Aus dem Institut für Geschichte, Theorie und Ethik der Medizin der
Universitätsmedizin der Johannes Gutenberg-Universität Mainz

Methoden und Anwendungen medizinethischen Argumentierens

Habilitationsschrift
zur Erlangung der venia legendi
für das Fach
Geschichte und Ethik der Medizin

Universitätsmedizin der Johannes Gutenberg-Universität Mainz

vorgelegt von

Dr. phil. Nils-Frederic Wagner, M.A.
aus Recklinghausen

Mainz, 2025

Wissenschaftliche Originalpublikationen der kumulativen Habilitationsschrift

I. Wagner, Nils-Frederic & Northoff, Georg; A Fallacious Jar? The Peculiar Relation between Norms and Facts in Neuroethics; Theoretical Medicine and Bioethics; 2015; 36(3), 215-235

II. Wagner, Nils-Frederic, Chaves, Pedro & Wolff, Annemarie; Discovering the Neural Nature of Moral Cognition? Empirical, Theoretical, and Practical Challenges in Bioethical Research with Electroencephalography (EEG); Journal of Bioethical Inquiry; 2017; 14(2), 299-313

III. Wagner, Nils-Frederic; Against Cognitivism about Personhood; Erkenntnis; 2019; 84(3); 657-686

IV. Wagner, Nils-Frederic; Personal Identity, Possible Worlds, and Medical Ethics; Medicine, Health Care and Philosophy; 2022; 25(3), 429-437

V. Wagner, Nils-Frederic; Doing Away with the Agential Bias: Agency and Patiency in Health Monitoring Applications; Philosophy and Technology; 2019; 32(1), 135-154

Inhaltsverzeichnis

Zusammenfassung der kumulativen Habilitationsschrift 4
Einleitung 4
I. Die Struktur naturalistischer Argumente in der Medizinethik 5
I.1 Methodische Fallstricke überwinden 7
II. Die neuronale Basis moralischer Kognition 9
II.1 Die normative Bedeutung von EEG-Studien zur moralischen Kognition . 11
III. Die anthropologischen Voraussetzungen von Personsein und ihre normative Signifikanz 14
III.1 Der Normative Fehlschluss des Kognitivismus 15
III.2 Die Berücksichtigung ontologischer und normativer Herausforderungen 16
IV. Methodische und empirische Plausibilität von Theorien diachroner personaler Identität und ihre medizinethische Relevanz 18
IV.1 Orthodoxe Commitments und reale Personen 19
IV.2 Gedankenexperimente und personale Identität 20
IV.3 Theoretische Überzeugungen, empirische Studien und praktische Anwendbarkeit 25
V. Akteurschaft und Autonomie bei KI-gestütztem Gesundheitsmonitoring 27
V.1 Autonomie, Persuasion und Paternalismus 27
V.2 Die handlungstheoretische Verzerrung von Akteurschaft 30
V.3 Stärkung der Handlungskompetenz 31
VI. Konklusion und Ausblick 33
Originalpublikationen 37
Literatur 133

Zusammenfassung der kumulativen Habilitationsschrift

Einleitung

Die empirisch-informierte Medizinethik oszilliert zwischen deskriptiven, medizinischen Fakten und normativen Ansprüchen ethischer Tragfähigkeit. Die in der Publikationslandschaft vorherrschende Vielfalt von Methoden und Anwendungen erschwert es jedoch, methodische Standards zu identifizieren, die ein kongruentes Bild des Fachs zeichnen.

Im thematischen Zusammenhang der vorgelegten Originalpublikationen werden in der kumulativen Habilitationsschrift je vier zentrale Elemente medizinethischen Argumentierens diskutiert, die sowohl argumentationslogisch als auch anhand konkreter Anwendungsbeispiele exemplarisch aufzeigen, wie thematisch mehrdimensionale medizinethische Analysen methodisch plausibel miteinander verbunden werden können.

Auf argumentationslogischer Abstraktionsebene werden zunächst die folgenden vier Elemente herausgearbeitet:

1. Empirisch-informierte Syllogismen: Eine normativ unvoreingenommene Argumentation wird durch eine bidirektionale Verbindung von Fakten und Normen ermöglicht
2. Explanativer, empirischer Zugang: Ermöglicht ergebnisoffene, normativ tragfähige Ableitungen aus empirischen Daten
3. Hypothetische und kontrafaktische Gedankenexperimente: Dienen als Prüfstein medizinethischer Theorien und deren praktischer Anwendung
4. Schlüsselkonzepte – Personsein, Akteurschaft und Autonomie: Beeinflussen zentrale theoretische Debatten in der Medizinethik und haben einen mittelbaren Einfluss auf Entscheidungen in der klinischen Ethik

Auf praxisbezogener Ebene werden diese argumentationslogischen Elemente anhand folgender Anwendungsfälle konkretisiert:

a) EEG-basierte Studien zu moralischer Kognition: Zeigen, dass moralische Kognition ein globaler Hirnprozess ist und deuten darauf hin, dass tugendethische Theorien besonders gut mit den, wenngleich tentativen, empirischen Daten zu moralischer Kognition kohärieren

b) Anthropologische Voraussetzungen von Personsein: Diskutieren die Vermengung ontologischer und normativer Grundlagen von Personsein und schlagen einen präkognitiven Ansatz vor
c) Gedankenexperimente zur personalen Identität: Ermöglichen eine kritische Analyse der Validität von Patientenverfügungen bei neurodegenerativen Erkrankungen
d) Ethik des KI-gestützten Gesundheitsmonitorings: Untersuchen die Auswirkungen von mHealth Anwendungen auf die Autonomie der Anwender und die Unterstützung deren autonomer Ziele durch ethisch fundierte mHealth Technologie

Im Folgenden werden, unter Berücksichtigung des genannten Referenzrahmens, wesentliche Ergebnisse der kumulativen Habilitationsschrift kondensiert dargestellt und deren thematischer Zusammenhang diskutiert sowie Bezug zu aktuellen Entwicklungen der Forschung hergestellt.

I. Die Struktur naturalistischer Argumente in der Medizinethik

Den theoretischen Ausgangspunkt der kumulativen Habilitationsschrift bildet eine argumentationslogische Analyse typischer medizinethischer Argumente (Wagner & Northoff, 2015). Der Fokus dabei liegt auf der validen argumentativen Verbindung von empirischen Prämissen und normativen Konklusionen. Dieser Zusammenhang wird sowohl konzeptuell als auch illustrativ anhand konkreter, empirischer Untersuchungen zur moralischen Kognition kritisch untersucht. Die hier erarbeiteten Ergebnisse werden in den praxisbezogenen Anwendungsfällen und theoretischen Debatten, die in den weiteren Publikationen der kumulativen Habilitationsschrift diskutiert werden, fortwährend aufgegriffen.

Ausgangspunkt der Analyse ist die Beobachtung, dass, inspiriert durch den rapiden Fortschritt moderner bildgebender Verfahren, traditionelle ethische Fragen in zunehmend akzentuiertem Zusammenhang mit empirischen Daten stehen. In der andauernden internationalen Debatte werden mittels eigens zu diesem Zweck konzipierter fMRT-Studien zur moralischen Kognition Theorien normativer Ethik gleichsam einer empirischen Plausibilitätsprüfung unterzogen. Ein wesentliches Merkmal dieser Art von Medizinethik ist es, aus der empirischen Untersuchung neuronaler und psychologischer Mechanismen, die moralischem Urteilen zugrunde

liegen, normative Schlussfolgerungen zur Plausibilität ebendieser Theorien abzuleiten. Dieses Unterfangen birgt jedoch einige methodische Fallstricke, die im Folgenden analysiert werden.

Ein medizinethischer Ansatz, in dem neurowissenschaftliche Daten nicht lediglich dazu dienen, moralpsychologisch aufzuzeigen, wie Menschen konfrontiert mit moralischen Stimuli *de facto* handeln, sondern dezidiert Prinzipien normativ-ethischer Theorien empirisch offengelegt und hinterfragt werden, kann als *robuster moralischer Naturalismus* bezeichnet werden. Es sollen gültige Beziehungen zwischen neurowissenschaftlichen Beobachtungen und normativ-ethischen Theorien hergestellt werden – mit dem Ziel, konkrete Handlungsweisen in der angewandten Medizinethik daraus ableiten zu können. Eine so verstandene, robust naturalistische Medizinethik untersucht sowohl die Bedeutung neurowissenschaftlicher Erkenntnisse für das Verständnis von Moral als auch die Relevanz der Ethik für die Bestimmung der normativen Bedeutung von Erkenntnissen aus den Neurowissenschaften. Erstere befasst sich mit neuronalen und psychologischen Mechanismen, die ethischen Konzepten und Urteilen zugrunde liegen; Letztere befasst sich hingegen mit den Implikationen dieser Erkenntnisse für die moralische Praxis.

Nach einer eingehenden Analyse der Funktionsweise derartiger robust naturalistischer Argumente wird gezeigt, dass valide, belastbare normative Schlussfolgerungen aus empirischen Daten nur gelingen können, wenn ‚normativ-indeterminierte' Prämissen eine verzerrte Interpretation der empirischen Daten verhindern und eine ergebnisoffene Argumentation ermöglichen. Zentral ist dabei die Frage, auf welche Weise empirische Daten herangezogen werden, um eine tragende Rolle bei normativen Schlussfolgerungen zu spielen. Es wird gezeigt, dass die Rolle, die empirische Daten spielen, insofern tragend (und nicht lediglich illustrativ) ist, als dass auf die Daten rekurriert wird, um die ethische Plausibilität konkurrierender normativ-ethischer Theorien zu bewerten. Um die methodischen Probleme eines solchen Ansatzes aufzudecken, werden die folgenden drei Untersuchungsschritte unternommen:

(1) Zunächst wird untersucht, wie sich der naturalistische Fehlschluss auf robust naturalistische Argumente in der Medizinethik anwenden lässt: Dabei wird herausgestellt, dass diese Art von Argumenten nicht direkt von ‚Sein' auf ‚Sollen' schließen, sondern ihre argumentative Überzeugungskraft vielmehr auf gleichsam eingeschmuggelten ‚semi-normativen Prämissen' beruht, die ihrerseits ihre

Plausibilität weniger aus den erhobenen Daten als vielmehr aus axiomatischen normativen Annahmen gewinnen.

(2) Anschließend wird gezeigt, dass diese Art von Argumenten zwar keinem genuinen naturalistischen Fehlschluss unterliegen, jedoch ein ‚normativer Fehlschluss' vorliegt, indem von ‚Sollen' auf ‚Sein' geschlossen wird und insofern die angestrebte Konklusion bereits in den Prämissen vorausgesetzt wird. Diese Form der Argumentation wird als ‚ergebnis-geschlossene Argumentation' bezeichnet.

(3) Abschließend wird ein Lösungsansatz für die diagnostizierten methodischen Probleme skizziert: zunächst wird konstatiert, dass medizinethische Argumentationen mit einer ‚normativen Unbestimmtheit' beginnen sollten, die es ermöglicht, ‚ergebnis-offene Argumente' zu formulieren. Das Ziel besteht darin, fundierte medizinethische Argumente zu ermöglichen, die aufgrund einer soliden ‚Norm-Fakt-Verknüpfung' am besten mit empirischen Daten kohärieren, wobei den empirischen Daten kein normativer Vorrang eingeräumt wird.

Es liegt auf der Hand, dass eine besondere methodische Herausforderung bei solchen medizinethischen Argumenten in der Verbindung zwischen Fakten und Normen liegt. Demnach müssen argumentative Schritte unternommen werden, um von empirischen Behauptungen, die mittels bildgebender Verfahren aus beobachteter Gehirnaktivität abgeleitet werden, die moralischen Intuitionen und Fähigkeiten zugrunde liegen, zu normativen Schlussfolgerungen darüber zu gelangen, was moralisch richtig oder falsch ist. Dies gilt sowohl prinzipiell, also nach normativ-ethischen Theorien, als auch in konkreten Kontexten der angewandten Medizinethik. Solche Ansätze, in denen die hier diagnostizierten Probleme aufgegriffen werden, werden in der aktuellen Literatur weiterhin kontrovers diskutiert (Malatesti & McMillan, 2024; Dubljević, 2021; Rueda, 2021).

I.1 Methodische Fallstricke überwinden

Vertreter einer robust naturalistischen Auffassung von Argumenten in der Medizinethik befürworten häufig eine reduktionistische Form normativer Schlussfolgerungen aus empirischen Daten. Dabei wird die normative Ebene auf die empirische Ebene reduziert; ethische Normen werden somit durch neuronale Fakten ersetzt. Ziel ist es, das ethische Konzept gleichsam zu ‚neuronalisieren', es also auf neuronale Fakten zu reduzieren. Dabei besteht jedoch die Gefahr einer unkritischen Akzeptanz empirischer

Voraussetzungen und daraus resultierender reduktionistischer Definitionen zentraler medizinethischer Konzepte. Traditionell wird diese Form der Schlussfolgerung von ‚Sein' auf ‚Sollen' als naturalistischer Fehlschluss bezeichnet. Die umfangreiche gegenwärtige Literatur zu diesem wichtigen metaethischen Thema beschäftigt sich auch mit der Frage danach, ob es prinzipiell einen validen Weg geben kann, ein ‚Sollen' aus einem ‚Sein' zu schließen (Sinhababu, 2024); bislang ist diese Frage nicht abschließend gelöst worden.

Diese Art der robust naturalistischen Argumentation in der gegenwärtigen Medizinethik birgt die Gefahr, in methodischen Vorurteilen festzustecken, bei denen sowohl von Fakten auf Normen als auch von Normen auf Fakten geschlossen wird, ohne jedoch dabei die empirische und normative Ebene argumentativ schlüssig miteinander zu verbinden. Die diagnostizierten methodischen Fallstricke gefährden allerdings nicht das Gesamtprojekt der empirisch-informierten Medizinethik. Es ist evident, dass moralische Überlegungen über das tatsächliche menschliche Leben empirisch begründete Antworten erfordern. In diesem Sinne kann die Konsultation moderner bildgebender Verfahren (und anderer empirischer Methoden) äußerst wertvoll sein.

Um empirische Befunde gewinnbringend in den Kontext normativer Überlegungen einzubetten, ist die Interpretation der gewonnenen Daten notwendig; diese Interpretationen haben zwangsläufig eine normative Dimension. Neben prä-empirischen normativen Commitments werden häufig die Kontingenzen, die die faktischen Normen und sozialen Strukturen des Alltagslebens bestimmen, nicht adäquat berücksichtigt. Bei dem Versuch, die normative Bedeutung neurowissenschaftlicher Erkenntnisse zu bestimmen, dürfen die individuellen und idiosynkratrischen sozialen Gesichtspunkte nicht außer Acht gelassen werden. Normative Konzepte wie moralisches Urteilen und damit verbundene theoretische Überlegungen wie der moralische Status von Personen (auf den dezidiert in Wagner, 2019a und 2022 eingegangen wird) können nicht umfassend analysiert werden, wenn der soziale und politische Kontext dieser Konzepte nicht berücksichtigt wird. Dies erfordert eine explizite Diskussion der oft impliziten Annahmen dieser kontextgeprägten Konzepte und die Notwendigkeit, diese sowohl in ihrem sozialen als auch in ihrem politischen Referenzrahmen zu verorten.

Werden normativ-ethische Überzeugungen als selbstverständlich vorausgesetzt und als nicht verhandelbare Bezugspunkte betrachtet, besteht die Gefahr des normativen

Fehlschlusses. Die Erkenntnis, dass normative Überzeugungen immer in Beziehung zu sozialen und politischen Kontexten stehen, liefert andererseits auch einen Grund dafür, warum eine Reduktion normativer Konzepte auf deskriptive Fakten zu kurz greift.

Positiv betrachtet weisen diese Unzulänglichkeiten einmal mehr auf die Notwendigkeit hin, robust naturalistische Formen der Medizinethik durch eine gründlich argumentierte konzeptuelle Analyse zu ergänzen, die einer sorgfältigen Verortung ethischer Konzepte in ihren relevanten sozialen und politischen Kontexten (als Quelle von Normen) gerecht wird (vgl. Everett & Kahane, 2020).

Um die Barriere zwischen dem normativen und dem empirischen Bereich zu überwinden – in beiden Richtungen, d. h. von Normen zu Fakten und von Fakten zu Normen – ist ein Weg notwendig, der eine wechselseitige Bezugnahme zwischen den beiden Bereichen herzustellen vermag. Dieses Desiderat bezeichne ich als ‚Norm-Fakt-Verknüpfung'. Dieses methodische Vorgehen setzt eine Fokussierung auf die Schnittstelle zwischen dem empirischen und dem normativen Bereich voraus, der die Entwicklung einer wechselseitigen Verknüpfung ermöglicht. Dies ist notwendig, um zu Theorien zu gelangen, die sowohl normativ als auch empirisch plausibel sind, ohne einer der beiden Dimensionen den Vorrang zu geben.

Im folgenden Abschnitt wird aufgezeigt, wie anhand konkreter, empirischer Untersuchungen zur moralischen Kognition mittels EEG eine solche Verbindung von empirischer und normativer Ebene gelingen kann.

II. Die neuronale Basis moralischer Kognition

Wie im vorangegangenen Abschnitt beschrieben, werden normativ-ethische Theorien mit verschiedenen kognitiven Prozessen in Verbindung gebracht. Diese kognitiven Prozesse werden Hirnprozessen zugeordnet, die in unterschiedlichen Netzwerken und Regionen angesiedelt sind und experimentell untersucht werden können. Hierbei wird deutlich, wie bildgebende Verfahren Hinweise auf den Zusammenhang von messbarer Hirnaktivität und moralische Kognition geben können.

Dabei verwenden die prominentesten und einflussreichsten neurowissenschaftlichen Studien zur moralischen Kognition einen neurologischen Lokalisierungsansatz (Greene et al. 2001, 2004; Koenigs et al., 2007). Lokalisierungstechniken wie fMRT untersuchen, welche Hirnregionen während eines Prozesses aktiv sind, indem sie

Veränderungen des Blutflusses messen (Hüttel et al., 2008). Um die elektrophysiologischen Grundlagen der moralischen Kognition zu erforschen werden mittels EEG die elektrischen Ströme gemessen, die kontinuierlich von kortikalen Schichten erzeugt werden; so kann bestimmt werden, wann etwas im Gehirn passiert. Auch wenn EEG die hohe räumliche Auflösung der fMRT nicht erreicht, ermöglicht diese nicht-invasive Technologie einen Einblick in sekundenschnelle Prozesse, die die Informationsverarbeitung im Zusammenhang mit der Präsentation von Stimuli widerspiegeln (Schomer & Lopes da Silva, 2012). Moralische Kognition, die an sekundenschnellen Entscheidungen beteiligt ist, ist ein solcher Prozess und eignet sich daher besonders für EEG-Studien.

Bei Experimenten zur moralischen Kognition wird der moralisch relevante Stimulus, z. B. ein Bild oder ein Wort, innerhalb einer Reihe anderer Stimuli (eine Abfolge von Bildern zu einer Handlung oder linguistische Stimuli) präsentiert, die ihm einen gewissen kontextuellen Bezug verleihen und moralische Relevanz haben. Die Reaktion auf den moralischen Stimulus wird anschließend analysiert, um die Mechanismen zu ermitteln, die für seine Verarbeitung verantwortlich sind. Auf diese Weise können der zeitliche Ablauf des Entscheidungsprozesses und seine verschiedenen Unterphasen untersucht werden. Bei moralischen Dilemmata kann es zum Beispiel nützlich sein, zu unterscheiden, ob die Entscheidungen auf niedrigschwelligen Merkmalen des Szenarios beruhen (wie z. B. der emotionalen Bedeutung) oder durch kognitive Anstrengung und Bezugnahme auf ethische Normen erreicht werden. Dies ermöglicht es, Rückschlüsse zu ziehen oder jedenfalls Hypothesen darüber aufzustellen, welche Art von normativ-ethischen Theorien am besten mit den empirischen Erkenntnissen kohärieren.

Aus methodischer Sicht ermöglicht die Verwendung von EEG eine gute Bewertung schneller Hirnprozesse, die sich auf moralische Kognition auswirken. Trotz der hohen zeitlichen Auflösung leidet die genaue Lokalisierung der Aktivität immer noch unter einigen Problemen und Einschränkungen. Der Großteil der erfassten Aktivität stammt aus kortikalen Strukturen und einer bestimmten Untergruppe von Schichten. Das bedeutet, dass große Gruppen von Regionen (z. B. die meisten subkortikalen Strukturen) und ihre Prozesse für das EEG undurchdringlich bleiben und nur durch den Einsatz anderer bildgebender Verfahren wie der fMRT zugänglich sind. Idealerweise sollten daher die Ergebnisse von EEG-Studien durch komplementäre fMRT Untersuchungen ergänzt werden, um der geringen räumlichen Auflösung des

EEG Rechnung zu tragen. Hierauf weisen auch neuere Untersuchungen zu neurowissenschaftlichen Studien moralischer Kognition hin (Angioletti & Balconi, 2024; Sackris & Rosenberg Larsen, 2022).

II.1 Die normative Bedeutung von EEG-Studien zur moralischen Kognition

Eine grundsätzliche Schwierigkeit, die sich bei der Interpretation von EEG-Daten ergibt, ist die häufig angenommene Objektivität, die mit der Visualisierung von Hirnaktivität einhergeht. Es ist verlockend anzunehmen, dass die statistische Analyse von Hirnstromaufzeichnungen gleichsam einen direkten Einblick in die Psyche der Probanden ermöglicht. Doch die Beziehung zwischen subjektiven mentalen Zuständen und elektromagnetischen Signalen ist nicht direkt beobachtbar (Poldrack, 2006). Trotzdem neigen viele Neurowissenschaftler zu einer reduktiven Sichtweise der menschlichen Psyche und gehen davon aus, dass von neurowissenschaftlichen Beobachtungen mehr oder weniger direkte Rückschlüsse auf subjektive mentale Zustände gezogen werden können. Dabei wird jedoch übersehen, dass bildgebende Verfahren auf probabilistischen Kovarianzen und nicht auf kausalen Beziehungen beruhen, was direkte Rückschlüsse schwierig macht. Dieses methodische Dilemma weist Parallelen und gewisse Überschneidungen zu dem s. g. „reverse inference problem“ (ebd.) auf: eine induktive Methode, bei der von der beobachteten Hirnaktivität auf bestimmte kognitive Prozesse rückgeschlossen wird, die jedoch nicht direkt getestet werden. Aus diesen Schwierigkeiten bei der Interpretation von EEG-Daten ergeben sich methodische Schwierigkeiten bei der Unterstützung oder Entkräftung normativ-ethischer Theorien, die sich stark auf die empirischen Ergebnisse neurowissenschaftlicher Studien stützen.

Ein großer Teil der experimentellen Parameter in EEG-Studien verweist auf intraindividuelle Unterschiede. Es wird also die Variabilität der Probanden in ihren Reaktionen auf moralische Dilemmata untersucht. Diese Verzerrung lässt sich zum Teil dadurch erklären, dass Studiendesigns mit wiederholten Messungen (d. h. Designs, bei denen dieselben Probanden unter verschiedenen Bedingungen stimuliert werden, z. B. bei konsequentialistischen oder deontologischen Ansätzen) im Vergleich zu Designs mit unabhängigen Stichproben (d. h. Designs, die auf interindividuelle Unterschiede zwischen verschiedenen Gruppen abzielen, z. B. beim Vergleich von

gesunden Kontrollpersonen und Psychopathen) kostengünstiger und einfacher durchführbar sind.

Unter Berücksichtigung der zuvor erörterten Vorzüge und Nachteile von EEG-basierten Studien zur moralischen Kognition können folgende Ergebnisse festgehalten werden.

(1) Die Erforschung neuronaler Grundlagen der moralischen Kognition kann dazu beitragen, zu empirisch fundierten, differenzierteren normativ-ethischen Theorien zu gelangen. In dieser Hinsicht kann die zuvor diskutierte experimentelle Literatur, die EEG als Marker für moralische Kognition einsetzt, ein neues Licht auf traditionelle Fragen der normativen Ethik werfen.

(2) Das übergeordnete Ziel, Ethik auf diese Weise zu naturalisieren, besteht darin zu zeigen, wie unsere moralischen Praktiken und die zugrundeliegenden moralischen Theorien auf der Komplexität des menschlichen Gehirns beruhen und als solche wissenschaftlich untersucht werden können. Ein derart hohes Maß an Komplexität macht jedoch die Suche nach Kausalität (derzeit) unmöglich; der Ausweg kann dann in der Suche nach Plausibilität bestehen, die freilich ihrerseits von lebensweltlichen Vornahmen getragen und gleichsam normativ vorbelastet ist.

(3) Entscheidend dabei ist, dass die Studiendesigns so angelegt sind, dass sie die zuvor erwähnte Kontextsensitivität der moralischen Kognition berücksichtigen und die ökologische Validität maximieren – eine Schwierigkeit, die darin besteht, dass die Experimente unter künstlichen Laborsituationen durchgeführt werden, die wenig Ähnlichkeit mit der moralischen Entscheidungsfindung im wirklichen Leben haben, was durch eine möglichst realitätsnahe Gestaltung der Experimente und die Verwendung von Anreizen mit realer Bedeutung, wie sie in Experimenten zur ökonomischen Entscheidungsfindung praktiziert werden, gemildert werden kann.

(4) Die Identifizierung neuronaler Korrelate moralischer Kognition hängt zum Teil davon ab, dass bereits eine normativ-ethische Theorie vorausgesetzt wird, wenn nach den neuronalen Grundlagen moralischer Kognition gesucht wird. Dies hat den Vorteil, dass bestimmte Theorien als neurowissenschaftlich unrealistisch ausgeschlossen werden können. Wenn zum Beispiel eine Theorie einen hohen Anspruch an Vernunft und Deliberation stellt, Experimente aber zeigen, dass moralisch gebotene Entscheidungen durch schnelle Heuristiken getroffen werden, würde eine vernunftbasierte Theorie im Widerspruch zu den Daten stehen.

(5) Bestimmte kognitive Merkmale wie exekutive Funktionen (z.B. Zielorientierung), die von verschiedenen normativ-ethischen Theorien betont werden, werden experimentell untersucht, um die Grundzüge dieser verschiedenen Theorien aufzudecken. Wenn, wie etwa deontologische Theorien annehmen, die moralische Erkenntnis von Vernunft über die Forderungen von z.B. Kants kategorischem Imperativ gesteuert wird, sind exekutive Funktionen, die hauptsächlich aus frontalen Hirnregionen stammen, am relevantesten. Nach dem von Mill vertretenen Utilitarismus besteht die wichtigste Fähigkeit eines moralisch Handelnden darin, Nutzenfunktionen zu erkennen und zu berechnen; dementsprechend ist eine Integration präfrontaler, limbischer und sensorischer Regionen für die Manipulation numerischer Werte und die Kodierung von Werten selbst unerlässlich. In den Theorien der Tugendethik, die auf Aristoteles zurückgehen, wird moralische Erkenntnis grob als die Fähigkeit betrachtet, gut darüber zu urteilen, welche Zustände für das menschliche Wohlbefinden am förderlichsten sind. Moralische Belange beziehen sich also auf das, was wir tun und denken sollten, um als menschliche Wesen gut zu funktionieren, was eine angemessene Koordinierung gut funktionierender kognitiver Untereinheiten erfordert, die über das gesamte Gehirn verteilt sind.

(6) Bei der Abwägung der Theorien ist es wichtig, nicht in den zuvor beschriebenen naiv-naturalistischen Reduktionismus zu verfallen oder eine bereits vorbestehende Agenda einzuschmuggeln, indem man versucht, die neuronalen Korrelate genau der Form der moralischen Kognition zu identifizieren, die am besten zu derjenigen Theorie passt, die von vornherein als am plausibelsten angenommen wurde.[1] Es ist also zusätzliche argumentative Arbeit erforderlich, um unabhängige Gründe dafür zu liefern, welche Theorie konzeptuell am überzeugendsten ist.

(7) In Anbetracht der diskutierten Herausforderungen scheint es offensichtlich, dass die aktuellen neurowissenschaftlichen Erkenntnisse nicht hinreichend sind, um ein konklusives Urteil darüber zu erlauben, welche normativ-ethische Theorie am besten mit der Funktionsweise des Gehirns übereinstimmt. Vorläufig festzuhalten ist allerdings, dass die Tugendethik, die weithin als die moralpsychologisch reichhaltigste Theorie gilt (da Rationalität, Deliberation, Emotionen und Affekte für ein gutes und tugendhaftes Leben von wesentlicher Bedeutung sind) gut mit den in Wagner et al., (2017) diskutierten EEG-Studien kohäriert, die zeigen, dass moralische Kognition ein

[1] Dieser argumentationslogische Fehlschluss wurde in I.1 als normativer Fehlschluss beschrieben.

groß angelegter, globaler Hirnprozess ist, der sich nicht auf bestimmte Bereiche oder Funktionen reduzieren lässt. Darüber hinaus scheinen derzeit die ökologisch validesten Studiendesigns die Hypothese zu stützen, dass moralische Kognition eine distributive Angelegenheit des Gehirns ist, die von der angemessenen Koordination vieler Bereiche abhängt (Hirstein, 2022; Jiang et al., 2022).

Nachdem argumentationslogische und methodische Schwierigkeiten bei der Datenerhebung zu moralischer Kognition mittels EEG besprochen wurden, wird im folgenden Abschnitt das in der Medizinethik zentrale Konzept des Personseins, das gemeinhin als Grundlage für den besonderen moralischen Status von Menschen angeführt wird, auf seine anthropologischen Voraussetzungen hin abgeklopft. Unter Zuhilfenahme des in I. besprochenen methodischen Instrumentariums wird gezeigt, dass die in der Medizinethik orthodoxen Theorien des kognitiven Ursprungs von Personsein, auf deren Grundlage normative Praxen – etwa im Zusammenhang mit Patientenverfügungen – erklärt und begründet werden, auf einem normativen Fehlschluss basieren.

III. Die anthropologischen Voraussetzungen von Personsein und ihre normative Signifikanz

Aus ontologischer Sicht werden Personen gemeinhin als Wesen verstanden, die komplexe kognitive Fähigkeiten haben. Die Art und Weise, wie die Ontologie des Personseins interpretiert und angewendet wird, hat dabei einen signifikanten Einfluss auf normative Fragen; insbesondere auf medizinethische Kontroversen, die mit dem moralischen Status von Personen zusammenhängen. Personen wird, jedenfalls prima facie, ein besonderer moralischer Status zugesprochen, demzufolge Personen ein Recht auf Leben und weitere moralische Interessen haben, die nicht verletzt werden dürfen. Außerdem wird Personsein als Quelle der moralischen Verantwortlichkeit angesehen und als Grundlage für autonome Entscheidungsfähigkeit, die eng mit der Einwilligungsfähigkeit von Patienten verknüpft ist. Der konzeptuelle Anspruch ontologischer Theorien von Personsein ist es demnach, Personen von Nicht-Personen zu unterscheiden – zunächst losgelöst von normativen Erwägungen. Nichtsdestotrotz werden ontologische Theorien von Personsein in medizinethischen Kontroversen häufig vorausgesetzt. So dreht sich z.B. ein großer Teil der Debatte über

die moralische Zulässigkeit von Abtreibung um die Frage der ontologischen Bedingungen des fötalen Personseins (Tooley, 1972).

Da ein wesentliches Merkmal des Personseins darin besteht, den besonderen moralischen Status von Personen auf der Grundlage ihrer integralen ontologischen Bedingungen sowohl zu erklären als auch zu rechtfertigen, wird deutlich, wie eng Ontologie und Normativität miteinander verwoben sind. Dennoch sollten beide Bedingungen des Personseins konzeptuell entwirrt werden. Auf diese Weise können einige der Probleme, die sich aus der Verquickung von Ontologie und Normativität ergeben, besser angegangen werden.

