten kennen und gerade deshalb strittige Fragen sind.

Diese Überlegungen eröffnen ein breites Spektrum an Fragen, die hier nicht weiter erörtert werden können. Festzuhalten ist, dass Methoden empirischer Moralforschung nicht auf Festlegungen moralischer wie ethischer Art verzichten können und die Kritik an bestimmten Methoden oft auf die Zulässigkeit dieser Festlegungen zielt. Nachfolgend wird anhand dreier methodischer Ansätze, die in den für diese Studie relevanten Arbeiten genutzt werden, aufgezeigt, welche Festlegungen getroffen werden. Auf den Problemkomplex „moralischer Stimulus“ wird in Abschnitt 3.2. ausführlicher eingegangen.

2.2.2 Absolute Moralskalen

Ein klassisches Verfahren zur Bestimmung von *moral agency* ist das Stufenmodell von Kohlberg (1995). Diese in den 1960er Jahren entwickelte Methode kann auf eine breite Rezeption zurückblicken, so dass an dieser Stelle nur wenige Bemerkungen folgen sollen. Es wird überdies in den heutigen Studien meist kritisch betrachtet und nur selten angewendet.

Man kann die Methode als ein Verfahren verstehen, das eine absolute Moralskala (die Stufen 1-6) voraussetzt, wobei zusätzlich postuliert wird (was mittels empirischen Untersuchungen auch untermauert wurde), dass diese Stufen während der Ontogenese eines *moral agent* in einer festen Sequenz durchlaufen werden, bis dann der *agent* auf einer bestimmten Stufe verharrt. Das Messverfahren besteht darin, dass dem Probanden Dilemmas[17] vorgelegt werden und in einem strukturierten Interview erfasst wird, aus welchen Gründen der Proband sich für welche der vorgeschlagenen Varianten entscheidet. Das Auswertungsprotokoll macht dabei sowohl Festlegungen moralischer (durch Beurteilung der durch den Probanden verwendeten Normen) als auch ethischer (durch Bewertung, welche Gründe als „gut“ gelten) Art. Es erstaunt deshalb nicht, dass die am Stufenmodell geäußerte Kritik sich an diesen Festlegungen reibt. So wandte beispielsweise Gilligan (1977) unter anderem ein, dass das Stufenmodell jene Aspekte, die in einer (weiblich geprägten) *care ethics* zentral sind, nur ungenügend erfasst. Dies ist ein Beispiel einer Kritik an ethischen Festlegungen.

Generell dürften Methoden, die eine absolute Moralskala benötigen, in einer pluralistischen Gesellschaft auf Skepsis stoßen. Dennoch ist für die meisten empirischen Untersuchungen der Bezug auf eine solche Skala nur schwer vermeidbar – nicht im Hinblick auf eine direkte Bewertung des *moral agent* selbst, aber im Hinblick auf die verwendeten moralischen Stimuli, die vorgängig einem *rating* oder zumindest einer Klassifikation in „moralisch“ bzw. „nichtmoralisch“ unterzogen werden müssen (dazu mehr unter Abschnitt 3.2).

2.2.3 Klassifikation dilemmatischer Entscheidungen anhand normativer Ethiken

Die theoretische Schärfung bestimmter normativer Ethiken mit Hilfe von Dilemmas ist ein Standardverfahren der Ethik. Das Verfahren beruht darauf, ein Dilemma derart zu konstruieren, dass eine Handlungsalternative X des Dilemmas einer normativen Ethik A (z.B. utilitaristischer Art) und die andere Alternative Y einer normativen Ethik B (z.B. deontologischer Art)

[17] Ein bekanntes Beispiel ist das Heinz-Dilemma. Darin 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 bezahlen kann.

zugeordnet werden kann. „Zugeordnet“ meint, dass die jeweils aufgeführten Gründe für X oder Y sich auf Prinzipien oder Normen der normativen Ethiken A oder B stützen. Je nachdem, ob A oder B verteidigt werden soll, werden X oder Y so gewählt, dass die der „gegnerischen“ normativen Ethik zugeordnete Alternative kontraintuitiv ist, so dass die eigene Position geschärft wird.

Es liegt nahe, solche Dilemmas auch für empirische Untersuchungen zu verwenden, indem diese Probanden vorgelegt werden, die dann eine Wahl (X oder Y) zu treffen haben. Damit kann zweierlei untersucht werden: Zum einen erlauben die Antworten eine Zuordnung des *moral agent* zu gewissen normativen Ethiken, zum anderen kann untersucht werden, was genau mit „kontraintuitiv“ gemeint ist bzw. mit welchen inneren Prozessen diese (Kontra-) Intuition beim *moral agent* verbunden ist. Solche Dilemmas sind in mehreren der hier untersuchten Arbeiten verwendet worden und werden im Abschnitt 3.2 genauer vorgestellt.

Hier sollen zwei grundlegende Bemerkungen hinsichtlich der mit diesen Dilemmas verbundenen Festlegungen folgen. Erstens ist gar nicht so klar, inwiefern aufgrund eines solchen Experiments die „Zuordnung“ des *moral agent* zu einer bestimmten Ethik (z.B. der *agent* ist aufgrund seiner Antworten „Utilitarist“) gerechtfertigt ist. Offensichtlich nährt sich die Gleichsetzung einer Dilemma-Alternative X mit der normativen Ethik A aus einer langen, in der wissenschaftlichen Ethik geführten Tradition hinsichtlich der Ausgestaltung und Verteidigung dieser Ethiken. Der *moral agent* selbst hingegen verfügt möglicherweise (oder vermutlich mit ziemlicher Sicherheit) gar nicht über diese Ethiken im Sinn, dass er beispielsweise seine Wahl von X als „utilitaristisch“ versteht. Kritiker verweisen darauf, dass die vorgängige Zuordnung von Alternativen ganzer Batterien solcher Dilemma-Tests zu bestimmten normativen Ethiken aus diesem Grunde gar nicht so eindeutig ist (Kahane & Shackel 2007).

Zweitens ist mit der Feststellung, dass bestimmte Dilemma-Alternativen mit bestimmten Intuitionen verknüpft sind, die messbar das Wahlverhalten des *moral agent* ändern, das Phänomen der *moral agency* nicht ausreichend erfasst worden. Dies deshalb, weil damit noch nicht der Umgang des *moral agent* mit diesen Intuitionen erfasst worden ist. Das experimentelle Setting verläuft auf einer kurzen Zeitskala (d.h. die Entscheidungszeit ist vergleichsweise kurz) – und das muss so sein, will man diesen Effekt messen. Deliberation ist nicht erwünscht, was ebenfalls eine Festlegung ist. Entsprechend bleibt es eine Frage der Interpretation, was vom Auftreten solcher Intuitionen zu halten ist: ist ihr Auftreten eine argumentative Stütze für bestimmte Ethiken (gemäß dem instrumentellen Einsatz solcher Dilemmas in der ethischen Debatte), oder eher ein Argument gegen solche Ethiken (wie beispielsweise von Singer, 2005, als Interpretation der Experimente von Greene et al., 2001/2004, vorgeschlagen wurde)? Die Festlegung, diesen (an sich messbaren[18]) Aspekt der Ethizität des *moral agent* offen zu lassen, führt also zu Debatten, die innerhalb der ethischen Theorie geklärt werden müssen.

2.2.4 Methoden der experimentellen Ökonomie

Da moralisches Verhalten im Alltag zahlreiche Ausdrucksformen finden kann (bsp. in Form von Verhaltensdispositionen und -formen wie Empathie, Fairness, Kooperation, Vertrauen, etc.), ist *moral agency* auch durch die Erfassung solcher „moralnaher Verhaltensweisen“ empirisch messbar. Diese können einerseits explizit erfasst werden (beispielsweise durch Befragungen des *moral agent*) oder aber implizit (durch das Verhalten des *moral agent* in einer definierten Situation). Von den zahlreichen Möglichkeiten soll hier nur eine Klasse von Methoden eingeführt werden, die auch in den hier untersuchten Arbeiten zunehmend Anwendung

[18] Man könnte sich beispielsweise ein Szenario vorstellen, in dem zwei *moral agents* argumentieren, um gemeinsam ein solches Dilemma beurteilen.

finden: experimentelle Spiele, mit denen man mittels ökonomischen Anreizen das Verhalten von Spielern untersucht, um Konzepte wie Vertrauen und Kooperation genauer zu analysieren (Fehr & Schmidt 1999).

Mit solchen aus der experimentellen Ökonomie stammenden Spielen sollen quantitative Aussagen über die Motive der Versuchspersonen gewonnen werden (beispielsweise messbar durch die Geldbeträge, welche in einem Spiel ausgetauscht werden). Die Struktur der Spiele erlaubt es zudem, den Effekt von Institutionen einzubeziehen – beispielsweise einer strafenden Partei. An dieser Stelle soll eine kurze Übersicht über einige Grundtypen solcher Spiele gegeben werden, bei denen die Spieler nur ein einziges Mal (*one-shot games*) oder aber mehrfach aufeinander treffen (Quellen: Camerer & Fehr 2002, Falk & Fischbacher 2000, Fehr & Schmidt 1999, McCabe et al. 2001):

- **Diktator-Spiel:** In diesem 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.

- **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 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.

- **Vertrauens-Spiel:** In diesem zweistufigen Zwei-Personen-Spiel, das man als eine Variante des Gefangenen-Dilemmas[19] auffassen kann, treffen 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.

- **Drittperson-Bestrafungs-Spiele:** Die genannten Spiele lassen sich so abwandeln, dass eine dritte Person C das Verhalten der beiden anderen Spieler beobachten und diese 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 kostspielig ist, d.h. im zweiten Fall muss C für den Akt der Bestrafung selbst etwas bezahlen.

Derartige Experimente haben den Vorteil, dass sie vergleichsweise einfach zu realisieren sind, quantifizierbare Resultate (z.B. in Form der ausgetauschten Geldbeiträge) liefern und mit anderen Untersuchungsmethoden (z.B. Bildgebung) kombiniert werden können. Fraglich ist hingegen, inwieweit der Gehalt der Motive bei den Spielern erfasst wird. So könnte bezüglich des Motivs „Vertrauen“ beispielsweise bemerkt werden, dass in Vertrauens-Spielen eher so

[19] Das Gefangenendilemma ist das klassische Paradigma der Spieltheorie: Zwei Spieler A und B sind mit dem Problem konfrontiert, eine binäre Entscheidung zwischen den Varianten „kooperieren“ (K) oder „den anderen verraten“ (V) zu treffen. Je nachdem wer kooperiert bzw. verrät, werden die Belohnungen T (für den Verräter, wenn der andere kooperiert), R (wenn beide kooperieren), P (wenn beide einander verraten), und S (für den Kooperierenden, der verraten wurde) wie folgt festgelegt: $T > R > P > S$ und $2R > T + S$. Jeder Spieler würde also jeweils dann am meisten erhalten, wenn er selbst den anderen verrät, der andere aber kooperieren würde. Insgesamt würde das System aber profitieren, wenn beide kooperieren.

etwas wie Prognosesicherheit gemessen wird und den Begriff damit nur unvollständig abdeckt.[20] Eine vertiefende Diskussion über diese methodischen Fragen kann hier aber nicht erfolgen.

2.3 Moral als Thema der Neurowissenschaft

2.3.1 Bibliometrische Analyse

Welchen Stellenwert hat die *neuroscience of ethics* innerhalb der gesamten Neurowissenschaft? Um einen Eindruck über die Forschungstätigkeit in den relevanten Gebiete zu erhalten, wurde eine bibliometrische Untersuchung für den Zeitraum 1975 bis 2008 basierend auf den Einträgen in den Datenbanken *PubMed* und *Science Citation Index expanded* (SCI) durchgeführt.[21] Untersucht wurde, wie sich der Anteil der Forschung in den Sozialen Neurowissenschaften, Arbeiten mit Bezug auf Moral/Ethik und die Bedeutung der Bildgebung in den jeweiligen Gebieten relativ zur Publikationstätigkeit in den gesamten Neurowissenschaften (die sich zwischen 1975 bis 2008 mehr als versechsfacht hat) verändert hat.

ABBILDUNG 6:

Abbildung 6: Bibliometrische Untersuchung neurowissenschaftlicher Arbeiten zu sozialem/moralischem Verhalten in PubMed (1975-2008) und SCI expanded (1991-2008): a) Anteil von Arbeiten aus dem Bereich *Social Neuroscience* und *Moral Neuroscience* relativ zur Gesamtzahl aller publizierter *Neuroscience*-Arbeiten. b) Anteil von „Methoden-Arbeiten" relativ zur Gesamtzahl aller Arbeiten der Mengen *Neuroscience* bzw. *Social Neuroscience*. Die kleine Abbildung zeigt den Anteil von Arbeiten mit „klassischen ethischen Themen" sowie von „Methoden-Arbeiten" relativ zur Gesamtzahl aller Arbeiten der Menge *Moral Neuroscience*.