III.1 Der Normative Fehlschluss des Kognitivismus

Diejenigen Theorien, die traditionell die meiste Aufmerksamkeit und Zustimmung finden, heben typischerweise höhere kognitive Fähigkeiten als notwendige (oder gar hinreichende) Bedingung für Personsein hervor;[2] wenngleich die aktuelle Tendenz in der Medizinethik die in Wagner (2019a) aufgeworfenen Probleme dieses Ansatzes beginnt zu berücksichtigen (Boddington, 2024).

Kognitivisten beanspruchen, Personen zunächst ontologisch zu kategorisieren, unabhängig von normativen Bedingungen. Gleichzeitig werden diese ontologischen Bedingungen jedoch herangezogen, um den vollen moralischen Status von Personen zu rechtfertigen, und werden so zu normativen Konzepten. Diese normativen Konzepte sehen die moralische Bedeutung von Personen jedoch genau auf der Grundlage dessen, was sie zunächst als ihre ontologische Bedingung angeben: höhere kognitive Fähigkeiten. Da Kognitivisten konstatieren, dass Personen ein Recht auf Leben (als Teil des besonderen moralischen Status) haben, wird nach Entitäten gesucht, die eine Ontologie aufweisen, die diesen vollen moralischen Status rechtfertigen kann. Ein naheliegender Kandidat sind höhere kognitive Fähigkeiten, die sodann zur Begründung einer ontologischen Kategorie herangezogen werden; wobei Kognitivisten dabei außer Acht lassen, dass diese Ontologie insgeheim durch eine normative Überzeugung motiviert ist. Es ist also die Normativität, die die Ontologie bestimmt – und nicht, wie behauptet, die Ontologie, die die Normativität legitimiert. Die Überzeugung, dass es Personen mit bestimmten kognitiven Merkmalen geben *sollte*,

[2] Ich subsumiere derartige Theorien unter dem Label ‚Kognitivismus', wohlweislich, dass Kognitivismus in der Metaethik eine gänzlich andere Bedeutung hat.

aufgrund derer Kognitivisten ihnen den besonderen moralischen Status zuerkennen wollen, bedeutet dabei jedoch nicht, dass es ontologisch tatsächlich Personen *gibt*, die diesen normativen Annahmen entsprechen. Vielmehr ist die ontologische Überzeugung, dass es de facto Personen gibt, die durch höhere kognitive Fähigkeiten konstituiert sind, durch die verdeckte normative Überzeugung motiviert, ein Mittel zu haben, mit dem ethische Kontroversen gelöst werden können.

Um den besonderen moralischen Status von Personen zu etablieren, müssen demnach Bedingungen gefunden werden, die Personen ontologisch einzigartig macht. Die bereits bestehende normative Verpflichtung des besonderen moralischen Status wird dabei herangezogen, um die ontologische Einzigartigkeit von Personen zu erklären. Personsein wird als normativ bedeutsam angesehen, weil die in Frage stehenden höheren kognitiven Fähigkeiten nur bei physiologisch entwickelten Menschen vorhanden sind, die ohnehin und prätheoretisch als Paradebeispiel für Wesen mit besonderem moralischen Status gelten. Dieses Argument beruht auf dem in I.1 eingeführten normativen Fehlschluss, da es keine zwingenden Gründe dafür liefert, warum die ontologische Konklusion der Existenz von kognitiv komplexen Personen aus normativen Prämissen (hier insbesondere das konstatierte Recht auf Leben) gültig abgeleitet werden kann.

III.2 Die Berücksichtigung ontologischer und normativer Herausforderungen

Das Konzept der Anthropologischen Konstanten entstammt der Philosophischen Anthropologie (Mittelstraß, 2003) und zielt darauf ab, die essentiellen Merkmale menschlicher Existenz offenzulegen. Anthropologische Konstanten verbinden dabei auf plausible Weise die Ontologie und die Normativität von Personsein, weil sie das sowohl ontologisch als auch normativ konstitutive Element von Personen herausgreifen. Ontologisch beschreiben anthropologische Konstanten das empirisch Grundlegendste an Personen und dienen als ontologische Taxonomie. Entitäten gehören zur ontologischen Kategorie der Personen, weil sie bestimmte universelle ontologische Bedingungen aufweisen: gegebene Eigenschaften, auf die weder kulturelle noch historische Kontingenzen relativierend einwirken. In normativer Hinsicht erfassen die anthropologischen Konstanten, warum Entitäten dieser Art in einer bestimmten Weise normativ funktionieren und daher sowohl für sich selbst als auch für andere moralisch von Bedeutung sind. Sie legen nahe, dass die

grundlegendste Bedingung von Personen es erforderlich macht, dass diese Entitäten sich einander entsprechend ihrer grundlegenden Natur behandeln, um ihr kollektives Überleben zu sichern.

Die anthropologische Konstante der Sozialität erklärt die Verbindung der normativen Bedingung des Personseins, die die normative Struktur des Alltagslebens bestimmt, mit der angeborenen Fähigkeit von Personen zur sozialen Interaktion als ihrer grundlegendsten ontologischen Bedingung. Im Gegensatz zur kognitivistischen Ontologie ist soziale Eingebundenheit dabei ein wichtiger Teil unseres Lebens vor und zunächst unabhängig von höheren kognitiven Fähigkeiten.

Die intrinsisch soziale, präreflexive menschliche Natur wird durch empirische Studien plausibilisiert, die zeigen, dass Säuglinge, sobald sie die Emotionen, Ziele und Aufmerksamkeiten anderer verstehen, nicht nur in der Lage sind, sondern auch hoch motiviert werden, ihre eigenen Emotionen, Ziele und Aufmerksamkeiten mit anderen zu teilen (Tomasello et al., 2005). Säuglinge beginnen im Alter von etwa 9 Monaten, ihre Aufmerksamkeit mit anderen auf Objekte von gemeinsamem Interesse außerhalb der direkten Verbindung mit dem Interaktionspartner zu koordinieren (Carpenter et al., 1998). Bereits im frühen Säuglingsalter, nehmen Säuglinge mit Freude an direkten Interaktionen mit ihrer Bezugsperson teil (Trevarthen, 1980) und wenden sich freiwillig von interessanten Objekten ab, mit denen sie zuvor beschäftigt waren. Darüber hinaus zeigen Neugeborene zwischen 2 und 5 Tagen nach der Geburt eine Vorliebe für das Betrachten von Gesichtern (oder Bildern von Gesichtern), deren Augen direkt auf sie gerichtet sind. Noch früher, innerhalb weniger Minuten nach der Geburt, zeigen Säuglinge ein beträchtliches Interesse an selbstgesteuerten Gesichtsbewegungen, vor allem an auffälligeren Handlungen wie dem Vorschieben der Zunge und dem Öffnen des Mundes und reagieren entsprechend interaktiv darauf (Kugiumutzakis, 1998; Meltzoff & Moore, 1977; Nagy & Molnar, 2004).

Eine notwendige Voraussetzung sozialer Interaktion ist die Fähigkeit, die Existenz anderer Mitglieder sozialer Gemeinschaften vorreflektiert zu erkennen. In Wagner (2019) wird diese vorreflektierte Selbst- und Fremdwahrnehmung als die Fähigkeit, eine ‚Zweite-Person-Perspektive' einzunehmen beschrieben, die soziale Interaktionen bereits präkognitiv ermöglicht. Ein solcher sozialer Ansatz entgeht dem normativen Fehlschluss, weil er das Personsein nicht auf vorbestehende normative Verpflichtungen stützt, aus denen ontologische Bedingungen abgeleitet werden, sondern die grundlegendste ontologische Bedingung von Personen normativ

begründet. Die paradigmatische Struktur personalen Lebens ist normativ, da sie die Art und Weise, wie wir wechselseitig miteinander umgehen, regelt. Personen sind gleichsam von Natur aus sozial und müssen ihr kollektives Leben entsprechend organisieren. In Übereinstimmung mit der grundlegenden ontologischen Natur von Personen ergibt sich die Normativität des Personseins aus der faktischen sozialen Organisation personalen Lebens und muss als solche auf einschränkende ontologische Fakten der natürlichen Welt reagieren. Während die kognitivistische Ontologie des Personseins das Normative voraussetzt, wird eine soziale Ontologie durch ihre Normativität konstituiert.

Neben den synchronen Bedingungen des Personseins ist für medizinethische Fragen, besonders im Zusammenhang mit der Validität von Patientenverfügungen, wichtig, unter welchen Bedingungen Personen über die Zeit fortbestehen – die Frage nach diachroner personaler Identität. Im folgenden Abschnitt werden Theorien personaler Identität, die tragende Teile ihrer argumentativen Kraft aus kontrafaktischen Gedankenexperimenten generieren, auf ihre methodische und empirische Plausibilität hin überprüft und ihre medizinethische Relevanz besprochen.

IV. Methodische und empirische Plausibilität von Theorien diachroner personaler Identität und ihre medizinethische Relevanz

Kontroversen über diachrone personale Identität spielen eine wichtige Rolle in der Medizinethik, wie etwa bei der Diskussion um Abtreibung (McInerney, 1990; Warren, 1977; Oderberg, 1997), bei Patientenverfügungen, insbesondere im Hinblick auf Neurodegeneration (Buchanan, 1988; DeGrazia, 1999; Vollmann, 2001; Limbaugh, 2016; Shelton & Geppert, 2024; Hart, 2024), und bei der Tiefen Hirnstimulation (DBS) (Lipsman & Glannon, 2013; Nyholm & O'Neill, 2016; Müller et al., 2017).

DBS, die als therapeutische Intervention bei neurodegenerativen Erkrankungen eingesetzt wird, hat das Potenzial, die psychologische Verfassung der Patienten erheblich zu verändern. DBS kann daher Auswirkungen auf den ontologischen, moralischen und rechtlichen Status von Patienten haben. Bei Patientenverfügungen im Zusammenhang mit Neurodegeneration müssen Patienten antizipatorisch Entscheidungen für ihr zukünftiges Selbst treffen, das jedoch unter Umständen nicht mehr mit dem ‚ursprünglichen' Selbst, das die Patientenverfügung unterzeichnet hat, identisch ist. Die medizinethische Zulässigkeit von Schwangerschaftsunterbrechung

ist eng mit dem moralischen Status von Föten und denjenigen Personen verbunden, zu denen sie potenziell werden: Wenn ein Fötus als eine werdende Person verstanden wird, ist es medizinethisch unzulässig, ihm eine „Zukunft wie die unsere“ zu verwehren, so jedenfalls das prominente Argument von Marquis (1989).

Diese und ähnliche Diskussionen über potenzielle Disruptionen personaler Identität in der Medizinethik stützen sich häufig implizit (gelegentlich auch explizit) auf philosophische Theorien diachroner personaler Identität. Diese Theorien werden wiederum paradigmatisch unter Berufung auf kontrafaktische Gedankenexperimente verteidigt, die zwar logisch möglich sind, aber häufig im Widerspruch zu relevanten Tatsachen der natürlichen Welt stehen. Diese Art von Gedankenexperimenten haben jedoch, wie gezeigt wird, keine argumentative Tragfähigkeit in Bezug auf reale Fälle diachroner personaler Identität im Zusammenhang mit medizinethischen Erwägungen in der klinischen Ethik.

IV.1 Orthodoxe Commitments und reale Personen

Wie in Abschnitt III besprochen, sind die synchronen Bedingungen von Personsein umstritten. Das Konzept des Personseins lässt prinzipiell auch nicht-menschliche Personen sowie künstliche Personen zu. Menschliche Personen sind jedoch derzeit der einzig unumstrittene Fall von Personsein. Es liegt also nahe, mit der Konstitution von menschlichen Personen als Paradigma zu beginnen. Dementsprechend stützen sich gegenwärtige Versuche, die synchronen Bedingungen von Personsein zu benennen, auf die folgenden drei Propositionen, die in Wagner (2022) als ‚Orthodoxe Commitments‘ zum Personsein bezeichnet werden: *Realismus*: ‚Person‘ ist eine natürliche Art, die Lebewesen bezeichnet, die de facto in der natürlichen Welt existieren. *Naturalismus*: Personen sind biologische Wesen, deren Existenz eine Frage empirischer Fakten ist. *Kognitivismus*: Personen sind mit höheren kognitiven Fähigkeiten ausgestattet, die ein diachrones Selbstbewusstsein ermöglichen. Zusammengefasst sind Personen reale, biologische Wesen, die sich selbstreflexiv via höherer kognitiver Fähigkeiten erstpersonal als über die Zeit fortbestehend begreifen (im Folgenden: ‚reale Personen‘).

Aus diesen orthodoxen Commitments folgt, dass Personsein untrennbar mit Tatsachen über die natürliche Welt verbunden ist und durch diese eingeschränkt wird. Diese einschränkenden Tatsachen der natürlichen Welt haben die konzeptuelle

Genese von Personsein entscheidend geprägt und ihre praktische Anwendung bestimmt. Wäre die natürliche Welt anders gewesen (bestünde beispielsweise die Möglichkeit, dass sich Menschen in zwei gleichwertige Nachfolger aufspalten), so hätte sich auch das Konzept des Personseins anders entwickelt. Die enge Beziehung zwischen Personsein und den einschränkenden Tatsachen der natürlichen Welt wird daher als ‚Intrinsische Verbindung' bezeichnet (ebd.).

Auch wenn die orthodoxen Comitmments weithin anerkannt sind, sind sich Theoretiker uneinig darüber, welche Bedeutung die synchronen Bedingungen des Personseins für die Beantwortung der Frage nach diachroner personaler Identität haben.

Die folgenden beiden Theorien sind die am ausführlichsten diskutierten Kandidaten: *Psychologische Kontinuitätstheorien* gehen davon aus, dass Person *x* zum Zeitpunkt t_1 notwendigerweise mit Person *y* zu t_2 identisch ist, genau dann, wenn[3] *x* und *y* durch immerwährende Ketten psychologischer Kontinuität verbunden sind.

Der *Animalismus* geht davon aus, dass Person *x* zu t_1 notwendigerweise mit Person *y* zu t_2 identisch ist, genau dann, wenn *x* und *y* durch immerwährende Ketten biologischer Kontinuität verbunden sind.

Wie psychologische und biologische Kontinuität ausbuchstabiert wird, unterscheidet sich zwischen den verschiedenen Vertretern dieser Theorien im Detail, ist allerdings für unsere Zwecke unerheblich.

IV.2 Gedankenexperimente und personale Identität

Unterschiede zwischen rivalisierenden Theorien personaler Identität treten häufig erst in Gedankenexperimenten zutage, die nicht selten zu bizarren kontrafaktischen Propositionen führen. Derek Parfit (1984) hat mit seinen paradigmatischen Arbeiten zur personalen Identität diese Art von Gedankenexperimente entscheidend mitangestoßen. Gedankenexperimente über mögliche Welten sind also nicht nur Illustrationen von Theorien und deren Implikationen. Vielmehr werden sie als adäquate, seriöse Versuche bemüht, um unser konzeptuelles Verständnis von personaler Identität in Bezug auf reale Personen zu schärfen. Der Hauptgrund für den Einsatz von Gedankenexperimenten besteht also darin, das konzeptuell Wesentliche an Personen und ihrer Identität herauszupräparieren, indem zentrale konzeptuelle Merkmale isoliert werden. Nach der vorherrschenden Auffassung erfordert dies, dass

[3] ‚Genau dann, wenn' wird im strikten Sinne einer logischen Äquivalenz verwendet.

man irrelevante ontologische Kontingenzen der Welt, in der sich Personen im wirklichen Leben befinden, ausblendet.

Hierbei geht es nicht um eine grundsätzliche Kritik an Gedankenexperimenten, die darauf abzielen, ontologisch irrelevante Kontingenzen der natürlichen Welt auszublenden; geschweige denn um eine Kritik an Gedankenexperimenten per se. Vielmehr wird die weitverbreitete Tendenz kritisiert, Gedankenexperimente zu verwenden, die mögliche Welt Modalitäten evozieren, um Schlussfolgerungen über Welten hinweg zu ziehen, und dabei die Intrinsische Verbindung zwischen dem Konzept Personsein und einschränkenden Tatsachen der natürlichen Welt außer Acht lassen. Dementsprechend ist es sinnvoll, zwischen zwei verschiedenen Arten von Gedankenexperimenten zu unterscheiden.

Eine Familie von Gedankenexperimenten, die ich für methodisch angemessen halte, nenne ich *hypothetisch*: Gedankenexperimente, die mit den Tatsachen der natürlichen Welt übereinstimmen, die sich auf reale Personen und deren Identität auswirken. Hypothetische Gedankenexperimente werden sowohl in der Philosophie als auch in den Wissenschaften häufig verwendet.

Eine Familie von Gedankenexperimenten, die ich für methodisch inadäquat halte, nenne ich *kontrafaktisch*:[4] Gedankenexperimente, die zwar logisch möglich sind, aber im Widerspruch zu den Tatsachen der natürlichen Welt stehen, die reale Personen und ihre Identität betreffen. Gedankenexperimente dieser Art sind in der Literatur zur personalen Identität besonders weit verbreitet; Teletransportation sowie Gehirntransplantationen spielen hierbei eine besonders wichtige Rolle.

Es ist weithin anerkannt, dass reale Personen ständigen biologischen Veränderungen ausgesetzt sind, die keine Gefahr für ihre personale Identität darstellen. Die Zellen des menschlichen Körpers werden fortwährend ersetzt, ohne dass dies eine Disruption weder der biologischen noch der psychologischen Kontinuität personalen Lebens zur Folge hat. Wenn psychologische Kontinuität gegeben ist, so ist in allen realen Fällen auch biologische Kontinuität gegeben, aber nicht umgekehrt. Das apallische Syndrom ist ein offensichtliches Beispiel dafür, dass biologische Kontinuität erhalten bleibt, psychologische Kontinuität jedoch nicht mehr gegeben ist.

[4] Ich verwende die Begriffe ‚hypothetisch' und ‚kontrafaktisch' ausschließlich, um faktisch mögliche bzw. faktisch unmögliche Gedankenexperimente in Bezug auf Personen und ihre Identität zu bezeichnen. Bei kontrafaktisch, spreche ich demnach nicht von Konditionalen, die Aussagen über Umstände machen, die eingetreten wären, wenn die tatsächliche Abfolge der Ereignisse anders gewesen wäre.

Um die Theorien psychologischer Kontinuität mit dem Animalismus zu vergleichen, ist es naheliegend sich vorzustellen, was passieren würde, wenn psychologische Kontinuität vorhanden wäre, biologische Kontinuität aber nicht. Gedankenexperimente zu Großhirntransplantationen scheinen hier besonders geeignet, da sie dem Naturalismus in Bezug auf das Personsein gerecht werden. Eine typische Darstellung einer Großhirntransplantation sieht wie folgt aus: Stellen wir uns vor, das Großhirn von *A* wird erfolgreich in den Schädel von *B* transplantiert, wobei der Hirnstamm und die Mittelhirnregionen von *A* intakt bleiben, so dass der Organismus von *A* am Leben bleibt. Stellen wir uns weiter vor, dass der daraus resultierende *B* psychologisch mit *A* identisch ist; die mentalen Zustände von *A* sind während des gesamten Prozesses physisch realisiert, und es gibt keine störenden konkurrierenden Kandidaten (Olson, 2016). Wer wacht nun nach dem Eingriff auf? Die scheinbar natürliche Intuition ist, dass Person *A* gleichsam mit ihrem Großhirn transferiert werden würde. Shoemaker (1963) stellt solche kontrafaktischen Großhirntransplantationen als entscheidenden Beweis für psychologische Kontinuitätstheorien gegen den Animalismus dar. Moderne Animalisten wie Snowdon (2014) sind jedoch anderer Meinung und bestreiten die Überzeugungskraft der Transplantationsintuition. Bei der Erörterung von Intuitionen, die aus möglichen Welt Modalitäten gewonnen werden, ist es wichtig zu bedenken, dass die so gewonnenen Ergebnisse dazu dienen, Differenzen zwischen rivalisierenden Theorien von personaler Identität in Bezug auf reale Personen beizulegen. Es geht nicht um die Behauptung, dass nur Personen in einer möglichen Welt, in der Großhirntransplantationen stattfinden, mit ihrem Großhirn transferiert werden. Sondern es geht darum, dass die Betrachtung dieser Kontrafaktizitäten zeigen soll, dass psychologische Kontinuität die ontologisch korrekte Theorie personaler Identität in der natürlichen Welt ist.

Angenommen, die Transplantationsintuition ist ein hinreichend zwingender Grund, den Animalismus zu verwerfen. Wir sind also nicht identisch mit dem lebenden Organismus, der bei einer Großhirntransplantation zurückbleibt. Vielmehr hören wir auf zu existieren, sobald unsere Psyche verschwunden ist; zumindest existieren wir nicht mehr in dem Organismus, dem das Großhirn entnommen wurde (Parfit, 2012). In der Praxis könnte dies bedeuten, dass angesichts der Verschlechterung des autobiografischen Gedächtnisses bei Alzheimer, die oft mit dem Verlust des Identitätsgefühls einhergeht (El Haj et al., 2017), Patientenverfügungen für Personen im Spätstadium einer Alzheimer-Erkrankung, die wenig bis gar keine psychologische

Kontinuität mit dem ursprünglichen Unterzeichner aufweisen, nicht als verbindlich angesehen werden sollten. Ebenso scheint es bei fehlenden Patientenverfügungen wenig Sinn zu machen, nahe Angehörige zu befragen, um den mutmaßlichen Willen des Patienten zu rekonstruieren. Schließlich ist der Patient, der derzeit behandelt wird, nicht mehr mit der ursprünglichen Person um deren mutmaßlichen Willen es geht, identisch ist. Bestenfalls kann argumentiert werden, dass die ursprüngliche Person diejenige mit der ‚engsten Kontinuität' zum aktuellen Patienten ist und somit aus pragmatischer Sicht am besten geeignet, über praktische Belange der ursprünglichen Person zu entscheiden.[5]

Es gibt jedoch mehrere einschränkende Tatsachen der natürlichen Welt, die verhindern, dass Großhirntransplantationen jemals realisierbar sein werden. Zum einen wird die zugrundeliegende Annahme, dass das Großhirn allein die psychologische Kontinuität einer Person aufrechterhält, durch Erkenntnisse der Kognitionswissenschaft in Frage gestellt. Die Theorien der Embodied Cognition (Clark 1997, 1999; Lakoff & Johnson, 1999) betonen die Interdependenz von Gehirn und Körper: die menschliche Psyche ist stark von körperlichen Eigenschaften abhängig und zu einem gewissen Grad vice versa. Dabei spielen Aspekte des Körpers einer Person, die über das Gehirn hinausgehen, eine bedeutende kausale oder physisch konstitutive Rolle bei der kognitiven Verarbeitung (Wilson et al., 2021). Selbst wenn es gelänge, ein ganzes, funktionstüchtiges Gehirn (oder auch nur das Großhirn) zu transplantieren, würde die psychologische Beschaffenheit der daraus resultierenden Person durch die Beschaffenheit eines völlig anderen Körpers geprägt und beeinflusst werden. Auch wenn sich der alte und der neue Körper sehr ähnlich wären, würden sie sich zwangsläufig immer noch geringfügig voneinander unterscheiden, und das würde sich auch auf die psychologische Konstitution der resultierenden Person auswirken.

Eine weitere Linie empirischer Forschung legt nahe, dass es eine starke ‚Gehirn-Körper-Historizität' (Munzer, 1994) gibt, die auf immunologischen Mechanismen beruht, die bei Transplantationen von Hirngewebe beobachtet werden. Das Immunsystem unterscheidet zwischen körpereigenem und fremdem Gewebe nur

[5] Die s.g. 'Closest Continuer Theory' wurde von Nozick (1981) entwickelt. Hierbei handelt es sich um eine externalistische metaphysische Theorie, die vorgibt, ein Problem zu lösen, das innerhalb der Grenzen einer internalistischen Metaphysik der personalen Identität unüberwindbar ist.

anhand der Qualität (nicht der Quantität) des eingebrachten Materials. Selbst wenn die Menge des eingebrachten Fremdmaterials gering ist, kann es zu Abstoßungsreaktionen kommen. Aus immunologischer Sicht scheint es also keine prinzipiellen Unterschiede zwischen der Transplantation von Hirngewebe und der Transplantation des gesamten Großhirns zu geben: Beide unterliegen der engen Interdependenz zwischen Gehirn und Körper. Man kann also nicht davon ausgehen, dass die Transplantation eines Großhirns in den Schädel einer anderen Person dazu führen würde, dass die Psyche der ursprünglichen Person transferiert wird. Vielmehr sind die lebenswichtigen Funktionen des gesamten Körpers, einschließlich, aber nicht nur, diejenigen Funktionen des Großhirns notwendig, um die ausgeprägte psychische Verfassung einer Person aufrechtzuerhalten – was darauf hindeutet, dass psychologische Kontinuität nicht nur kontingent von biologischer Kontinuität abhängt sondern eine konstitutive Beziehung besteht.

Wenn man sich vorstellt, dass psychologische Kontinuität von biologischer Kontinuität getrennt ist, wie es in den kontrafaktischen Fällen der Großhirntransplantation von uns verlangt wird, um psychologische Kontinuität als abhängige Variable zu isolieren und biologische Kontinuität durch unabhängige Variablen (oder andere Ursachen für psychologische Kontinuität) zu ersetzen, wird die nomologisch notwendige Interdependenz von psychologischer und biologischer Kontinuität verletzt. Wenn psychologische Kontinuität in allen tatsächlichen Fällen von biologischer Kontinuität abhängt, kann die bloße begriffliche Möglichkeit, dass beide sich voneinander lösen könnten, nicht als verlässliche Quelle von Intuitionen dienen, die Fälle von realen Personen informieren kann. Die Transplantationsintuition ist nicht nur unzuverlässig, wenn sie für Urteile über die Identität von realen Personen verwendet wird, sondern auch weitgehend irrelevant. Wenn wir in die kontrafaktische Perspektive eintreten, generieren wir Intuitionen über Wesen, deren erdachte physiologische und biologische Konstitution sich entscheidend von realen Personen unterscheidet, so dass die Rückübertragung dieser Intuitionen auf die natürliche Welt einen Kategorienfehler darstellt. Es werden realen Personen Eigenschaften zugesprochen, die Wesen dieser Art grundsätzlich nicht haben.

IV.3 Theoretische Überzeugungen, empirische Studien und praktische Anwendbarkeit

Das gegenwärtige Bemühen experimenteller Methoden in medizinethischen Diskussionen über personale Identität ist ein wichtiger Schritt, um die Signifikanz von Theorien personaler Identität, die auf kontrafaktischen Gedankenexperimenten beruhen, für den Transfer in die klinische Ethik zu hinterfragen.

Strohminger und Nichols (2015) haben Veränderungen der personalen Identität bei Patienten mit verschiedenen Arten neurodegenerativer Erkrankungen (Demenz, Alzheimer und amyotrophe Lateralsklerose) untersucht, wie sie von den Angehörigen der Patienten wahrgenommen werden. Den Probanden wurde mitgeteilt, dass der Zweck der Studie darin besteht, zu untersuchen, wie sich die neurodegenerative Erkrankung auf die persönlichen Beziehungen der Patienten auswirkt. Die Ergebnisse der Studie deuten darauf hin, dass die Beeinträchtigung des moralischen Vermögens die personale Identität der Patienten in der Wahrnehmung der Angehörigen besonders bedroht. Neurodegenerative Erkrankungen wie die frontotemporale Demenz, die die moralische Verarbeitungsfähigkeit des Gehirns beeinträchtigen, hatten die größten Auswirkungen auf die wahrgenommene Veränderung der personalen Identität, während neurodegenerative Erkrankungen wie die Amyotrophe Lateralsklerose, die hauptsächlich die kognitive Verarbeitung beeinträchtigen, die geringsten Auswirkungen auf die wahrgenommene Veränderung haben. Freilich lässt sich die personale Identität der Patienten, wie sie von den Angehörigen in der dritten Person wahrgenommen wird, nicht unmittelbar auf das übertragen, was die Patienten selbst in der ersten Person erleben, sodass nur begrenzte theoretische Rückschlüsse aus diesen Ergebnissen gezogen werden können. Aber wenn diese Studien überhaupt einen Hinweis darauf geben, was Personen für sich selbst in Bezug auf die medizinische Behandlung neurodegenerativer Erkrankungen wünschen, dann sollten die Ergebnisse ernsthaft in Betracht gezogen werden, wenn es um Patientenverfügungen geht.

Im Hinblick auf die potenzielle Bedrohung personaler Identität durch DBS haben Bluhm et al. (2020) kürzlich dafür plädiert, empirische Daten aus Patientenberichten zu nutzen, um sowohl die Patientenversorgung zu verbessern als auch Theorien personaler Identität zu untermauern. Ihre Ergebnisse deuten darauf hin, dass die tatsächlichen Erfahrungen von Patienten, die sich einer DBS unterzogen haben, mehr mit einem relationalen Verständnis von personaler Identität übereinstimmen,

demzufolge personale Identität innerhalb eines Netzes sozialer Beziehungen konstituiert wird (Schechtman, 2014), als mit einem metaphysischen Reduktionismus bei dem Personen vollständig auf die Existenz bestimmter psychologischer und/oder biologischer Zustände und ihrer verschiedenen Beziehungen reduziert werden. Bluhm et al. (2020) berichten, dass die Mehrheit der Patienten, die sich einer DBS unterziehen, nicht das Gefühl haben, sich danach grundlegend verändert zu haben; zumindest nicht mehr als die Veränderungen, die sie aufgrund ihrer Krankheit oder pharmakologischer Behandlungen erfahren haben. Nimmt man diese Patientennarrative ernst, so führt dies zu einer differenzierteren Lesart dieser Erfahrungen, die konkrete, praktische und theoretische Auswirkungen haben. Die Kontextualisierung der tatsächlichen Erfahrungen der Patienten erfordert die Frage, wie es ist eine Person zu sein, die mit DBS behandelt wird, anstatt zu fragen, ob DBS eine Bedrohung für die personale Identität darstellt. Dies ist nicht nur ein semantischer Taschenspielertrick, sondern kann zu einem besseren Verständnis dessen beitragen, was tatsächlich mit der personalen Identität von DBS-Patienten geschieht, und so helfen, ein maßgeschneidertes Advance Care Planning zu entwickeln, letztlich also die Patientenversorgung verbessern.

Es liegt nahe, dass die Beweislast bei den Befürwortern kontrafaktischer Gedankenexperimente liegt, aufzuzeigen, wieso diese Methode ein adäquates Instrumentarium darstellt, um Kontroversen personaler Identität mit Bezug zu Anwendungen in der klinischen Ethik zu lösen. Anstatt über bizarre Kontrafaktizitäten nachzudenken, scheint es angemessen, die kontroversen Fälle von personaler Identität in der natürlichen Welt auch theoretisch ernster zu nehmen. Was mit der Identität von Personen geschieht, die an Bewusstseinsstörungen (wie dem Wachkoma) oder neurodegenerativen Erkrankungen (wie dem Spätstadium von Alzheimer) leiden, wird beispielsweise in der bereits erwähnten medizinethischen Literatur zu Patientenverfügungen ausgiebig diskutiert. In Anbetracht der ontologischen Abhängigkeit zwischen Personsein und den einschränkenden Tatsachen der natürlichen Welt verdienen diese und andere Fälle aus der klinischen Ethik mehr theoretische Aufmerksamkeit, als ihnen derzeit zuteilwird.

In engem theoretischen und praktischen Zusammenhang mit den synchronen Bedingungen von Personsein und diachroner personaler Identität steht das in der Medizinethik zentrale Konzept der Autonomie, das vor allem durch aktuelle

Entwicklungen der Künstlichen Intelligenz neuen normativen Fragen ausgesetzt ist, die im folgenden Teil der kumulativen Habilitationsschrift besprochen werden.

V. Akteurschaft und Autonomie bei KI-gestütztem Gesundheitsmonitoring

Aktuelle Entwicklungen im Bereich des KI-gestützten Gesundheitsmonitorings ermöglichen es einer immer größeren Anzahl von Nutzern, eine große Menge personenbezogener Gesundheitsdaten zu sammeln und zu analysieren. Mobile Gesundheitsanwendungen (mHealth-Apps), die in der Regel auf Smartphones betrieben werden, sind die am weitesten verbreitete Form solcher Anwendungen. Zu den Kernfunktionen dieser Apps gehören die Überwachung von Bewegung, Ernährung und sportlichen Aktivitäten. Physiologische Parameter wie Herzfrequenz und Blutdruck, von denen bekannt ist, dass sie mit emotionalen Zuständen und Wohlfühlfaktoren wie Schlafqualität und sozialer Interaktion korrelieren, werden ebenfalls überwacht, erfordern aber möglicherweise Wearables wie Brustgurte oder Pulsuhren.

Solche mHealth Anwendungen werfen eine Reihe medizinethischer Fragen auf, vor allem in Bezug auf Datensicherheit, Verantwortung, Paternalismus und Autonomie (Krieger, 2013; Owens & Cribb, 2019; Davies, 2021; Wieczorek & Rossmaier, 2023) sowie mögliche Interessenkonflikte zwischen verschiedenen Stakeholdern. Die Bedenken hinsichtlich der Autonomie der Nutzer werden immer dringlicher, da eine wachsende Zahl solcher Anwendungen nicht nur Daten sammelt, sondern auch darauf abzielt, die Nutzer davon zu überzeugen, ihren Lebensstil zum Besseren zu verändern, d. h. gesünder und aktiver zu leben. Um dieses Ziel zu erreichen, werden eine Vielzahl von Persuasionsstrategien eingesetzt, die potenziell die Autonomie der Nutzer untergraben und so ihre Akteurschaft erodieren können.

V.1 Autonomie, Persuasion und Paternalismus

Beauchamp und Childress (2019) verstehen Respekt vor Autonomie mindestens als eine Selbstbestimmung, die frei ist von kontrollierenden Eingriffen durch andere und von Beschränkungen, wie z.B. unzureichendem Verständnis, das sinnvolle Entscheidungen verhindert. Die Semantik von Persuasion spielt somit eine zentrale Rolle bei der Beurteilung der Frage, ob ihre Anwendung in mHealth-Apps eine Bedrohung für die Autonomie der Nutzer darstellt. Persuasion bedeutet in diesem

Kontext, jemanden dazu zu bringen, etwas zu tun (oder zu unterlassen), indem die Überzeugungen und Absichten, die zu einer Handlung führen, verändert werden. Während Persuasion in der Sozialpsychologie und den Gesundheitswissenschaften eine weitgehend positive Konnotation hat (Cialdini et al., 2005), ist ihr Ruf in der Handlungstheorie eher zwielichtig; manchmal wird Persuasion als ähnlich zu (oder bestenfalls zwischen) Manipulation und Überzeugung gesehen (O'Keefe, 2012). Persuasive Technologien sind in der Regel so konzipiert, dass sie technologische Mittel bereitstellen, um das Verhalten der Nutzer über ihre Überzeugungen und Intentionen absichtlich und oft dauerhaft zu verändern, indem sie fortwährend Rückmeldungen über als unangemessen empfundenes Verhalten geben und auf verschiedene Weise Anreize für erwünschtes Verhalten schaffen (Fogg, 2003). Persuasion ist somit inhärent normativ, da ihr Hauptziel darin besteht, nicht nur Gesundheitsdaten zu beschreiben oder die Nutzer darüber zu informieren, sondern Verhaltensveränderungen herbeizuführen. Häufig wird argumentiert, dass solche persuasiven Interventionen gegen den Grundsatz des Respekts vor der Autonomie verstoßen, da sie eine Form des Paternalismus darstellen und in den autonomen Willen des Akteurs eingreifen, auch wenn sie durch die Überzeugung motiviert sind, dass es dem Nutzer bessergehen wird, wenn er dazu gebracht wird, sein Verhalten entsprechend zu ändern (Enoch, 2016).

In zahlreichen Schriften legt Gerald Dworkin (1972, 2005, 2015, 2017) eine Definition paternalistischer Interventionen vor, die zu verstehen hilft, warum Persuasion als Bedrohung für autonomes Handeln im Kontext von mHealth angesehen werden kann. Dworkin schlägt folgende Bedingungen für eine Analyse von X handelt paternalistisch gegenüber Y, indem er Z tut (unterlässt) vor: (1) Z (oder deren Unterlassung) greift in die Freiheit oder Autonomie von Y ein. (2) X tut dies ohne die Zustimmung von Y. (3) X tut dies nur, weil X der Meinung ist, dass Z das Wohlergehen von Y verbessert oder in relevanter Weise die Interessen, Werte oder das Wohl von Y fördert. Auch wenn Dworkin nicht explizit sagt, dass X und Y zwei unterschiedliche Akteure sind, die jeweils ihre eigenen Intentionen haben, legt seine Analyse implizit nahe, dass er genau das im Sinn hat. Paternalismus liegt also vor, wenn ein Akteur einem anderen Akteur seinen Willen aufzwingt, und zwar in der wohlwollenden, wenn auch bevormundenden Absicht, das Wohlergehen oder die Interessen des anderen im Allgemeinen zu fördern (oder zumindest zu wahren).

Da bei paternalistischen Eingriffen à la Dworkin zwei verschiedene Akteure beteiligt sind, stellt (2) in Verbindung mit (1) eine Verletzung der Autonomie von Y eo ipso dar. Man kann nicht einerseits die Autonomie einer Person respektieren und andererseits in ihre Autonomie eingreifen, indem man auf sie einwirkt, ohne zuvor ihre informierte Einwilligung eingeholt zu haben. Die Verbindung von (1) und (2) lässt jedoch offen, ob ein Eingriff in die Autonomie einer Person ethisch zulässig wäre, wenn der Bevormundete seine Zustimmung gegeben hätte. (3) untergräbt die Autonomie von Y insofern, als es impliziert, dass X in einer besseren epistemischen Position ist, um zu urteilen (oder auf andere Weise kompetenter ist) als Y selbst, wenn es darum geht, herauszufinden, welche Handlungen und Absichten dem Wohlergehen von Y dienen; damit wird ihre Entscheidungsfähigkeit und praktische Rationalität in Frage gestellt. (1) beruht ganz offensichtlich auf der Annahme, dass die paternalistische Handlung Z von X durchgeführt oder initiiert wird, wobei X ein eigenständiger Akteur ist, der Überzeugungen und intentionale Zustände hat, die er Y aufzwingen will. Da (2) davon ausgeht, dass X an Y ohne dessen Zustimmung Z durchführt, werden X und Y offensichtlich als zwei verschiedene Akteure betrachtet. Während dies in Standardfällen von Paternalismus zutreffen mag, ist es alles andere als klar, ob im Kontext von mHealth tatsächlich zwei verschiedene Akteure im Spiel sind. Es stellt sich also die Frage: Sind mHEalth Apps ‚verdeckte Akteure', die mit den Mitteln der technischen Persuasion die Agenda eines tatsächlichen, anderen Akteurs vorantreiben, oder sind diese Apps vielleicht nur eine Erweiterung des eigenen Willens des Akteurs?

Dworkins drei Bedingungen für Paternalismus beruhen auf der prima facie plausiblen Annahme, dass X und Y unterschiedliche Akteure sind, die jeweils ihre eigenen Absichten vertreten. Paternalismus scheint also nur dann vorzuliegen, wenn X auf Y einwirkt und mindestens eine der drei Bedingungen von Dworkin erfüllt ist. Während dies in (2) vernünftig sein mag, da diese Bedingung ipso facto einen anderen Akteur voraussetzt, ist dies in (1) und (3) nicht unbedingt zutreffend, denn (1) und (3) müssen nicht immer einen anderen Akteur voraussetzen. Ich kann im Prinzip paternalistisch auf mich selbst einwirken, indem ich zum Beispiel distale Absichten formuliere und Maßnahmen ergreife, die mein zukünftiges Ich dazu bringen, diese zu erfüllen – in gewisser Weise kann man Patientenverfügungen als ein solches Instrumentarium verstehen. Eine Situation, in der X auf Y einwirkt, wobei X kein eigenständiger Akteur ist, sondern ein technologisches Gerät, das im Dienste von Y handelt (d. h. als

Erweiterung des Willens von Y und nicht im Gegensatz zu diesem), stellt möglicherweise keinen Fall von Paternalismus dar – es könnte sogar ein genuiner Ausdruck der Autonomie des Handelnden sein.

V.2 Die handlungstheoretische Verzerrung von Akteurschaft

Der scheinbare Gegensatz von Paternalismus und Akteurschaft legt nahe, dass paternalistische Interventionen Nutzer in erster Linie als passive Rezipienten von Intentionen anderer ansprechen und damit ihre Autonomie untergraben. In Wagner (2019b) wird diesem fehlgeleiteten Fokus entgegengewirkt, indem darauf hingewiesen wird, dass Paternalismus nicht zwangsläufig autonomiefeindlich sein muss, sondern unter bestimmten Umständen auch autonomieerhaltend oder gar autonomiefördernd sein kann.

Autonomie wird für gewöhnlich durch die eigenen Handlungen zum Ausdruck gebracht und nicht durch das, was uns widerfährt. Die zugrundeliegende Dichotomie zwischen Handlungen und Dingen, die uns zustoßen, spielt eine entscheidende Rolle für den Ruf des Handelns als eines inhärent aktiven Merkmals des Lebens von Akteuren, greift aber zu kurz. Die Konzepte ‚patient' und ‚patiency' als Fachbegriffe in der Handlungstheorie haben eine andere Konnotation als der enge Begriff des Patienten im medizinischen Kontext. Handlungstheoretisch beschreibt der Begriff ‚patient' die Passivität von jemandem, der sich einer Handlung ausgesetzt sieht oder dem etwas geschieht. Diese Passivität wird vor allem mit der Aktivität von Akteuren kontrastiert. Der philosophische Gegensatz zwischen Akteuren und ‚patients' lässt sich grob mit dem Slogan zusammenfassen: ‚Akteure tun etwas aktiv, während patients etwas passiv geschieht'. Akteurschaft genießt in der Handlungstheorie ein erhebliches Privileg, ebenso wie das moralische Handeln in der Ethik. Philosophen betonen oft, dass unser Leben aufgrund dessen, was wir tun, gut verläuft, und nicht aufgrund dessen, was uns widerfährt (Lott, 2016).

Um die Verbindung von Akteurschaft und ‚patiency' zu verstehen, ist es wichtig zu erkennen, dass es sich hierbei um Korrelate und nicht um Gegensätze handelt. Reader (2007) beschreibt treffend, wie die aktiven Merkmale des Lebens von Akteuren, die als exklusiv für die Akteurschaft angesehen werden, eine entsprechend komplementäre, andere Seite haben.

Im medizinischen Kontext besteht eine Besonderheit im Verhältnis zwischen Akteurschaft und ‚patiency': Die besondere Beziehung zwischen medizinischen Patienten und medizinischem Personal beruht auf zwei Annahmen. Zum einen verlassen sich Patienten auf medizinische Expertise und stützen ihre Entscheidungen auf die Empfehlungen der Ärztinnen und Ärzte. Andererseits sollen Mediziner so handeln, dass sie den ihnen anvertrauten Patienten medizinisch nützen und gleichzeitig die Autonomie der Patienten respektieren. Wenn Mediziner allein auf der Grundlage des Patientenwohls handeln, und dabei die Autonomie der Patienten ignorieren, handeln sie paternalistisch. Da mHealth Apps nicht nur von gesunden Menschen genutzt werden, sondern auch zur Überwachung von Krankheiten genutzt werden, verschwimmt die Unterscheidung zwischen dem handlungstheoretischen und dem medizinischen Patientenbegriff und damit auch die Frage, ob eine medizinische Form von Paternalismus, die möglicherweise technologische Persuasion beinhaltet, in solchen Fällen akzeptabel ist. Eine weitere Komplikation, die diese Entscheidung erschwert, ist die Tatsache, dass sich Nutzer in ihrer Gesundheitskompetenz stark unterscheiden, was sich auf die Autonomie und Entscheidungskompetenz auswirkt (Barello et al., 2020).

V.3 Stärkung der Handlungskompetenz

Die Autonomie der Nutzer von mHealth Apps zu erhalten und letztlich zu stärken, kann durch das erreicht werden, was als volitionale Hilfen (Wagner, 2019b) beschrieben wird. Wenn die übergeordneten Ziele der Verhaltensänderung von den Nutzern selbst (d. h. autonom) festgelegt werden, ist kein anderer Akteur beteiligt und somit keine Bedrohung der Autonomie gegeben. Die App unterstützt lediglich die Realisierung der ursprünglichen Intentionen des Nutzers. Um diesen Gedanken weiter zu vertiefen, unterscheide ich zwischen Autonomie erster und zweiter Ordnung: Autonomie erster Ordnung kann in diesem Zusammenhang als die ausgeübte Fähigkeit eines Akteurs beschrieben werden, autonom Entscheidungen auf niedriger Ebene zu treffen. Zum Beispiel kann er aus einer Laune heraus entscheiden, was er zu Mittag isst, wie er zur Arbeit kommt oder ob er heute ins Fitnessstudio geht. Autonomie zweiter Ordnung kann in diesem Zusammenhang als die ausgeübte Fähigkeit eines Akteurs beschrieben werden, autonom Entscheidungen auf einer höheren Abstraktionsebene

zu treffen. Zum Beispiel die Entscheidung, die eigene Ernährung zu verbessern oder ein aktiveres Leben zu führen.

In Anlehnung an Harry Frankfurts (1971) hierarchisches Modell der Autonomie sind Akteure in ihren Handlungen genau dann autonom, wenn ihre autonome Entscheidungsfähigkeit erster Ordnung im kleinen Maßstab von ihrer autonomen Entscheidungsfähigkeit zweiter Ordnung im großen Maßstab gebilligt (oder sanktioniert) wird. Die vorgeschlagene Sichtweise der Autonomie zweiter Ordnung als die ausgeübte Fähigkeit eines Akteurs, autonome Entscheidungen auf einer höheren Abstraktionsebene zu treffen, die den gelegentlichen Verzicht auf autonome Entscheidungen erster Ordnung einschließen kann, wirft in Verbindung mit der Hypothese, dass mHealth Apps als ‚extended mind' (Clark & Chalmers, 1998) angesehen werden können, ein neues Licht auf Spahns (2012) zweites Prinzip der ethisch vertretbaren Persuasion, das letztlich darauf abzielt, Persuasion zu beenden. Die Extended-Mind Hypothese schlägt vor, dass kognitive Prozesse nicht auf das Gehirn beschränkt sind, sondern sich auch auf die externe Welt erstrecken können. Kognitive Prozesse können dabei externe Ressourcen wie Smartphones einbeziehen. Diese externen Hilfsmittel sind Teil des kognitiven Systems, wenn sie effektiv die gleichen Funktionen wie interne Gedächtnis- oder Denkprozesse erfüllen. Wenn mHealth Apps in diesem Sinne als volitionale Hilfen verstanden werden, besteht keine Notwendigkeit, auf ein Ende der Persuasion hinzuarbeiten. Durch die Verwendung solcher volitionalen Hilfen wird die innere Willenskraft gleichsam an die Umwelt ausgelagert, die dennoch Teil des eigenen Willensapparats des Akteurs bleibt, solange sie auf einer autonomen Entscheidung zweiter Ordnung beruht.

Sowohl in der Freizeit als auch im medizinischen Kontext werden mHealth Apps hauptsächlich zu Tracking- und Überwachungszwecken oder zusätzlich zur Unterstützung der Nutzer bei der Änderung ihres Verhaltens eingesetzt. Wenn sie mit dem expliziten Ziel eingesetzt werden, bei der Änderung des eigenen Verhaltens unterstützt zu werden, erwarten die Nutzer, dass solche Geräte sie zu einem gewünschten Ergebnis überzeugen. Unter diesen Umständen kann Persuasion kaum als Bedrohung der Autonomie gelten, sondern eher als Ausdruck der autonomen Entscheidung zweiter Ordnung, solche Geräte zu nutzen. Wichtig ist, dass autonomiefördernde persuasive mHealth Apps den Nutzer als vernunftfähigen Akteur behandeln, indem sie ihm Gründe und motivierende Anreize für selbstinitiierte Verhaltensänderungen präsentieren. Weintraub & Barilan (2001) konstatieren, dass

der Wert der Autonomie auf das Recht der Menschen zurückgeht, als Akteure respektiert zu werden, die in Angelegenheiten von größter persönlicher Bedeutung, wie z. B. Entscheidungen über die medizinische Versorgung, argumentieren, überzeugen und überzeugt werden können. Autonomie wird demnach erst dann respektiert, wenn ein solcher Persuasionsversuch stattgefunden hat.

Bei der Suche nach praktischen Lösungen für persuasive mHealth Apps müssen jedoch einige Punkte beachtet werden. Wichtig ist, dass die von den autonomen Entscheidungskapazitäten zweiter Ordnung gesetzten Ziele selbstmotiviert sein müssen, um die beabsichtigte Wirkung zu erzielen. Nach der Selbstbestimmungstheorie (Ryan & Deci, 2000) ist autonome Motivation wesentlich effektiver, um das eigene Verhalten nachhaltig zu verändern, als das, was Ryan und Deci als kontrollierte Motivation bezeichnen. Wenn Akteure autonom motiviert sind, erhalten sie Selbstunterstützung und Verstärkung durch ihre eigenen Handlungen; die Motivation ist intrinsisch und das Verhalten ist somit selbstbestimmt (Hager et al., 2014). Im Gegensatz dazu ist die kontrollierte Motivation eine externe, introjizierte Regulierung des eigenen Verhaltens (z. B. Vermeidung von Strafe oder Schuldgefühlen). Es gibt zahlreiche Belege dafür, dass die autonome Motivation die stärksten Auswirkungen auf Verhaltensänderungen hat, insbesondere auf gesundheitsbezogenes Verhalten (ebd.). Pavey und Sparks (2010) zeigen außerdem, dass autonome Motivation einen gesünderen Lebensstil fördert, indem sie die Absicht unterstützt, gesundheitsschädliches Verhalten zu reduzieren.

Persuasive mHealth Anwendungen stehen demnach nicht prinzipiell im Widerspruch zur Autonomie der Nutzer. MHealth stellen unter den genannten Bedingungen keine invasiven, externen Eingriffe in das Leben von Akteuren dar, sondern sind vielmehr volitionale Hilfen, die Akteuren bei der Bewältigung schwieriger Aufgaben helfen, die andernfalls aufgrund eines vorübergehenden oder dauerhaften Mangels an den entsprechenden internen Ressourcen schwer zu bewältigen gewesen wären.

VI. Konklusion und Ausblick

Konkludierend lassen sich die folgenden, zentralen Ergebnisse der vorliegenden Habilitationsschrift „Methoden und Anwendungen medizinethischen Argumentierens“ festhalten:

- Es wurden zunächst vier zentrale, argumentationslogische Elemente medizinethischen Argumentierens identifiziert und kritisch diskutiert
- Anschließend wurde anhand konkreter Anwendungsbeispiele exemplarisch aufgezeigt, wie anhand der erarbeiteten argumentationslogischen Elemente thematisch mehrdimensionale medizinethische Analysen methodisch plausibel miteinander verbunden werden können

Auf argumentationslogischer Abstraktionsebene wurden die folgenden vier Elemente medizinethischen Argumentierens herausgearbeitet (Wagner, 2015):

(1) Empirisch-informierte Syllogismen: Valide Verbindung von Fakten und Normen via ‚normativ-indeterminierter' Prämissen, die eine ergebnisoffene Argumentation ermöglichen

(2) Explanativer, empirischer Zugang, der ergebnisoffene, normative Ableitungen ermöglicht

(3) Hypothetische und kontrafaktische Gedankenexperimente, die, in Übereinstimmung mit den restriktiven Fakten der natürlichen Welt, als Prüfstein medizinethischer Theorien dienen

(4) Schlüsselkonzepte – Personsein, Akteurschaft und Autonomie – die sowohl zentrale theoretischen Debatten im Fach bestimmen als auch konkrete Entwicklungen in der klinischen Ethik maßgeblich beeinflussen

Auf praxisbezogener Konkretisierungsebene wurden die o.g. vier Elemente medizinethischen Argumentierens anhand der folgenden Anwendungsfälle näher bestimmt:

a) In der Anwendung funktioneller Bildgebung mittels Elektroenzephalographie (EEG) als Prüfstein moralischer Kognition, die es ermöglicht, kognitive, affektive und exekutive Hirnfunktionen sofort nach oder unmittelbar vor der Verarbeitung moralisch relevanter Stimuli zu messen. Anhand der Analyse solcher EEG-basierter Studien wurden konzeptuelle und experimentelle Herausforderungen diskutiert. Dabei wurde die Hypothese aufgestellt, dass tugendethische Theorien am besten mit den erhobenen Daten kohärieren, was nahelegt, dass moralische Kognition ein globaler Hirnprozess ist, der nicht auf einzelne Hirnregionen oder Funktionen beschränkt ist (Wagner et al., 2017).

b) Im Bereich der anthropologischen Voraussetzungen von Personsein und ihrer normativen Signifikanz. Die Hinwendung zur Analyse einer weitverbreiteten

Vermengung ontologischer und normativer Grundlagen von Personsein hat gezeigt, dass die in der Medizinethik orthodoxen Theorien des kognitiven Ursprungs von Personsein, auf deren Grundlage normative Praxen erklärt und begründet werden, auf dem in Wagner & Northoff (2015) eingeführten ‚normativen Fehlschluss' basieren. Hierbei wird zur Begründung des exklusiven moralischen Status von Personen die Ontologie und Normativität von Personsein unzulässig vermengt, indem das vermeintliche ontologische Primat höherer kognitiver Fähigkeiten herangezogen wird, um präexistente moralische Überzeugungen post hoc zu rechtfertigen – es wird also von Sollen auf Sein geschlossen. Alternativ wurde aufgezeigt, wie ein Ansatz der Personsein präkognitiv, d.h. anhand der sozialen Einbettung von Personen, konstituiert, diesen normativen Fehlschluss vermeidet (Wagner, 2019a).

c) Anhand einer konzeptuellen Analyse von Gedankenexperimenten zu Theorien diachroner personaler Identität, die zur medizinethischen Evaluation der Validität von Patientenverfügungen vor allem bei neurodegenerativen Erkrankungen herangezogen werden. Es wurde gezeigt, dass kontrafaktische Gedankenexperimente, die regelmäßig bemüht werden, um die Plausibilität konfligierender theoretischer Ansätze diachroner personaler Identität zu verhandeln, ein methodisch fehlerhaftes Instrumentarium darstellen. Personsein ist, bei Lichte besehen, untrennbar mit restriktiven Fakten der natürlichen Welt verbunden. Ein modaler Skeptizismus legt daher nahe, dass das Bemühen von Intuitionen, die aus kontrafaktischen Gedankenexperimenten generiert werden, zu ‚konzeptuellen Inkongruenzen' führt. Diese konzeptuellen Inkongruenzen stellen die Adäquatheit der aus kontrafaktischen Gedankenexperimenten gewonnenen Theorien für potenziell reale Anwendungen in der Medizinethik erheblich in Frage (Wagner, 2022).

d) Im Bereich der Ethik des KI-gestützten Gesundheitsmonitorings. Die Untersuchung des Zusammenhangs zwischen digitalem, KI-gestützten Gesundheitsmonitoring (mHealth) und Akteurschaft legt initial nahe, dass die Anwendung persuasiver Technologien die Autonomie des Akteurs unzulässig erodiert. Unter Bezugnahme der Theorien des Extended Mind und Extended Will wurde demgegenüber gezeigt, dass mHealth Anwendungen, unter bestimmten, medizinethisch fundierten Voraussetzungen, Akteure dabei

unterstützen können, autonom gesetzte Ziele mit technologischer Hilfe effektiver und effizienter zu erreichen (Wagner, 2019b).

Zukünftige Forschungsvorhaben im Bereich der Methoden und Anwendungen medizinethischen Argumentierens könnten die folgenden Ebenen der vorgelegten Habilitationsschrift aufgreifen und weiter vertiefen.

Auf **theoretischer Ebene** ist eine Weiterentwicklung argumentationslogischer Grundlagen in der Untersuchung alternativer konzeptueller Strukturen zur Modellierung moralischer Urteile denkbar. Hierbei sollte die Verbindung von empirischer Unvoreingenommenheit und normativer Relevanz zentral sein.

Auf **empirischer Ebene** sind multimodale Studiendesigns zur Erforschung moralischer Kognition, die fMRT, EEG und aktuelle Anwendungen generativer Künstlicher Intelligenz kombinieren, vielversprechend. Zentral wird hierbei sein, ein möglichst hohes Maß an ökologischer Validität der Studiendesigns mit konzeptueller Plausibilität zu verknüpfen.

Auf **klinisch-ethischer Ebene** ist die Weiterentwicklung empirisch-informierter Theorien personaler Identität (z.B. für neurodegenerative Erkrankungen), die praxisnahe Lösungen für die Gültigkeit von Patientenverfügungen ermöglichen, zentral. Dabei ist die Integration von empirischen Daten (z. B. Krankheitsverläufe) für die Evaluation diachroner Identitätsfragen entscheidend.

Auf normativer Ebene erscheint die weiterführende Untersuchung der Balance zwischen weiter fortschreitenden technologischen Eingriffen in die autonome Entscheidungsfindung und menschlicher Entscheidungsfreiheit von zentraler Bedeutung. Dabei ist die Entwicklung ethischer Standards für die Gestaltung KI-gestützter Technologien, die Akteurschaft stärken statt untergraben, besonders relevant.

Originalpublikationen

Theor Med Bioeth (2015) 36:215–235
DOI 10.1007/s11017-015-9330-z

A fallacious jar? The peculiar relation between descriptive premises and normative conclusions in neuroethics

Nils-Frederic Wagner[1] · **Georg Northoff**[1]

Published online: 20 May 2015

Abstract Ethical questions have traditionally been approached through conceptual analysis. Inspired by the rapid advance of modern brain imaging techniques, however, some ethical questions appear in a new light. For example, hotly debated trolley dilemmas have recently been studied by psychologists and neuroscientists alike, arguing that their findings can support or debunk moral intuitions that underlie those dilemmas. Resulting from the wedding of philosophy and neuroscience, neuroethics has emerged as a novel interdisciplinary field that aims at drawing conclusive relationships between neuroscientific observations and normative ethics. A major goal of neuroethics is to derive normative ethical conclusions from the investigation of neural and psychological mechanisms underlying ethical theories, as well as moral judgments and intuitions. The focus of this article is to shed light on the structure and functioning of neuroethical arguments of this sort, and to reveal particular methodological challenges that lie concealed therein. We discuss the methodological problem of how one can—or, as the case may be, cannot—validly infer normative conclusions from neuroscientific observations. Moreover, we raise the issue of how preexisting normative ethical convictions threaten to invalidate the interpretation of neuroscientific data, and thus arrive at question-begging conclusions. Nonetheless, this is not to deny that current neuroethics rightly presumes that moral considerations about actual human lives demand empirically substantiated answers. Therefore, in conclusion, we offer some preliminary reflections on how the discussed methodological challenges can be met.

✉ Nils-Frederic Wagner
nils-frederic.wagner@web.de

[1] Mind, Brain Imaging and Neuroethics Research Unit, University of Ottawa Institute of Mental Health Research, Ottawa, Canada

Keywords Neuroethics · Ethical theory · Neuroscience of morality · Moral psychology · Naturalistic fallacy · Normative fallacy · Normative indeterminacy · Norm-fact linkage

Introduction

There is a contemporary zeitgeist reflecting the tendency to bring together philosophy and the empirical sciences, in particular neuroscience. Inspired by the rapid advance of modern brain imaging techniques, such as functional magnetic resonance imaging (fMRI), philosophers who engage with neuroscience have begun to work on what was long believed (and is still believed by some) to be the last secure stronghold of a purely conceptual discipline: ethics.[1] So far as this is concerned, scholars from different disciplines look toward neuroscience for guidance in moral questions. This often naturalistically minded endeavor has led to the rise of neuroethics[2] as a fairly novel branch of interdisciplinary research.[3] Yet, there is a difficulty in giving a straightforward all-encompassing definition of neuroethics, since the work done under this label is not monolithic in its methodology. What we are concerned with in what follows is a naturalistic form of the neuroscience of ethics which is the predominant, though not only, school within neuroethics. In further characterizing the burgeoning field of neuroethics, we are following the lead of Adina Roskies, who asked,

> Will the biologizing of the moral undermine its status as moral? … It is clear that as such questions are approached scientifically, the answers we get will shape our ethical views and, thus, will affect how we approach the ethics of neuroscience. As we learn more about the neuroscientific basis of ethical reasoning, as well as what underlies self-representation and self-awareness, we may revise our ethical concepts. [3]

Along those lines, Michael Gazzaniga affirms: 'Cognitive neuroscience has valuable information to contribute to the discussion of certain topics that have traditionally been taken up by bioethicists, namely, those issues in which brain science has relevant knowledge that should impact the ethical questions being debated' [4].

Such approaches to neuroethics can be labeled as 'robustly naturalististic' in their methodology. Attending to neuroscientific evidence, then, is believed not only to be expository as to how human beings de facto act but also to reveal tenets of normative ethical theories. Doing neuroethics in this way can be roughly delineated

[1] Recently, this conviction has been opposed by experimental approaches to philosophy in general and experimental ethics in particular. Major proponents endorsing experimental philosophy are, amongst others, Joshua Knobe and Shaun Nichols (for a theoretical and methodological justification of their approach, see [1]).

[2] For a discussion of the theoretical and methodological hallmarks of neuroethics, see [2].

[3] The same holds for the ever-growing research conducted under the umbrella term of moral psychology. Unlike in neuroethics, however, moral psychologists are less concerned with making normative claims, but rather aim at describing human functioning in moral contexts. This is not to ignore that the findings of moral psychology are frequently invoked into ethical controversies; sometimes they are consulted to serve as a 'tie-breaker' between conflicting theories.

as the attempt to draw conclusive relationships between neuroscientific observations and normative ethical theories—consequently aiming to suggest concrete prescriptions in applied ethics. The general aim is to merge scientific descriptions and normative evaluations of human (and occasionally non-human) life. Neuroethics so defined investigates both the significance of neuroscientific findings for the understanding of morality, and the relevance of ethics for determining the normative import of evidence from neuroscience.[4] The former is concerned with neural and psychological mechanisms underlying ethical concepts and judgments; whereas the latter is concerned with the implications of those findings for moral practice. Accordingly, Adina Roskies introduced the clarifying distinction between 'neuroscience of ethics' and 'ethics of neuroscience' [3].

On a related note, neuroessentialist views asserting that, 'for all intents and purposes, we are our brains' [6] have become increasingly popular. Such views have found their way into ethical discourses, creating new methodological difficulties emerging from a neuroscientific naturalism that materializes in the structure of neuroethical arguments that, more often than not, rely effusively on empirical evidence. While there is much debate over the 'is-ought divide' in empirically informed ethics in general [7], the underlying structure of neuroethical arguments in particular remains elusive.