Bibliometrische Studien dienen dazu, durch Wahl geeigneter Suchbegriffe quantitative Trends in der Forschung, ausgedrückt durch Publikationstätigkeit zu erkennen.[22] Zu diesem Zweck wurde die Menge *Neuroscience* (Arbeiten mit Worten, die auf Neuro/Hirn-Themen schließen lassen) als Referenz, sowie innerhalb dieser die Mengen *Social Neuroscienc*e (Arbeiten, die Begriffen mit den Wortstämmen „social", „emotion", „cultura", „econom" enthalten) *Moral*

[20] In einem Vertrauensspiel hat die Variable „Vertrauen" mathematisch die Struktur einer stetigen Funktion. Vertrauen (insbesondere bei längeren Interaktionsbeziehungen) hingegen dürfte mathematisch eher die Struktur einer *step-funktion* haben (man vertraut einem Gegenüber zunächst auch, wenn dieser dieses Vertrauen missbraucht, bis ein Punkt erreicht ist, an dem das Vertrauen zusammenbricht).

[21] Die öffentlich zugängliche Datenbank *PubMed* (http://www.ncbi.nlm.nih.gov/pubmed) wird von der *US National Library of Medicine* unterhalten und enthält in erster Linie Zeitschriften aus der Biomedizin und verwandten Bereichen. Die weltgrösste Zitations-Datenbank *Science Citation Index expanded* (kostenpflichtiger Zugang) wird von der Firma Thomson Reuters unterhalten und enthält Zeitschriften aus den Bereichen Naturwissenschaft und Psychologie. Durch den Rückgriff auf zwei unterschiedliche Datenbanken können robustere Ergebnisse über Trends erhalten werden, wenngleich die absoluten Zahlen nicht übereinstimmen, weil letztere Datenbank umfassender ist. Hingegen werden im SCI expanded erst seit 1991 auch Abstracts von Zeitschriften erfasst, so dass aus Vergleichsgründen diese Datenbank erst ab diesem Zeitpunkt untersucht wurde.

[22] In Titel und Abstracts von Publikationen wurden die jeweiligen Zahlen in jedem Jahr mit den folgenden booleschen Suchausdrücken abgeschätzt. *Neuroscience* insgesamt: neuro* OR neural OR brain* OR amygdala OR cerebellum OR cortical OR cortex OR hippocampus. Innerhalb der Menge Neuroscience wurde *Social Neuroscience* abgeschätzt mit: social* OR socio* OR cultura* OR emotion* OR econom*. *Moral Neuroscience* wurde abgeschätzt mit: ethic* OR moral*. Die Zahl der Paper mit *Imaging*-Methoden (und TMS/EEG) innerhalb der verschiedenen Grundmengen wurde abgeschätzt mit: "brain imaging" OR „functional magnetic resonance imaging" OR „functional MRI" OR fMRI OR „magnetic resonance imaging" OR MRI OR „positron emission tomography" OR PET OR „Electroencephalography" OR EEG OR „transcranial magnetic stimulation" OR TMS. Innerhalb der Menge *Moral Neuroscience* schliesslich wurde die Zahl der Arbeiten zu „klassischen Themen" abgeschätzt mit: fetal OR fetus OR transplant* OR "organ donor" OR "brain death" OR "informed consent" OR "stem cell". Die Untersuchung wurde am 28. Januar 2009 durchgeführt.

Neuroscience (Arbeiten, die Begriffen mit Wortstämmen „ethic“, „moral“ enthalten) und *Methoden* (Arbeiten, die Begriffe zu den meisten unter Abschnitt 2.2.2 und 2.2.3 beschrieben Methoden enthalten) definiert. Für jedes Jahr wurde in beiden Datenbanken die jeweilige Zahl an Publikationen für die jeweilige Menge bestimmt. Die Resultate finden sich in Abbildung 6.

Aus der Abbildung lassen sich insbesondere folgende Schlüsse ziehen: 6.a zeigt, dass Arbeiten mit einer „sozialen“ bzw. „moralischen“ Begrifflichkeit erst ab Mitte der 1990er Jahre ihren Anteil an der gesamten Publikationstätigkeit erhöht haben – in letzterem Fall gemäß den PubMed-Daten erst um 2000. Beide Datenbanken zeigen eine deutliche Steigerung der Zahl der *Social Neuroscience* und *Moral Neuroscience* Arbeiten (in der Größenordnung von knapp 4- bis knapp 5-mal mehr).[23] Abbildung 6.b zeigt den zunehmenden Einfluss der Bildgebung (und TMS) innerhalb der Neurowissenschaften und insbesondere innerhalb der Sozialen Neurowissenschaften. In absoluten Zahlen zeigen sich zwar Unterschiede zwischen den Daten von PubMed und SCI (u.a. deshalb, weil PubMed spezifischer die Lebenswissenschaften einschließt), hingegen zeigt sich deutlich, dass die Steigerungsrate dieser Methoden in der Menge *Social Neuroscience* jeweils klar höher ist, als in *Neuroscience* insgesamt. Zu den Arbeiten der Menge *Moral Neuroscience* lässt sich zudem sagen, dass der Anteil von Arbeiten, die sich „klassischen“ ethischen Themen der Hirnforschung zuwenden (z.B. Hirntod, neuronale Stammzellen, Transplantation, *informed consent* bei Hirnkranken) tendenziell eher abnehmen, Arbeiten unter Nutzung der genannten Methoden aber ab ca. 2000 zunehmen. Dies verweist auf das zunehmende Interesse an der Untersuchung der neurobiologischen Grundlagen der Moral (*neuroscience of ethics*) bzw. der so genannten Neuroethik.[24]

Zusammengefasst bestätigt diese Untersuchung den (qualitativen) Befund, dass Arbeiten zu sozialem/moralischem Verhalten erst in jüngster Zeit Thema der Neurowissenschaft geworden sind und die Bildgebung bei diesen Untersuchungen eine zentrale Rolle spielt. Andere bibliometrische Untersuchungen stützen diese Aussage: So finden Illes et al. (2003) 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 2000). So genannte *higher-order cognition* und Emotionen bilden einen zunehmenden Anteil in diesen fMRI-Studien (er beträgt gegen 20% im Jahr 2001). Eine auf psychologische Literatur beschränkte Untersuchung von Haidt (2003) basierend auf der PsycINFO-Datenbank zeigte ebenfalls ein deutlich gesteigertes Interesse an der Erforschung bestimmter Arten von Emotionen im Zeitraum 1975 bis 1999 – vorab solcher Emotionen, welche Haidt „moralisch“ nennt (siehe Abschnitt 2.3.2).

2.3.2 Das wissenschaftliche Umfeld der neuroscience of ethics

Gerade in den Neurowissenschaften zeigt sich das Phänomen einer vergleichsweise raschen Ausdifferenzierung neuer Disziplinen (so genannte „Bindestrich-Disziplinen“ wie Neuroethik, Neuroökonomie, Neuropädagogik, Neuromarketing). Dieses „disziplinäre Feld“ soll hier etwas genauer dargestellt werden, bevor in Abschnitt 3 die *neuroscience of ethics* im Detail vorgestellt wird. Inwieweit diese disziplinäre Neuschöpfungen als Differenzierungsprozess getrieben durch spezifische Fragestellungen oder aber als Prozess der Deklaration neuer Disziplinen (möglicherweise als Folge der geänderten Karriere- und Finanzierungsbedingungen des wissenschaftlichen Arbeitens) zu werten ist, kann hier nicht untersucht werden.[25]

[23] Zu der Ausprägung der Sozialen Neurowissenschaft siehe Matusall et al., in Vorbereitung.
[24] Neuroethik bezeichnet jenes Gebiet der angewandten Ethik, das Auswirkungen der Hirnforschung auf ethische und rechtliche Fragen untersucht (Roskies 2002).
[25] Die nachfolgenden Ausführungen sind Zwischenresultate des Forschungsprojektes „Disciplinary dynamics in emerging social neurosciences and neuroeconomics”, das der Autor zusammen mir Ina Kaufmann (Graduierten-

Wie im vorangegangenen Abschnitt beschrieben, ist die disziplinär deklarierte neurowissenschaftliche Untersuchung von *moral agency* bzw. sozialem Verhalten generell ein junges Phänomen. Dennoch können zahlreiche Fragestellungen, welche als konstitutiv für diese Disziplinen bezeichnet werden, auf eine deutlich längere Geschichte zurückblicken.[26] Diese wiederum lassen sich vier Feldern zuordnen, die nachfolgend nur skizzenhaft beschrieben werden. Ein umfassenderes Bild müsste zudem die Verhaltensforschung (insbesondere an Primaten) mit einbeziehen, was in dieser Skizze fehlt. Die *neuroscience of ethics* und deren wissenschaftliches Umfeld „nährt" sich aus folgenden, sich überlappenden disziplinären Feldern:

- **Kognition-Entscheidung-Handlung:** Mit diesen drei Stichworten ist ein Komplex angesprochen, der die Neurowissenschaft, Verhaltensforschung und Psychologie seit der so genannten „kognitiven Revolution", deren Beginn man auf die Mitte des 20. Jahrhundert (Gardner 1985) lokalisieren kann, machtvoll beherrscht. Keinesfalls kann hier dieses riesige disziplinäre Feld genauer beschrieben werden; es folgen lediglich zwei Bemerkungen, die für die nachfolgenden Überlegungen wichtig sind. Zum einen gibt es innerhalb der Neurowissenschaft seit vielen Jahren detaillierte Untersuchungen zum *decision making* anhand einfacher senso-motorischer Aufgaben, denen sich Primaten im Tierexperiment zu stellen haben und bei welchen die Aktivität einzelner Neuronen bzw. Neuronengruppen unterschiedlicher Hirnregionen gemessen werden kann (für einen aktuellen Übersichtsartikel siehe Gold & Shadlen 2007). Heutige Forschungen zum *social decision making* (die man der Neuroökonomie zuordnen kann, Fehr & Camerer 2007) stützen sich wesentlich auf diese Untersuchungen, die freilich von einem sehr einfachen Entscheidungsbegriff ausgehen, ab. Zum anderen haben sich (in Überlappung mit der Emotionsforschung) mehrere so genannte *dual-processing-theories* des *decision making* ausgebildet, die zwei grundsätzlich verschiedene Formen der Entscheidungsfindung unterscheiden (für eine Übersicht siehe Evans 2008): eine rasche, automatisch und unbewusst ablaufende Form gegenüber einer langsamen, deliberativen und kognitiven Form. Diese Unterscheidung findet sich auch in der *neuroscience of ethics* an prominenter Stelle (siehe Abschnitt 3.5).

- **Emotionen:** Seit den 1990er Jahren sind Emotionen wieder ein bevorzugtes Thema der Neurowissenschaft geworden (eine kompakte Übersicht bietet Vaas 2000), nachdem lange Zeit das „kognitive Paradigma" herrschte, unter welchem neuronale Prozesse unter dem Blickwinkel von Informationsverarbeitung untersucht wurden (LeDoux 2000). 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. Trotz des anhaltenden Booms der Emotionsforschung in verschiedensten, auch geisteswis-

---

programm für interdisziplinäre Ethikforschung, Universität Zürich) und Svenja Matussall (Wissenschaftsforschung, ETH Zürich) durchführt.