Current debates in neuroethics mostly focus on the application of neuroscientific findings to concrete ethical questions, such as free will [8], moral responsibility and psychopathy [9], and impacts of brain interventions on personal identity [10]. Little, however, has been said about the general structure of how these arguments operate. In addressing this important and underappreciated methodological concern, the focus of this article is to reveal the implicit metaethical premise that underlies a great deal of neuroethical arguments and to showcase specific methodological predicaments that result thereof. In that way, we seek a better grip on assessing merits and demerits of neuroethical work and point to possible limitations and fruitful applications.

In what follows, we will focus on two particular methodological challenges in the neuroscience of ethics. First, we will tackle the methodological problem of how one can—or, as the case may be, cannot—validly infer normative conclusions from neuroscientific observations. Second, we will shed some light on how preexisting normative ethical convictions threaten to invalidate the interpretation of neuroscientific data, and, by so doing, arrive at question-begging conclusions. Thereby, the guiding methodological questions will be: Can neuroscience contribute to definitions of moral reasoning and moral behavior? Can empirical findings that explain how we actually behave or reason, support claims about how we ought to behave and reason? Are neuroethical investigations unconsciously biased by preexisting normative convictions? If so, is it feasible to invoke those theoretical assumptions to justify neuroethical conclusions?

In targeting these issues, our goal is not to give an exhaustive overview of work done in neuroethics but, rather, to examine paradigmatic examples of important

[4] In current medical ethics, decision making and informed consent are hotly debated topics. For a recent investigation of these issues from a neuroethical perspective, see [5].

neuroethical arguments as present in the work of Joshua Greene and Patricia Churchland.[5] Here, a note of caution is appropriate: the aim of our article is not to assess Greene's or, for that matter, anyone else's *arguments* for their normative ethical convictions—this has been done thoroughly elsewhere.[6] Our aim, rather, is to reveal the *underlying structure* of how neuroethical arguments work when they appeal to empirical evidence from neuroscience; hereby, Greene's and Churchland's views are dealt with as paradigmatic examples, not as specific targets. The central point we want to address is the issue of how empirical evidence is invoked to play a decisive role in reaching neuroethical conclusions claiming to be morally significant. The role that empirical evidence plays is believed to be decisive insofar as it is appealed to in order to tip the balance in favor of one position or another. In order to reveal the methodological predicaments in such an approach, we proceed in three main steps:

(1) Examine how the naturalistic fallacy applies to robust naturalistic neuroethical arguments: do these arguments directly infer an 'ought' from an 'is'—or is their force rather based on, as it were, smuggled-in normative assumptions? We will call these assumptions 'semi-normative' claims.
(2) Introduce the normative fallacy: do these arguments further infer an 'is' from an 'ought'—thus presupposing the conclusion in the premises? We will call these 'result-closed' arguments.
(3) Sketch some possible solutions to the diagnosed methodological predicaments: namely, avoid biased interpretations of empirical data by starting neuroethical investigations with a normative indeterminacy leading to what we will describe as 'result-open' arguments. The aim here is to achieve sound neuroethical arguments that cohere best with empirical evidence due to what we indicate as a 'norm-fact linkage'.

The structure of neuroethical arguments: naturalism and methodological predicaments

To be clear from the outset, we believe that current neuroethics rightly presumes that moral considerations about actual human lives demand empirically substantiated answers. Yet, the naturalistic conjecture that inquiry into the natural world can increase moral knowledge in just the same way as it increases scientific knowledge seems rather contentious. Along these lines, a good deal of current neuroethics proposes (sometimes implicitly, i.e., neither defended nor even stated as such) a naturalistic form of moral realism, according to which there are objective moral truths, or moral facts and moral properties. And these moral facts, as

[5] There are, of course, a great many other scholars doing important work in neuroethics and moral psychology that we are not discussing here; see, e.g., [11–15].

[6] For an ingenious analysis of Greene's normative ethical arguments, see [16].

naturalists believe, are, at the same time, natural facts and properties.[7] This is frequently called 'moral naturalism', which is, at least in its most robust form, committed to a rejection of the fact-norm distinction.[8] Neuroethics thus seems to be grounded in some form or other of 'metaphysical naturalism', which is the conviction that all facts and properties are natural, even if we are unable to presently recognize them as such. Accordingly, non-naturalistic forms of normativity are seen, in this view, as unfounded or founded upon illusory beliefs. Neuroethical arguments of this sort typically appear in a form like this:

(1) (Implicit) metaethical premise: there are (depending on the metaethical conviction) absolute (context independent) or relative (context dependent) normative truths which are, in either case, natural facts or properties that can be discovered scientifically.[9]
(2) Empirical claim: observation of some fact x about brain activity during a morally significant judgment or conduct.

Therefore,

(3) Normative conclusion: according to the empirical evidence about fact x, the moral judgment or conduct in question is inferred to be either right or wrong (good or bad).

It is quite evident that a particular methodological challenge in these kinds of neuroethical endeavors lies in the relation between facts and norms. More precisely, there are crucial steps to be taken to get from empirical claims, derived from observations of brain activity underlying moral intuitions and capacities, to normative conclusions of how to get things right. This holds both in principle, i.e., according to normative ethical theories, and in concrete settings of applied ethics. In this regard, Guy Kahane emphasizes the difference between an investigation into underlying psychological and neural mechanisms that may account for moral

[7] Many 'traditional' contemporary moral philosophers endorse one form or another of non-naturalistic moral realism (most prominently Derek Parfit, Tim Scanlon, and Thomas Nagel). On the other hand, proponents of naturalistic neuroethics are mostly either neuroscientists, like Sam Harris, who endorses a naturalistic form of moral realism, or neurophilosophers—some of whom hold opposing metaethical views, such as Jesse Prinz's non-cognitivism and Patricia Churchland's naturalistic moral realism. Josuha Greene is metaethically agnostic and lately considers himself to be a moral skeptic. Although he has frequently asserted the supremacy of utilitarianism as a normative view based on neuroimaging studies that he interprets according to a coherentist moral epistemology.

[8] For recent reflection on the plausibility of moral naturalism, see [17].

[9] Henceforth, we assume this metaethical premise to be present in the exemplary neuroethical arguments we consider. We are aware of the fact that not all neuroethical arguments are ipso facto committed to moral naturalism. However, the arguments we are using as paradigmatic examples throughout this article attempt to draw more or less direct normative conclusions from descriptive premises that describe brain activity, and such arguments are likely to presuppose moral naturalism. Our claim is just that the argumentative force of endeavors like this depends on moral naturalism, not that these are the only sorts of arguments available or that all proponents of neuroethics explicitly endorse or implicitly commit themselves to a robust form of moral naturalism.

competence and normative ethical theories that seek to provide principled correct answers to moral questions. Kahane writes:

> [T]he aim of ethical theory is surely not to investigate moral competence, or people's psychology or capacities, but to answer substantive normative questions. Moral intuitions may be evidence for or against possible answers to these questions, but they are not data that moral theories seek to causally explain. The aim of ethical theory is to get things right, not to explain why we have a certain set of beliefs (let alone of intuitions). [18]

Kahane's reflections call into question the feasibility of inferences from factual descriptions (based on the observation of brain activity underlying moral beliefs and intuitions) to normative conclusions that purport to either vindicate or impugn certain forms of judgment or conduct.

We now turn to shed some light on how neuroethical arguments that deal with this issue are threatened by two different (albeit intertwined) forms of fallacies: the 'naturalistic fallacy' and the 'normative fallacy'.

The threat of naturalistic and normative fallacies

Naturalistic proponents of neuroethics frequently endorse a reductionist form of inferring normative conclusions from factual descriptions. In so doing, the normative level is mostly disregarded; that is, the normative level is reduced to the factual level. Ethical norms are thus unilaterally replaced by neural facts. Hence, the aim is to, as it were, 'neuronalize' the ethical concept, which means to reduce it to neuronal facts.[10] This, however, involves the danger of an uncritical acceptance of empirical presuppositions and resulting terminological definitions. Traditionally, this form of inference has even been presented as fallacious.

Ever since David Hume [19] and George Edward Moore [20], philosophers take issue with drawing normative conclusions from factual descriptions. These famous non-cognitivists argue, in slightly different ways, against naturalistic forms of moral realism and ethical rationalism by stating that ethical conclusions cannot be drawn validly from premises which are in themselves non-ethical. It is not valid, as they say, to infer an 'ought' from an 'is'; to infer from fact to value, from descriptive to normative ethical propositions, from visibility to desirability. In a nutshell, evaluative conclusions cannot be drawn from non-evaluative premises. This has been labelled as the naturalistic fallacy. Moore puts it as follows: 'I have thus appropriated the name Naturalism to a particular method approaching Ethics. … This method consists in substituting for "good" some property of natural object or of a collection of natural objects' [20, pp. 91f.]. In another passage of *Principia Ethica*, Moore writes, 'But if [one] confuses "good," which is not … a natural object, with any natural object whatever, then there is a reason for calling that a naturalistic fallacy' [20, p. 65]. Norms and values are regarded as distinctively

[10] For a discussion of this issue, see [2]. Another form of neuroethical investigation (with which we are not concerned here) is merely to reveal the relevance of ethical concepts for neuroscientific research.

different properties than facts and descriptions; so each set of properties belongs to a different realm. The implication of this view is the invalidity of unilateral inferences from one realm to the other. For example, it is a non sequitur to infer from the premise that if someone observes his fellow students skipping class today that she is morally right to skip class herself as well. What is at stake as to norms and facts, Hume and Moore say, is a difference in kind, not merely a difference in degree.

There is much philosophical controversy about the naturalistic fallacy and accordingly a substantial body of literature concerning this major metaethical issue.[11] One can, and justifiably, be undecided about whether or not there can in principle be a way to derive an 'ought' from an 'is'. However that may be, if a robust naturalistic neuroethical argument that aims at prescribing how to get things right is pursued, a methodological step to bridge the gap between the different realms is needed. This holds true even if one assumes that there is no such thing as a naturalistic fallacy, since one still has to explain how to get from observations to prescriptions, for fallacies are not the only source of gaps. As long as the debate over the divide between 'is' and 'ought' is not dissolved, the claim that there is a peculiar relation (or a possible gap) between facts and norms, stands. Ignoring this would be to confuse metaphysical with epistemic and linguistic differences and, therefore, neglecting the need for distinct methodological approaches when relating 'is' and 'ought' claims, either in a metaphysical or in an epistemological or in a linguistic sense. That is to say, even if one accepts the contentious assertion that there is no deep metaphysical difference between facts and norms, and thus believes that it is not fallacious to infer from 'is' to 'ought', it does not follow that there cannot be crucial methodological differences in the ways that conclusions are inferred from these two sorts of statements. For the epistemic differences between facts and norms—accompanied by the linguistic differences in uttering 'is' and 'ought' sentences (for examples of the linguistic formulation of the is-ought gap, see [33, 34])—cannot be denied. In other words, granted for the sake of the argument that there is no such thing as a naturalistic fallacy, it would still be unclear how facts relate to norms for the simple epistemic reason that 'ought' statements make claims about how the world *should be*, whereas 'is' statements are descriptions of how the world *is*. Here, it is worth emphasizing the different ways in which evidence is gathered in support of these two sorts of statements. Evidence for 'is' claims is gathered by observation, whereas evidence for 'ought' claims is gathered by arguments from principles that, more often than not, appeal to consequences. Thus, even if one denies the metaphysical difference between facts and norms, or is agnostic as to how 'is' and 'ought' claims are metaphysically related, the epistemic and linguistic difference still holds and has to be methodologically accounted for.

There are intelligible arguments on either end of the 'is-ought' debate, and, as we will attempt to illuminate, there are also some viable intermediate positions. Without trying to resolve the naturalistic fallacy (which seems to be a matter of trying to square the circle), what the concerns initiated by Hume and Moore illustrate is the *peculiarity* in the metaphysical relation between descriptive

[11] For pertinent recent discussions, see [21–31]. Since Frankena [32], some people have called into question the existence of a naturalistic fallacy.

statements and normative conclusions[12] or, more broadly construed, between facts and norms.[13] In current neuroethics, however, this relation is often implicitly assumed to be a straightforward naturalistic obviousness. Normative facts are believed to be, both in principle and in particular, more or less directly inferable from neuroscientific facts.

Keeping these methodological considerations in mind, we will propose that it is important to ask the following two questions when confronted with naturalistic neuroethical arguments that attempt to draw normative conclusions from descriptive claims.

(1) Are the descriptive claims correct? This involves asking whether the experimental designs of neuroscientific studies are actually significant and thus able to capture what they aim to investigate.[14]
(2) Do the normative conclusions really follow? This involves asking whether the interpretation of the empirical data, given that the designs are significant, is sound.

Granting for the moment that there might be a principled way to overcome the naturalistic fallacy, one can nonetheless be skeptical about how descriptive claims could have, even prima facie, any normative significance whatsoever. Considering the following basic structure of a naturalistic neuroethical argument shall help to further illustrate this point.

Descriptive claim: the amygdala *is* firing when φing.

Therefore,

Neuroethical conclusion: we *ought* (or, depending on the interpretation of the descriptive claim, *ought not*) to φ.

Left in this basic structure (admitted, of course, that this presentation is quite simplified; yet, for our purposes, there is no need to consider empirical details at this point), arguments of this sort appear hopelessly flawed, for they are based on what we will call a *strong* form of the naturalistic fallacy. By 'strong', we here refer to a direct, i.e., unilateral normative inference from a descriptive claim.

To avoid such a strong form of the naturalistic fallacy, it seems as though an *extra premise* is needed that aims at connecting the descriptive claim to the

[12] As we said before, if one is skeptical about that, at least the epistemic and linguistic difference between facts and norms is rather uncontentious and needs to be methodologically accounted for.

[13] The relation between facts and norms is peculiar because there is no straightforward, i.e., direct, way of drawing conclusive arguments from one realm to the other. That is, neither from facts to norms, nor from norms to facts. We will further explain this point in what follows. The general idea is that in current neuroethics, investigators frequently form conclusions according to a suite of convictions that covertly inherits some content from the investigators' preexisting normative convictions.

[14] Asking whether the empirical claim is correct is, of course, an empirical question and can thus be tackled through the assessment of the data. More important for present purposes is the second claim, which questions the normative significance of the experimental design. In the debate on free will, for example, an often invoked criticism of the Libet experiments is that they do not properly investigate free will, but rather, as Markus Schlosser puts it, 'freedom of indifference' [35].

normative conclusion. It is difficult, however, to find extra premises that can support valid inferences from descriptive claims to normative conclusions; especially without presupposing the result, or, for that matter, begging the question. Consider the following integration of what we will call a *semi-normative* extra premise (for lack of a better term)[15] into an exemplary neuroethical argument:

(1) Empirical claim: the amygdala is known to be integral for emotional processing.
(2) Semi-normative claim: emotions are an unreliable source for moral judgments.[16]

Therefore,

(3) Neuroethical conclusion: when the amygdala is firing during moral decision making, the resulting moral judgments—including whatever implications may follow for particular normative ethical theories—should not be trusted.[17]

When evaluating this argument, what immediately comes to mind is the question of where exactly the justification of the semi-normative claim (2) comes from. And what, after all, is a semi-normative claim? As we will argue hereafter, in answering this question, there may be concealed a source of yet another form of a fallacious inference.

Empirical claim (1)—that the amygdala is firing during emotional processing—aims to causally explain psychological states and not to evaluate them. That much is clear. Whereas the semi-normative claim (2)—that emotions are an unreliable source of moral judgments—bears in itself a normative assumption. It is, therefore, not a purely empirical statement. This is so because the assumption that emotions/sentiments are an unreliable source of moral judgments lies at the core of certain ethical theories and is as a result bound to their normative implications, as in various forms of consequentialism. Therefore, semi-normative claims aim to justify—and consequently to prescribe—morally correct principles that are in accordance with (or follow from) said normative assumptions.

A paradigmatic example of this kind of neuroethical reasoning can be found in the teaming up of neuroscientist Joshua Greene and philosopher Peter Singer. The team argues that the fact that consequentialist moral theories sometimes prescribe actions that are at odds with widely held moral intuitions/emotions cannot by itself count against these theories. On the contrary, it rather increases the plausibility of consequentialist moral theories, they say, since when following consequentialists

[15] In what follows, we will elaborate more on this and aim at making the point that this can also be thought of as a kind of 'meta-normative' conviction. These sorts of claims take a stand on the viability of certain forms of normativity (e.g., emotional arousal in moral judgments), but without really providing any metaethical reasoning for this conviction.

[16] This, again, is merely illustrative and not meant to be a statement that we endorse or believe to be true.

[17] When considering this example it is not our purpose to argue for or against any particular normative ethical theory. Rather, we attempt to show how the persuasiveness of neuroethical arguments with this sort of structure is threatened by the particular methodological challenges we discuss.

frameworks, we arrive at conclusions through, as it were, 'cold' (prudential) reasoning, which is a more reliable source with a higher likelihood to arrive at sound moral judgments than emotions are.[18] This claim has for a long time been invoked in moral philosophy by proponents of consequentialism arguing against deontology, among others, initially detached from any empirical considerations. Greene argues along those lines, as we will show in what follows, but supplements his case with empirical evidence from fMRI studies he conducted on trolley dilemmas.

Neuroscience of morality and trolley dilemmas

Trolley dilemmas consist of a series of thought experiments originally developed by Philippa Foot [41] to trigger and analyze moral intuitions. Ever since, these hypothetical cases are extensively discussed in normative ethics and have recently attracted much attention in experimental approaches to ethics. The popularity of these dilemmas is particularly due to the debate ignited by Judith Jarvis Thomson, whose widely cited formulation of the trolley dilemma is as follows:

> Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. [42]

A slight variation of this case in which you are a bystander that can detour the trolley by pulling a switch, is often referred to as 'pulling the switch case', or, for short, 'switch'. Another vexed variant is the so called 'footbridge case'. As before, a trolley is running down a track towards five people. You are on a footbridge under which it will pass, and you can stop the trolley by pushing a heavy weight person onto the track. As it happens, there is a person next to you that appears to be sufficiently weighty to stop the trolley—that is, to the best of your knowledge (however that may have come about). Your only way to stop the trolley is to push him over the bridge onto the track, killing him in order to save five, thereby using him as a means to an end.

According to Greene (see [36] and elsewhere), it is especially the footbridge scenario that showcases how the often observed emotionally based moral judgment

[18] Admittedly, Greene's picture is a bit more complex than we describe it here for the sake of simplicity. Greene argues for what he calls a 'dual-process theory of moral thinking', according to which there are two distinct psychological systems forming moral judgments. The 'automatic mode system' generates moral judgments based on emotional responses to morally significant situations and leads to rapid ad hoc reactions. In contrast, the 'manual mode system' generates moral judgments based on conscious deliberation, thereby possibly overriding the automatic mode system in favor of what reason tells us to be the morally right conduct. For further details, see [36–39]. For Singer's treatment of moral intuitions, see [40].

to refrain from pushing the weighty man over the bridge is problematic. The reason is that, as Greene puts it, an 'up close and personal' involvement in directly harming a stranger likely leads to an emotionally based moral judgment—in this case, the deontological conclusion that it is intrinsically wrong to push the stranger over the bridge onto the track even if it would save five workmen. This is so because it would be to use the weighty man as a means to an end, rather than an end in himself. In contrast, according to the consequentialist's calculus, it is morally right to push the stranger in order to save the five workmen, since it leads to a favorable consequence: saving five lives at the cost of sacrificing one person's life. From a consequentialist point of view, whether or not this involves up close and personal harming, or treating someone merely as a means to an end, is neither here nor there. Greene asserts that based on the fact that there is no inherent moral difference in either pulling a switch or pushing a weighty man in the different versions of the trolley dilemma, one can, and in fact *ought to*, infer that it is unreasonable and unjustified to conceive that pushing the weighty man in the footbridge case is morally wrong while pulling the switch is morally right.[19] After all, Greene believes, the only difference between the footbridge and the switch case lies in a 'morally insignificant' fact: the involvement of up close and personal harming of a stranger. Accordingly, he believes that this 'debunks' deontological moral judgments.[20] For these deontological judgments are based on moral emotions which are contingent reactions shaped by the details of our evolutionary history, and thus *should* be divested of their moral significance, as far as this is possible.

Here, it becomes apparent that the earlier discussed semi-normative claim that emotions are an unreliable source for moral judgments, although somewhat disguised as an empirical statement, is for the most part a presupposed hidden normative conviction. Considering that the empirical part of this claim appears as a rather idiosyncratic attempt to support the normative claim, Greene's claim ought to be seen as contentious.[21] The inference between the invoked empirical evidence and the normative conclusion is often very difficult to draw, and should not be subject to an author's undefended normative convictions. Even if these normative convictions are defended, it is important—though often omitted—to show exactly what role the invoked empirical evidence plays in reaching the conclusion. If the invoked empirical evidence was not the backbone of the argument, then why not refrain from it altogether? In the most recent defense of his view, Greene explicitly asserts that 'moral psychology matters, not because it can generate interesting normative conclusions all by itself, but because it can play an essential role in generating interesting normative conclusions' [44]. In footnote 68 of the same paper, Greene further opposes Selim Berker's claim that 'nonempirical normative assumptions "do all the work" in the above arguments, rendering the science "normatively

[19] Here, we should add that this judgment is based on the outcome, thereby presupposing a consequentialist framework. In light of this, one wonders whether this begs the question; however this is a question for another day.

[20] Elsewhere, Greene elaborates more on why he believes Kantian ethics to be grounded in emotions and why this is not a plausible basis for ethical reflection on his view [43].

[21] For a similar claim, see Berker's critique of Greene's fMRI based discussions of said trolley cases [16].

insignificant"', arguing instead: 'Normative assumptions do some work, but empirical evidence does essential work as well'.

How the 'essential work' that empirical evidence does in these arguments is contentious can be made particularly clear by looking at opposing interpretations of similar empirical data. This phenomenon is most plausibly explained by differences in preexisting normative commitments that consequently lead to different conclusions, and thus, they undermine the point of empirical evidence being crucial in reaching the conclusion. Gerd Gigerenzer, for example, holds a view that is exactly opposed to Greene's and Singer's consequentialist convictions. He asserts that people should not rely on conscious deliberation when making moral judgments but, rather, on moral intuitions, or, as Gigerenzer calls them, 'fast and frugal heuristics' [45, 46].

With that said, the attempt to support normative ethical views by invoking empirical evidence is not prevalent in consequentialists' theories alone, but also in deontologists' theories. For example, contrary to Greene's (seemingly) empirically supported endorsement of consequentialism when faced with trolley dilemmas, Patricia Churchland believes there is empirical evidence in similar studies supporting deontology.[22] A simplified version of her argument goes as follows:

(1) Empirical claim: the amygdala is known to be integral for emotional processing.
(2) Semi-normative claim: emotions are a reliable source for moral judgments.

Therefore,

(3) Neuroethical conclusion: when the amygdala is firing during moral decision making, we should trust the resulting moral judgments—including the supportive implications that may follow for deontology.

Result-closed arguments

Neuroethical arguments of the kind discussed above might be described as *result-closed* arguments. The purpose of adding a semi-normative claim is to provide a 'smuggled-in-assumption' that presupposes the intended conclusion in a question-begging way, thereby trying to bring home the preexisting commitment. Arguably, the addition of a semi-normative claim intenerates the gap between descriptive statements and normative conclusions because it is empirically more sophisticated, but ultimately the gap remains. It remains because there is no bridging connection between the normative assertion and the empirical data. The conclusion is already present in the semi-normative claim and therefore the result of the argument is predetermined, or closed. The empirical data are merely used to justify the preexisting normative commitment by asserting that the normative force of the claim is present in a sound interpretation of the empirical data. This interpretation, however, presupposes the normative commitment and thereby begs the question. In

[22] See [47] for details.

contrast to the aforementioned strong naturalistic fallacious neuroethical arguments that directly infer normative conclusions from descriptive premises, the empirically more sophisticated variant that draws on a semi-normative claim involves what can be described as a *weak* naturalistic fallacy. Those arguments still derive an 'ought' from what they posit to be an 'is' but do so by combining the normative and the empirical premise into one, albeit question-begging, claim.

Normative fallacy

We now turn to arguing that these forms of argument beg the question in yet another way that is more difficult to see. Besides the danger of invalidly inferring from facts to norms, there is an additional danger involved in making inferences in the other direction: the drawing of seemingly factual conclusions from hidden normative premises. In what follows, we first shed light on how such reasoning begs a methodological question, demonstrating that some neuroethical arguments make empirical claims that are covertly shaped by undefended normative positions. In a second step, we raise these methodological considerations in further addressing the result-closed quality of neuroethical arguments as discussed above.

The general methodological issue of drawing factual conclusions from normative premises has first been discussed by Tom Campbell [48] and has ever since been underappreciated.[23] We call this methodological difficulty in neuroethical arguments, in line with Campbell's general conceptual analysis, a 'normative fallacy'. Campbell argues that philosophical analyses of concepts and the logic of discourse are, despite denials, either descriptive or normative in nature. He believes this is at least in part due to the fact of an indistinct boundary between philosophy and social science. If philosophy is seen as a merely conceptual endeavor (being solely committed to following the rules of logic and internal coherence), and therefore believed to be independent of empirical facts, then there is seemingly no need to involve empirical evidence in arguing for theoretical conclusions or normative commitments. Thus, some philosophers believe themselves to be guarded from the necessity to consider empirical evidence. In Campbell's words, 'this frees [philosophers] from the responsibility of providing empirical evidence to support their conclusions and also wards off the accusation that they are parading subjective preferences as if they were rationally justifiable propositions' [48]. This belief has led to the undesirable consequence that some philosophical analyses contain empirical generalizations that reflect hidden normative assumptions or convictions of the philosopher rather than carefully interpreted empirical evidence. Campbell states further:

> In doing this they may be said to reverse the naturalistic fallacy, and, by arguing from 'ought' to 'is', commit what I shall call the normative fallacy. This fallacy consists of arguing from propositions which are themselves normative, or could count as evidence only for normative propositions, to conclusions which contain factual assertions. The error of such reasoning is

[23] Despite the fact that a lot of philosophical discussions are affected by some sort of a normative fallacy, as of yet, Campbell's paper is the only published work elaborating on this issue.

> obvious, but it is frequently masked by confusions about the precise nature of philosophical analysis. [48]

The same general point made about the naturalistic fallacy can thus in reverse be applied to the normative fallacy. Nothing can appear in the conclusion of a valid deductive inference which is not already implicit in the conjunction of the premises.

Now, of course, by their very nature, neuroethical arguments rely heavily on empirical evidence; but taken closely into consideration, these supposedly factual generalizations sometimes turn out to contain hidden preexisting normative assertions. According to Campbell, it is a confusion about what is actually going on in many philosophical arguments that leads to the sort of conjunction of normative arguments and factual conclusions that constitutes the normative fallacy. Campbell posits that it is easy enough for philosophers to consider their analyses of concepts to constitute factual discoveries based on a certain understanding of philosophical activity. Needless to say, there is a rival view of what it is to do conceptual analysis that derives from Wittgenstein [49]. Rather than discovering the nature of a given concept, it can be said that philosophers are in the business of recapitulating often elusive and occasionally profound, but nevertheless normative assertions about acceptable meanings of words, or the way concepts feature in ordinary discourse. The normative arguments that sometimes invisibly buttress philosophers' apparently factual conclusions about concepts are normative claims about those meanings or uses of words and concepts of which the philosopher approves. What follows is that in many cases, the techniques of inquiring after what is meant by certain phrases or arguing about what they should mean can be regarded as appealing to normative opinions rather than empirical evidence.

So, Campbell convincingly argues that we cannot make valid claims about facts of the world by means of inferring from claims that are based on normative grounds. Descriptive conclusions cannot be drawn validly from normative premises.

Result-closed neuroethical arguments

How does this general methodological point apply specifically to neuroethical arguments? The above discussed example that we *ought* to follow a certain normative ethical theory, which is seemingly supported by neuroscientific evidence describing brain functions, is a case in point threatened by the normative fallacy. For it does not entail that there really *is* anything normatively significant in the invoked empirical evidence that can support these normative assumptions, unless the normative commitment that leads to the conclusion is already implicit in the premise and therefore presupposed.

In order to illustrate what we mean by that, consider again the neuroethical argument based on fMRI studies on trolley dilemmas that have been consulted to debunk deontological moral judgments. Recall Greene's argument regarding the footbridge case, which roughly says that moral judgments resulting from moral intuitions based on emotional reactions are unreliable because they are at odds with prudential reasoning (which in turn is believed to be a reliable source of moral judgments). However, nothing in the empirical observation of the amygdala firing

when people are presented with the footbridge case provides in itself any evidence for the rightness or wrongness of the resulting moral judgment. What the observation of the amygdala firing merely shows, is that people are emotionally aroused when confronted with these scenarios. The claim leading to the conclusion that deontological judgments are unreliable because of emotional involvement is based on the preexisting normative assertion that moral emotions (or intuitions) should not be trusted because they may be at odds with what reason recommends. However, this normative assertion in itself cannot be directly supported by the observation of brain activity, since those empirical observations can only reveal emotional arousal and do not themselves render evaluations as to the reliability of the resulting moral judgments. The force of the argument, therefore, must come in elsewhere; that is, it must derive from its normative presupposition. In Greene's case, the presupposition seems to be that moral judgments based on emotions are wrong because one *ought* not to trust in moral emotions according to consequentialism. This presupposition is a case in point: it shows that Greene's consequentialism predates and animates his conclusions from the fMRI studies he conducted on trolley dilemmas. (It goes without saying that these convictions have been asserted long before people even dreamt about modern brain imaging techniques.)

Clearly, this is a form of a normative fallacy and amounts to what we earlier called a result-closed argument. How so? In starting from the normative assumption that emotions are an unreliable source of moral judgments, then going on to conclude that one therefore ought not to trust them, the result of the neuroethical argument, which is supposed to be based on empirical evidence, is already pre-established by the hidden normative commitment. As a consequence, the pre-established normative commitment significantly determines the interpretation of the data. From the normative conviction that, say, emotions are an unreliable source of moral judgments, Greene and others infer that there de facto must be something empirically observable (namely, in this example, the firing of the amygdala during moral reasoning) corresponding to the unreliability of emotions. Thereby, an 'is' is inferred from an 'ought'. The fallacious inference runs as follows: one 'ought' not to trust in emotions as a source of moral judgments, thus there must be some empirical observable 'is' that affirms this 'ought'. This, then, reverses the naturalistic fallacy, since it consists in inferring from the 'ought' of not trusting moral judgments when based on emotions to the 'is' of the amygdala firing that is invoked to support the earlier mentioned 'ought'. One ought not to trust moral judgments based on emotions, consequentialists posit, therefore, something in empirically observable brain functioning shows that emotions are unreliable when it comes to moral reasoning.

It is, however, highly contentious whether there is any equivalent of the normative commitment to be discovered in the data itself. This can, for example, be seen in opposing normative interpretations of the same data. If one happens to be a proponent of deontological theories, believing that moral intuitions actually are a reliable source of moral judgments, then the very same data appear in a completely different light and, as a consequence, are interpreted in the opposite way. Some deontologists argue, as we observed earlier, that one really ought to follow one's moral intuitions in the footbridge case precisely because the amygdala is firing,

which indicates emotional involvement and therefore, since deontologists believe in the reliance on moral emotions, not pushing the weighty man is the right thing to do. This normative interpretation of the amygdala firing in the footbridge case, by the same token, would also be a result-closed argument, equally based on a normative fallacy. Therefore, it seems as though conclusions of neuroethical arguments of this sort are crucially based on preexisting normative convictions, and the empirical data in their own right play no significant role in the conclusion; rather, the empirical data is consulted in order to bring home the preexisting normative ethical commitment.