[26] Die hier vorgestellte Differenzierung stützt sich auf eine bibliometrische Untersuchung ab, die den Zeitraum 1991 bis 2007 umfasste (mit anderen Worten: auch diese Untersuchung betrifft lediglich die jüngste Geschichte). Hierzu wurde in einem ersten Schritt in einer qualitativen Untersuchung zentrale *keywords* identifiziert, die Arbeiten der angesprochenen Disziplinen (*social/moral neuroscience, neuroeconomics*) charakterisieren. In einem zweiten Schritt wurde untersucht, wie häufig Arbeiten, die diese *keywords* enthalten, pro Jahr in den Mengen *Neuroscience* und *Social Neuroscience* (siehe Fussnote 22) auftauchen. Das ergab eine Klassifikation dieser *keywords* wie folgt: solche, die im genannten Zeitraum keinen markante Zunahme erlebten (z.B. *violence*), solche, bei denen diese Zunahmen in den 1990er Jahren erfolge (z.B: der Wortstamm *psychopath**) und solche, die erst in diesem Jahrzehnt eine deutliche Zunahme zeigten (z.B. der Wortstamm *neuroeconom** sowie die Begriffe *cooperation* und *fairness*). Aufgrund dieser Klassifikation (und weiterer qualitativer Untersuchungen) ergab sich die hier vorgestellte Differenzierung.

senschaftlichen Disziplinen wird innerhalb der Neurowissenschaft bemängelt, dass der Emotionsbegriff unscharf und deshalb experimentell schwierig zugänglich ist (LeDoux 2000). Dennoch wird heute nicht bestritten, dass die Erklärung komplexer kognitiver Fähigkeiten und des Verhaltens von Tieren und Menschen eine Theorie der Emotionen verlangt – beispielsweise als Element der oben genannten *dual-processing-theories.* Zahlreiche Bildgebungs-Studien liefern Hinweise, welche Hirnregionen bei der Verarbeitung emotionaler Stimuli bzw. der Erzeugung von Emotionen besonders aktiv sind (für eine Meta-Studie siehe beispielsweise Phan et al. 2002). Solche Arbeiten sind mit dem Problem konfrontiert, dass es heute noch keine allgemein anerkannte Klassifikation der verschiedenen Arten von Emotionen gibt – insbesondere, wenn man ein ausdifferenziertes Begriffsfeld mit einer feinen Unterteilung von Emotionen untersuchen will (siehe z.B. Roberts 2003). Natürlich gibt es zahlreiche Versuche, eine solche Klassifizierung zu liefern; für diese Studie wichtig sind die so genannten moralischen Emotionen (siehe Abschnitt 3.3.3).

- **Empathieforschung:** Der Begriff der Empathie ist ähnlich schillernd wie jener der Emotion und steht in der Interpretation „Empathie = Fähigkeit zu fühlen, was eine Person A fühlt" diesem auch nahe. Bevor Empathie ab der zweiten Hälfte der 1990er Jahre zunehmend in der Neurowissenschaft untersucht wurde, wurde dieses Phänomen vorab in der Psychologie erforscht und dabei teilweise eng mit moralischem Verhalten in Beziehung gesetzt (Hoffman 2000). Auch innerhalb dieser Forschungstradition wird oft bemerkt, dass es keine allgemein akzeptierte und präzise Definition von Empathie gebe; zuweilen wird vermutet, dass der Begriff Empathie gar kein kohärentes Phänomen beschreibe, sondern ein *umbrella term* für eine Reihe verschiedener Phänomene sei (siehe den Beitrag von Davis in Preston & De Waal 2002: 32-33). Vreeke und Vandermark (2003:178) beispielsweise definieren Empathie 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*). Da Empathie in der allgemeinen Formulierung als ein *shared-state*-Phänomen aufgefasst wird (Preston & De Waal 2002b: 287), bietet sich *Imaging* als Ansatz für das Erkennen solcher *shared states* im Sinne gleichartiger neuronaler Aktivierungsmuster an. Entsprechend finden sich in jüngster Zeit zahlreiche *Imaging*-Studien in diesem Feld, die heute meist den *social neurosciences* zugeordnet werden (für eine Übersicht siehe Singer 2006). Wichtig in diesem Kontext ist, dass die Empathieforschung (und die *social neuroscience* generell, siehe Rizzolatti & Fabbri-Destro 2008) hinsichtlich der neuronalen Grundlagen der Empathie (kontrovers diskutierte) Bezüge zu den Spiegelneuronen setzt. Diese Neuronen gehören zu den so genannten *visuomotor neurons* und wurden ursprünglich im Areal F5 des prämotorischen Kortex von Affen entdeckt (Rizzolatti & Craighero 2004). Der Begriff „visuomotor" sagt aus, dass diese Neuronen sowohl dann aktiv sind, wenn der Affe bestimmte Bewegungen (unabhängig von der Art des Objektes) vollzieht, wie auch, wenn der Affe dieselbe Bewegung bei anderen Affen beobachtet. Die Entdeckung der Spiegelneuronen gab Anlass zu unterschiedlichen Hypothesen (Rizzolatti & Craighero 2004): So dürften sie eine wichtige Rolle für die Imitation und damit das Erlernen von Handlungen spielen. 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 verwendete Methode der elektrischen Ableitung von Nervenzellen nicht verwendet werden kann. Es

gibt eine Reihe von *Imaging*-Experimenten, welche mit der Existenz von Spiegelneuronen beim Menschen kompatibel sind. Dies gilt aber noch nicht als Beweis für die Existenz von Spiegelneuronen bei Menschen, sondern nur für die Existenz eines *mirror systems* (zuweilen auch *mirror neuron system* genannt) im menschlichen Gehirn, das möglicherweise von solchen Neuronen gebildet wird.

- **Psycho-/Soziopathologie:** Die Erforschung psychopathologischer Zustände und ihrer neuronalen Grundlage bildet ebenfalls ein weites Feld, das hier ebenfalls nur sehr rudimentär abgedeckt werden kann. Wichtig ist, dass die Untersuchung von Personen mit abnormen moralischem Verhalten (moralische Pathologien) nicht nur eine wichtige Erkenntnisquelle für die *neuroscience of ethics* bildet, sondern auch in ihrer historischen Rekonstruktion als disziplinäres Feld bedeutsam ist. Vorab in der US-Literatur wird der Fall des Eisenbahnarbeiters Phineas Gage, bei welchem 1845 durch einen Sprengunfall eine Eisenstange durch den vorderen Schädelbereich getrieben wurde (Karnath & Thier 2003: 515-516), als historischer Referenzfall genannt. Klare Fälle von moralischen Pathologien, bei denen sich Hirnschädigungen mit moralischen Defiziten in Verbindung bringen lassen, haben die mediale Verbreitung der *neuroscience of ethics* (wie der *social neuroscience* generell) gefördert – etwa mit den Arbeiten von Damasio (Damasio 2003, Damasio 1999, Damasio 1994). Diese Forschung fokussiert die Entscheidungsfähigkeit der Betroffenen in moralischen und/oder sozialen Kontexten und postuliert insbesondere eine Dissoziation zwischen der Fähigkeit einer abstrakten moralischen Kognition und einer emotional unterlegten moralischen Kognition (bzw. den Verlust von Empathie beim Psychopathen, verstanden als Fähigkeit zu fühlen, was das Opfer fühlt und dies als Handlungsmotivation zu nehmen). Darauf wird in Abschnitt 3.4.1 weiter eingegangen.

ABBILDUNG 7:

Abbildung 7: Eine Übersicht über die disziplinären Felder, die wesentliche Impulse zur sozialen Neurowissenschaft geliefert haben (sowohl an den Überlappungsfeldern als auch in exemplarischen Entwicklungen innerhalb der Felder). Dargestellt ist ebenfalls eine erste Binnendifferenzierung der sozialen Neurowissenschaft mit Einbettung der *neuroscience of ethics* und einigen zentralen Konzepten, die in mehreren Bereichen untersucht werden.

Diese vier Felder können aufgrund der bibliometrischen und qualitativen Analyse (siehe Fußnote 26) als wesentliche Vorläufer bzw. Impulsgeber für die sich entwickelnde soziale Neurowissenschaft identifiziert werden (siehe Abbildung 7). Diese *social neuroscience* mit der so genannten Neuroökonomie als bedeutsamste Binnendifferenzierung sollen nun noch kurz skizziert werden:

- **Soziale Neurowissenschaft:** Die Soziale Neurowissenschaft (*social neuroscience*, zuweilen auch *social cognitive neuroscience*) hat die neuronalen Grundlagen des Sozialverhaltens sowohl von Tieren wie Menschen sowie die Rückwirkung dieses Verhaltens auf das Gehirn (untersucht unter dem Gesichtspunkt der Ontogenese des Gehirns, der neuronalen Plastizität etc.) als Forschungsgegenstand (Adolphs 2003, Blakemore et al. 2004). Ansätze zur Ausdifferenzierung einer solchen Disziplin finden sich seit Beginn der 1990er Jahre (Cacioppo & Berntson 1992). Unterschiedliche Grundfragestellungen charakterisieren dieses Feld. Adolphs (2003) nennt deren drei: 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. (2004: 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 Fernald (2004) 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? Lieberman (2007: 259) schließlich nennt vier Forschungsgebiete, welche die soziale Neurowissenschaft charakterisieren sollen: „(a) understanding others, (b) unerstanding oneself, (c), controlling oneself, (d) the processes that occur at the interface of self and others". Angesichts dieser unterschiedlichen Fragestellungen erstaunt es nicht, dass die ersten Bücher zum Thema ein eher breit gefasstes Verständnis dieses disziplinären Feldes vorschlagen (Harmon-Jones & Winkielman 2007). Da sich grundsätzlich bei jedem sozialen Phänomen die Frage nach den dabei involvierten neuronalen Prozessen stellen lässt (einmal abgesehen davon, ob man dies auch methodisch sinnvoll untersuchen kann), ist in der sozialen Neurowissenschaft ein reiches Potential an Ausdifferenzierungen vorhanden, die Gegenstand der in Fußnote 26 beschrieben Studie ist.

- **Neuroökonomie:** Die so genannte Neuroökonomie (oder auch Neuroökonomik, *neuroeconomics*) dürfte die bislang am deutlichsten ausdifferenzierte Binnendisziplin der Sozialen Neurowissenschaft sein. Das oben bereits angesprochene *decision making* in ökonomischen (oder generell: sozialen) Kontexten ist der zentrale Fokus des Erkenntnisinteresse dieser (Sub-)Disziplin. Ob unter „ökonomischen Kontexten" relativ abgrenzbare Fragen der Ökonomie als Wissenschaft (Kenning & Plassmann 2005) oder gar jegwelcher Umgang von Organismen mit Knappheit (Montague 2007) verstanden werden soll, wird derzeit noch offen diskutiert.[27] Hier widerspiegelt sich möglicherweise die auch in anderen Bereichen feststellbare Diskussion um die Reichweite des ökonomischen Paradigmas (z.B. welche Bereiche der sozialen Organisation einer Marktdynamik unterworfen werden sollen). Motivierender Gedanke der Neuroökonomie ist die Frage, inwieweit die in der klassischen Ökonomie verwendeten Modelle von Entscheidungsfindung und Rationalität dem realen Entscheidungsverhalten unter Unsicherheit entsprechen (siehe z.B. Loewenstein et al. 2008). Diese Fragestellung wurde bereits innerhalb der Ökonomie und Psychologie (unter anderem mit den unter Abschnitt 2.2.4 genannten Methoden) seit längeren untersucht (erinnert sei an die *prospect theory* von Daniel Kahnemann) und insofern nicht typisch für die Neuroökonomie. Diese erfuhr erst durch den Einbezug neurowissenschaftlicher Methoden ihr eigenständiges Gepräge.

Es liegt nahe, die *neuroscience of ethics* ebenfalls eine Binnendifferenzierung innerhalb der sozialen Neurowissenschaft anzusehen, wobei insbesondere zur Neuroökonomie eine unscharfe Abgrenzung bestehen dürfte, sobald Konzepte wie Fairness, Kooperation und Vertrauen zum Untersuchungsgegenstand werden. Im Hinblick auf die vier genannten disziplinären Feldern, aus denen die heutige soziale Neurowissenschaft zentrale Impulse erhalten hat, zieht die *neuroscience of ethics* ebenfalls aus allen vier Bereichen Erkenntnisse für die Theoriebildung heran. Dies soll im folgenden Abschnitt 3 nun im Detail vorgestellt werden.

[27] Interessant in diesem Zusammenhang ist auch, dass zunehmend eine ökonomische Begrifflichkeit für die Beschreibung neuronaler Prozesse Anwendung findet. So fragen Montague und Berns (2002) beispielsweise, ob es eine Art „Währung" im Gehirn gebe, mit welcher man den im Gehirn ablaufenden Prozess einer Entscheidung universell messen oder bewerten könne. Glimcher (2002) postuliert, dass das Gehirn dazu da sei, effiziente Entscheide im Sinn der Ökonomie zu treffen.

3 Die Erforschung neuronaler Grundlagen der Moral

Dieser Abschnitt bietet eine Übersicht über Forschungsarbeiten, die in den vergangenen zehn Jahren entstanden sind und der *neuroscience of ethics* zugeordnet werden können. Die Einführung zeigt grundlegende Zielsetzungen dieser Forschungsrichtung, gibt einen Überblick über die unterschiedlichen Theorien, die sich bislang ausdifferenziert haben und schätzt den Einfluss dieser Forschungen auf andere wissenschaftliche Disziplinen ab. Danach erfolgt eine Analyse dieser Forschungen anhand der in Abschnitt 1.3.3 vorgestellten Gliederung moralischer Stimulus, *decision making*, Handlung und Begründung.