Concluding remarks and proposed solutions: normative indeterminacy and norm-fact linkage

The main aim of this article has been to reveal methodological challenges that threaten to invalidate conclusions of robust naturalistic neuroethical arguments. Both the naturalistic and normative fallacies suggest that there are methodological predicaments in neuroethical arguments that seem like solid brick walls, difficult to overcome. This, as the title of the article suggests, can also be seen as, figuratively speaking, a 'fallacious jar' in which one is stuck between methodological predicaments that involve the danger of falsely inferring both from facts to norms and from norms to facts. Nevertheless, we believe that the overall project of neuroethics can be fruitful and is indeed needed, since moral considerations about actual human lives demand empirically substantiated answers. With this in mind, the consultation of modern brain imaging techniques (and other empirical methods) can be extremely valuable. For this reason, we do not want to remain entirely destructive, but also want to provide a few pertinent (admittedly preliminary) ideas for how some of the discussed problems can be tackled and might lead to a means of escaping the fallacious jar.

Pertinent to our reflections on how to tackle some of the diagnosed problems that occur in robust neuroethical research, there is a substantial body of thought coming from philosophy of science that addresses the complex ways in which scientific practices, and the products of science, are interwoven with values. Since the 1950s and 1960s, it has been argued that science is inevitably, at least to some extent, governed by value judgments [50]. For one, the application of scientific methods is value-laden. In the quest for empirical discoveries, methods are restrained according to (often implicitly presupposed) normative ethical convictions. For example, invasive or potentially harmful experiments on healthy human participants are disallowed even at the expense of potentially finding a cure for cancer. Accordingly, it has been argued that besides logic and evidence, science is in need of additional guidance for theory choice [51]. In order to account for this, the term 'epistemic values' was introduced to encompass the values that were seen as acceptable in guiding scientific research and theory building [52].

More recently, Heather Douglas [53] has opted to abandon the ideal of value free science, particularly if value free science is meant to include the rejection of epistemic values. We agree that there need not be (and maybe cannot be) a science

that is freed of epistemic values, and we concur that it would be good for science to allow for 'more open discussion of the factors that enter into scientific judgments and the experimental process' [53]. Douglas acknowledges the methodological predicament of the naturalistic fallacy, alongside the difference between descriptive and normative statements, but she contends that 'this does not mean … that a descriptive statement is free from values in its origins. Value judgments are needed to determine whether a descriptive label is accurate enough and whether the errors that could arise from the description call for more careful accounts or a shift in descriptive language. Evidence and values are different things, but they become inextricably intermixed in our accounts of the world' [53]. In order to make sense of empirical discoveries, we are dependent on scientific interpretations of the acquired data, and these interpretations inevitably have a normative dimension.

Now, how does this apply to the discussed methodological predicaments in robust naturalistic forms of neuroethical research? It is not only the threat of preexisting normative commitments that makes these sorts of neuroethical arguments contentious, but also the frequent neglect of the contingencies governing the de facto norms and social structures of everyday life. Any scientific endeavor inevitably presupposes certain epistemological and metaphysical commitments since agents are shaped by a particular context—perceiving and interpreting the world around themselves in a great many different ways. In trying to determine the normative significance of neuroscientific evidence, the relatively austere individual and idiosyncratic social points of view cannot be altogether disregarded. In other words, the social and political contexts in which neuroethical questions are posed and answers are proposed need to be taken into consideration.

Normative concepts like moral judgment and related theoretical reflections, such as the moral status of persons, cannot be comprehensively understood if the social and political context of these concepts is not accounted for.[24] This calls for an explicit discussion of the often implicit assumptions of these context-shaped concepts and the need to situate them in both their social and political contexts. If normative ethical convictions are taken for granted and seen as non-negotiable points of reference, the threat of falling for the normative fallacy arises. If, however, these normative ethical convictions are explicitly discussed and critically engaged, there is warranted hope for minimizing this predicament.

On the other hand, the recognition that normative convictions always stand in relation to social and political contexts, provides also a reason for why a 'neuronalization' (i.e., a reduction of ethical concepts to descriptive facts) falls short. This is so because descriptive facts are, presumably, context-independent; whereas, normative convictions are context-dependent. Such a 'neuronalization' may then even have fatal implications, especially within applied questions of neuroethics, as they are inevitably shaped by their current social and political context. On a more positive note, these shortcomings suggest once more the need to

[24] For an illuminating theory of how personhood and personal identity are fundamentally based on dynamic interactions among biological, psychological, and social attributes and functions of a person's life that are mediated through a social and cultural infrastructure, see Marya Schechtman's 'person life view' [54]. A great merit of this view is that it coheres very well with recent evidence from social neuroscience and developmental psychology.

complement robust naturalistic forms of neuroethics, and, for that matter, all sorts of empirically informed ethics, with a thoroughly argued conceptual analysis that does justice to carefully situating ethical concepts within their relevant social and political contexts (as a source of norms).

We have argued that there are systematic methodological predicaments in robust naturalistic neuroethical endeavors. For this reason, the development of a methodological tool that enables the avoidance of fallacious inferences from facts to norms and vice versa is needed. One obvious answer to the above described challenge of the normative fallacy is to start neuroethical investigations without preexisting normative convictions, thereby keeping the ultimate result open (rather than result-closed arguments that derive from certain presupposed normative convictions right from the beginning). This helps to avoid biased interpretations of empirical data that are invoked, as it were, to justify these preexisting convictions 'postmortem'.

Instead, we propose that starting with what we call a *normative indeterminacy* leads to a result-open argument. A normative indeterminacy, of course, does not imply a normative indifference. On the contrary, such a starting point asks for a thorough assessment of the normative ethical convictions that are at stake. In other words, without having a certain normative ethical theory in mind (either in order to support or to debunk said theory), all sorts of neuroethical investigations should at a first stage remain uncommitted to any theory while considering a variety of possible candidates. In so doing, it is crucial, first, to explicitly reveal possible candidates, and second, to discuss the merits and demerits of these normative ethical convictions within the relevant social and political context. This methodological strategy, however, does not remain on the level of revealing merely the particular normative ethical convictions themselves—say, the consequentialists take on endorsing rational moral judgments—but also the implicit presuppositions of these normative ethical convictions that need to be uncovered and discussed. That is to say, the role particular normative ethical convictions play within their social and political context needs to be taken into consideration.

Western societies may, for example, be more driven by consequentialist convictions, whereas eastern societies may be more driven by deontological or virtue ethical convictions. Acknowledging this, then, might also call for further interdisciplinary collaborations with scholars from the social sciences. Such an approach leads to result-open arguments, because it is based on a context-sensitive normative pluralism. A commitment to discussing presuppositions of norms, as well as appreciating that facts have different normative implications depending on which norms are presupposed, allows for comparing and matching of facts and norms in a bilateral way. In following such a methodological strategy, these kinds of neuroethical arguments are more likely to remain open to what happens to result from bringing together ethical questions and empirical evidence. Not only are such neuroethical investigations likely to lead to unbiased results, but they can be reasonably anticipated to have a higher explanatory value since they would be based on a more fine-grained and context-sensitive methodology that captures better the complexities of moral reasoning.

More generally, in order to overcome the barrier between the normative and descriptive realms—in either direction, i.e., from norms to facts and from facts to

norms—some way to introduce a reciprocity of influence between the two realms is needed. We call this desideratum a *norm-fact linkage*.[25] Such a notion presupposes a vantage point from within the interface between the descriptive and the normative realm that allows for a development of a mutual linkage between the two. This is needed in order to arrive at theories that are both normatively and empirically plausible without giving precedence to either dimension of the neuroethical endeavor. In other words, a normatively plausible empirical foundation is needed to show that one is neither falsely inferring an 'ought' from an 'is', nor an 'is' from an 'ought'. How can such a norm-fact linkage that makes possible a close intertwining of normative concepts and neuroscientific observations get off the ground?

The aforementioned examples of moral judgments in trolley dilemmas point to the fact that descriptive change may entail normative change, and vice versa. Therefore, a methodological strategy needs to be sensitive to the close interdependency between norms and facts, while not neglecting their difference. One way is the just mentioned norm-fact linkage; consisting in going back and forth between normative concepts and neuroscientific findings. The usual starting point of empirical neuroethics is to scrutinize an ethical concept that is linked to neuroscientific observations. As said before, the aim is to either 'neuronalize' the ethical concept, or to nail down its philosophical content in order to showcase its relevance for neuroscientific research. One neglected aspect of first encounters between normative concepts and neuroscientific observations is the matter of what neuroscientific observations imply for the normative concept itself. What, for example, is implied for the concept of moral judgments if it is shown empirically to be driven by emotions rather than by reason? What does the fact of the involvement of emotions or intuitions in moral judgments imply for the norms inherent in said judgments? Does the linkage between norms and facts in moral judgments need to be conceptualized differently if emotions or intuitions, rather than reason, are predominant? This may lead to conceptual modifications in neuroethical theories depending on the neuroscientific findings; thus, the initial ethical concept becomes neuroethical in a way that is sensitive to the context-dependent norm-fact linkage.

An interdependent linkage between neuroscientific findings and normative concepts, however, needs to go a step further. What is needed is a method that details the different dynamics of influence between facts and norms at different points in the neuroethical pursuit of such kind, clarifying when and why the friction goes one way or the other. This sort of pursuit cries out for a much greater focus on methodology than is presently typical in the different approaches to neuroethics. However divergent these neuroethical pursuits may be, in any case, there is a need to be self-consciously agile, capable of forging and refining diverse linkages between norms and facts in which both levels are accountable to the other. By following such a method, one may be able to find normative concepts that cohere best with empirical evidence and, thereby, to further the understanding of how scientific observations and normative evaluations of human life are tied together.

[25] For further details, see [2].

Acknowledgments We owe special thanks to Pedro L. Chaves, Gabriele Contessa, Lucas Jurkovic, Daniel T. Kim, Gregory J. Walters, Katherine Wayne, Annemarie Wolff, and three anonymous referees for valuable feedback on earlier drafts of this article. For financial support, we are grateful to the Canadian Institutes of Health Research (CIHR), Michael Smith Chair in Neurosciences and Mental Health (EJLB-CIHR), Michael Smith Foundation, and the Hope of Depression Foundation (HDRF/ISAN) to Georg Northoff, and for a Postdoctoral Fellowship from the University of Ottawa and a research grant from Taipei Medical University-Shuang Ho Hospital, Brain and Consciousness Research Center to Nils-Frederic Wagner.

References

1. Knobe, J., and S. Nichols. 2007. An experimental philosophy manifesto. In *Experimental philosophy*, ed. J. Knobe and S. Nichols, 3–14. Oxford: Oxford University Press.
2. Northoff, G. 2009. What is neuroethics? Empirical and theoretical neuroethics. *Current Opinion in Psychiatry* 22(6): 565–569.
3. Roskies, A. 2002. Neuroethics for the new millenium. *Neuron* 35(1): 21–23.
4. Gazzaniga, M.S. 2005. Facts, fictions and the future of neuroethics. In *Neuroethics: defining the issues in theory, practice, and policy*, ed. J. Illes, 141–148. Oxford: Oxford University Press.
5. Northoff, G. 2006. Neuroscience of decision making and informed consent: An investigation in neuroethics. *Journal of Medical Ethics* 32(2): 70–73.
6. Reiner, P. 2011. The rise of neuroessentialism. In *Oxford handbook of neuroethics*, ed. J.I.B. Sahakian, 161–177. New York: Oxford University Press.
7. Schleidgen, S., M.C. Jungert, and R.H. Bauer. 2010. Mission: Impossible? On empirical-normative collaboration in ethical reasoning. *Ethical Theory and Moral Practice* 13(1): 59–73.
8. Libet, B.W. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8(4): 529–566.
9. Maibom, H.L. 2014. To treat a psychopath. *Theoretical Medicine and Bioethics* 35(1): 31–42.
10. Lipsman, N., and W. Glannon. 2013. Brain, mind and machine: What are the implications of deep brain stimulation for perceptions of personal identity, agency and free will? *Bioethics* 27(9): 465–470.
11. Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4): 814–834.
12. Haidt, J. 2007. The new synthesis in moral psychology. *Science* 316(5827): 998–1002.
13. Greene, J., and J. Haidt. 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6(12): 517–523.
14. Huebner, B., S. Dwyer, and M.D. Hauser. 2009. The role of emotion in moral psychology. *Trends in Cognitive Science* 13(1): 1–6.
15. Hauser, M.D. 2008. When your moral organ is right! *Think* 7(19): 17–21.
16. Berker, S. 2009. The normative insignificance of neuroscience. *Philosophy and Public Affairs* 37(4): 293–329.
17. Kitcher, P. 2014. Is a naturalized ethics possible? *Behaviour* 151(2–3): 245–260.
18. Kahane, G. 2013. The armchair and the trolley: An argument for experimental ethics. *Philosophical Studies* 162(2): 421–445.
19. Hume, D. 1978. *A treatise of human nature*. Oxford: Oxford University Press.
20. Moore, G.E. 1993. *Principia ethica*, ed. T. Baldwin. Cambridge: Cambridge University Press.
21. Tanner, J. 2006. The naturalistic fallacy. *Richmond Journal of Philosophy* 13. http://www.richmond-philosophy.net/rjp/rjp13_tanner.php. Accessed May 11, 2015.
22. Hudson, W.D. (ed.). 1969. *The is-ought question: a collection of papers on the central problem in moral philosophy*. London: Macmillan.
23. Searle, J.R. 1964. How to derive "ought" from "is". *Philosophical Review* 73(1): 43–58.
24. Wilson, D.S., E. Dietrich, and A.B. Clark. 2003. On the inappropriate use of the naturalistic fallacy in evolutionary psychology. *Biology and Philosophy* 18(5): 669–681.
25. Harman, O. 2012. Is the naturalistic fallacy dead (and if so, ought it be?). *Journal of the History of Biology* 45(3): 557–572.
26. Dodd, J., and S. Stern-Gillet. 1995. The is/ought gap, the fact/value distinction and the naturalistic fallacy. *Dialogue* 34(4): 727–746.

27. Walsh, F.M. 2008. The return of the naturalistic fallacy: A dialogue on human flourishing. *Heythrop Journal* 49(3): 370–387.
28. Baumrin, B.H. 1968. Is there a naturalistic fallacy? *American Philosophical Quarterly* 5(2): 79–89.
29. Landeweerd, L. 2004. Normative-descriptive and the naturalistic fallacy. *Global Bioethics* 17(1): 17–23.
30. Sinnott-Armstrong, W. 2008. *The neuroscience of morality: emotion, brain disorders, and development.* Vol. 3, *Moral philosophy*. Cambridge, MA: MIT Press.
31. Rescher, N. 1990. How wide is the gap between facts and values? *Philosophy and Phenomenological Research* 50: 297–319.
32. Frankena, W.K. 1939. The naturalistic fallacy. *Mind* 48(192): 464–477.
33. Stevenson, C.L. 1944. *Ethics and language*. New Haven: Yale University Press.
34. Hare, R.M. 1991 [1952]. *The Language of Morals*. Oxford: Oxford Clarendon Press.
35. Schlosser, M.E. 2014. The neuroscientific study of free will: A diagnosis of the controversy. *Synthese* 191(2): 245–262.
36. Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–2108.
37. Greene, J.D., L.E. Nystrom, A.D. Engell, J.M. Darley, and J.D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44(2): 389–400.
38. Greene, J.D., S.A. Morelli, K. Lowenberg, L.E. Nystrom, and J.D. Cohen. 2008. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107(3): 1144–1154.
39. Greene, J.D., F.A. Cushman, L.E. Stewart, K. Lowenberg, L.E. Nystrom, and J.D. Cohen. 2009. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* 111(3): 364–371.
40. Singer, P. 2005. Ethics and intuitions. *Journal of Ethics* 9(3–4): 331–352.
41. Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5: 5–15.
42. Thomson, J.J. 1985. The trolley problem. *Yale Law Journal* 94: 1395–1415.
43. Greene, J.D. 2008. The secret joke of Kant's soul. In *The neuroscience of morality: emotion, brain disorders, and development*, vol. 3 of *Moral psychology*, ed. W. Sinnott-Armstrong, 35–79. Cambridge, MA: MIT Press.
44. Greene, J.D. 2014. Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics* 124(4): 695–726.
45. Gigerenzer, G. 2008. Moral intuition = fast and frugal heuristics? In *The cognitive science of morality: intuition and diversity*, vol. 2 of *Moral psychology*, ed. W. Sinnott-Armstrong, 1–26. Cambridge, MA: MIT Press.
46. Gigerenzer, G. 2010. Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science* 2(3): 528–554.
47. Churchland, P.S. 2012. *Braintrust: What neuroscience tells us about morality*. Princeton: Princeton University Press.
48. Campbell, T.D. 1970. *The normative fallacy. Philosophical Quarterly* 20(81): 368–377.
49. Wittgenstein, L. 2009 [1953]. *Philosophical Investigations*. Ed. and trans. P.M.S. Hacker and J. Schulte. Oxford: Wiley.
50. Rudner, R. 1953. The scientist qua scientist makes value judgments. *Philosophy of Science* 20(1): 1–6.
51. Churchman, C.W. 1956. Science and decision making. *Philosophy of Science* 23(3): 247–249.
52. McMullin, E. 1982. Values in science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982(4): 3–28.
53. Douglas, H. 2007. Rejecting the ideal of value-free science. In *Value-free science? ideals and illusions*, ed. H. Kincaid, J. Dupr'E and A. Wylie, 120–141. Oxford: Oxford University Press.
54. Schechtman, M. 2014. *Staying alive: Personal identity, practical concerns, and the unity of a life*. Oxford: Oxford University Press.

Bioethical Inquiry (2017) 14:299–313
DOI 10.1007/s11673-017-9780-2

CRITICAL PERSPECTIVES

Discovering the Neural Nature of Moral Cognition? Empirical, Theoretical, and Practical Challenges in Bioethical Research with Electroencephalography (EEG)

Nils-Frederic Wagner · Pedro Chaves · Annemarie Wolff

Received: 13 January 2016 / Accepted: 27 November 2016 / Published online: 28 February 2017

Abstract In this article we critically review the neural mechanisms of moral cognition that have recently been studied via electroencephalography (EEG). Such studies promise to shed new light on traditional moral questions by helping us to understand how effective moral cognition is embodied in the brain. It has been argued that conflicting normative ethical theories require different cognitive features and can, accordingly, in a broadly conceived naturalistic attempt, be associated with different brain processes that are rooted in different brain networks and regions. This potentially morally relevant brain activity has been empirically investigated through EEG-based studies on moral cognition. From neuroscientific evidence gathered in these studies, a variety of normative conclusions have been drawn and bioethical applications have been suggested. We discuss methodological and theoretical merits and demerits of the attempt to use EEG techniques in a morally significant way, point to legal challenges and policy implications, indicate the potential to reveal biomarkers of psychopathological conditions, and consider issues that might inform future bioethical work.

Keywords Electroencephalography (EEG) · Brain imaging · Moral cognition · Normative ethics · Legal, practical and policy implications · Biomarkers

N.-F. Wagner (✉)
Department of Philosophy, University of Duisburg-Essen, Lotharstr 65, 47057 Duisburg, Germany
e-mail: nils-frederic.wagner@uni-due.de

P. Chaves
Faculty of Medicine, University of Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
e-mail: pedromlchaves@gmail.com

A. Wolff
Royal Ottawa Mental Health Centre, Mind, Brain Imaging and Neuroethics Research Unit, University of Ottawa Institute of Mental Health Research, 1145 Carling Avenue, Ottawa, ON K1Z 7K4, Canada
e-mail: awolff02@gmail.com

Introduction

The study of moral cognition via neuroscientific methods has recently become of great interest to moral psychologists, neuroscientists, bioethicists, and philosophers alike. Due to the rapid advance of modern brain-imaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), novel means to access moral cognition are available. These techniques promise to shed new light on traditional moral questions by helping us to understand how effective moral cognition is embodied in the brain. As Michael Gazzaniga puts it: "Cognitive neuroscience has valuable information to contribute to the discussion of certain topics that have traditionally been taken up by bioethicists, namely, those issues in which brain science has relevant knowledge that should impact the ethical questions being debated" (2005, 141). However, the exact way in which neuroscience impacts our understanding of moral cognition remains unclear.

One of the challenges in empirically studying moral cognition lies in the fact that different normative ethical

theories emphasize conflicting convictions of how to get things right. Of course, the question of what constitutes a moral judgement is itself in dispute. These normative ethical convictions can be roughly linked to different cognitive features and can, in a broadly conceived naturalistic attempt, be associated with different brain processes that are rooted in different networks and regions. Casebeer (2003), for example, suggests that different normative ethical theories can be linked to specific functions and networks in the brain. The role of particular cognitive features implied by different normative ethical theories (for example, the use of executive functions in goal-orientation), can be experimentally investigated. Now, if moral cognition is, for example, believed to be first and foremost a rational endeavour, being committed to "reason purely" about the demands of, say, Kant's categorical imperative, executive functioning stemming mostly from frontal brain regions is most pertinent. For advocates of utilitarianism such as Mill, on the other hand, a moral agent's most important ability is to recognize and compute utility functions; accordingly, an integration of pre-frontal, limbic, and sensory regions is essential to the manipulation of numerical values, as well as the coding of value itself. Yet others, such as virtue ethicists Plato and Aristotle, think of moral cognition roughly as the ability to reason well about which states of being would be most conducive to general human flourishing and to the individual good life in particular. This more deflationary process, then, focuses on the appropriate coordination of properly functioning cognitive sub-entities and can be seen as involving the whole brain.[1] One of the most strongly held contemporary neuroethical accounts is Neo-Humean or sentiment-based in nature, emphasizing the pivotal role of emotions and affects rather than reason and deliberation for moral cognition (Greene and Haidt 2002). Another, more inclusive, view takes moral cognition to be an evolved neural adaptation to social interactions, integrating reason and emotion. The moral brain is characterized as an integrative cognitive system, using a dynamic responsive equilibrium where reality and imagination shape human behaviour and experience (Gillett and Franz 2014).

When looking at such a mapping of moral cognition associated with different brain processes (cognitive or affective) which can be linked to certain brain regions, it becomes apparent how neuroimaging methods can yield evidence as to how measurable brain activity and moral cognition are related.

This linking of brain activity in certain regions to moral cognition suggests that neuroscientific evidence is not only expository as to how human beings de facto act but also reveals tenets of normative ethical theories. Doing bioethics in this way constitutes an attempt to draw conclusive relationships between neuroscientific observations and normative ethical theories—consequently, it is aimed at suggesting concrete prescriptions in bioethical debates. The general aim is to merge scientific descriptions and normative evaluations of human (and occasionally non-human) life. This branch of bioethics investigates both the significance of neuroscientific findings for the understanding of morality and the relevance of ethical considerations for determining the normative import of evidence from neuroscience. The former is concerned with neural and psychological mechanisms underlying ethical concepts and judgements, whereas the latter is concerned with the implications of those findings for moral practice.

In this article, we briefly review the neural mechanisms of moral cognition that have been studied via the paradigmatic imaging modality electroencephalography (EEG), discuss methodological merits and demerits of this technique, point to legal and practical applications, and consider issues that could potentially inform future experimental work. Our focus on EEG is designed to remedy the lack of an organized review of this technique (in contrast to work already done in fMRI: Mendez 2009; Pascual et al. 2013; Raine and Yang 2006) and attempts to mitigate the understandable bias, particularly when it comes to a non-specialized audience, against research using neuroimaging techniques.

Experimental Challenges: Using EEG Measures in the Study of Moral Decision-Making

While the most widely known neuroscientific studies of moral cognition make use of a neurological localization approach (Greene et al. 2001, 2004; Koenigs et al. 2007), work exploring the electrophysiological basis of moral cognition has found less exposure—both within academia and with the general public. Localization techniques such as fMRI attempt to determine which brain regions are active during a task by measuring changes in blood flow (Huettel et al. 2008); EEG measures the electrical currents continuously generated by

[1] See Casebeer (2003) for further details.

cortical layers, thus aiming at discovering when something happens in the brain. Although lacking the spatial resolution of fMRI, non-invasive techniques like EEG enable us to peek into sub-second processes, reflecting information processing related to stimulus presentation (Schomer and Lopes da Silva 2012). Moral cognition, involved in sub-second decision-making, is one such process. Studying moral cognition has thus led neuroscientists to take advantage of techniques with a high temporal resolution.

Event-Related Potentials and Moral Cognition

The relationship between deflections of the electrical field that are captured at the scalp and time-locked to a specific event (so called event-related potentials, or ERPs) and cognition has been known for over fifty years (Walter et al. 1964). Recently, there has been a surge of interest in studying ERPs related to moral cognition and decision-making in general. That is, both morally relevant and morally irrelevant decisions have been studied. Given its high temporal resolution, EEG allows for probing and describing the attentional, cognitive, and affective mechanisms immediately following or preceding the processing of, in the present case, a morally relevant stimulus. Furthermore, ERP techniques have increasingly gained importance as a means of studying consciousness itself or levels thereof. For example, EEG is used to determine the degree to which patients with so-called "disorders of consciousness" are in fact conscious, even in the absence of behavioural responses (Beukema et al. 2016). Considering that moral cognition is most likely dependent on conscious awareness—hence sharing fundamental mechanisms such as working memory or multisensory integration—it seems fit to use EEG to explore such types of processes. With increasing evidence of the importance of distributed semantic space (Huth et al. 2012) and oscillating activity with functional synchronization across the anterior and posterior cortices (Mogilner et al. 1993; Moll et al. 2005; Lenartowicz et al. 2016) in regards to awareness and attention, it seems likely that moral cognition would be similarly dependent upon some degree of binding and synchrony across these diverse areas of the brain, given its dependence on memory, (both working memory space and long-term stores) reasoning, and emotional state.[2]

[2] We are grateful to an anonymous reviewer for this point.

A typical ERP experimental setup involves the presentation of discrete (one at a time) and short (from milliseconds up to a few seconds) stimuli within a certain context. The stimuli to be processed are then presented several times in order to obtain a sufficiently high signal-to-noise ratio (Luck 2014). Given the noisy nature of the EEG signal, a significant number of samples is needed in order to obtain analysable data. In the case of moral decision-making experiments, the morally relevant stimulus, such as a picture or a word, is presented within a batch of other stimuli (a sequence of pictures pertaining to an action or a sentence depicting a behaviour) that give it a certain contextual relevance and that bear moral significance. The response to the morally laden stimulus is then subsequently analysed in order to infer the mechanisms subserving its processing. This allows examining the time-course of the decision-making processes and its different sub-stages. In the case of moral dilemmas, for example, it can be useful to disentangle whether decisions are based upon low-level features of the scenario (such as emotional salience) or reached through cognitive effort and referral to moral rules. This allows for drawing inferences or generating hypotheses about what kind of normative ethical theories cohere best with empirical evidence generated from recording EEG signals during a moral dilemma task.

Most of the ERP components have been thoroughly described and analysed. These can be roughly divided into early activity (50-200ms), such as the P50 (Nathan et al. 2007) or the N1-P2 complex (Näätänen and Picton 1987), and later activity, such as the P300 (Polich 2007), N400 (Kutas and Federmeier 2011), or the Late Positive Potential (LPP) (Cacioppo et al. 1994).

The former are usually, though not exclusively, associated with automatic processes and modulated by physical, low-level properties of the stimuli, while the latter are seen as a product of higher-order cognitive processes (Macnamara et al. 2009; Weinberg and Hajcak 2010). An illustration of some of the described components is sketched, for exemplification purposes, in Fig. 1. In this figure, the X-axis indicates the timing, measured in milliseconds, of each data point—typical recordings collect data anywhere between 200 and 1000 times per second. The recorded amplitudes are measured in microvolts and are indicated in the Y-axis. Different experimental manipulations impact different dimensions of the components (e.g., amplitude or timing).

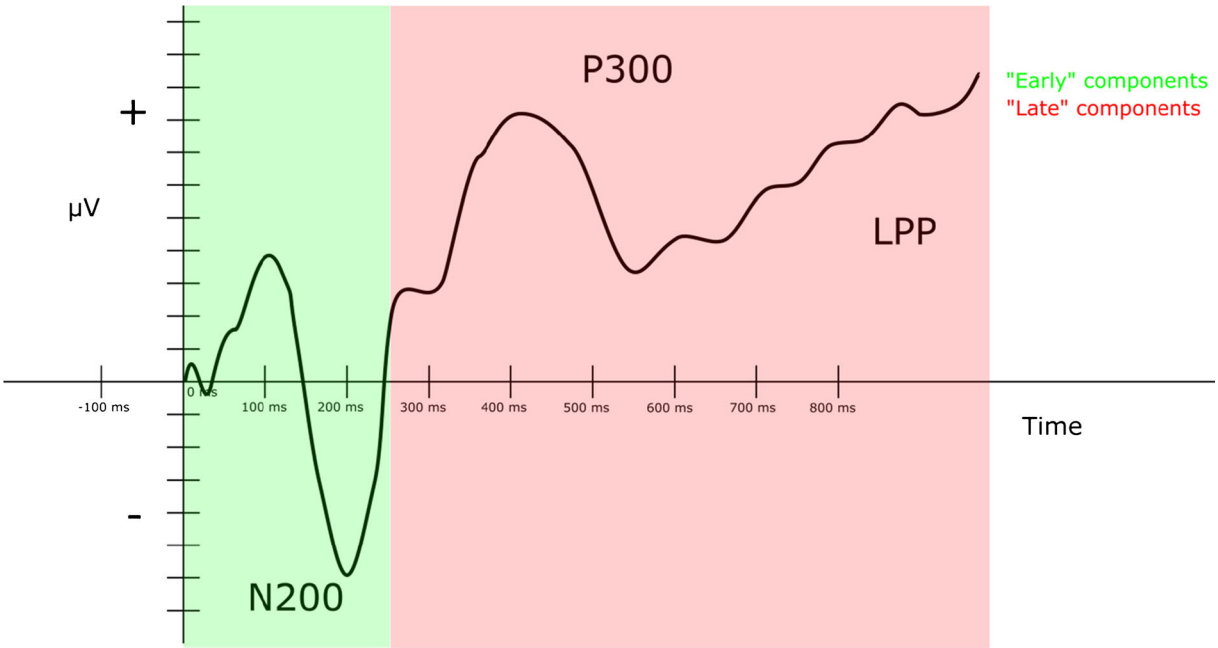

Fig. 1 Event-related Potentials (ERPs): Some of the Typical ERP Components Analysed in EEG Experiments

Early ERP Components

Several of these ERP components have been linked to, and studied within, the context of moral decision-making. It has been shown that the early component N200—a negative deflection occurring approximately 200 milliseconds after stimulus presentation—is larger for words reflecting a moral violation (independent of authority or rules) rather than a conventional (contingent on social constraints) violation (Lahat et al. 2012). This activity is usually considered to be a reflex of conflict monitoring and saliency processing—the automated orientation of perceptual and cognitive resources to sensory data (Botvinick et al. 2000; Harsay et al. 2012). At the same time, the N200 is also reported to differentially react to fair and unfair proposals in the ultimatum game: an economic reasoning experiment in which participants have to accept or reject the proposal of other participants regarding how to split a certain sum of money. Reactions to fair or unfair proposals have been interpreted as consistently reflecting moral attitudes. It has also recently been demonstrated that this early activity is modulated by prosocial and antisocial actions, pointing towards a rapid classification of the emotional saliency of the presented moral scenarios (Boksem and de Cremer 2010; Yoder and Decety 2014). The involvement of this very rapid differential processing of morally relevant stimuli has also been suggested by another research group (Sarlo et al. 2014), where the P260 component was shown to be sensitive to the affective distress felt during the decision-making process in "footbridge-style" moral dilemmas. This indicates that the electrophysiological activity is sensitive to the unpleasantness of the decision-making process. The same component seems to be sensitive to the consideration of legal consequences of moral behaviour, being larger in the case where there are no such consequences, indicating heightened affective conflict (Pletti et al. 2015). These authors have also shown how, in the condition where legal considerations matter, motor preparation is enhanced, as measured by the pre-response readiness potential ("Bereitschaftspotential"), a negative deflection occurring prior to pressing the button, reflecting readiness to act upon the scenario.