3.1 Einführung: Ziele, Modelle und Resonanz der *neuroscience of ethics*

In jüngerer Zeit findet sich eine zunehmende Anzahl neurowissenschaftlicher Studien, welche explizit Moral bzw. moralisches Verhalten zum Gegenstand haben. In den letzten zehn Jahren sind über 60 Arbeiten publiziert worden, die den Gegenstand dieser Untersuchung bilden und Grundlage dieses Abschnittes bilden.[28] Bereits 2003 haben Casebeer und Churchland in einer der ersten Übersichtsarbeiten festgehalten, dass diese Arbeiten vor allem moralische Emotionen, moralische soziale Kognition und abstraktes moralisches Denken zum Gegenstand haben. Sie zeichneten sich durch sehr vereinfachende Annahmen über *moral reasoning* aus. Casebeer und Churchland stellen fest, dass bereits durch die frühen Arbeiten klar geworden sei, dass es kein abgrenzbares *moral module* oder „Moralzentrum" im Gehirn gebe, was im Übrigen auch unplausibel sei. Spätere Arbeiten (siehe Abschnitt 3.2.2) widersprechen dieser Schlussfolgerung bis zu einem gewissen Grad.

Casebeer und Churchland sehen – nebst einem genuinen Erkenntnisinteresse an empirisch erfassbaren Aspekten von Moral – drei weiterführende Motive für solche Forschungen: Erstens sollen sie Hilfestellung zur Identifizierung moralischer Pathologien mit biologischer Ursache liefern; zweitens sollen Hinweise für eine Verbesserung der Moralerziehung gewonnen werden; drittens sollen solche Studien Beiträge für die Lösung normativ-ethischer und metaethischer Fragen liefern.[29] Diese Motive finden sich auch in anderen Arbeiten, wobei der sozialtechnologische Impetus zuweilen deutlich hervortritt – beispielsweise in den Worten von Moll et al.: „Understanding the neural basis of moral cognition will help to shape environmental, psychological and medical interventions aimed at promoting prosocial behaviours and social welfare" (Moll et al. 2005).

Diese Motive sind durchaus Gegenstand kritischer Einwände Seitens anderer Disziplinen. Grundsätzlich lässt sich sicher kaum bestreiten, dass die Diskussion (mancher) ethischer Fragen von solchen empirischen Erkenntnissen profitieren kann. Es ist aber auch klar, dass die Reichweite neurowissenschaftlicher Erkenntnisse hinsichtlich der Erreichung der durch diese Motive implizierten Ziele kontrovers diskutiert wird. Sowohl „moralische Pathologien" als auch „Moralerziehung" sind Phänomene, die in einem sozialen Raum stattfinden. Entsprechend sind beispielsweise die Bedingungen, unter denen solche Phänomene naturwissenschaftlich untersucht werden (z.B. Wahrnehmung von Patienten mit frontalen Hirnschäden im

[28] Aus diesen Arbeiten wird nachfolgend nur eine Auswahl genauer vorgestellt.

[29] Casebeer beispielsweise argumentiert, dass die Erkenntnisse der Neurowissenschaft die aristotelische Tugendethik gegenüber deontologischen und utilitaristischen Konzeptionen auszeichnen würde (Casebeer 2003). Roskies (2005) plädiert dafür, dass Studien an Patienten mit Schäden am ventromedialen Frontalkortex den so genannten *motive internalism* (die These, dass „a moral belief or judgment is intrinsically motiating", S. 22) als metaethische Position widerlegen würden. Zu solchen und anderen Thesen finden sich natürlich Gegenpositionen, die nicht Gegenstand dieser Studie sind.

Labor vs. in der klinischen Praxis), ebenso einzubeziehen, um solche Fragen zu klären. Eine kritische Auseinandersetzung mit diesen Motiven – auch hinsichtlich der die normative Ethik und Metaethik betreffenden Thesen – ist aber nicht Thema dieser Studie und wird teilweise in den anderen Beiträgen dieses Bandes aufgegriffen.

Obgleich sich zuweilen weit greifende Ausführungen zur Relevanz der erzielten Ergebnisse in den *Conclusion*-Abschnitten der untersuchten Arbeiten finden, so sind sich die Forscher im Bereich der *neuroscience of ethics* durchaus klar über die Komplexität des Phänomens, das sie untersuchen. Moll et al. (2005) nennen beispielsweise folgende 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 von Moralsystemen – zweifellos bedeutende und auch bekannte Schwierigkeiten für empirische Moralforscher.

Insgesamt betrachtet sind die Forschungen im Bereich *neuroscience of ethics* als Teil einer größeren Entwicklung in der Ethik insgesamt anzusehen; einer zunehmenden Hinwendung zu empirischen Aspekten der Ethik (*empirical ethics*, siehe Musschenga 2005). Die zunehmenden Bemühungen zur Vernetzung solcher Forschungen findet unter anderem Ausdruck im kürzlich erschienenen Dreibänder „Moral Psychology“ (Sinnott-Armstrong 2008), der vorab die amerikanischen Forscherinnen und Forscher (rund 80 Personen) vereint. In der aktuellen Diskussion (2009) können drei „Theorien“ unterschieden werden, die sich etwa Mitte dieses Jahrzehnts formierten, (idealtypisch) drei Zugangsweisen zum Problem aufzeigen und (von der Tendenz her) aus unterschiedlichen wissenschaftlichen Traditionen und Fragestellungen stammen (Abbildung 8). Gewiss finden sich Überschneidungen und Einflüsse anderer Forschergruppen, die nachfolgend nur unvollständig wiedergegeben werden können. Dennoch gibt diese Übersicht eine erste Orientierung, ohne dass hier eine detaillierte Darstellung der Modelle und deren Kritik erfolgen können:

- **Universal Moral Grammar Model:** Dieses Modell geht von der Intuition aus, dass Chomsky’s Theorie einer generativen Grammatik auch für das Verständnis von Moral hilfreich ist. „Verständnis von Moral“ meint hier insbesondere eine Erklärung für die Generierung bestimmter moralischer Urteile (und Handlungen) aufgrund von Wahrnehmungen und die Aneignung dieser Fähigkeit durch den *moral agent*. Kurz gefasst behauptet dieses Modell die Existenz einer *moral grammar* als Ergebnis eines evolutiven Prozesses, die es dem *moral agent* erlauben soll, Urteile hinsichtlich der moralischen Angemessenheit bestimmter Verhaltensweisen zu fällen, wobei Emotionen wie rationale Deliberation dem Wahrnehmungs- und Urteilsprozess *nachgeschaltet* sind. Die *moral grammar* besteht dabei aus einem Set unterschiedlicher Typen von hierarchisch strukturierten Regeln (siehe Hauser 2006, Mikhail 2007). Protagonisten dieses Modells sind Marc Hauser und John Mikhail, wobei Hauser aus der Primatologie stammt und sich stark für „Vorformen“ von Moral in anderen Spezies interessiert (relevante Autoren sind hier u.a. Frans de Waal, siehe den Beitrag von Carel van Schaik und Claudia Rudolf von Rohr in diesem Band) und er im Zug der Entwicklung dieses Modells auch entsprechende Bezüge setzt (Hauser 2006b). Neurowissenschaftliche Erkenntnisse bildeten zu Beginn keine große Bedeutung bei der Theorieentwicklung, werden derzeit aber (unter anderem) durch Liane Young (2007) vermehrt einbezogen. Kritiker an diesem Modell bezweifelt unter anderem, dass moralisches Verhalten eine der Linguistik vergleichbaren Grammatik unterliegt und dass dieses Modell faktische Diversität im moralischen Verhalten erklären kann (Dupoux & Jacon 2007).

- **Dual-Processing Model:** Dieses Modell basiert auf der klassischen Unterscheidung zwischen Emotion und Kognition (im Sinne rationaler Deliberation), die als gegenwirkende Kräfte moralische Urteilsbildung beeinflussen. Ausgangspunkt dieses Modells war ein scheinbar paradoxes Entscheidungsverhalten bei verschiedenen Dilemmas (siehe Abschnitt 3.2). Kurz gefasst behauptet das Modell, das moralische Entscheidungssituationen in unterschiedlicher Stärke Emotionen hervorrufen, wobei starke Emotionen gegenüber der rationalen Deliberation überwiegen, außer es werden zusätzliche kognitive Ressourcen rekrutiert (was die erhöhte Aktivität spezifischer Hirnregionen bedingt). Das Modell lehnt sich an zahlreiche weitere *dual-processing*-Ansätze in der Psychologie (Evans 2008) an, insbesondere am *social intuist model* von Jonathan Haidt (siehe Abschnitt 3.5). Protagonist dieses Modells ist Joshua Greene (Greene 2001/2004/2007), der mit diesem Modell Unterschiede zwischen deontologischen und utilitaristischen Entscheidungsmustern erklären will – eine Erklärung, die von Kritiker angezweifelt wird (Timmons 2008).

- **Event-Feature-Emotion-Complex:** Dieses Modell bezweifelt die Möglichkeit einer strikte Trennung zwischen emotionalen und rationalen Aspekten in der moralischen Urteilsbildung. Demgegenüber seien die Komponenten *event knowledge* (kognitive Wissensinhalte über das zu beurteilende Ereignis), *features* (relevante Eigenschaften von Wahrnehmungen, z.B. Gesichtsausdrücke) und *emotions* (emotionale Grundzustände wie z.B. Aggressivität) die relevanten Einflussfaktoren, die sich jeweils unterschiedlichen Hirnregionen zuordnen ließen. Protagonisten dieses Modells sind Jorge Moll und Mitarbeiter (für eine Übersicht siehe Moll et al. 2005/2008), wobei nebst *Imaging* auch die Untersuchung von Personen mit spezifischen Hirnschäden bei der Theoriebildung einbezogen werden (De Oliveira-Souza 2008) – eine Erkenntnisquelle, der sich auch die anderen Modelle bedienen. Kritiker dieser Modells sehen darin vorab eine Zusammenstellung zahlreicher Einzelfaktoren ohne klares theoretisches Gerüst und bemängeln ein unklares (bzw. reduziertes, rein am Altruismus orientiertes) Verständnis von Moral (z.B. Hynes 2008).

ABBILDUNG 8:

Abbildung 8: Eine Übersicht über die wichtigsten theoretischen Ansätze in der *neuroscience of ethics*.

Idealtypisch geben diese drei Ansätze den „klassischen" Komponenten moralischer Urteilsbildung, Emotion und Kognition, unterschiedliche Rollen: Entweder sind beides nachgelagerte Phänomene, oder stehen in einem Konkurrenzverhältnis oder aber sind zwar zentrale, aber nicht klar separierbare Einflussfaktoren. Diese Darstellung dürfte keine abschließende Klassifikation des Feldes möglicher Theorien sein, sie benennt aber die derzeit dominierenden Ansätze. Unabhängig davon ist festzuhalten, dass zahlreiche weitere Forschungsgruppen Beiträge liefern, die in der Theoriebildung dieser drei Ansätze eine Rolle spielen. Abbildung 8 zeigt die wichtigsten davon. Einerseits Untersuchungen, die heute der Neuroökonomie zugeordnet werden und moralnahe Konzepte wie Fairness, Kooperation und Vertrauen zum Thema haben; andererseits Arbeiten, die mit verschiedenen Mitteln Kontext-Effekte genauer untersuchen. Hier kann die Arbeit von Philosophen wie Prinz (2007) und Nichols (2002) genannt werden, die sich stark auf Arbeiten empirischer Forscher beziehen.[30]

[30] Es sei erwähnt, dass gewisse Forschungsfragen, die beispielsweise in der Moralpsychologie seit längerem Thema sind, in der *neuroscience of ethics* praktisch nicht untersucht werden. Ein Beispiel ist die „moralische Scheinheiligkeit" (*moral hyporisy*), das unter anderem von Batson (2003) untersuchte Phänomen, wonach Versuchspersonen gegen Aussen hin moralisch korrekt erscheinen wollen, die Kosten des moralischen Verhaltens aber zu vermeiden versuchen.