Although somewhat out of the scope of this article, it should be noted that the readiness potential has attracted major attention in the free-will debate, ever since Benjamin Libet first observed its importance for moral reasoning in the early 1980s. The fact that behaviour can be predicted from brain states preceding the perception of the conscious decision to act is crucial to the conceptualization of morality, given its dependency on intentionality and agency.

Later ERP Components

Despite the evidence for this early modulation of stimulus processing based on the stimulus characteristics and experimental conditions, the aforementioned and other studies have also shown how later stimulus processing, reflecting higher-level cognitive mechanisms, is involved in morally relevant scenarios. These include allocation of attentional resources to a particular stimulus, cognitive appraisal (i.e., evaluation of the stimulus following a set of rules), and re-appraisal (DeCicco et al. 2012; Shafir et al. 2015), or even the processing of semantic content and the detection of incongruences, that is, words that do not fit the overall meaning of the sentence.

In a way similar to the results found in the earlier components, the aforementioned LLP (a positive deflection occurring 300 milliseconds post-stimulus) seems to differentiate between prosocial and antisocial scenarios and is sensitive to the type of scenario presented, being modulated by perceived unpleasantness. Interestingly, cognitive empathy as well as guilt perception appears to modulate the amplitude of the LPP (Yoder and Decety 2014). Although this suggests that individual differences in empathic processing might underlie and modulate moral reasoning, further work presages that a disentanglement and clarification of the concept of empathy is needed in order to assess its specific role. More precisely, it seems that different facets of empathy (emotional and cognitive empathy) predict behaviour in different ways, producing conflicting data; therefore, generalizations should be taken with reservations (Decety and Cowell 2014).

These findings are consistent with previous studies demonstrating an N400 effect, typical of incongruence detection, to words inserted in objectionable statements in line with—or in conflict with—participants' reported values, followed as well by a modulation of the LPP (Van Berkum et al. 2009). The same overall pattern of modulation of both early and late onset components that is prevalent in the evaluation of words in moral transgression scenarios was again found in another recent study (Leuthold et al. 2015). In particular, the LPP was found to exhibit a higher amplitude for words considered to be morally unacceptable when compared to acceptable words in a given sentence. These findings replicate and extend an earlier work where prototypical social scenarios were used (Leuthold et al. 2012).

In sum, it seems clear that regardless of the stimulus type (linguistic or visual content), both early and late ERP components are modulated by different dimensions considered to be morally relevant. This seems to favour the idea that moral cognition implies both automatic and fast processing of presented stimuli as a means of rapidly categorizing them, as well as subsequent active appraisal and cognitive effort. This may present evidence in favour of viewing moral reasoning as a particularly complex type of processing with distinct and distributed spatio-temporal patterns, involving multiple regions (e.g., prefrontal, limbic, and parietal regions) and timescales (e.g., immediately and long after stimulus presentation).

Curiously, and following from the previous consideration, it is interesting that some particular components, such as the P300, have also been extensively used as a means to study consciousness itself, acting as a marker of conscious awareness under the "global workspace theory" and similar information-integration-based models (Baars et al. 2013; Dehaene and Changeux 2011; Edelman and Tononi 2000). More precisely, it has been suggested that this and later components are tightly linked to the binding problem, acting as a correlate of integration of information from different regions in the brain (Franklin et al. 2012). Using stimulus-locked paradigms such as the ones described above, it has also become possible to identify further potential mechanisms underlying information integration through the synchronization of firing patterns across neuronal populations. These firing patterns, although not strictly ERPs, reveal how oscillatory activity in several frequency bands can organize and support conscious events by supporting the propagation and integration of information across the brain (Steriade 2006).

Setbacks of EEG-Based Studies on Moral Cognition

From a methodological point of view, the use of EEG allows for a robust assessment of fast brain processes that affect moral cognition and decision-making. As such, and given its low cost and non-invasiveness, it would seem as if most experimental research ought to use EEG. However, it must be noted that despite its high temporal resolution, the precise localization of the activity itself (i.e., which particular neuronal populations are contributing to the signal) still suffers from some problems and limitations. Furthermore, as we have noted, most of this captured activity comes from cortical

matter and a particular subset of layers. This means that large sets of regions (most of the subcortical structures, for example) and their processes remain impenetrable to EEG, being only accessible through the use of other neuroimaging techniques such as fMRI. Ideally, then, results from EEG studies should be complemented with relevant fMRI investigations to account for the low spatial resolution of EEG. Nonetheless, as can be seen in some of the aforementioned studies on moral reasoning (e.g., Yoder and Decety 2014), some techniques such as LORETA allow for the partial reconstruction of cortical sources of the signal captured at the scalp level. When used carefully, this can contribute to more spatial information regarding the regions involved in a particular type of processing, linking it to other studies resorting to techniques such as fMRI. A further method of solving what has become known as the EEG inverse problem of source localization is the weighted minimum-norm method (Song et al. 2006) which, like LORETA, aims to determine spatially the source (group of neurons) which are the source of the specific EEG activity being studied. Together, these methods attenuate the drawbacks of EEG, namely its poor spatial resolution.

The nature of the EEG signal makes it a rather noisy source of information and thus demands strict experimental setup and data-processing procedures. As a safe and conservative solution, the selection of the experimental technique must take into account the kind of problem that is being tackled and follow a rigorous experimental procedure in order to avoid collecting futile or potentially misleading data.

Aside from the fundamental implications of disentangling the neural mechanisms underlying moral decision-making—due to the attempt of naturalizing it through empirical data collection and modelling—both more theoretical and practical questions arise. These questions will be addressed in what follows.

Theoretical Challenges: The Normative Significance of EEG Studies on Moral Cognition

The discussed research on the neurological basis of morality has led to vigorous conclusions regarding, for example, the act of killing (Greene 2003) or punishment (Sunstein and Vermeule 2005). Much, however, has yet to be clarified both in terms of methodology and empirical data interpretation (Levy 2007). Moreover, the theoretical framework upon which such research is conducted needs further development.

Plea for More Theoretical Work

More theoretical work is needed in order to show how normative ethical theories relate to actual and hypothetical human behaviour and what role neuroscientific evidence can play in such an endeavour. For example, in the previously mentioned debate between deontology and utilitarianism, it is far from clear whether consulting hypothetical moral dilemmas in general, and the neuroscientific study of participant's reactions to such dilemmas in particular, can yield relevant conclusions as to what kind of normative ethical judgement actually underlies the decision-making processes in question (Kahane 2015), let alone determine which normative ethical theory ought to be favoured. Kahane suggests that when participants deliberate about, say, the infamous trolley dilemma, they are in fact not deciding between opposing utilitarian and deontological solutions, but they are engaging in a richer process of weighing opposing moral reasons. These may point to virtue ethical theories, though this is not Kahane's interpretation. Whether this broader process of moral reasoning can be tracked down by EEG experiments remains elusive.

Drawing Normative Conclusions from Descriptive Claims

A principle difficulty that arises from the interpretation of EEG data—likewise when interpreting data obtained from other neuroimaging techniques—is the frequently assumed objectivity that comes with visualizing brain activity. It is tempting to assume that statistical analysis of brain wave recordings allows the researcher, as it were, a direct glimpse into participants' minds. But the relationship between subjective mental states and electromagnetic signals is far from straightforward (Poldrack 2006).

The majority of neuroscientists lean towards a reductive view of the human mind, presuming that more-or-less direct inferences can be drawn from neuroscientific observations to subjective mental states.[3] On this robust

[3] This is by no means to assert that all neuroscientists are reductionists, or that neuroscience is ipso facto committed to reductionism. We merely claim that the prevalent view in neuroscience about the nature of the human mind is naturalistic.

naturalistic view, mental states are reducible to brain states and, as such, are susceptible to empirical investigation. It is thereby sometimes overlooked, however, that imaging studies are based on probabilistic covariances, not on causal relations, making direct inferences contentious—a problem that has been duly noted, methodologically speaking, by the neuroimaging community itself. This methodological predicament parallels and somewhat overlaps with what has been called the "reverse inference problem" (ibid.): an inductive method that entails extrapolating backwards from the observed brain activity to particular cognitive processes which are, however, not directly tested. Furthermore, the interpretation of EEG data depends on theoretical assumptions of the researcher and inevitable idiosyncrasies of the study design; results may also be subject to social and cultural contingencies.

From these difficulties in interpreting EEG data, methodological predicaments arise in making strong cases for supporting or debunking normative ethical theories that lean heavily on the empirical results of such neuroimaging studies. In light of this, it is important to be aware of the following two considerations when confronted with naturalistic arguments that attempt to draw normative conclusions from descriptive claims.

- Are the descriptive claims correct? This involves asking whether the experimental designs of neuroscientific studies are actually significant and thus able to capture what they aim to investigate.[4]
- Do the normative conclusions really follow? This involves asking whether the interpretation of the empirical data, given that the design is adequate, is sound.

Epistemic Values in Scientific Research

Pertinent to the aforementioned methodological predicaments in neuroscientific research on moral cognition, there is a substantial body of thought coming from philosophy of science that addresses the complex ways in which scientific practices, and the products of science, are interwoven with values. Since the 1950s and 1960s, it has been argued that science is inevitably to some extent governed by value judgements (Rudner 1953). For one, the application of scientific methods is value-laden. In the quest for empirical discoveries, methods are restrained according to (often implicitly presupposed) normative ethical convictions. For example, invasive or potentially harmful experiments on healthy human participants are disallowed even at the expense of potentially impeding the progress of finding a cure for cancer. Accordingly, it has been argued that besides logic and evidence, science is in need of additional guidance for theory choice (Churchman 1956). In order to account for this, the term "epistemic values" was introduced to encompass the values that were seen as acceptable in guiding scientific research and theory building (McMullin 1982).

More recently, Douglas has opted to abandon the ideal of value-free science altogether, particularly if value-free science is meant to include the rejection of epistemic values. We agree that there need not be (and perhaps cannot be) a science that is freed of epistemic values, and we concur that it would be good for science to allow for "more open discussion of the factors that enter into scientific judgements and the experimental process" (Douglas 2007, 121). Douglas acknowledges the methodological predicament of the naturalistic fallacy, alongside the difference between descriptive and normative statements, but she contends that:

> [t]his does not mean … that a descriptive statement is free from values in its origins. Value judgments are needed to determine whether a descriptive label is accurate enough and whether the errors that could arise from the description call for more careful accounts or a shift in descriptive language. Evidence and values are different things, but they become inextricably intermixed in our accounts of the world (ibid., 126).

So, in order to make sense of empirical discoveries, we are dependent on scientific interpretations of the acquired data, and these interpretations inevitably have a normative dimension.

Context-sensitivity

As we have argued, there is a peculiar relation between values and facts in empirically driven ethical research.

[4] Several philosophers have taken issue with Libet's insinuation that neuroscience suggests free will to be an illusion. Even though Libet himself remained moderate regarding the evidence against free will gathered from his studies, some of his successors have made much stronger claims, positing that neuroscience has shown that free will is but an illusory trick that the brain plays on us.

This calls for paying close attention to the contingencies governing the de facto norms and social structures of everyday life, since these are at issue in empirically driven research on moral cognition (Wagner and Northoff 2015). Any scientific endeavour inevitably presupposes certain epistemological and metaphysical commitments because agents are shaped by a particular context—perceiving and interpreting the world around themselves in a great many different ways. In trying to determine the normative significance of neuroscientific evidence, the relatively austere individual and idiosyncratic social points of view cannot be altogether disregarded. In other words, the social, cultural, and political contexts in which ethical questions are posed and empirically driven answers are proposed need to be taken into account.

When aiming at drawing normatively significant conclusions from neuroscientific data, what we call a "content–context relationship" needs to be considered. Merely looking at EEG data in a vacuum, collected during a moral cognition task (thus focusing on the content), as investigated in the neuroscience of ethics, does not reveal how these empirical observations stand in relation to particular moral values (the relevant context), as discussed in moral philosophy. For example, observing an N200 effect tells us something about the cognitive components implicated in a particular moral judgement, such as the automated orientation of perceptual and cognitive resources to sensory data. Hence, the EEG data are considered and interpreted within a neuroscientific context. However, these empirical interpretations do not by themselves reveal anything about the normative significance of the data when weighing, say, utilitarianism and deontology. Here the gap needs to be bridged between the respective content that arises when inferring from the purely descriptive level of cognitive functions to the normative context within which the former are set and occur. To reduce normative ethical theories to neural and cognitive functions would thus be to absorb the normative realm within the descriptive level and ultimately entails a reduction of context to content.

Most of the discussed EEG-based studies on moral cognition focus exclusively on particular stimuli that bear moral significance, thereby, however, disregarding larger questions such as: in which kind of social and cultural setting are the stimuli presented, and how do they stand in relation to generally accepted sociocultural norms in a given setting? Such omission calls for paying closer attention to context-sensitivity as one criterion for the possibility of inferring normative conclusion from neuroscientific evidence; this has to be accounted for both experimentally and conceptually.

Furthermore, a great proportion of the experimental conditions in EEG studies focuses on intra-individual differences. They examine the variability between participants in their responses to moral dilemmas. This bias can somewhat be explained by the fact that repeated-measures designs (i.e., designs where the same subjects are stimulated with different conditions, reflecting, for example, consequentialist or deontological approaches) are experimentally more economical and easy to set up compared to independent-sample designs (i.e., designs targeting inter-individual differences between distinct groups, such as when comparing healthy controls and psychopaths). Nonetheless, the latter are still present in the literature and make important contributions that remain undetected in intra-individual experiments. For example, a recent study found that, contrary to the prevailing view that people who endorse non-utilitarian judgements to moral dilemmas are committing an error, there is evidence suggesting that participants who respond in a utilitarian fashion possess personality traits that are widely seen to be highly immoral. Participants responded to a battery of personality assessments and a set of dilemmas that pit utilitarian and non-utilitarian options against each other. Interestingly, those participants who indicated greater endorsement of utilitarian solutions also had higher scores on measures of psychopathy, Machiavellianism, and life meaninglessness (Bartels and Pizzaro 2011). The authors rightly suggest that these results indicate a need to be methodologically wary of favouring an experimental structure that equates the quality of moral judgements with responses that are endorsed primarily by individuals who are likely perceived as less moral as they possess traits such as callousness and manipulativeness. Adopting such a method can lead to the counterintuitive inference that "correct" moral judgements are most likely to be made by individuals that are simultaneously most likely to possess character traits generally perceived as immoral.

The aforementioned shortcomings of EEG studies on moral cognition suggest once more the need to complement robust naturalistic forms of empirically driven ethical research with a thoroughly argued conceptual-normative analysis that does justice to carefully situating ethical concepts within their relevant social, cultural, and political contexts.

Keeping in mind the previously discussed merits and demerits of EEG-based studies on moral cognition, we reach the following broad intermediate results. (1) Exploring the neural underpinnings of moral cognition is justifiably believed to lead to empirically-informed, more sophisticated moral theories. In that regard, the previously discussed experimental literature employing EEG as a marker for moral cognition can shed new light on traditional issues in normative ethics. (2) The overall goal to naturalize ethics in this way is to show how our moral practices and the underling moral theories are features based on the complexity of the human brain and can as such be scrutinized scientifically. (3) Crucial in this regard is that experiments are designed so as to take into account the aforementioned context-sensitive nature of moral cognition, as well as to maximize ecological validity—a difficulty that consists in the nature of experiments conducted under artificial lab situation that bear little resemblance to real-life moral decision-making. This can be alleviated by designing experiments as realistically as possible and by implementing incentives that have real-life significance, as is practiced in economic decision-making experiments. (4) The identification of neural correlates of moral cognition is partly dependent on—certainly connected to—having a moral theory in place when searching for its neural underpinnings. This has the advantage of enabling the researcher to eliminate certain moral theories as neurobiologically unrealistic. If, for example, a theory requires a high demand of reason and deliberation, but experiments consistently show that morally praiseworthy decisions are reached via fast heuristics, a reason-based theory would appear to be at odds with the evidence. (5) Particular cognitive features like executive functions, (e.g., goal-orientation), emphasized by different normative ethical theories, are investigated experimentally in order to reveal tenets of these different theories. If, as deontological theories assume, moral cognition is governed by "reason purely" about the demands of, say, Kant's categorical imperative, executive functioning stemming mostly from frontal brain regions is most pertinent. According to utilitarianism as proposed by Mill, a moral agent's most important ability is to recognize and compute utility functions; accordingly, an integration of prefrontal, limbic, and sensory regions is essential to the manipulation of numerical values and to the coding of value itself. Virtue ethics theories going back to Aristotle think of moral cognition roughly as the ability to reason well about what states of being would be most conducive to human flourishing. Moral concerns thus relate to what we ought to do and think so as to function well as human beings, involving an appropriate coordination of properly functioning cognitive sub-entities that are distributed throughout the whole brain. (6) When weighing these theories, it is important not to fall prey to a naive naturalistic reductionism or to smuggle in an agenda by trying to identify the neural correlates of exactly that form of moral cognition that best fits the theory that was assumed to be most plausible in the first place. So, additional argumentative work is needed in order to offer independent reasons as to which theory is conceptually most persuasive. (7) The discussed neuroscientific evidence thus far suggests that moral cognition is a process widely distributed throughout the entire brain and not restricted to either reason or emotion. To a first approximation, then, when seeking a theory that coheres best with current evidence, we do well to cast our net widely enough so as to include all aspects of moral reasoning.

Practical Implications

Apart from the impact that EEG studies have on normative ethical theory, what might be the role that evidence from such studies can play regarding moral questions that figure relevant in clinical or legal settings? Statistically speaking, inter-individual differences in basic information processing or brain functioning, such as sensory processing and integration (Stevenson et al. 2012), for example, can account for a large amount of variance in moral decision-making which falls outside the norm. This, in turn, could theoretically and practically inform decisions on legal transgressions and mental health.

Legal Challenges

Historically, there have been attempts to introduce lie-detection techniques based on both physiological measures and neuroimaging evidence into the court room. The former has precedents of being accepted legally, although seldom and restrictedly, but the latter has encountered vivid opposition—for reasons which seem obvious. While the philosophical definition of lying is far from settled (Kagan 1998), its legal role is rather clear-cut and its reliable detection something that the

justice system would immensely benefit from. However, the possibility of false positives due to an over-reliance on technology is unappealing. In a similar way, as seen in the recent hype of psychopathy research, demonstrating that a particular suspect has brain lesions or malformations that correspond to impaired emotional processing or inhibitory mechanisms places a huge burden on the demonstration of imputability (Hughes 2010), or lack thereof as the case may be.

As it is generally the case with findings of this nature, ethical considerations surrounding the implications for responsibility and privacy arise. To begin with, issues of criminal responsibility and free will, specifically in the arena of the criminal court, have been challenged by recent neuroscientific findings on moral decision-making (Aharoni et al. 2011; Koenigs et al. 2012). This can be illustrated with the treatment of psychopathy and criminal responsibility. Psychopaths have been characterized as persons with a strong lack of empathy or guilt, shallow affect, manipulative behaviour with superficial charm, delusions of grandeur, a parasitic use of others, early onset of antisocial behaviour, and other similar traits (Hare 2003). Recent imaging studies have associated the ventromedial prefrontal cortex with interpersonal deficits as well as an antisocial lifestyle and actions, while the mirror neuron network has been linked to interpersonal deficits and antisocial lifestyle (Contreras-Rodriguez et al. 2014; Li et al. 2014; Motzkin et al. 2011). In addition, EEG studies have associated forebrain circuit dysfunction with psychopathy (Cummings 2015); whereas fMRI studies have shown amygdala dysfunction in psychopaths (Thompson et al. 2014). These studies show dysfunction in anatomical and physiological components in the psychopathic brain. The practical implication, however, is the issue of how these empirical findings bear on responsibility in the legal context. Many support the use of the evidence in the courtroom to negate, or at least mitigate, the psychopaths' *mens rea* with respect to crime. The opposite view, affirming that despite clear abnormalities in the brains of these offenders, a reliable step from neuroscientific findings to criminal responsibility has not yet been made, has many supporters. The debate continues.

The general implications of these empirical findings on criminal and moral responsibility and free will are disputed. In one view, a more careful, logical link from neuroscientific findings to conclusions about responsibility is needed, without, as Stephen Morse states, falling prey to a "brain overclaim syndrome" in which unsustained claims about neuroscientific implications are made. "Brains do not commit crimes;" Morse claims, "people commit crimes" (Morse 2005). One possible response to this claim states that although people commit crimes, it is their brain that determines their behaviour, and so an abnormal brain produces, as it were, abnormal behaviour. To separate the brain from behaviour would imply a second seat of criminal responsibility, positing the existence of an immaterial mind of sorts. According to most neuroscientists and philosophers alike, this dualistic view is false; some even claim that science has proven it to be so (Martell 2009; Kendler 2005).

Policy Issues

A further ethical problem with implications for policymakers arises from data privacy and security. Regulations are needed as to who will have access to generated data and who will be granted access to collect such data, and for what purpose. As has been shown in cases of psychopathy (Aharoni et al. 2011; Koenigs et al. 2012) and alcohol intoxication (Duke and Bègue 2015), certain groups produce more consequentialist responses than controls. These studies have induced much discussion on underlying reasons for these differences: is the more consequentialist response pattern due to an increased reasoning capacity or due to a decreased emotional capacity? The prevailing opinion relates to impaired emotional activity. From this hypothesis, these specific populations are being branded with an impairment which may have practical consequences in their everyday lives. For example, would a potential employer be within their rights to require an applicant to undergo an EEG moral dilemma study, similar to the ones mentioned above, in the hope of determining psychopathic brain patterns? Will these neuroscientific findings relieve a person with a high *Psychopathy Checklist-Revised* score from criminal responsibility, while also burdening them with decreased personal and professional opportunities? Along with the possible implications of these neuroscientific findings described above, how will certain people be differentiated from the majority, and how will possible issues of discrimination be countered?

Related to the point of impaired criminal responsibility, this method of research applied to moral dilemmas may pose risks to the general decision-making capacity

of patients and at-risk populations. Recently, several studies have provided evidence to support the hypothesis that intact emotional activity in the brain is necessary for decision-making in daily life (Bechara 2004; Chang and Sanfey 2008; Heilman et al. 2010). These studies show that intact emotional processing is important for decision-making in financial matters and in social settings, but also on the appraisal of the consequences of a decision at the level of a "gut reaction" (Chang and Sanfey 2008). Since these neuroscientific studies of moral dilemmas expose processing in the brain of emotions related to these dilemmas, and also since emotional processing has been shown to be vital to everyday decision-making, the potential for the exploitation of individuals through its use exists. For example, it is possible that this experimental paradigm could be used to justify the classification of an individual, such as an elderly parent, as incompetent regarding financial decisions by their greedy adult child. Such cases of strategizing to seize control of an estate for one's one ends have been commonplace. However, the ability of moral dilemma experiments to expose deficits in emotional activity and its consequences creates a new tool in the exploitation of vulnerable populations, as mentioned. For this reason, as well as the ones previously stated regarding criminal responsibility, close attention must be paid towards the implications of the neuroscientific findings of moral dilemmas on personal responsibility and autonomy.

Biomarkers

EEG studies also have the potential to discover biomarkers of mental illnesses by empirically measuring biological processes that can be quantified in a precise and reproducible fashion (Leiser et al. 2011). These indicators have been used to diagnose diseases and to predict clinical outcomes of treatments (Kuhlmann and Wensing 2006). Neuroimaging biomarkers have, for the most part, focused on identifying neural functions associated with psychopathology. FMRI and positron emission tomography (PET) have been used to identify biomarkers of psychopathology, though the interest in and capacity for EEG to do the same have recently soared. The reasons for this are as follows. To begin with, recent advances in software and computer systems for the processing of EEG signals and their visualization have improved the spatial information drawn from this continuous signal, which was always considered the weakness of EEG. With these ongoing advances in its supporting technology (most significantly, the aforementioned LORETA), EEG is now able to provide data with both high temporal and somewhat improved spatial resolution. EEG becomes increasingly more effective in detecting where in the cortex the activity on the scalp comes from, which enables it to identify fast changes in brain activity. Nevertheless, there is still a certain degree of spatial constraint (McLoughlin et al. 2014). Secondly, as EEG is a much lower-cost method than fMRI—costs per participant are approximately $45 versus $650 for fMRI—larger sample sizes are possible, which are required for identifying biomarkers. Finally, EEG is the most portable and non-invasive of all the neuroimaging modalities and allows for participants with metal in their body, which fMRI, for example, does not. Also, with new developments in EEG sensors and systems, the feasibility of testing individuals previously considered difficult, such as children or infants, has improved. Therefore, due to low cost, portability, and advances in EEG equipment and computational software, the use of EEG for discovering biomarkers of psychopathology has become promising for the future diagnosis and treatment of neurological and psychiatric conditions.

The above-described challenges (summarized in Table 1) are of an interdisciplinary nature and strongly suggest that policymakers, healthcare providers and the legal system need not only be well-informed about the empirical foundations and limitations of brain-imaging techniques such as EEG, but are also in need of consulting complementary work from bioethicists that enables an instructive assessment of these techniques from an ethical point of view.

Conclusion

In this article, we pointed to EEG as an informative neuroimaging technique bearing relevance to bioethical research. EEG allows for probing and describing the cognitive, affective, and attentional mechanisms immediately following or preceding the processing of morally relevant stimuli. When compared to other neuroimaging techniques (such as fMRI or PET), EEG has methodological and practical advantages; it is mostly noninvasive, has superb temporal resolution and lower costs, and it is relatively easily accessible to participants and patients alike. However, despite EEG's temporal precision, there are limitations in its spatial resolution, leading to problems with precisely localizing the measured

Table 1 Empirical, Theoretical, and Practical Challenges in EEG-Based Studies on Moral Cognition

Level	Challenges
Empirical	Noisy EEG signal
	Lack of precise localization of brain activity
	Reverse inference problem
	Inter-individual differences
	False positives
Theoretical	Link between ethical theory and human behavior
	Drawing normative conclusions from descriptive claims
	Epistemic values in research
	Context-sensitivity
Practical	Lie detection
	Criminal responsibility
	Data privacy and security
	Discrimination against minority groups
	Biomarkers

brain activity. Since most of the captured brain activity comes from a very specific set of neuronal populations, large groups of regions and their processes remain inaccessible to EEG; these regions can only be mapped through the use of other techniques such as fMRI. Furthermore, the EEG signal in itself is inherently noisy, demanding a strict experimental setup and careful data-processing procedures. The selection of the experimental technique must also take into account the kind of problem the researcher is tackling and follow a thorough experimental procedure in order to avoid unsustainable conclusions. When these measures are taken, informative evidence can be gathered from EEG studies on moral cognition that might be able to open up new perspectives that can contribute to successfully shedding new light on bioethical problems.

However, an inherent predicament when relying on neuroscientific evidence in studying moral cognition is the contentious way in which norms and facts are related. EEG studies are based on probabilistic covariances, not on causal connections, making direct inferences from objective brain activity to subjective mental states problematic. In trying to determine the normative significance and import of evidence gathered from EEG studies, the relatively austere individual and idiosyncratic social and cultural points of view must be taken into account. Therefore, the social and cultural contexts in which ethical questions are posed and empirically driven answers are proposed must theoretically be taken into consideration and practically accounted for when designing experimental paradigms.

In light of the discussed challenges, it seems obvious that current neuroscientific findings are not conclusive enough to allow for a decisive verdict as to which moral theory coheres best with how the brain works. Nevertheless, there is considerable evidence suggesting that an empirically-informed Aristotelian virtue theory coheres best with what is so far known about brain functions during moral cognition. Virtue ethics theories are widely believed to have the richest moral psychology of the major moral theories, as reason, deliberation, emotion, and affect are all integral to living good and virtuous lives. This fits well with the discussed EEG studies indicating that moral cognition is a large-scale distributed brain process that cannot be reduced to certain areas or functions. Furthermore, it currently appears that the most ecologically valid experimental designs support the hypothesis that moral cognition is a large-scale brain affair depending on the appropriate coordination of many areas. Casebeer and Churchland (2003) point to research at a range of levels of organization from synapses to neurons to brain areas and systems, indicating that the organism which best triangulates norms will be one that uses (1) multi-modal signals (2) conjoined with appropriately cued executive systems that (3) share rich connections with affective and conative brain structures (4) which draw upon conditioned memories (5) and insight into the minds of others so as to (6) think about and actually behave in a manner enabling it to function as best it can. They emphasize that these capacities do have neural correlates, but generally such correlates will be multifaceted high-order functional relationships distributed throughout the brain. The discussed evidence also indicates that, contrary to deontology, there is most likely no such thing as a "pure reason" capacity and that emotion is integral to moral cognition. Contrary to utilitarianism, it appears that utility calculations by themselves are insufficient for moral cognition.

Tackling bioethical issues with the help of consulting evidence from EEG-based studies on moral cognition can be a valuable complementary source for normative ethical theory and bears the potential of having profound implications on legal, health care, and policy issues. These include determining criminal responsibility, the potential discovery of biomarkers for psychopathological conditions, and concerns on data privacy. To account

for the methodological and theoretical complexity of studies with such wide-ranging conclusions, there is a strong need of further interdisciplinary research in which bioethicists can play a pivotal role.

References

Aharoni E., O. Antonenko, and K.A. Kiehl. 2011. Disparities in the moral intuitions of criminal offenders: The role of psychopathy. *Journal of Research in Personality* 45(3): 322–327.

Baars, B., S. Franklin, and T. Ramsoy. 2013. Global workspace dynamics: Cortical "binding and propagation" enables conscious contents. Frontiers in Psychology 4: 200.

Bartels, D., and D. Pizzaro. 2011. The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. Cognition 121(1): 154–161.

Bechara, A. 2004. The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. Brain Cognition 55(1): 30–40.

Beukema, S., L.E. Gonzalez-Lara, P. Finoia, et al. 2016. A hierarchy of event-related potential markers of auditory processing in disorders of consciousness. NeuroImage: Clinical 12: 359–371.

Boksem M.A., and D. De Cremer. 2010. Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. Social Neuroscience 5(1): 118–128.

Botvinick, M., J.D. Cohen, and C.S. Carter. 2000. Conflict monitoring and anterior cingulate cortex: An update. Trends in Cognitive Sciences 8(12): 539–546.

Cacioppo, J., S.L. Crites, and W. Gardner. 1994. Attitudes to the right: Evaluative processing is associated with lateralized late positive event-related brain potentials. Personality and Social Psychology Bulletin 22(12): 1205–1219.

Casebeer, W.D. 2003. Moral cognition and its neural constituents. Nature Reviews Neuroscience 4(10): 840–847.

Casebeer, W.D., and P. Churchland. 2003. The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. Biology and Philosophy 18(1): 169–194.

Chang L.J., and A.G. Sanfey. 2008. Emotion, decision-making and the brain. In *Neuroeconomics (Advances in Health Economics and Health Services Research, Volume 20)* edited by D. Houser and K. McCabe, 31–53. Emerald Group Publishing.

Churchman, C.W. 1956. Science and decision making. Philosophy of Science 23(3): 247–249.

Contreras-Rodriguez O., J. Pujol, I. Batalla, et al. 2014. Functional connectivity bias in the prefrontal cortex of psychopaths. Biological Psychiatry 78(9): 647–655.

Cummings, M.A. 2015. The neurobiology of psychopathy: Recent developments and new directions in research and treatment. CNS Spectrums 20(3): 200–206.

Decety, J., and J.M. Cowell. 2014. Friends or foes: Is empathy necessary for moral behavior? Perspectives on Psychological Science 9(4): 525–537.

DeCicco, J., B. Solomon, and T. Dennis. 2012. Neural correlates of cognitive reappraisal in children: An ERP study. Developmental Cognitive Neuroscience 2(1): 79–80.

Dehaene S., and Changeux J.-P. 2011. Experimental and theoretical approaches to conscious processing. Neuron 70(2): 200–227.

Douglas, H. 2007. Rejecting the ideal of value-free science. In Value-free science? Ideals and illusions, edited by H. Kincaid, J. Dupré, and A. Wylie, 120–141. Oxford: Oxford University Press.

Duke, A., and L. Bègue 2015. The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. Cognition 134: 121–127.

Edelman G.M., and G. Tononi. 2000. A universe of consciousness: How matter becomes imagination. New York: Basic Books Inc.

Franklin S., S. Strain, J. Snaider, R. McCall, and U. Faghihi. 2012. Global workspace theory, its LIDA model and the underlying neuroscience. Biologically Inspired Cognitive Architectures 1: 32–43.

Gazzaniga, M.S. 2005. Facts, fictions and the future of neuroethics. In Neuroethics: Defining the issues in theory, practice, and policy, edited by J. Illes, 141–148. Oxford: Oxford University Press.

Gillett, G., and E. Franz. 2014. Evolutionary neurology, responsive equilibrium, and the moral brain. Consciousness and Cognition. 45: 245–250.

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

Greene, J. 2003. From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? Nature Reviews Neuroscience 4(10): 847–850.

Greene, J.D., L.E. Nystrom, A.D. Engell, J.M. Darley, and J.D Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. Neuron 44(2): 389–400.

Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. Science 293(5537): 2105–2108.

Hare, R.D. 2003. The psychopathy checklist–Revised, 2nd ed. Toronto: Multi-Health Systems.

Harsay, H.A., M. Spaan, J.G. Wijnen, and K.R. Ridderinkhof. 2012. Error awareness and salience processing in the oddball task: shared neural mechanisms. Frontiers in Human Neuroscience 6: 246.

Heilman R.M., L.G. Crişan, D. Houser, M. Miclea, and A.C. Miu. 2010. Emotion regulation and decision making under risk and uncertainty. *Emotion* 10(2): 257–265.

Huettel, S., A. Song, G. McCarthy. 2008. Functional magnetic resonance imaging. Sunderland, MA: Sinauer Associates, Inc.

Hughes, V. 2010. Science in court: Head case. Nature 464(7287): 340–342.

Huth, A., A. Nishimoto, and J. Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76(6): 1210–1224.

Kagan, S. 1998. Normative Ethics. Oxford: Westview Press.

Kahane, G. 2015. Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. Social Neuroscience 10(5): 1–10.

Kendler, K.S. 2005. Toward a philosophical structure for psychiatry. American Journal of Psychiatry 162(3): 433–440.

Koenigs, M., M. Kruepke, J. Zeier, and J.P. Newman. 2012. Utilitarian moral judgment in psychopathy. Social Cognitive and Affective Neuroscience 7(6): 708–714.

Koenigs, M., L. Young, R. Adolphs, et al. 2007. Damage to prefrontal cortex increases utilitarian moral judgments. Nature 446(7138): 908–911.

Kuhlmann J, Wensing G. 2006. The applications of biomarkers in early clinical drug development to improve decision-making processes. Current Clinical Pharmacology 1(2): 185–191.

Kutas, M., and K. Federmeier. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology 62: 621–647.

Lahat, A., C. Helwig, and P. Zelazo. 2012. An event-related potential study of adolescents' and young adults' judgments of moral and social conventional violations. Child Development 84(3): 955–969.

Lenartowicz, A., S. Lu, C. Rodriguez, et al. 2016. Alpha desynchronization and frontoparietal connectivity during spatial working memory encoding deficits in ADHD: A simultaneous EEGfMRI study. NeuroImage: Clinical 11: 210–223.

Leiser S., J. Dunlop, M. Bowlby, and D. Devilbiss 2011. Aligning strategies for using EEG as a surrogate biomarker: A review of preclinical and clinical research. Biochemical Pharmacology 81(12): 1408–1421.

Leuthold H., R. Filik, K. Murphy, and I.G. Mackenzie. 2012. The on-line processing of socio-emotional information in prototypical scenarios: inferences from brain potentials. Social Cognitive and Affective Neuroscience 7(4): 457–466.

Leuthold, L., A. Kunkel, I.G. Mackenzie, and R. Filik. 2015. Online processing of moral transgressions: ERP evidence for spontaneous evaluation. Social Cognitive and Affective Neuroscience 10(8): 1021–1029.

Levy, N. 2007. Neuroethics: Challenges for the 21st century. Cambridge, MA: Cambridge University Press.

Li W., X. Mai, and C. Liu. 2014. The default mode network and social understanding of others: what do brain connectivity studies tell us? Frontiers in Human Neuroscience 8: 74.

Luck, S. 2014. An introduction to the event-related potential technique. Cambridge, MA: MIT Press.

Macnamara, A., D. Foti, and G. Hajcak. 2009. Tell me about it: Neural activity elicited by emotional pictures and preceding descriptions. Emotion 9(4): 531–543.

Martell, D.A., 2009. Neuroscience and the law: Philosophical differences and practical constraints. Behavioral Sciences & the Law 27(2): 123–136.

McLoughlin G., S. Makeig, and M.T. Tsuang. 2014. In Search of biomarkers in psychiatry: EEG-based measures of brain function. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 165(2): 111–121.

McMullin, E. 1982. Values in science. PSA: Proceedings of the biennial meeting of the Philosophy of Science Association 1982(2): 3–28.

Mendez, M. 2009. The neurobiology of moral behavior: Review and neuropsychiatric implications. CNS Spectrums 14(11): 608–620.

Mogilner A., J.A. Grossman, U. Ribary, et al. 1993. Somatosensory cortical plasticity in adult humans revealed by magnetoencephalography. Proceedings of the National Academy of Sciences 90(8): 3593–3597.

Moll, J., R. Zahn, R. de Oliveira-Souza, F. Krueger, and J. Grafman. 2005. The neural basis of human moral cognition. Nature Reviews Neuroscience 6(10): 799–809.

Morse, S. 2005. Brain overclaim syndrome and criminal responsibility: A diagnostic note. Ohio State Journal of Criminal Law: 397–412.

Motzkin J.C., J.P. Newman, K.A Kiehl, and M. Koenigs. 2011. Reduced prefrontal connectivity in psychopathy. Journal of Neuroscience 31(48): 17348–17357.

Näätänen, R., and Picton, T. 1987. The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. Psychophysiology 24(4): 375–425.

Nathan, Z., D. Katz, and R. Zafonte. 2007. Brain injury medicine: Principles and practice. New York, NY: Demos Medical Publishing.

Pascual, L., P. Rodrigues, and D. Gallardo-Pujo. 2013. How does morality work in the brain? A functional and structural perspective of moral behavior. Frontiers in Integrative Neuroscience 7: 65.

Pletti, C., M. Sarlo, D. Palomba, R. Rumiati, and L. Lotto. 2015. Evaluation of the legal consequences of action affects neural activity and emotional experience during the resolution of moral dilemmas. Brain Cognition 94: 24–31.

Poldrack, R. 2006. Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences 10(2): 59–63.

Polich, J. 2007. Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology 118(10): 2128–2148.

Raine, A., and Y. Yang. 2006. Neural foundations to moral reasoning and antisocial behavior. Social Cognitive and Affective Neuroscience 1(3): 203–213.

Rudner, R. 1953. The scientist qua scientist makes value judgments. Philosophy of Science 20(1): 1–6.

Sarlo, M., L. Lotto, R. Rumiati, and D. Palomba. 2014. If it makes you feel bad, don't do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas. Physiology & Behavior 130: 127–134.

Schomer, D., and F. Lopes da Silva. 2012. Niedermeyer's Electroencephalography. Philadelphia, PA: Lippincott Williams & Wilkins.

Shafir, R., N. Schwartz, J. Blechert, and G. Sheppes. 2015. Emotional intensity influences pre-implementation and implementation of distraction and reappraisal. Social Cognitive and Affective Neuroscience 10(10): 1329.

Song, C.Y., Q. Wu, and T.G. Zhuang. 2006. Hybrid weighted minimum norm method a new method based LORETA to solve EEG inverse problem. Conference Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society 2005, 1079–1082.

Steriade M. 2006. Grouping of brain rhythms in corticothalamic systems. Neuroscience 137(4): 1087–1106.

Stevenson, R.A., R.K Zemtsov, and M.T. Wallace. 2012. Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. Journal of Experimental Psychology: Human Perception and Performance. 38(6): 1517–1529.

Sunstein, C., and A. Vermeule. 2005. Is capital punishment morally required? The relevance of life-life tradeoffs. University of Chicago Law & Economics Olin Working Paper No. 239.

Thompson D.F., C.L. Ramos, and R.K. Willett. 2014. Psychopathy: clinical features, developmental basis and therapeutic challenges. Journal of Clinical Pharmacy and Therapeutics 39(5): 485–495.

Van Berkum, J., B. Holleman, M. Nieuwland, M. Otten, and J. Murre. 2009. Right or wrong? The brain's fast response to morally objectionable statements. Psychological Science 20(9): 1092–1099.

Wagner, N.-F., and G. Northoff. 2015. A fallacious jar? The peculiar relation between descriptive premises and normative conclusions in neuroethics. Theoretical Medicine and Bioethics 36(3): 215–235.

Walter, W.G., R. Cooper, V.J. Aldridge, W.C. McCallum, and A.L. Winter. 1964. Contingent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. Nature 203 (4943): 380–384.

Weinberg, A., and G. Hajcak. 2010. Beyond good and evil: The time-course of neural activity elicited by specific picture content. Emotion 10(6): 767–782.

Yoder, K., and J. Decety. 2014. The good, the bad, and the just: Justice sensitivity predicts neural response during moral evaluation of actions performed by others. Journal of Neuroscience 34(12): 4161–4166.

Erkenn (2019) 84:657–686
https://doi.org/10.1007/s10670-018-9976-9

Against Cognitivism About Personhood

Nils-Frederic Wagner[1]

Received: 8 September 2016 / Accepted: 1 February 2018 / Published online: 21 February 2018

Abstract The present paper unravels ontological and normative conditions of personhood for the purpose of critiquing 'Cognitivist Views'. Such views have attracted much attention and affirmation by presenting the ontology of personhood in terms of higher-order cognition on the basis of which normative practices are explained and justified. However, these normative conditions are invoked to establish the alleged ontology in the first place. When we want to know what kind of entity has full moral status, it is tempting to establish an ontology that fits our moral intuitions about who should qualify for such unique normative standing. But this approach conflates personhood's ontology and normativity insofar as it stresses the primacy of the former while implicitly presupposing the latter; it thereby suffers from a 'Normative Fallacy' by inferring from 'ought' to 'is'. Following my critique of Cognitivism, I sketch an alternative conception, contending that, whereas the Cognitivist ontology of personhood presupposes the normative, a social ontology is constituted by it. In due consideration of evidence from developmental psychology, the social embeddedness of persons—manifested in the ability of taking a 'second-person stance'—is identified as a key feature of personhood that precedes higher-order cognition, and is directly linked to basic normative concerns.

✉ Nils-Frederic Wagner
nils-frederic.wagner@web.de

[1] Department of Philosophy, University of Duisburg-Essen, Forsthausweg 2, Room LE 329, 47057 Duisburg, Germany

1 Introduction

Personhood is a remarkably versatile concept that aims to uncover what persons have that non-persons don't have. The philosophical focus has largely been on defining what constitutes persons synchronically. Resulting theories are multifarious and spark ongoing debates in philosophy, law, the social sciences, and neuroscience. As such, personhood has both ontological and normative significance. A major source of confusion is the ambiguity of whether personhood is first and foremost an ontological or a normative concept. This ambiguity is fostered by a variety of different philosophical conceptions of what personhood encompasses, some of which are only loosely connected. For example, questions of corporate personhood (Kusch 2014) that discuss the metaphysical and moral standing of abstract entities, likely talk about something very different from questions of whether cognitive disability is an impediment to moral personhood in humans (Kittay 2005).

In this article, I am concerned only with a narrow conception of personhood that takes persons to be a set of concrete entities who form a distinct ontological category, and possess a unique moral status. There have been various, sometimes opposing attempts to link persons' ontology with their normativity. Some of the resulting theories are so diverse (for example Animalism and Psychological Continuity conceptions), suggesting that there might be different senses at play as to what constitutes a person. Shoemaker (2007, 2016) disentangles some of these senses in the context of personal identity and practical, normative concerns. While it is worth bearing in mind that there are different ways to conceive of what a person is, depending on what one aims to track (e.g., the metaphysical persistence conditions of persons or their moral responsibility in some particular case), what inherently unites these conceptions is that all of them make claims regarding the way ontological and normative conditions of personhood are connected. This, then, is not to deny that there are different senses of what a person is, but to narrow down the target of the following analysis to one of the main families of philosophical theories of personhood: views that take persons to form a distinct ontological category whose members possess a unique moral status.

In what follows, I begin by arguing against a particularly widespread class of such views of personhood: 'Cognitivist Views' (henceforth: Cognitivism) that see persons constituted by complex mental capacities; on this view, the necessary condition of personhood is higher-order cognition. The problem with Cognitivism is that such views fall prey to what I call a 'Normative Fallacy'. Cognitivists begin with a set of presuppositions about the unique moral status of persons, look for an ontological category that maps onto these presuppositions, and then draw normative conclusions from the ontological category they have put forward. The problem, then, is that Cognitivism begs the question by attempting to draw normative conclusions from an ontological condition that is covertly based on normative presuppositions.

Following my critique of Cognitivism, I argue in favor of an account that sees personhood constituted by its social ontology; on this view, the necessary condition of personhood is a pre-reflective capacity to engage in social relations which I take

to be a form of implicit first-order cognitive awareness that occurs prior to, and independently of, explicit second-order cognitive reflection. I contend that we should look to the social embeddedness of persons; specifically, to our innate ability of engaging in social interactions, i.e., persons' urge to adopt what I shall call a 'second-person stance'. In due consideration of recent evidence from developmental psychology, I argue that human infants display such pre-reflective social tendencies from a very early age on. An account of personhood grounded in social embeddedness evades the Normative Fallacy, I shall argue, by inherently linking a person's ontology to their normative significance, and therefore does not beg the question in the way Cognitivism does. Unlike Cognitivism, a social ontology of personhood does not presuppose normativity, but is rather constituted by it. Finally, I vindicate a social ontology of personhood considering possible objections.

2 Conceptual Richness of Personhood: Ontological or Normative Primacy?

Ontologically, personhood has largely been seen as a concept that categorizes entities according to particular mental features they share. Then again, how we interpret and apply personhood's ontology has a great impact on normative applications; particularly on bioethical and neuroscientific questions as well as on legal issues. Integral to the normativity of personhood is that by identifying someone as a person, we accord what is frequently called "full moral status" to them; most fundamentally, we believe that persons have a *right to life*.[1] And personhood is seen as the source of moral responsibility and legal accountability; implying that persons are an essential part of everyday life.

An ontological theory of personhood distinguishes persons from non-persons, supposedly *detached* from normative conditions. Nonetheless, such theories are often presupposed and invoked in ethical controversies. For example, a great deal of the debate on the moral permissibility of abortion and infanticide is centered around the issue of fetal and infantile personhood (Tooley 1972). Much of this controversy is owing to the conceptual vagueness and linguistic ambiguity of personhood. This, in turn, is grounded in the inherently contestable application of personhood in ordinary language.[2] Despite of its forceful normative application, there is no clear-cut and non-contentious way in which personhood is deployed in everyday life.

2.1 Disentangling Ontology and Normativity

On the one hand, personhood is *anthropocentric* in its actual application. Apart from counterfactuals, the only uncontroversial case of persons is human persons.[3]

[1] In Section 3.2, I detail how personhood is widely seen as the grounds of full moral status.

[2] For an analysis of this problem see English (1975) and DeGrazia (1996).

[3] So far as morality is believed to directly follow from personhood, this assertion has recently been called into question by de Waal's (2014) research suggesting that at least rudimentary levels of morality are present in apes and monkeys. Other members of the animal kingdom, such as dolphins, have also been suggested as candidates whose lives are governed by moral rules. Revisiting the theory of mind debate,

On the other hand, in its intension, personhood is *eo ipso not anthropocentric* (Kemmerling 2014). Nothing intrinsically restricts personhood to human beings; the designators 'human animal' and 'person' are clearly not coextensive. Conceptually, personhood allows both for human beings that don't qualify as persons because they lack person-constitutive conditions, and for the possibility of non-human persons that fulfill person-constitutive conditions but aren't members of our species.

Now, since an essential hallmark of personhood is to simultaneously signify and justify persons' unique moral status based on their integral ontological conditions, it becomes apparent how ontology and normativity are deeply interlocked. Nonetheless, both conditions of personhood can, and as I argue in what follows, should be disentangled. By so doing, some of the problems that arise from conflating ontology and normativity can be tackled more effectively.

Here is a first approximation of the two conditions of personhood:

(1) Ontology: ontological, non-normative conditions that distinguish persons from both human and non-human animals, and possibly others. *An entity* A *belongs to the distinct ontological category of persons if and only if it possesses the set of ontologically person-constitutive conditions* x. *Species membership has no bearing on this ontological taxonomy. That is, entities belonging to the same natural kind could, in principle, have a different moral status.*[4]

(2) Normativity: normative, non-ontological conditions that account for persons' special moral status. *An entity* A *has the same distinct moral status persons have if and only if it possesses the set of normatively person-constitutive conditions* y. *Species membership has no bearing on this normative taxonomy. That is, entities belonging to different natural kinds could, in principle, have the same moral status.*

Independent of whether the set of person-constitutive ontological conditions *x* and the set of person-constitutive normative conditions *y* turn out to be the same in (1) and (2), they nonetheless serve different purposes; a potential overlap can thus safely be ignored here. For in (1), *x* serves the purpose of an ontological categorization, initially detached from normative conditions; whereas in (2), *y* accounts for a unique moral status of its possessor, initially independent of their ontological conditions.

Footnote 3 continued
Andrews (2012) argues that some of the mental features that are by most believed to be uniquely human may also be present in great apes.

[4] Insisting that someone can only be a person by way of having features that are *ipso facto* human ultimately collapses into 'Speciesism': the doctrine that just by virtue of being human there is a good enough reason to have a superior moral status to non-human animals.

3 Cognitivism's Conflation of Ontology and Normativity

The relation between ontology and normativity may turn out to play a pertinent role in the conceptual richness of personhood. Views that have attracted both most attention and affirmation typically highlight higher-order cognition as the necessary condition possessed by all and only persons. The aim is to distinguish entities that lack higher-order cognition from persons: entities that are essentially constituted by having a complex mental life. This fits well with the widespread pre-theoretical intuition that a person is, most fundamentally, a mental being.

3.1 Traditional and Modern Approaches to Cognitivism

Cognition as the necessary condition of personhood is fairly well established ever since Aristotle told us that we are 'rational animals'. Modern versions of Cognitivism reach back to John Locke who famously regarded a person as a "thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places" (Locke 1975, 335). Peter Singer, a prominent contemporary Cognitivist, sees the special value of a person's life conferred by and preserved in four features: (1) Being rational and self-consciously aware of itself as an extended body existing over an extended period of time. (2) Having desires and making plans. (3) Containing a necessary condition for the right to life that it desires to continue living. (4) Being autonomous (Singer 1979, 78–84). Frankfurt's (1971) seminal analysis of the higher-order cognitive capacity to form 'second-order volitions' is often taken to be necessary for personhood. Persons, so defined, fundamentally act from reasons, which inevitably requires a certain degree of rationality, presupposing higher-order cognition. This, in turn, constitutes a distinct ontological category, setting persons apart from all the others.

3.2 Cognition as the Grounds of 'Full Moral Status': Stringent Presumption Against Interference

If interests matter morally to some degree for an entity's own sake such that it can be wronged, that entity has a moral status. Since moral status is frequently taken to come in degrees, Cognitivist reserve the notion of persons to designate entities with the highest degree of moral status; namely *full moral status* (Tannenbaum and Jaworska 2013). The main aspect of possessing full moral status is a "*very stringent moral presumption against interfering* with the being in various ways"—most importantly, it is morally impermissible to take their life or directly cause its suffering (ibid.). The stringent presumption against interference holds even when the life and interests of another valued creature are at stake, or for the sake of any other value. The presumption against interference is mainly cashed out in terms of rights: Singer talks about the "right to life" (Singer 1979, 81–83), Feinberg prefers the term "right not to be killed" (Feinberg 1980, 98–104).

Grounds for attributing full moral status to all and only persons have been offered most frequently based on higher-order cognition. On this view, an entity has full moral status if and only if it has higher-order cognition. Michael Tooley, perhaps most prominently, puts it as follows: "What properties must something have in order to be a person, i.e., to have a serious right to life? An organism possesses a serious right to life only if it possesses the concept of a self as a continuing subject of experiences and other mental states, and believes that it is itself such a continuing entity" (Tooley 1972, 44). Temporally extended self-consciousness of the sort Tooley requires for personhood is, of course, cognitively highly demanding.

Tannenbaum and Jaworska (2013) emphasize that according to Cognitivism the condition that grounds full moral status is not relational, but an intrinsic feature of persons:

> [T]he source of moral status is neither a relation the individual stands in (e.g., membership in a species) nor a capacity whose exercise requires active participation of another (e.g., the capacity to relate to others in certain mutually responsive ways). … Individuals have FMS [full moral status] solely because they can engage in certain cognitively sophisticated acts or responses on their own. Moreover, any being that has these sophisticated cognitive capacities has FMS, and so the accounts avoid anthropocentrism (ibid., 4.1).

This intrinsic higher-order cognition condition commits Cognitivism to what might be called 'Moral Intrinsicalism': the view holding that the final value of a person supervenes solely on features intrinsic to that entity—resulting in entities whose interests matter morally in their own right. Persons, so goes the argument, enjoy full moral status because they are the only beings possessing higher-order cognition: an ontological condition intrinsic and exclusive to persons.

I now turn to swiftly reconstruct two recent paradigmatic versions of Cognitivism in order to analyze in more detail what such views take to constitute personhood ontologically, and how this is believed to endow persons with a unique moral status. Subsequently, I spell out mistakes besetting these approaches.

3.3 Persons as Self-Conscious, Planning Agents

At the outset of *Epistemic Dimensions of Personhood*, Simon Evnine says that "[w]e think and reason at a far richer and more sophisticated level than any other beings with which we are acquainted" (Evnine 2008, 1). But Evnine doesn't take personhood's ontology to be merely descriptive: "Other animals cannot be moral or immoral and they are, one supposes, severely limited in what they can value and in what forms that valuing can take. Reason, an epistemic feature, thus lies at the foundations of all the things that make us (for better or worse) special" (ibid.). So, reason is, on Evnine's view, the ontological condition grounding persons' unique moral status.

When fleshing out his version of Cognitivism, Evnine posits four necessary conditions of personhood: finitude, belief, agency, and second-ordinality (ibid., 10ff.). Accordingly, persons are seen as finite, spatiotemporally extended, self-reflective agents endowed with concepts and beliefs. Evnine takes agency and

second-ordinality not only to refer to the performance of individual intentional actions, but also to the engagement in relatively long-term plans and projects, and the ability to deliberate about such actions, plans and projects. Whether someone qualifies as a person depends on their intrinsic cognitive features which are, *ipso facto*, ontologically independent of extrinsic factors. Someone endowed with higher-order cognition is thus a person in all possible worlds, even if one such world does not include anyone else, let alone other persons.

So, for Evnine, persons are entities constituted by higher-order cognition who belong to a distinct ontological category. This sets persons apart from all other beings and therefore entails a special moral status. The conclusion is familiar: planning agency, an inherently prospective outlook with the intention of executing long-term plans, equips persons with their right to life.

3.4 Persons as Cognitive, Social Entities

While most versions of Cognitivism base personhood exclusively on intrinsic conditions, Lynne Baker's approach is a notable exception. Baker acknowledges the social condition of personhood, averring that persons are intrinsically social entities. She does so, however, on Cognitivist grounds: for Baker, higher-order cognition is at the core of what makes persons social entities. She situates her view as part of a social ontology that includes "all social entities, social kinds and social properties that are irreducible and ineliminable. A social property is one for which social or linguistic communities are necessary for its instantiation. A community is one whose members bear significant intentional relations to one another" (Baker 2015, 78). Baker sees language and intentionality as necessary conditions for persons' social ontology. This is further characterized by what Baker calls a 'robust first-person perspective': the capacity to first-personally conceive of oneself as oneself. She locates this feature at the roots of personhood because it defines person-characteristic activities and concerns. So, Baker identifies persons' linguistic capacity as the key ingredient of their robust first-person perspective—language makes persons social and, on her view, both ontologically and normatively unique.

The special moral status Baker ascribes to persons becomes evident when she contends that "[w]e share with other species the property of having a rudimentary first-person perspective, but only we persons develop a robust first-person perspective that makes us moral and responsible agents" (ibid., 86). In further characterizing the moral importance of planning agency, Baker claims that "[w]ith respect to the range of what we can do (from trying to control our destinies to fantasizing about the future) and with respect to the moral significance of what we can do (from assessing our goals to confessing our sins), it is obvious that beings with robust first-person perspectives are unique" (ibid.). Elsewhere, Baker adds: "Clearly, again, nothing can be a moral agent without a robust first-person perspective. Since only persons can have robust first-person perspectives, only persons can be rational or moral agents" (Baker 2013, 192–193).

Baker grants a rudimentary first-person perspective, the ability to recognize the distinctiveness of one's own viewpoint, to non-human animals, but insists that only a linguistically-grounded robust first-person perspective enables sociality proper.

Baker is convinced, then, that having a robust first-person perspective makes persons ontologically and normatively unique: "Whereas our rudimentary first-person perspectives tie persons to the seamless animal kingdom, our robust first-person perspectives set us apart from everything else in the natural world" (Baker 2015, 87).

All this goes to show that, according to the two prototypical examples of Cognitivism just sketched, persons are the only beings that display higher-order cognition—an exclusive ontological category. They, therefore, so the argument goes, deserve full moral status. I now turn to argue that this way of reasoning falls prey to a Normative Fallacy: a conflation of personhood's ontology and normativity, ultimately rendering these views implausible.

4 Cognitivism's Normative Fallacy

The brief reconstruction of Evnine's and Baker's view has illustrated that Cognitivists typically base the desired moral status of persons on their allegedly defining ontology. In so doing, Cognitivists disregard that the primary, though mostly hidden, motivation for this ontology is in effect based on a pre-existing normative conviction. The uniqueness of self-reflective conscious persons, I argue, is invoked to ascribe the exclusive moral status that Cognitivists believe persons to have *in the first place*.

Here is a rapid outlook on how my argument runs: Cognitivists claim to initially categorize persons ontologically, independent of normative conditions. At the same time, however, these views are employed to justify persons' full moral status, thus becoming normative concepts.[5] Yet these normative concepts see the moral significance of persons based on precisely what they initially claim to be their ontological condition: higher-order cognition. Since Cognitivists believe persons to have a unique right to life, they look for entities displaying an ontology that can serve to justify this unique moral status. The most obvious contender to do the trick is higher-order cognition, which is then employed to found an ontological category. Cognitivists thereby disregard that this ontology is covertly motivated by a normative conviction. So, it's 'normativity *explaining* ontology'—and not, as asserted, 'ontology *justifying* normativity'. Cognitivism thus begs the question, employing unaccounted-for inferences from normative premises to seemingly ontological conclusions. These inferences are neither explicitly identified as such nor is a rationale given for their validity.

Before I turn to map out the argumentative structure underlying the Normative Fallacy, and to arguing in detail how Cognitivism falls prey to such reasoning, it should be noted that there is a potentially valid way of inferring ontological claims

[5] An anonymous reviewer pointed out that there is a difference between a 'descriptive concept that plays a normative role' and a straightforwardly 'normative concept'. Since Cognitivists claim that their descriptive, or ontological concept of personhood justifies persons' superior moral status, it becomes even more important to show on what grounds the descriptive concept does that normative work. And if it turns out, as I argue in what follows, that the grounds for so doing are wobbly, there is all the more reason to be wary of Cognitivism.

from normative premises.[6] For example, imagine that the 'Club of Wealthy Torontonians' stipulates as one of its main principles that 'every club member ought to give 10% of their income to charity'. Inferring from this normative principle to the ontological claim that 'every club member does, in fact, give 10% of their income to charity' is a valid inference (granted that the board makes sure all members comply, and bans those who refuse to follow the normative principle). Arguably, these sort of inferences are a frequent method of moral psychology: identifying normatively significant behavior as ontological facts about the world, and in that sense inferring the descriptive part of the normatively significant behavior from the normative premise guiding that behavior. But such an inference is not thereby committed to affirming the normative *content* of the normative premise; e.g., by claiming as an ontological fact about the world that all 'Club of Wealthy Torontonians' members give 10% of their income to charity, we are not, by inference, committed to saying that it is morally right to do so.

There is a crucial difference, then, between inferring ontological conditions from normative premises without thereby affirmatively carrying over the normative content of these premises (as a common practice in moral psychology), and inferring ontological conditions from normative premises that, at the same time, affirm the normative content of that premise. Such normative content affirming inferences call for additional corroboration, and failing to do so, as I detail in what follows, constitutes the Normative Fallacy.

4.1 The Normative Fallacy

The general methodological claim of falsely inferring ontology from normativity was first discussed by Campbell (1970), and has been underappreciated ever since. Campbell argues that philosophical analyses of concepts are, despite denials, either ontological or normative. He believes this to be at least in part due to the indistinct boundary between philosophy and social science, leading to methodological difficulties. If philosophy is seen as a merely conceptual endeavor (being solely committed to following the rules of logic and coherence) and therefore believed to be independent of empirical considerations, there is seemingly no need to invoke empirical evidence in arguing for normative commitments. As a consequence, some philosophers believe themselves to be guarded from the need to consider empirical evidence. In Campbell's words:

> This frees [philosophers] from the responsibility of providing empirical evidence to support their conclusions and also wards off the accusation that they are parading subjective preferences as if they were rationally justifiable propositions (Campbell 1970, 368).

This has led to the undesirable consequence that some philosophical analyses contain empirical generalizations that reflect hidden normative assumptions or convictions of the philosopher rather than carefully interpreted empirical evidence.

[6] I am grateful to an anonymous reviewer for pointing this out.

> In doing this they may be said to reverse the naturalistic fallacy, and, by arguing from 'ought' to 'is', commit what I shall call the normative fallacy. This fallacy consists of arguing from propositions which are themselves normative, or could count as evidence only for normative propositions, to conclusions which contain factual assertions (ibid.).