Die Forschungen in der *neuroscience of ethics* stoßen durchaus auf Resonanz, wie eine Impact-Analyse[31] aufzeigt. Diese erfolgt anhand der Arbeiten von Joshua Greene und Jorge Moll – jenen Wissenschaftlern, die am spezifischsten im Bereich der *neuroscience of ethics* publiziert haben. Abbildung 9 zeigt einen deutlichen Unterschied zwischen Publikationen und Zitationen: Beide Autoren publizieren vorab in den Neurowissenschaften, werden aber am meisten in sozial- und geisteswissenschaftlichen Arbeiten zitiert. Im Vergleich zu anderen Themen der sozialen Neurowissenschaft hat „Moral" damit den weitaus deutlichsten Widerhall in den Sozial- und Geisteswissenschaften.

ABBILDUNG 9:

Abbildung 9: Eine Impact-Analyse der Arbeiten der beiden Autoren Joshua Greene und Jorge Moll belegt, dass die *neuroscience of ethics* in den Sozial- und Geisteswissenschaften auf großes Interesse stoßen.

3.2 Moralische Stimuli

Die experimentelle Untersuchung von moralischem Verhalten verlangt nach der präzisen Definition eines moralischen Stimulus, der dieses Verhalten auslöst. In *Imaging*-Experimenten gehen die Forscher davon aus, dass man solche moralischen Stimuli definieren und gegenüber anderen Stimuli auszeichnen könne.[32] Bei gewissen fMRI-Experimenten muss man (je nach Art der Durchführung des Experiments) für die Referenzmessung andere Stimuli mit einer vergleichbaren Wahrnehmungskomplexität aber ohne moralischen Gehalt finden.[33] Dazu 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 Ausgestaltung dieser Stimuli soll im Folgenden ausgeführt werden. Ihre Verwendung in Experimenten zwecks Suche nach neuronalen Korrelaten von moralischen Entscheidungen wird in Abschnitt 3.3.2 erläutert.

- **Bilder:** Fotografien von (in vielen Fällen) Szenerien mit Menschen, die einen moralischen Gehalt zum Ausdruck bringen sollen, sind ein oft verwendeter Stimulus. Moll et al. (2002) charakterisieren moralische Bilder als „portraying emotionally charged, unpleasant social scenes, representing moral violations". Beispiele sind Bilder, die physische Bedrohungen zum Ausdruck bringen, Kriegsszene, oder Bilder verlassener Kinder. Als Kontrollstimuli 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 ohne spezifischen Gehalt. Weiter können auch Bilder verwendet werden, die nicht einen direkt zugänglichen, sondern einen gelernten moralischen Gehalt aufweisen. Singer et al. (2004) verwendet in einer Studie Gesich-

[31] Zu diesem Zweck werden die Publikationen von Autoren und deren Zitationen klassifiziert anhand verschiedener wissenschaftlicher Bereiche miteinander verglichen. Ausgangspunkt sind die Daten des *ISI Web of Science* (siehe Fussnote 21). Die Methodik wird in Matussall et al (in Vorbereitung) erläutert.

[32] Gemäss Moll et al. (2002b: 697) soll dies bereits in psychologischen Studien der 1980er und 1990er Jahre gezeigt worden sein; siehe auch der Beitrag von Eslinger et al. in Preston & De Waal (2002: 34-35).

[33] Eine Möglichkeit ist es, die Unterscheidung zwischen moralischen Regeln und konventionellen Regeln zu verwenden (*moral-conventional-distinction*, siehe dazu Blair, 1995). Doch muss hier der Einbezug von Emotionen bei der Beurteilung konventioneller Regeln berücksichtigt werden (Nichols 2002). Die Verletzung von Konventionen, welche einen affektiven Charakter haben (z.B. Ekel erzeugen wie auf den Tisch spucken) dürften anders beurteilt werden als Konventionen, welche keine solche affektive Komponente haben und demnach auch als Referenz-Stimuli unterschiedlich wirken.

ter von Personen, die zuvor in einem Spiel (Gefangenendilemma) als kooperierende oder als defektierende Personen aufgetreten sein sollen. Hier figurieren die Bilder als Platzhalter von moralisch wünschenswertem oder anzulehnendem Verhalten.

Die Verwendung von Bildstimuli ist mit einer Reihe von Problemen behaftet: Generell dürfte es schwierig sein, den moralischen Gehalt von Bildern präzise zu charakterisieren – auch wenn die Bilder vorab durch Testgruppen klassifiziert werden. So weisen Heekeren et al. (2003) 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. Dieser Zusammenhang wurde experimentell geprüft, indem Bilder verwendet wurden, die Körperverletzungen zeigten (Heekeren et al. 2005). 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 (Phelps 2001). Diese Beispiele zeigen, dass unerkannte Korrelationen in den verwendeten Bildern die Resultate beeinflussen können, da eben diese Korrelationen zu den in den *Imaging*-Studien beobachteten neuronalen Aktivierungen führen können.

- **Sätze:** Eine Möglichkeit, dieses Problem zu umgehen, bietet die Verwendung einfacher Sätze, weil deren Inhalte im Unterschied zu Bildern explizit gegeben 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 werden Sätze verwendet, die semantisch zu bewertende Sachverhalte zum Ausdruck bringen (z.B. „Das Viereck ist rund“) und dann als „korrekt“ oder „inkorrekt“ bewertet werden. Ein solcher Ansatz verwendete Heekeren et al. (2003). Moll et al. (2002b) 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 Fragen offen. Ist beispielsweise das erste Beispiel (das Nichtbenutzen eines Sicherheitsgurtes einhergehend mit einer Selbstgefährdung) 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 wurde zwar in einer Begleituntersuchung validiert. Dennoch erscheint es seltsam, warum in der Studie derart mehrdeutige Beispiele genannt werden. Spätere Studien (Zahn et al. 2009) lassen hingegen auf eine erhöhte Sensibilität auf solche Einflussfaktoren (z.B. psycholinguistische Aspekte) schließen.

- **Dilemmas:** Die bislang komplexesten Stimuli bestehen aus Beschreibungen moralischer Dilemmas, üblicherweise mit längeren Texten, teilweise kombiniert mit Bildinhalten. Diese Dilemmas vereinen miteinander konfligierende moralische Werte und verlangen von der Versuchsperson eine Entscheidung darüber, welcher Wert verletzt werden soll. Greene et al. (2001/2004), unterscheidet dabei zwei Arten von Dilemmas:

personal und *impersonal* Dilemmas. In ersteren ist die Versuchsperson aufgefordert sich vorzustellen, durch Einsatz seines Körpers in das Szenario einzugreifen, wobei drei Kriterien ein persönliches Dilemma definieren: Erstens ist einer der beiden im Dilemma involvierten moralischen Werte von solcher Art, dass seine Verletzung zur Schädigung eines menschlichen Körpers führt; zweitens betrifft diese Schädigung eine konkrete Person oder Personengruppe; und drittens ist die Schädigung keine Folge eines Abwehrverhaltens. Unpersönliche Dilemmas erfüllen diese Kriterien nicht.

Greene verwendet unter anderem folgende persönliche Dilemmas: 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, dass die Versuchsperson eine beleibte Person von einer Brücke auf das Geleis 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 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 unpersönlichen Dilemmas werden 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 ein Teenager 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 gut untersuchten Unterschied, dass persönliche Dilemmas intuitiv mehrheitlich 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 mit neurobiologischen Mitteln zu verstehen.

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 in beiden Fällen in die involvieren Personen hinein versetzen und damit vergleichbare emotionale Empfindungen haben kann – demnach könnte die Vorausklassifikation da-

von Abhängen, unter welchen Bedingungen die Versuchspersonen diese vornehmen müssen.[34] Auch die Klassifikation der Optionen, die durch solche Dilemmas vorgegeben werden, ist keine einfache Angelegenheit (siehe die Bemerkungen unter Abschnitt 2.2.3). Schließlich sind, wie bei Bildstimuli, unterkannte Korrelationen zwischen den verschiedenen Dilemmas denkbar, was die Auswertung der Resultate der *Imaging*-Studie erschwert.

Zur generellen Charakterisierung moralischer Stimuli ist zudem vorgeschlagen worden, zwischen einfachen und komplexen moralischen Stimuli zu unterscheiden (Heekeren et al. 2003). Einfache Stimuli zeichneten sich durch einen eindeutigen, nicht-dilemmatischen Charakter aus, während komplexe Stimuli einen starken emotionalen Charakter haben. Auch hier ist unklar, ob diese Unterscheidung für alle Klassen Stimuli gleichermaßen zu bewerkstelligen ist. Wie wollen Bilder einen moralischen Sachverhalt zum Ausdruck bringen, gleichzeitig aber keine emotionale Reaktion auslösen? Bei Sätzen wiederum scheint diese Unterscheidung eher möglich, was wohl der Grund ist, dass Heekeren et al. für ihre Untersuchung Sätze verwendet hatten.

Hinsichtlich der Charakterisierung des moralischen Gehalts solcher 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 (Takahashi et al. 2004). 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. Schließlich ist es auch möglich, dass sich bei der Klassifizierung von Stimuli als „moralisch" *gender*-Unterschiede zeigen. Moll et al. (2002) sprechen dieses Problem an und verweisen darauf, dass dies offenbar noch nicht untersucht worden sei. In ihren Daten ließen sich aber bisher keine deutlichen Hinweise auf solche Unterschiede nachweisen.

3.3 *Moral decision making:* Kognition und Emotion

3.3.1 Moralische Kognition und Intuition

Im Zug der Forschungen in der *neuroscience of ethics* stellt sich die Frage, ob *moral reasoning*, *moral cognition* bzw. *moral decision making* (diese Begriffe finden sich gleichermaßen in der Literatur und werden meist synonym gebraucht) ein genuines Phänomen darstellen und wie deutlich sich dieses von anderen Formen sozialer Kognition abgrenzen lässt. Grundsätzlich sind es die Inhalte, die Kognition zu *moralischer Kognition* machen, wie etwa Casebeer und Churchland (2003) und Moll et al. (2008) festhalten. Die spezifischen moralischen Inhalte werden aber meist nur vage definiert. So schreiben Casebeer und Churchland: „Moral reasoning deals with cognitive acts and judgments associated with norms, or with facts as they relate to norms" (2003: 171). Moll et al. schreiben: "its [moral cognition] most distinctive feature is the ability to altruistically motivate social behaviour" (2008: 162). Casebeer (2003) nennt sechs charakteristische Eigenschaften moralischer Kognition: Sie involviert Emotionen

[34] So ist beispielsweise bekannt, dass vorab geäusserte Anweisungen, wie stark man sich in die in Versuchsszenarien vorkommenden Personen einfühlen muss, einen Einfluss auf das Entscheidungsverhalten haben kann (Batson et al. 2003).

(ist *hot*), bezieht sich auf einen sozialen Aspekt, ist abhängig vom Kontext, orientiert sich auf ein Ziel hin, ist *distributed* und *genuine*. Angesichts dieser inhaltlich eher vagen Bestimmung erwarten Casebeer und Churchland keine klaren Unterschiede zwischen *practical reasoning* und *moral reasoning* hinsichtlich der involvierten neuronalen Prozesse.

Ähnlich vage bleibt der Intuitionsbegriff, der in der moralischen Kognition als Platzhalter für jene Prozesse im Verlauf des *moral decision making* steht, die nicht der deliberativen Kontrolle des *moral agent* unterliegen. Dabei ist festzuhalten ist, dass der Begriff „intuition" in der neurowissenschaftlichen Literatur erstaunlich selten auftaucht[35] – auch bei den Arbeiten, die explizit die neuronalen Grundlagen moralischen Verhaltens untersuchen Greene (2003) spricht von Intuitionen (*intuitions, gut-feelings*) und verwendet den Begriff weitgehend im Sinne der von Haidt (2001) angesprochenen Unterscheidung zwischen rationalen versus intuitiven Entscheidungsstrategien. Arbeiten aus der Emotionsforschung sehen in Intuitionen eine Art Entscheidungshilfe-Gefühle, die verhältnismäßig wenig kognitive Ressourcen beanspruchen sollen (Turnbull et al. 2005). Eine explizite, mit dieser Auffassung vereinbare Definition moralischer Intuitionen liefert Haidt: „Moral intuition appears to be the automatic output of an underlying, largely unconscious set of interlinked moral concepts (2001: 825). Lieberman (2000) schlägt schließlich vor, Intuition 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" (Lieberman 2000: 111). Damit wären (im Gegensatz zu einem Instinkt) bestimmte Aspekte von Intuitionen – wie etwa die Fähigkeit, solche zu erfassen und zu nutzen – durch Lernen verbesserbar.

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. Diese Studien stehen unter der Leithypothese, dass eine Reihe von Teilprozessen, die zum *moral decision making* gehören, unbewusst ablaufen und diesen Prozess entscheidend prägen (Greene & Haidt 2002). Solche quasi automatisierten Prozesse führen bei Personen, die mit moralischen Dilemmas konfrontiert sind, zu einer unmittelbaren Einschätzung der Situation als „moralisch gut" oder „moralisch schlecht". Diese Hypothese wird im direkten Gegensatz zum Kohlbergschen Modell der moralischen Entwicklung gesetzt (Kohlberg 1995), das – bei einem ausgereiften moralischen Bewusstsein – rationalen Prozessen der moralischen Entscheidungsfindung eine zentrale Rolle einräumt. Im Folgenden soll nun eine Übersicht über einige Studien gegeben werden, die nach neuronalen Korrelaten für moralische Kognition suchen:

- In einer ersten, für das Gebiet wegweisenden Studie von Greene et al. (2001) wurden 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. Geprüft wurde die Frage, inwiefern Dilemmas mit starkem emotionalen Gehalt (persönliche Dilemmas) andere Hirnregionen aktivieren als unpersönliche Dilemmas. Diese Vermutung wurde bestätigt, d.h. Hirnregionen, die mit *emotional processing* zu tun haben (der mediale frontale Gyrus, der posteriore *cingulate gyrus*, und der *angular gyrus*), sind bei persönlichen Dilemmas stärker aktiviert als bei unpersönlichen Dilemmas

[35] Innerhalb der Menge *neuroscience* (siehe Fussnote 22) findet sich der Ausdruck „intuition" ingesamt in 127 Arbeiten, 98 davon sind Arbeiten, die seit dem Jahr 2000 erschienen sind (Suche vom 2. April 2009).

sowie den Kontrolldilemmas. 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 (untypisch) 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.

- In einer weiteren Studie von Greene et al. (2004) wurden Versuchspersonen im Scanner (fMRI) mit einer Reihe verschiedener Dilemmas konfrontiert. In einem ersten Schritt mussten die Versuchspersonen *personal* und *impersonal* Dilemmas lösen, dies mit dem Ziel, die Ergebnisse der oben genannten 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 andere Unterscheidung wurde durch eine vorgängige Kategorisierung der verwendeten Dilemmas durch Testpersonen validiert. Ein schwieriges Dilemma ist dadurch charakterisiert, dass eine emotional problematische Komponente (z.B. Töten eines Babys 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. Ziel des Versuchs war es unter anderem, das unter Abschnitt 3.1 genannte *dual-processing model* zu prüfen (das damals noch nicht so genannt wurde). Die Autoren kamen 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.

- In einer fMRI-Studie von Moll et al. (2002) wurden Versuchspersonen mit verschiedenen Bildstimuli (moralische Bilder vs. Varianten unangenehmer Bilder, siehe oben) konfrontiert. Ziel war 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 wurden die Bilder von den Versuchspersonen hinsichtlich des emotionalen und moralischen Gehalts bewertet. Die Autoren kamen zum Schluss, dass das Betrachten moralischer Bilder insbesondere mit einer erhöhten Aktivierung des orbitalen und medialen präfrontalen Kortex und des superioren temporalen Sulcus einhergeht.

- In einer weiteren fMRI-Studie von Moll et al. (2002b) wurde zwischen emotionalen Stimuli in Form von Sätzen mit und ohne moralischem Gehalt unterschieden – und zwar unter Verwendung der Kategorien „nicht-moralisch neutral", „nicht-moralisch unangenehm" und „moralisch" (siehe oben). Die Autoren kamen 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. (2003) wurden Versuchspersonen mit kurzen Sätzen (einfachen moralischen Stimuli, siehe oben) konfrontiert, die entweder einen moralisch oder einen semantisch richtigen oder falschen Sachverhalt beschreiben. 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 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.

- In einer Studie von Singer et al. (2004) wurden Bilder von Gesichtern als moralischer Stimulus verwendet. 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 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 intentional oder nach einem vorgegebenen Schema handelt. In einem dritten Schritt wurde 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) – 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 stärker sind als jene, die bei Nicht-Kooperierenden auftreten.

Weitere vergleichbare Studien sind nach 2005 erschienen, in denen die Dilemmas ausgefeilter wurden (z.B. Schaich Borg et al, 2006), spezifische Aspekte moralischer Kognition untersucht wurden (z.B. *belief attribution*, Young & Saxe 2008) oder andere normative Konzepte wie die *justice-care*-Unterscheidung (Robertson et al. 2007) oder Tugenden (Takahashi et al. 2008) thematisiert wurden. Demnach steht heute eine große Zahl an Studien zur Verfügung, die neuronale Korrelate unterschiedlicher Formen moralischer Kognition aufzeigen, wobei sich die Unterschiede anhand der Inhalte, der verwendeten Stimuli und weiterer geänderter Kontext-Bedingungen manifestieren.

Inwieweit sich daraus ein einheitliches Bild eines „moralischen Gehirns“ gestalten lässt, ist aber keineswegs klar. Meta-Studien, die Resultate unterschiedlicher *Imaging*-Studien zu einem einheitlichen Bild formen wollen, stehen vor dem Problem, dass die Resultate unter Verwendung verschiedener experimenteller Paradigma gewonnen wurden, was eine Ver-

gleichbarkeit nur schon methodisch erschwert (Phan 2002). Dennoch sind bislang mehrere Skizzen eines „moralischen Gehirns“ publiziert worden (Greene & Haidt 2002, Moll et al. 2003/2005) – also eine Übersicht über jene Hirnregionen, die in moralischer Kognition involviert sein sollen (vgl. mit Abbildung 3). Die Vielfalt an Hirnregionen, die bei moralischer Kognition beteiligt sein sollen, verdeutlicht, dass eine Lokalisation der moralischen Kognition wie erwartet unsinnig ist. Vielmehr bestehe eine moralische Entscheidung aus einer Vielzahl unterschiedlicher Prozesse affektiver wie kognitiver Art, wie generell festgehalten wird. 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.

3.3.3 Moralische Emotionen

Im Kontext der Erforschung der neurobiologischen Grundlagen der Moral ist die These aufgekommen, dass es sogenannte „moralische Emotionen“ gebe.[36] Für Moll et al. (2003) gehören jene Emotionen in diese Klasse, die einen Bezug zur Wohlfahrt einer Gruppe oder eines anderen Individuums als der Betreffende selbst haben. Haidt 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“ (2003: 853). Moralische Emotionen beziehen sich also explizit auf soziale Interaktionen, während nichtmoralische Emotionen sich primär auf den emotionalen Zustand des *agent* als Folge 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 Gruppen:

- **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. Verachtung/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 Gruppe spielen gemäß Haidt für den Aufbau der (negativen) Reputation einzelner Individuen in einer sozialen Gemeinschaft eine 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

[36] Einige dieser Emotionen wurden früher auch „soziale Emotionen“ genannt, siehe Adolphs (2003: 116).

(*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 sozial höhergestellte 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 Gruppe zur Aufrechterhaltung bestimmter sozialer Regelsysteme und Ordnungen.

- **Die Mitleids-Emotionen (*other-suffering emotions*):** In diese Gruppe 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.[37] Für ihn ist Mitgefühl/Mitleid die zentrale Emotion dieser Familie, welche Verhaltensweisen wie Helfen und Unterstützen motiviert.

- **Die andere lobenden Emotionen (*other-praising emotions*):** Die Emotionen dieser Gruppe 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.

3.4 (Un-)moralisches Handeln

3.4.1 Moralische Pathologien

Eine in der *neuroscience of ethics* zunehmend genutzte Möglichkeit, Moral besser zu verstehen, ist die Untersuchung von Personen mit abnormen moralischem Verhalten aufgrund von (mehr oder weniger gut) spezifizierten Hirnschäden (moralische Pathologien). 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 Forschun-

[37] Wie bereits ausgeführt (Abschnitt 2.3.2) gibt es unterschiedliche Definitionen von „Empathie“.

gen die Frage nach den Abgrenzungskriterien, die ein 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 (Damasio 2003, Damasio 1999, Damasio 1994) einem breiteren Publikum (neu) bekannt geworden.

Die Untersuchung moralischer Pathologien bildet 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 (Karnath & Thier 2003: 515-516), als historischer Referenzfall genannt. Aber auch im Europa sind bereits im 19. Jahrhundert mehrere solche Fälle beschrieben worden (Welt 1888). Im Rahmen der *neuroscience of ethics* ist das Interesse an solchen Fällen deshalb neu geweckt worden, weil Menschen mit solchen Hirnschäden quasi als „Modelle“ für die Unterscheidung einer abstrakten moralischen Kognition von einer emotional unterlegten moralischen Kognition gelten. Moralische Pathologien würden sich dadurch auszeichnen, dass nur erstere, nicht aber letztere Fähigkeit genutzt werden könne. Unter anderem die folgenden Studien behaupten die Existenz einer solchen Differenzierung:

- Blair (1995) entwickelte ein Modell für die Erklärung von Psychopathologie basierend auf der Hypothese, dass bei Tiere mit ausgeprägten Sozialstrukturen ein Hemmmechanismus für aggressives Verhalten aktiviert wird, wenn einer von beiden Konfliktpartnern Unterwerfungssignale aussendet. Blair postuliert die Existenz eines vergleichbaren Mechanismus (*violence inhibition mechanism*) bei Menschen, welcher bei Psychopathen nicht mehr funktionieren soll. Das Modell (das hier nicht weiter vorgestellt wird) sagt voraus, dass Psychopathen die *moral/conventional* Unterscheidung – ein bei Menschen offenbar konsistent beobachtetes Verhalten, wonach moralische Regelverletzungen im Vergleich zu Konventionsverletzungen als gravierender beurteilt werden – 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 einhergeht, 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 anthropomorphen Interpretation natürlicher Vorgänge in der Welt. Ein klassisches Experiment 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 (2004) benutzten das Paradigma von Heider und Simmel für Experimente mit Versuchspersonen mit einer bilateralen Schädigung der Amygdala. 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 – sehr klein. Dies ist eine grundlegende Schwierigkeit bei solchen Untersuchungen, da es in der Regel nur wenige Patienten gibt, welche die entsprechende 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.

- Aufsehen hat die Studie von Anderson et al. (1999) erregt, in welcher über zwei Fälle frühkindlich erworbener Schäden im präfrontalen Kortex berichtet wird. 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. Erwachsene mit Schäden im präfrontalen Kortex zeigen durchaus auch, in unterschiedlichem Ausmaß, Anzeichen von abnormem sozialem bzw. moralischem Verhalten (Damasio 2002b). Dennoch vermögen diese Personen im Kohlberg-Paradigma zu bestehen. Dies wird so interpretiert, dass Menschen mit vorher normalen Gehirnen zwar neue Strategien für die Bewältigung sozialer und moralischer Probleme entwickeln können, diese aber nicht in Echtzeit (also im Alltag) anwenden können.

In jüngster Zeit findet sich ein deutlich gestiegenes Interesse an derartigen Fällen, die entweder in Meta-Studien im Hinblick auf die Bedeutung von Läsionen für (un)moralisches Handeln (Raine & Yang 2006) oder unter Nutzung der bereits erwähnten Methoden (Dilemmata etc, z.B. Ciaramelli et al. 2007, Koenigs et al. 2007) untersucht werden. Diese Studien zeigen, dass relativ klar abgrenzbare Hirnregionen für die Fähigkeit zur moralischen Kognition entscheidend sein können. In den Gesprächen mit Fachpersonen der Universität Zürich wurde aber zu einer vorsichtigen Interpretation dieser Resultate geraten. Anton Valavanis erklärte 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 das Problem der Operationalisierung von moralischem Verhalten, um moralische Pathologien angemessen beschreiben zu können. Zudem führe das Vorliegen einer solchen Schädigung keineswegs sicher zu entsprechenden moralischen Pathologien.

3.4.2 Verhaltensexperimente: Vertrauen und Kooperation als Beispiele

Dilemmatische moralische Entscheide, die als moralischer Stimulus Probanden in *Imaging*-Experimenten vorgelegt werden, können (aus ethischen Gründen) nicht real umgesetzt und in Verhaltensexperimenten geprüft werden.[38] Hingegen können moralnahe Verhaltensweisen mit Hilfe von Methoden der experimentellen Ökonomie (siehe Abschnitt 2.2.4) und unter Einbezug neurowissenschaftlicher Methoden untersucht werden. Auch hierzu sind in den letzten Jahren zahlreiche Studien erschienen, wobei nachfolgend nur eine kleine Auswahl vorgestellt wird, die die Begriffe Vertrauen und Kooperation betreffen.

Vertrauen kann als eine Verhaltensdisposition aufgefasst werden, die in einem weiten Sinn moralrelevant ist, weil Vertrauen beispielsweise ausdrückt, inwieweit sich ein (anderer) *moral agent* an getroffene Vereinbarungen moralischer Art hält. Unter Benutzung der Methoden der experimentellen Spieltheorie wird Vertrauen durch Investitions-Experimente erfasst und durch den Betrag des investierten Geldes quantifiziert. Gewiss wirft diese Operationalisierung bereits die Frage auf, inwieweit damit der Begriff „Vertrauen“ korrekt wiedergegeben wird (siehe Fußnote 20). Diese Problematik soll hier nicht weiter interessieren. Vielmehr soll anhand zweier Studien exemplarisch 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. (2005) wird diese Frage Mithilfe von Vertrauensspielen angegangen. 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

[38] Bereits deren Simulation dürfte Schwierigkeiten haben, vor einer Forschungethik-Kommission Akzeptanz zu finden. Die Milgram-Experimente beispielsweise können in ihrer ursprünglichen Form in den USA seit längerem nicht mehr durchgeführt werden, obgleich das Experiment kürzlich in einer adaptierten Fassung wieder durchgeführt wurde (Burger 2009).

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 die Stärke der Aktivierung des dorsalen Striatum die Absicht, in einem kommenden Spiel zu kooperieren, reflektiert (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. Unklar ist, warum das Signal in B als neuronales Korrelat von Vertrauen aufgefasst wird, zumal ja Person A der Person B vertrauen muss, dass sein Investitionsentscheid von B honoriert wird.

- Ein zweites Thema ist, inwieweit chemische Substanzen Vertrauen zu beeinflussen vermögen. Aufsehen hat diesbezüglich eine Studie von Kosfeld et al. (2005) erregt, welche in einer Variante des *trust game* den Einfluss des Neuropeptids Oxytocin auf der Veralten der Versuchspersonen untersucht hat. 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. Bei B wiederum hat die Gabe von Oxytocin keinen Einfluss auf dessen Entscheide gehabt. Interessant bei solchen Studien ist, dass sie einen Zusammenhang zwischen einem komplexen sozialen Verhalten und der Applikation einer einzigen Substanz schaffen, was später in mehreren weiteren solchen Oxytocin-Studien untersucht wurde.

Kooperation ist Gegenstand einer Vielzahl spieltheoretischer wie verhaltensbiologischer Untersuchungen und kann in einem weiteren Sinn (bei entsprechendem Kontext) als eine Form von moralischem Verhalten angesehen werden. Wird Kooperation mittels experimenteller Spiele untersuchen, so wird Kooperation als ein Signal aufgefasst, mit welchem sich ein Mitglied einer Gruppe gegenüber anderen Mitgliedern als zuverlässiger Partner präsentiert, was mitunter mit Kosten für dieses Mitglied verbunden ist.[39] Zahlreiche Studien unter Verwendung unterschiedlicher experimenteller Spiele haben in den letzten Jahren gezeigt, dass Probanden in solchen Spielen Kosten auf sich nehmen, um Nichtkooperierende zu bestrafen, was auch mit teilweise starken emotionalen Reaktionen bei den Versuchspersonen einher geht (die so genannte Aversion gegen Ungleichheit, Fehr & Schmidt 1999). In einer Verbindung solcher Spiele mit *Imaging*-Experimenten wurde dann untersucht, ob die in diesen Spielen auftretenden Emotionen sich auch in entsprechenden Hirnaktivierungen zeigen. Unter anderem wurden folgende Studien durchgeführt:

- Haselhuhn und Mellers (2005) untersuchen diesen Zusammenhang mit modifizierten Ultimatum- und Diktatorspielen. Die Spiele wurden derart verändert, dass die Versuchspersonen verschiedene Angebots-Szenarien hinsichtlich ihrer Präferenz und des damit erwarteten Lustgewinns (*pleasure*) beurteilten. Es zeigten sich zwei Gruppen

[39] Die daran anschliessenden Modelle wie z.B. *strong reciprocity* (Gintis 2000) oder *altruistic punishment* (Fehr & Gächter 2002) werden hier nicht weiter behandelt.

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 tatsächlich zeigen werden. Dies verweist auf die bekannte Tatsache, dass unterschiedliche Verhaltensweisen mit unterschiedlichen neuronalen Aktivierungen einhergehen.

- DeQuervain et al. (2004) untersuchten mit Hilfe eines Vertrauens-Spiels, in welchen Hirnregionen der Akt des Strafens eine erhöhte Aktivität auslöst. 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 einhergeht. Zum zweiten zeigte sich, dass diese Aktivität umso höher ist, desto mehr man für den Akt des Strafens investierte.

- Rilling et al. (2002) schließlich suchten nach neuronalen Korrelaten von kooperativem Verhalten, wobei sie Personen untersuchten, die das 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*.

In diesen Experimenten lag die Lokalisierung von Hirnregionen, die bei den verschiedenen *tasks* überdurchschnittlich aktiv sein sollen, bislang im Zentrum der Forschungsbemühungen. Auch hier wurden in den letzten Jahren weitere Experimente durchgeführt, zum einen unter Verwendung neuer Methoden wie TMS (Knoch et al. 2006) oder unter Einbezug weiterer moralnaher Konzepte wie Wohltätigkeit (Harbaugh et al. 2007).

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 falschen oder inkonsistenten 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 facto* Ereignisse zu werten sind, die moralische Handlungen nicht kausal verursachen, sondern erst nachträglich für deren Rechtfertigung konstruiert werden.

Wie Bargh und Chartrand (1999) feststellen, beinhaltet der Automatisierungsprozess in (moralischen) Entscheidungen mindestens drei Komponenten: 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 ausgeht, dass eine erhöhte Aktivierung in einem bestimmten Gebiet als bewusster oder eben automatischer Prozess zu gelten habe.

Im Folgenden soll ein Modell des Moralpsychologen Haidt vorgestellt werden, das die Bedeutung automatisierter Komponenten bei moralischen Entscheidung betont und Begründungen als post fakto Ereignisse charakterisiert (Haidt 2001). Dieses Modell wird in der Literatur über die neurobiologische Grundlagen von Moral oft zitiert. Haidt greift in seinem Modell kognitivistische Varianten des *moral decision making*, basierend etwa auf Kohlberg (1995), 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 consist of transforming given information about people in order to reach a moral judgment (Haidt 2001: 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 evaluations (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“ (Haidt 2001: 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“ (Haidt 2001: 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 individuell 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* kovariieren. 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 stellt sich eine Reihe von Fragen. So ist zum ersten nicht klar, was mit „alltäglichen“ moralischen Handlungen gemeint ist. In sozialen Interaktionen lassen sich 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 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 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 Gruppen oder in einem institutionellen Rahmen. Die interessante Frage ist nun, wie – unter der Voraussetzung der Korrektheit des Modells von Haidt – Entscheidungsprozesse auf dieser Stufe stattfinden.

4 Kritische Fragen an eine *neuroscience of ethics*

Die bisherigen Ausführungen belegen ein zunehmendes Interesse empirischer Disziplinen, *moral agency* unter Einbezug neurowissenschaftlicher Methoden zu untersuchen. Es existieren mehrere Theorieansätze, von denen aber noch keine den Status einer eigentlichen Theorie beanspruchen kann. Auch finden sich zahlreiche Ansätze für Kritik an einzelnen Ergebnissen wie an der Theoriebildung insgesamt. Dieser abschließende Abschnitt soll dazu dienen, diese Kritik zu strukturieren und wichtige Kritikinhalte deutlich zu machen, ohne aber einen umfassenden Gegenentwurf zu liefern. Auch wird an dieser Stelle keine philosophische Analyse einzelner Themen geleistet (z.B. zur Frage, inwieweit solche Studien überhaupt Beiträge zu metaethischen Fragen liefern können), dazu finden sich in den anderen Beiträgen dieses Bandes vertiefende Überlegungen.

Moral agency ist ein komplexes Phänomen, das in empirischen Studien notwendigerweise vereinfacht werden muss. Die Analyse dieser Vereinfachungen bildet den Kern der Einschätzung des Beitrags der Neurowissenschaft für das Verständnis moralischer Orientierung. Dass in dieser Studie methodische Aspekte stark betont wurden, ist kein Zufall, weil die verwendeten Methoden Teilaspekte aus einem komplexen Phänomen herausschälen und dieses damit für die weitere Theoriebildung prägen. Im Zentrum der Beurteilung ist demnach eine Methodenkritik, wobei hier drei Formen anhand der Träger der Kritik unterschieden werden:

- **Interne Methodenkritik:** Träger der Kritik ist die disziplineninterne Diskussion über die Nützlichkeit der verwendeten Methoden in Bezug auf die zu erreichenden Ziele. Was mit „disziplinenintern“ gemeint ist, ist dabei bei neuen, sich ausdifferenzierenden wissenschaftlichen Gebieten (wie die *neuroscience of ethics*) nicht einfach abgrenzbar. Mit Sicherheit gehören aber die unter Abschnitt 2.2.3 am Beispiel von fMRI genannten Probleme des *Imaging* dazu. Die Abschätzung der Bedeutung und Tragweite dieser methodischen Probleme bilden einen wichtigen Teil der Diskussion innerhalb der Sozialen Neurowissenschaft und betrifft damit auch die *neuroscience of ethics*. Für unsere Zwecke ist insbesondere festzuhalten, dass zentrale Fragen wie jene nach der genauen Ausgestaltung der „kognitiven Ontologie“, die für die Auswertung und Beurteilung der gewonnen Daten nötig sind, nicht gelöst sind. Die bislang erzielten empirischen Befunde haben demnach einen vorläufigen Charakter (und zwar in größerem Masse als in anderen empirischen Disziplinen).

- **Normative Methodenkritik:** Die normative Kritik hat zweierlei Träger. Zum einen andere Disziplinen, die eine unkritische Übernahme eigener Forschungsziele und Methoden durch die neue Disziplin kritisieren, zum anderen spezifisch jene (ethischen) Institutionen, welche die rechtliche oder ethische Legitimität der eingesetzten Methoden (und dabei auch die Ziele) beurteilen. Ersteres kommt in der Analyse von Abschnitt 2.2 zum Ausdruck, wo unter anderem auf den doppelten Charakter einer „Moralmessung“ hingewiesen wurde. Letzteres ist in den Forschungsprozess (Einholen von Zustimmungen von Ethikkommissionen für konkrete Experimente) eingebunden, drückt sich aber auch in weiterführenden Diskussionen über den Sinn dieser Forschung (z.B. Willensfreiheits-Debatte) aus, die in dieser Studie nicht weiter untersucht wurden. Angesichts des eingeschränkten Moralverständnisses (z.B. die Gleichsetzung von moralischen mit altruistischen Handlungen) mancher Vertreter der *neuroscience of ethics* findet die normative Methodenkritik zunehmend Beachtung, beispielsweise in Arbeiten von Philosophen (Kahane & Shakel 2007). Diese Kritik zielt vorab darauf, dass das Phänomen „Moral“ ungenügend differenziert dargestellt wird – etwa im Sinne einer fehlenden „phenomenology of morals“ (Horgan & Timmons 2008). Eine

normative Kritik im Hinblick auf die Ziele der *neuroscience of ethics* oder der möglichen Instrumentalisierung ihrer Resultate für gesellschaftliche Zwecke (z.B. *moral enhancement*, siehe Douglas 2008) wird in den nächsten Jahren sicher ebenfalls zunehmen geleistet werden.

- **Genealogische Methodenkritik:** Unter genealogischer Methodenkritik wird hier jene (z.B. wissenschaftssoziologische) Perspektive verstanden, die den Fokus auf die Genese von Methoden und damit verbundenen Forschungsfragen legt. Es geht also um das Zusammenspiel wissenschaftlicher mit anderen gesellschaftlichen Bereichen, was sich in der Ausbildung von wissenschaftlichen Trends, gesellschaftlichen Ansprüchen an die Wissenschaft oder auch den innerhalb einer Wissenschaft formulierten Motiven für bestimmte Forschungen widerspiegeln kann. Eine genealogische Methodenkritik der *neuroscience of ethics* würde beispielsweise die Tatsache des zunehmenden Interesses empirischer Forschung an einer Naturalisierung des „Guten" im Menschen (Forschung über Fairness, Kooperation etc.) in einen historischen Kontext stellen, zumal sich solche Entwicklungen in der Hirnforschung (das „kriminelle Gehirn", das „geniale Gehirn", siehe Hagner 2004) immer wieder finden. Dieser Bereich dürfte derzeit noch am wenigsten entwickelt sein. Von Interesse sollte beispielsweise sein, wie derzeit die Figur des „unmoralischen Patienten" als Modell der empirischen Moralforschung eingesetzt wird und wie sich die Beschreibung solcher Personen mit spezifizierten Hirnschäden von Beschreibungen in klinischen und lebensweltlichen Kontexten unterscheiden.

Anhand dieser Dreiteilung lassen sich kritische Fragen an eine *neuroscience of ethics* systematisieren. Dabei sollte aber nicht vergessen werden, dass die Forschungen in diesem Bereich durchaus auch das Potenzial haben, etablierte Denktraditionen in Frage zu stellen. Ein besonders interessantes Beispiel ist die Dichotomie Emotion-Vernunft. Diese wirkt einerseits theoriebildend (z.B. im *dual-processing*-Ansatz von Greene), doch manche Experimente werden auch so interpretiert, dass diese Dichotomie entweder keine Rolle spielt (der Ansatz von Hauser) oder aber moralische Urteile nicht als Resultat eines Wechselspiels dieser zwei Komponenten angesehen werden sollte (Moll). Für diese Frage ein *experimentum crucis* zu erwarten, dürfte der falsche Weg sein. Hingegen zeigt sich hier, dass der Einbezug eines empirischen Denkens die Theoriebildung in der Ethik im Sinne einer „experimentellen Ethik" (Appiah 2008) durchaus unterstützen kann.

Zum Autor

Markus Christen, Dr. sc. ETH, studierte Philosophie, Physik, Mathematik und Biologie an der Universität Bern und doktorierte in Neuroinformatik an der ETH Zürich. Derzeit arbeitet er am Graduiertenprogramm des Universitären Forschungsschwerpunkts Ethik in Zürich an einem Projekt über *moral agency*. Seine Forschungsinteressen umfassen methodische Fragen der Neurowissenschaft, Autonomie in sozialen Systemen und Neuroethik.

Die Forschungen von Markus Christen werden vom Projekt Nr. 100011-116725"Die neurobiologische Untersuchung des moral agent" des Schweizerischen Nationalfonds unterstützt.

Bibliografie

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

Amaro Jr. E, Barker GJ (2006): Study design in fMRI: Basic principles. *Brain and Cognition* 60: 220-232.

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

Appiah KA (2008): *Experiments in Ethics*. Harvard University Press, Cambridge.

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

Batson CD, Lisher DA, Carpenter A, Dulin L, Harjusola-Webb S, Stocks EL, Gale S, Hassan O, Sampat B (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.

Beauchamp TL, Childress JF (2001): *Principles of Biomedical Ethics* (fifth edition). Oxford, Oxford University Press.

Bergman MG, Jonides J, Nee DE (2006): Studying mind and brain with fMRI. *Social Cognitive and Affective Neuroscience* 1(2): 158-161.

Bestmann S (2008): The physiological basis of transcranial magnetic stimulation. *Trends in Cognitive Sciences* 12(3): 81-83.

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

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

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

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

Burger JM (2009): Replicating Milgram: Would people still obey today? *American Psychologist* 64: 1-11

Cacioppo JT, Berntson GG (1992): Social psychological contributions to the decade of the brain: Doctrine of multilevel analysis. *American Psychologist* 47: 1019-1028.

Camerer CF, Fehr E (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.

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

Casebeer WD (2003): Moral cognition and ist neural constituents. *Nature Reviews: Neuroscience* 4: 841-847.

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

Ciaramelli E, Muccioli M, Làdavas E, di Pellegrino G (2007): Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience* 2: 84-92.

Damasio AR (2003): *Looking for Spinoza. Joy, Sorrow, and the Feeling Brain*. Harvest Book, Orlando.

Damasio AR (1999): *The Feeling of What Happens – Body and Emotion in the Making of Consciousness*. Harcourt Brace, New York.

Damasio AR (1994): *Descartes' Error – Emotion, Reason, and the Human Brain*. Penguin Putnam, New York.

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

DeCharms RC (2008): Applications of real-time fMRI. *Nature Reviews: Neuroscience* 9(9): 720-729.

De Oliveira-Souza R, Hare RD, Bramati IE, Garrido GJ, Ignácio FA, Tovar-Moll F, Moll J (2008): Psychopathy as a disorder of the moral brain: Fronto-temporo-limbic grey matter reductions demonstrated by voxel-based morphometry. *NeuroImage* 40: 1202-1213.

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

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.

Douglas T (2008): Moral enhancement. *Journal of Applied Philosophy* 25(3): 228-245.

Dumit J (2004): *Picturing Personhood. Brain Scans and Biomedical Identity*. Princeton University Press, Princeton.

Düwell M, Hübenthal C, Werner MW (Hrsg.) (2002): *Handbuch Ethik*. Verlag J.B. Metzler, Stuttgart, Weimar.

Dupoux E, Jacob P (2007): Universal moral grammar: a critical appraisal. *Trends in Cognitive Science* 11(9): 373-378.

Evans JSBT (2008): Dual-processing-accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255-278.

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

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

Fehr E, Gächter S (2002): Altruistic punishment in humans. *Nature* 415: 137-140.

Fehr E, Camerer CF (2007): Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences* 11(10): 419-427.

Gardner H (1985): *The Mind's New Science. A History of the Cognitive Revolution*. Basic Books, New York.

Gilligan C (1977): In a Different Voice: Women's conceptions of self and morality. *Harvard Educational Review* 47: 481-517.

Gold JI, Shadlen MN (2007): The neural basis of decision making. *Annual Review of Neuroscience* 30: 535-574.

Greene J (2007): Why are VPPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences* 11(8): 322-323.

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

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

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

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

Greenstein B, Greenstein A (2000): *Color Atlas of Neuroscience*. Thieme, Stuttgart, New York.

Hagner M (2004): *Geniale Gehirne. Zur Geschichte der Elitegehirnforschung*. Wallstein, Göttingen.

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

Haidt J (2003): The moral emotions. In: Davidson RJ, Scherer KR, Goldsmith HH (eds.): *Handbook of Affective Sciences*. Oxford University Press, Oxford: 852-870.

Harbaugh WT, Mayr U, Burghart DR (2007): Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316: 1622-1625.

Harmon-Jones E, Winkielman P (2007): *Social Neuroscience. Integrating Biological and Psychological Explanations of Social Behavior*. The Guilford Press, New York, London.

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

Hauser MD (2006): The liver and the moral organ. *Social Cognitive and Affective Neuroscience* 1(3): 214-220.

Hauser MD (2006): *Moral Minds. How Nature Designed our Universal Sense of Right and Wrong*. HarperCollins, New York.

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

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

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

Hoffman ML (2000): *Empathy and Moral Development*. Cambridge University Press, Cambridge.

Horgan T, Timmons M (2008): Prolegomena to a future phenomenology of morals. *Phenomenology and Cognitive Science* 7: 115-131.

Hume D (1751/1998): *An Enquiry Concerning the Principles of Morals*. Oxford University Press, Oxford.

Hynes CA (2008): Morality, inhibition, and propositional content. In: Sinnott-Armstrong W (2008): *Moral Psychology* (3 volumes). MIT Press, Cambridge: 25-30.

Illes J, Kirschen MP, Gabrieli JDE (2003): From neuroimaging to neuroethics. *Nature Neuroscience* 6(3): 205.

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

Jäncke L. (2005): *Methoden der Bildgebung in der Psychologie und den kognitiven Neurowissenschaften*. Kohlhammer, Stuttgart.

Kahane G, Shackel N (2007): Do abnormal responses show utilitarian bias? *Nature* 452: E5.

Kandel ER, Schwartz JH, Jessell TM (2000): *Principles of Neural Science*. McGraw-Hill, New York.

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

Kenning P, Plassmann H (2005): NeuroEconomics: an overview from an economic perspective. *Brain Research Bulletin* 67: 343-354.

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

Knoch D, Nitsche MA, Fischbacher U, Eisenegger C, Pascual-Leone A, Fehr E (2008): Studying the neurobiology of social iteraction with transcranial direct current stimulation - The example of punishing unfairness. *Cerebral Cortex* 18: 1987-1990.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E (2006): Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314: 829-832

Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M, Damasio A (2007): Damage to the prefrontal cortex increases utilitarian morak judgements. *Nature* 446: 908-911.

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

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

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

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

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

Lieberman MD (2007): Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology* 58: 259-289.

Loewenstein G, Rick S, Cohen JD (2008): Neuroeconoomics. *Annual Review of Psychology* 59: 647-672.

Logothetis NK (2008): What we can do and what we cannot do with fMRI. *Nature* 453(7197): 869-78.

Matusall S, Kaufmann I, Christen M (in Vorbereitung): Disciplinary dynamics in emerging social neurosciences and neuroeconomics

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

Mikhail J (2007): Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Science* 11(4): 143-152.

Moll J, de Oliveira-Souza R, Zahn R (2008): The neural basis of moral cognition. *Annals of the New York Academy of Sciences* 1124: 161-180.

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

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

Moll J, de Oliveira-Souza R, Eslinger PJ, Bramati IE, Mourao-Miranda J, Andreiuolo PA, Pessoa L (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.

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

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

Musschenga AW (2005): Empirical ethics, context-sensitivity, and contextualism. *Journal of Medicine and Philosophy* 30:1467-490

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

Nida-Rümelin J (2005): *Über menschliche Freiheit.* Reclam, Stuttgart.

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006): Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* 10(9): 424-430.

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

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

Poldrack RA (2008): The role of fMRI in cognitive neuroscience: where do we stand? *Current Opinion in Neurobiology* 18: 223-227.

Poldrack RA (2006): Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10(2): 59-63.

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

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

Prinz JJ (2007): *The Emotional Construction of Morals*. Oxford University Press, Oxford.

Raine A, Yang Y (2006): Neural foundations to moral reasoning and antisocial behaviour. *Social Cognitive and Affective Neuroscience* 1: 203-213.

Ridding MC, Rothwell JC (2007): Is there a future for therapeutic use of transcranial magnetic stimulation? *Nature Reviews: Neuroscience* 8(7): 559-567.

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

Rizzolatti G, Fabbri-Destro M (2008): The mirror system and its role in social cognition. *Current Opinion in Neurobiology* 18: 179-184.

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

Roberts RC (2003): *Emotions. An Essay in Aid of Moral Psychology*. Cambridge University Press, Cambridge.

Robertson D, Snarey J, Ousley O, Harenski, DuBois Bowman F, Gilkey R, Kilts C (2007): The neural processing of moral sensitivity to issues of justice and care. *Neuropsychologia* 45: 755-766.

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

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

Schaich Borg J, Hynes C, van Horn J, Grafton S, Sinnot-Armstrong W (2006): Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *Journal of Cognitive Neuroscience* 18(5): 803-817.

Sellars W (1956): *Empiricism and the Philosophy of Mind*. Harward University Press, Cambridge.

Singer T (2006): The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience and Biobehavioral Reviews* 30: 855-863

Singer P (2005): Ethics and intuitions. *The Journal of Ethics* 9: 331-352.

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

Sinnott-Armstrong W (2008): *Moral Psychology* (3 volumes). MIT Press, Cambridge.

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

Stark CEL, Squire LR (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.

Takahashi H, Kato M, Matsuura M, Koeda M, Yahata N, Suhara T, Okubo Y (2008): Neural correlates of human virtue judgment. *Cerebral Cortex* 18: 1886-1891.

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

Takano T, Tian GF, Peng W, Lou N, Libionka W, Han X, Nedergaard M (2006): Astrocyte-mediated control of cerebral blood flow. *Nature Neuroscience* 9: 260-267.

Timmons M (2008): Toward a sentimentalist deontology. In: Sinnott-Armstrong W (2008): *Moral Psychology*, volume 3. MIT Press, Cambridge: 93-104.

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

Uttal WR (2001): *The New Phrenology. The Limits of Localizing Cognitive Processes in the Brain*. MIT Press, Cambridge.

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

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

Vul E, Harris C, Winkielman P, Pashler H (2008): Voodoo Correlations in Social Neuroscience. *Perspectives on Psychological Science*, in press.

Welt L (1888): Über Charakterveränderungen des Menschen infolge von Läsionen des Stirnhirns. *Deutsches Archiv der Klinischen Medizin* 42: 339-390.

Young L, Saxe R (2008): The neural basis of belief encoding and integration in moral judgment. *NeuroImage* 40: 1912-1920.

Young L, Cushman F, Hauser M, Saxe R (2007): The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences USA* 104(20): 8235-8240.

Zahn R, Moll J, Paiva M, Garrido G, Krueger F, Huey ED, Grafman J (2009): The neural basis of human social values: evidence from functional fMRI. *Cerebral Cortex* 19: 276-283.