So, nothing can appear in the conclusion of a valid deductive inference which is not already implicit in the conjunction of the premises. Closely considered, it turns out that these supposedly ontological generalizations contain covert pre-existing normative assertions. Campbell thinks that it is easy enough for philosophers to consider their analyses of concepts to constitute ontological discoveries based on a certain understanding of philosophical activity; namely, the conviction that philosophers can arrive at ontological discoveries through conceptual analysis alone. Needless to say, there is a rival view of what it is to do conceptual analysis deriving from Wittgenstein (1953): rather than discovering the nature of a given concept, philosophers are in the business of recapitulating often elusive and occasionally profound, but nevertheless normative assertions about acceptable meanings of words, or the way concepts feature in ordinary discourse. The normative arguments that sometimes invisibly buttress philosophers' apparently ontological conclusions about concepts are normative claims about those meanings or uses of words or concepts of which the philosopher approves. What follows is that, in many cases, the techniques of inquiring after what is meant by certain concepts, or arguing about what they *should* mean, can be regarded as appealing to normative opinions rather than ontological evidence.

4.2 Norms and Facts

Both the well-known Naturalistic and the lesser-known Normative Fallacy concern the way norms and facts are related, but they begin from opposite directions. The Naturalistic Fallacy takes issue with drawing normative conclusions from observations, while the Normative Fallacy raises concerns about drawing ontological conclusions from normative premises. Without digging too deep into this vexed debate, a couple of remarks are needed to show why it should suffice for my purposes to demonstrate that there is a metaphysically contentious relationship between norms and facts.

One may justifiably be undecided about whether there can, in principle, be a valid way of deriving an 'ought' from an 'is' or vice versa. All the same, if a theory of personhood is based on a normative assertion from which a claim about persons' ontology is derived, a way of bridging the ontology-normativity gap is needed. This holds true even if one thinks that there is no such thing as a Naturalistic and/or a Normative Fallacy, for fallacies are not the only source of gaps. One must still explain how facts and norms are interlocked, and how to get from here to there. As long as the debate over the divide between 'is' and 'ought' is not settled, the claim that there is a potential gap between ontology and normativity stands. Suggesting otherwise would be to confuse ontological with epistemic and linguistic differences, and would neglect the need for correspondingly distinct methodological approaches

when relating 'is' and 'ought' claims, either in an ontological or in an epistemological or a linguistic sense. Even when granting the contentious assertion that there is no metaphysical difference between facts and norms, and that it is not fallacious to infer from 'is' to 'ought' or vice versa, it doesn't follow that there cannot be crucial methodological differences in inferring conclusions from these two sorts of statements. For the epistemic differences between facts and norms—accompanied by the linguistic differences in uttering 'is' and 'ought' sentences—cannot be denied. Granted, for argument's sake, that there were no Naturalistic and/or a Normative Fallacy, it would still be unclear how facts relate to norms for the simple epistemic reason that 'ought' statements make claims about how the world *ought to be*, whereas 'is' statements are descriptions of how the world *is*. It is worth emphasizing the different ways in which evidence is gathered in support of these two sorts of statements. Evidence for 'is' claims is gathered by observation, whereas evidence for 'ought' claims is gathered by arguments from principles that, more often than not, appeal to consequences. Thus, even if one denies the metaphysical difference or is agnostic as to how 'is' and 'ought' claims are metaphysically related, the epistemic and linguistic difference still holds and has to be accounted for methodologically.

Before proceeding, it is worth bearing in mind that there are at least two different interpretations of what the Normative Fallacy is.[7] On one interpretation, it is, in principle, fallacious to allow normative facts to determine ontological facts. But this is not necessarily so. As the previously discussed example of moral psychological inferences from situations with normative content to ontological observations has illustrated, there are lots of cases where we make this kind of inference unproblematically. I have suggested that such inferences are unproblematically because the ontological conditions that are inferred from normative premises do not thereby affirmatively carry over the normative content of these premises. Problems begin once inferences from normative premises to ontological conditions at the same time—and without providing independent reasons for so doing—affirm the normative content of that premise.

Another interpretation of the Normative Fallacy to which I now turn is more closely related to personhood and suggests that Cognitivist accounts are question-begging because they implicitly and simultaneously take facts about personhood to be both dependent and independent justifiers of persons' unique moral status. This is particularly worrisome since it is unclear and often left ambiguous whether personhood is taken to be first and foremost and ontological or a normative concept; i.e., do normative practices elucidate ontological conditions of personhood or do ontological conditions justify normative practices? Either way, as I argue in what follows, there is a need for independently substantiating such ought-is inferences by coherently interlocking personhood's ontology and normativity. A need that Cognitivism does not adequately meet.

[7] I am grateful to an anonymous referee for drawing my attention to this.

4.3 Cognitivism and the Normative Fallacy

Cognitivism generally aims at discovering necessary and jointly sufficient conditions of personhood that hold across all possible worlds. An entity so defined is a person not only in this world, but anywhere else. To morally make sense of this doctrine when accounting for persons' full moral status—which most Cognitivists believe persons to possess necessarily, not merely contingently—requires a person to have intrinsic features that are independent of changing circumstances. This commits Cognitivism to what I previously called Moral Instrinsicalism: the claim that whatever determines someone's value must be based on that entity's intrinsic nature.

I argue that this is having things back to front. The conviction that there 'ought' to be persons with certain cognitive features, on the basis of which Cognitivists want to accord them full moral status, doesn't entail that there ontologically 'are' persons that fit these normative assumptions. Rather, the ontological conviction that there de facto are persons constituted by higher-order cognition is motivated by the covert normative urge to have a means by which ethical controversies shall be solved. However, this ontological conviction is question-begging since it derives its foundation from pre-existing normative commitments.

Since Cognitivists want to establish and adhere to an exclusive and intrinsic full moral status of persons, they must find something that makes persons ontologically unique. The pre-existing normative commitment of full moral status is thereby invoked to justify the ontological uniqueness of persons. Personhood is seen as normatively significant because of higher-order cognition which is perhaps uniquely—but not uniformly or universally—present in physiologically developed humans. For this reason, the widespread belief is that these higher-order cognitive functions must *at the same time* be necessary conditions for personhood. This argument is based on a Normative Fallacy since it fails to provide compelling reasons for why ontological conclusions can be validly derived from normative premises; nor is this argumentative move explicated. Simply asserting the Cognitivist ontology is not compelling since it is inferred from the solely normative conviction that persons must be constituted by higher-order cognition, on the basis of which their special moral status shall then be justified.

Take first Evnine's and then Baker's approach as examples of how two prototypical Cognitivist views fall prey to the Normative Fallacy.

Evnine, as mentioned earlier, tells us that persons' higher-order cognition "put us in touch with morality and value in quite distinct ways" (Evnine 2008, 1), endowing persons with their full moral status. In this regard, Evnine further asserts: "If something is a person, that obliges us to treat it with a certain respect and consideration that are not called for, and not appropriate for, things that are not persons. It also entitles us to expect a certain consideration and respect from it that we do not expect from non-persons" (ibid., 3). Ontology is invoked to justify persons' normative uniqueness—or so it seems. Now, why does Evnine assert that higher-order cognition is both an ontological category and a normatively significant property? The idea seems to be that persons have particular intrinsic, ontological features that determine how such entities ought to be treated. What matters

normatively is higher-order cognition which establishes a unique ontological category. But these ontological claims are based on the initial normative conviction of persons possessing full moral status that has been asserted in the first place. Evnine adopts this pre-existing normative conviction because he thinks this neatly explains moral intuitions that we have pre-theoretically. In order to account for this normative conviction, the search for a unique ontological category has led to the most obvious contender: higher-order cognition. This effectively serves the pre-existing normative conviction that Evnine adheres to, since it looks as if persons are the only beings displaying higher-order cognition—fittingly, in just the way Evnine believes this to be necessary for personhood.

Contrary to Evnine's suggestion, we do not, however, look for higher-order cognition in an entity and decide *on these grounds* whether that entity qualifies as a person with full moral status. It's the other way around. Having once decided, on normative grounds, that an entity is a person and has full moral status, we know that this makes it the kind of entity that is likely to display higher-order cognition. Furthermore, it comes in quite handy that if the cognitive demands are set high enough, only human persons qualify. This fits well with the initial normative conviction: persons ought to have a unique moral status. Seen in this light, it becomes clear how the 'ought' of personhood—the belief that persons are "morally special" (as Evnine has it)—calls for an ontological justification. The ontological 'is'—belief, agency, and second-ordinality—that Evnine takes to capture our moral intuitions about personhood, is thus inferred from the moral 'ought'. Not the other way around. Otherwise, one would imagine, a detailed rationale would be given, explaining why it is that the alleged ontological uniqueness comes with a special moral status. It might, however, be difficult to give such a rationale without begging the question since, despite appearances, the belief in an exclusive moral status of persons takes precedence. Whereas the unique ontological category comes second and is chosen so as to accommodate the initial normative conviction. The problem is that the concept of a person is allegedly taken to ontologically designate a certain kind of being; but actually, it firstly picks out persons' exclusive moral status, and only then looks for ontological features that fit best this normative conviction.

A closer look at Baker's normative account of personhood shall help strengthening the case against Cognitivism. In response to Animalism, Baker says that were we to take human animals as purely biological beings, "human organisms are no more morally or ontologically significant than cockroaches or dinosaurs. To hold that to be a person simply is to be a human organism is to stipulate a meaning of 'person' that has no connection with the historical or contemporary use of the term" (Baker 1999, 158). Having a robust first-person perspective is what "gives a reason to regard human animals as morally significant in ways that other kinds of things are not: The moral significance of human animals is rooted in their ontological role of constituting persons" (ibid., 159). Elsewhere, Baker says more about how the robust first-person perspective grounds moral status:

> Since all and only persons have a capacity for a first-person perspective, the question of the importance of being a person comes down to the importance of having a first-person perspective. … However the first-person perspective

> came about, it is unique and unlike anything else in nature, and it makes possible much of what matters to us. It even makes possible our conceiving of things *as* mattering to us. The first-person perspective—without which there would be no inner lives, no moral agency, no rational agency—is so unlike anything else in nature that it sets apart the beings that have it from all other beings. The appearance of a first-person perspective makes an ontological difference in the universe. Much of what is distinctive about us and much of what we care most deeply about—our ideals, values, life plans; our status as rational and moral agents—depends on our being persons. … Our moral agency, our rational agency, the cognitive and practical abilities that require a first-person perspective, and the ability to have an inner life are all unique to persons. And these things, I submit, are among the most significant things about us. (Baker 2000, 163–164).

It is evident that Baker is not only concerned with the ontology of personhood, but asserts that an exclusive moral status directly follows. One way of accounting for such superior moral status is Baker's emphasis on the social embeddedness of persons, supposedly based on possessing a robust first-person perspective. This might look like an advantage of her view since it seemingly offers an alternative to the earlier mentioned doctrine of Moral Intrinsicalism. But that is not so. Baker takes cognition to be indispensable for social embeddedness, thus claiming that without a robust first-person perspective an entity cannot achieve a level of sociality sufficient for personhood.[8]

In pressing the ontological uniqueness of persons, Baker describes the normative implications that come with acquiring a robust first-person perspective: "We share with other species the property of having a rudimentary first-person perspective, but only we persons develop a robust first-person perspective that makes us moral and responsible agents" (Baker 2015, 86). It becomes clear how deeply interlocked ontology and normativity are in her view. But Baker fails to account for *why* it is that a robust first-person perspective—the ontological 'is' of her theory of personhood—accounts for persons' unique moral status. Instead, Baker elaborates on what makes us ontologically different from our fellow creatures. This is done, explicitly, to ensure the exclusive, full moral status, the 'ought', of persons; a status that Baker believes persons to have in the first place. Her argument is thus subject to the Normative Fallacy: the assertion of the ontological uniqueness of having a robust first-person perspective is inferred from the primal normative conviction of persons being equipped with a unique moral status.

Two related features of Baker's account demand closer attention so as to demonstrate why her view falls prey to the Normative Fallacy. The robust first-person perspective is characterized by graduality and by a mere difference in degree from a rudimentary first-person perspective. Baker concedes that humans acquire a robust first-person perspective not all at once, but gradually. At the end of this development, humans become persons. Suggesting that what distinguishes a robust from a rudimentary first-person perspective is a difference in degree, not in kind.

[8] In the next section, I argue that it is not the exclusion of the social dimension of personhood that renders Cognitivism flawed, but the insistence on higher-order cognition as its necessary condition.

When pressed with these charges, Baker refers to 'ontological vagueness', insisting that ever so small a step can add up to a difference in kind (Baker 2007). It is up for debate (but a question for another day) whether tiny developmental steps add up to a difference in kind. Granted, for argument's sake, that they do, it does not easily follow that a unique moral status appears once the final developmental step has been taken. If nothing else, a separate argument is needed to demonstrate that all the normative difference depends on ever so small an ontological step. But Baker offers no such argument. Instead the full moral status of persons is simply asserted—indicating that the normative conviction comes first.

What might make Cognitivism initially attractive, its congruence with the pre-existing normative conviction that persons are morally unique, actually renders Cognitivism incoherent because of a mismatch with what is basic and fundamental about persons. For Evnine finitude, belief, agency, and second-ordinality are ontologically fundamental; whereas for Baker a robust first-person perspective is fundamental. Most Cognitivists have their own, slightly different, idea about which cognitive conditions are necessary for personhood. Be that as it may, Evnine and Baker refer to their respective version of Cognitivism, but agree—in line with a clear majority of Cognitivists—that higher-order cognition secures persons' full moral status. So, they uphold the traditional ontology with its clean normative demarcation between persons and everything else. However, this unique moral status that is allegedly based on higher-order cognition is not constituted by fundamental ontological conditions. Rather, the reverse is true: higher-order cognition helps constitute the normative ideal of personhood.

Now, why is this a mismatch with what is basic and fundamental about persons? Treating someone as a person is to engage with her as the kind of entity to which that normative ideal applies. But to treat her as a person based on higher-order cognitive features that she happens to display at some point in her life is not, at the deepest level, a response to her nature, but a response to a pre-existing normative ideal that Cognitivists have about persons.[9] Therefore, Cognitivists base the allegedly defining ontology of personhood on pre-existing normative convictions: the 'is' of higher-order cognition as a unique ontological category is fallaciously inferred from the 'ought' of the unique, full moral status that Cognitivists believe persons to have in the first place.

Along the lines of the earlier suggested distinction between normative and ontological conditions of personhood, it could be argued that Cognitivist accounts such as Baker's and Evnine's views are what is sometimes called 'practice-independent' accounts; i.e. accounts which settle the question of what persons are prior to, and independently of, considerations about what kinds of entities possess full moral status. Conversely, some Cognitivists such as Locke can be viewed as 'practice-dependent' accounts of personhood which often appeal to certain normative facts, e.g. rights and responsibilities to settle what entities are persons.[10]

Now, in principle, Cognitivists could formulate their views as solely ontological, practice-independent or solely normative, practice-dependent accounts. Cognitivist

[9] Chappell (2011) makes a similar remark with regards to the very idea of having criteria for personhood.

[10] I am thankful to an anonymous referee for drawing my attention to this.

could, or so it seems, settle the ontological category through a natural kind, and then point out that this natural kind is such that it confers a particular moral status.

However, as the discussion of their views has shown, neither Baker nor Evnine hold such practice-independent views. Both these authors repeatedly make clear that being an ontological person comes with certain normative demands, and accordingly with appropriate practices, and certain ways persons ought to be treated. This strategy is pursued by the vast majority of Cognitivists for at least two reasons. For one, personhood as a solely practice-independent concept would be deeply at odds with well-established ordinary language practices where person talk always has an at least implicit normative dimension, pointing to an inherent interlocking of personhood's ontology and normativity. It, thus, would be unclear if such views would at all track what persons are. Rather, such practice-independent accounts would, more likely, invent some ontological ideal of personhood that has no bearing on how things really are. Practice-independency can, moreover, hardly justify any claim about normative implications of such views. Unless Cognitivists were to confess that there naturally comes some normativity with the ontological status of persons. Baker and Evnine, as most Cognitivists, are essentialists about persons and try explaining the normative conditions via that route; i.e., persons are, by nature, beings with higher-order cognition that confers a special moral status. But now we are back to the original problem of the Normative Fallacy, and the lack of properly relating the alleged normative status back to the claim of what makes persons ontologically distinct. The weak point is, in particular, that Cognitivists do not offer a rationale as to how ontological and normative conditions are related such that one explains and justifies the other. It is by no means self-evident that beings with higher-order cognition have not only a higher moral status than animals, but that this status is so superior that persons so defined have certain inherent moral rights.

Or think of Locke's view as the perhaps most prominent practice-dependent account of personhood. Locke says that the capacity for moral agency is such a unique feature that it defines a special kind of thing; a thing worthy of unique moral status. Person is a different kind of thing than Man. So the question is what constitutes objects of the first kind; what defines the kind of thing which is worthy of a unique moral status. Locke determines it must be a thinking thing with reason and reflection. This, too, is the wrong methodology for defining an ontological kind, though. Beginning with normative assumptions and only then looking for ontological conditions that map onto these pre-established assumptions lacks a justification why, and a rationale for how ontology and normativity are interlocked. Now, Locke can, perhaps, get away with it because he is a nominalist. But for philosophers that are essentialist of any kind (as, e.g., Baker is), this move is highly suspect.

5 Overcoming the Fallacious Inferences of Cognitivism

To plausibly account for the interlocking of personhoods' ontology and normativity, a coherent view of what is basic and fundamental about persons is needed. Such a view must both categorize persons ontologically and, at the same time, offer an

explanation as to why persons' lives are governed by a certain normative structure that comes with these ontological conditions.

In order to do so, I first draw on the notion of 'Anthropological Constants', revealing fundamental features of the human condition, map out how such constants fulfill both normative and ontological challenges, and explicate how they might lead the way to evading the Normative Fallacy. Subsequently, I suggest that taking a 'second-person stance' might be the Anthropological Constant that constitutes personhood.

5.1 Anthropological Constants: Fulfilling Both Ontological and Normative Challenges

The notion of Anthropological Constants derives from 'Philosophical Anthropology', and aims at unraveling what is most essential to the human condition. Jürgen Mittelstraß puts it as follows: "The goal of philosophical anthropology—its main rationale—is to search for and to determine those anthropological constants, that, independent of concrete social, cultural and historical circumstances and developments, constitute the components or properties of being human" (Mittelstraß 2003, 484). There are various conceptions of what lies at the core of human nature that can be described in different ways, scientifically, historically or phenomenologically. But, "[w]hat is common to these different approaches to human nature," Mittelstraß emphasizes, "is that common or similarly applicable determinations are made that—as universal determinations—can be understood as conditions of every social, cultural, historical, and intellectual development in the form of anthropological constants" (ibid.).

Anthropological Constants plausibly connect personhood's ontology and normativity because they pick out what is most essential about persons. This, in turn, can make intelligible why there are such close ties between persons' ontology and normativity. Ontologically, Anthropological Constants describe what is empirically most basic about persons, serving as an ontological taxonomy. Entities belong to the ontological category of persons because they have certain universal ontological conditions: given properties on which neither cultural nor historical contingencies have relativizing effects. Normatively, Anthropological Constants capture why entities of that sort function normatively in a certain way and thus matter morally, both to themselves and to others. Suggesting that the most essential condition of persons makes it necessary for such entities to treat each other in accordance with their basic nature so as to ensure their collective survival.

Anthropological Constants must thus both be ontologically plausible and fit the normative characterization of persons in order to explain their moral significance. Accordingly, they must satisfy the normative urge to ascribe personhood, and thereby capture the conditions we believe to be important in this concept. Most importantly, only a notion of personhood that isn't covertly motivated by pre-existing normative convictions is able to escape the Normative Fallacy. What is needed is not either/or, but both an ontological and a normative condition of personhood. By so doing, a suitable, ontologically plausible concept can be discovered that can reasonably be agreed upon for normative purposes. Such a

notion of personhood must capture the practical importance we bind to personhood along with its ontological condition. Crucial in this regard is that, despite the fact that we ascribe personhood by finding ontologically plausible conditions that we agree upon, these constitutive conditions are no longer chosen arbitrarily, as they fulfill both ontological and normative challenges. They depict the ontological characteristics that lie at the core of what makes a person and, at the same time, the normative standards we apply in practical concerns of everyday life.

A normatively plausible ontological condition of personhood is necessary to not falsely infer an 'is' from an 'ought'. If Anthropological Constants are considered, personhood is both based on an ontologically plausible taxonomy and is, at the same time, able to capture the normative significance that this concept bears. The earlier described conflation of personhood's ontology and normativity manifested in the Normative Fallacy can thereby be evaded. While humans may not be the only kinds of persons, there is a reason to start with them as the paradigm. The whole point of Anthropological Constants is, then, that normativity arises out of human social organization, but must simultaneously be responsive to constraining ontological facts about the world.

In contrast to Cognitivism, social ontological approaches are better suited to overcome the Normative Fallacy, because they are likely based on Anthropological Constants. In order to demonstrate why and how such views escape the Normative Fallacy by presenting both an ontologically and a normatively plausible condition of personhood, I now turn to examine one such view in detail. This will subsequently allow me to pinpoint general characteristics and merits of this approach.

5.2 The Pre-reflective Sociality of Persons

Schechtman's (2010, 2014) well-defended theory sees personhood grounded in the social embeddedness of persons that precedes higher-order cognition. In a recent article and in her latest book, *Staying Alive*, Schechtman develops the "Person Life View"—a theory that is based on Anthropological Constants, able to connect ontological and normative dimensions of personhood, and as such well-suited to overcome the Normative Fallacy.

In contrast to Cognitivism, Schechtman aims to show that the Lockean distinction between ontological and normative conditions of personhood fails to appreciate basic practical concerns of persons that are directly linked to their social embeddedness. On Schechtman's account, "the interactions of everyday life are at the core of what we are" (Schechtman 2010, 283), suggesting that a large part of the practical importance of personhood is based on the social and cultural embeddedness of persons. Personhood is inherently constituted by social relations, understood as a place in a matrix of relationships embedded in social practices through which these relations acquire meaning.

Persons are defined in terms of the characteristic lives they lead, and seen as unified loci of practical interaction. "The duration of a single person is determined by the duration of a single person life" (Schechtman 2014, 110). Whereby a person life is made up of three elements: "individual capacities, typical activities and interactions, and a social infrastructure" (ibid., 115). The identity conditions of

persons are thus not so much set by ontological conditions, but by the typical structure of person lives. The key aspect of what it takes to be a person is social embededdness which is not initially grounded in higher-order cognition, but "begins with a period of social dependence, and relatively basic cognitive capacities which develop over time into the full range of personal capacities and activities" (ibid., 112). Referring to the development of interactions between newborns and caregivers, Schechtman contends that personhood does not initiate with higher-order cognition like moral responsibility and prudential reasoning. Once children develop these capacities, it is not that a different entity suddenly comes into existence, but rather "the repertoire of interactions merely becomes increasingly complex and varied" (ibid., 79).

Schechtman's 'Person Life View' entails a certain kind of developmental trajectory of personal existence on which persons start out as dependent infants. They gradually develop a variety of cognitive, psychological, and social capacities which Schechtman characterizes as an "expansion of a circle from its center" (Schechtman 2010, 279), at a certain point of which persons become complex psycho-physical agents. Implicit in her account is the idea that person-constitutive conditions are neither acquired all at once nor are they lost all at once. Central to her view is that person lives are necessarily socially embedded, as they are lives in a certain culture in interaction with other persons. Schechtman calls this the 'person space': a space in which persons, dependent on their state of development, obtain certain social roles.

So, a person's most essential condition is her social embeddedness which precedes higher-order cognition. As recent evidence from developmental psychology and social neuroscience suggests, persons engage into social interaction right from birth on. This sociality, then, can be seen as an Anthropological Constant; a condition that defines what persons are to the innermost centers of their being. Based on this social embeddedness, persons ascribe normative significance to each other. Since persons live in a person space, they are members of a community (loosely defined) that is governed by a normative structure in accordance with the social nature of persons. Persons' essential ontological condition of sociality and their normative need to treat each other in a way that accounts for their basic nature is thus interlocked, and the Normative Fallacy evaded.

Furthermore, Schechtman's view is better able to account for marginal cases, where Cognitivism is controversially committed to dispute personhood, which is often both ontologically and normatively contentious. On Schechtman's view, because of our animal nature, we can lose certain person-specific capacities. But this loss—for example in conditions such as vegetative state, as well as in more transient states like severe depression—does not simply erase our complex histories with others or remove us from the web of social relations that makes up our lives. The practical concerns of persons do not cease when those higher-order cognitive capacities to which Cognitivists adhere to wane. A human infant, for example, is already embedded in a web of social relations, even though he might still be unable to actively take part in linguistic communication. In other words, there is something more basic to persons than higher-order cognition. This essential condition of personhood consists in persons' social embeddedness and thus, in their broadly

conceived web of social relations. Accordingly, the lack of higher-order cognition may diminish personhood, but it does not erase it altogether.

It could be argued that this concept isn't based on intrinsic features of personhood, but rather on an extrinsic socially-dependent ascription thereof. For example, practices such as the social convention of treating someone in a vegetative state *as if* she were a person may not be the same as actually being a person. When personhood is constitutively grounded in higher-order cognition, patients with extensive cortical damage but functioning brain stems are, despite being biologically alive, no longer persons, because they appear to lack any mental life. This is an example of a peculiar consequence of the social ascription (or lack thereof, as the case may be) of Cognitivism. It can however not serve as a plausible condition for that ascription itself, since it is chosen to vindicate covert normative convictions. Schechtman claims in contrast that personhood, in the sense of our unique endowments, informs the whole of our lives and does not represent a tidy package of concerns and activities that can be removed from our basic nature. As such, social embeddedness is an Anthropological Constant, and, even though it inevitably has a measure of social ascription, it is not merely an arbitrary convention.

The Anthropological Constant of sociality explicates the connection of personhood's normative condition, governing the normative structure of everyday life, with a person's innate capacity for social interaction as their most fundamental ontological condition. Schechtman argues that lives of encultured persons, practical tasks, concerns, and activities that are shared with others are infused with (and infuse) higher and more person-specific cognitive functions. This infusion gives person lives their unique character, even though the specific details can differ greatly from person to person and from culture to culture; it does not, however, entail higher-order cognition as the necessary condition of personhood. Social embeddedness is an important part of our lives before, and independent of, higher-order cognition.

5.3 Pre-reflective Self-and-Other-Awareness: Taking a 'Second-Person Stance'

Abstracting from Schechtman's view, I am now in a position to sketch a broader picture of social ontological views of personhood. By so doing, I aim to demonstrate how the core of such views, what I shall call the ability of taking a 'second-person stance', allows to better account for the ontological and normative interlocking of personhood, and thereby escapes the Normative Fallacy.

Personhood is inevitable to some degree a matter of social ascription; persons don't exist in a vacuum but come into being because they enter a space of shared meaning with others. This is so because of our innate urge to engage into social interactions that takes place before, and independent of, higher-order cognition. So, personhood has an intrinsically social condition. But what are the essential conditions enabling persons to enter a space of shared meaning? A necessary condition to do so must be the ability to pre-reflectively appreciate the existence of other members of such communities. Only if one appreciates, in a way yet to be defined more precisely, that they are embedded in a social environment, can they

become a member of such a community. I call this pre-reflective self-and-other-awareness the ability of taking a 'second-person stance'. If a community consists of members that share certain essential features, then the most salient feature must allow for recognizing and, for that matter, *taking part* in such a community. This essential and most salient condition of persons, as I propose in line with Schechtman, is the innate human urge for sociality: the pre-reflective cognitive awareness by virtue of which a second-person stance is taken. Taking a second-person stance is what constitutes personhood, because it allows for entering a space of shared meaning. More precisely, this space emerges out of the interface between the first-person perspective and the second-person perspective—that is, between the 'I' of introspection and the 'You' of extrospection. When someone becomes able to transcend from the 'I' of their first-person perspective, realizing that they share a space with a 'You', they then take a second-person stance towards another and become a person; a social space is created.

The key to connect personhood's ontology and normativity that creates much trouble for Cognitivism lies in this form of sociality: persons' inherent condition to take a second-person stance. This is based on self-and-other-awareness as the pre-reflective, tacit understanding of the personal pronouns 'I' and 'You' which are bound together. Without a 'You' there can be no 'I', and vice versa, without an 'I', there can be no 'You'.[11] Out of this tacit understanding of personal pronouns personhood materializes in a space of shared meaning, tying together the most fundamental ontological condition of personhood with the need to govern a so evolved community by a normative structure that best fits persons' fundamentally social nature.

Contrary to what Cognitivism suggests, the necessary condition of personhood is likely based on the innate ability of taking a second-person stance which precedes higher-order cognition. Sociality, then, is an *ideae innatae* which is part of persons' lives before, and independent of, higher-order cognition. In order to further flesh out the notion of taking a second-person stance, and to show how this capacity of sociality takes place pre-reflectively, I now turn to briefly discuss some recent evidence from developmental psychology. Subsequently, I explicate how taking a second-person stance connects personhood's ontology and normativity, and thus evades the Normative Fallacy.

Developmental psychologists have shown that infants as young as 1 year of age already understand others as possessing a variety of mental contents, including goals and intentions, perception and attention, knowledge, ignorance, and even false beliefs (Carpenter 2010). One important finding of another behavioral study is the observation that 9-months-old infants respond more patiently when an adult social partner is unable to give them a toy, compared to their unwillingness to give them a toy (Behne et al. 2005)—suggesting that, despite their lack of higher-order cognition, infants can differentiate between the intentional stances that adults take toward them under these conditions. In another study, it has been shown that if an adult looks at something that is beyond an infants' range of vision, 1 year olds will

[11] In social neuroscience, Schilbach et al. (2013) coined the term 'second-person engagement' reporting behavioral and neural evidence for persons' pre-reflective self-and-other-awareness.

autonomously move to another position so as to see what the adult is looking at (Moll and Tomasello 2004). This study clearly indicates that the infant ascribes mental states and intentions to the adult (in this example the assumption that the adult sees something interesting) on a pre-reflective level.

One might infer from these studies that infants are indeed able to take a second-person stance in relating themselves as an 'I' to the 'You' of the adult by being pre-reflectively aware of the person space they share.

The idea of personhood being intrinsically social and pre-reflective is further empirically supported by studies showing that as soon as infants understand others' emotions, goals and attentions, they will not only be able, but also highly motivated (even urged) to share their own emotions, goals and attentions with others (Tomasello et al. 2005). Infants start to coordinate their attention with others to objects of mutual interest outside the direct connection with the interaction partner at the age of 9 months (Carpenter et al. 1998). Much before that, in early infancy, babies delightfully participate in face-to-face interactions with their caregiver (Trevarthen 1980), voluntarily turning away from interesting objects with which they were engaged previously. In addition, between 2 and 5 days after birth, neonates show a preference for looking at faces (or pictures of faces) with eyes directly looking towards them. Even earlier than this, within minutes of birth, infants show considerable interest in, and respond appropriately to, self-directed facial movements, primarily more noticeable actions such as tongue protrusion and opening the mouth wide (Kugiumutzakis 1998; Meltzoff and Moore 1977; Nagy and Molnar 2004).

Empirical evidence suggests, therefore, that the infants' compulsion to take a second-person stance in order to attend a room of shared attention and meaning, where both infant and adult are engaged in a shared environment, is very likely to be congenital. Furthermore, these studies show that three of the most important social skills that infants apply in their proto-social eagerness to participate in the person space, namely, joint-attention, a special type of communication, and collaboration, take place on a pre-reflective level. So, the most salient condition of personhood is independent of higher-order cognition. The ability to understand that when an 'I' shares mental states with a 'You', something new is created, a person space in which persons experience something together, comes into existence way before higher-order cognition is acquired.

5.4 Evading the Normative Fallacy

Developmental psychology provides empirical support for the idea that taking a second-person stance is an Anthropological Constant. A socially-based approach seems thus able to connect persons' ontology with their normativity, as it accounts for the innateness of a basic form of sociality in accordance with which person lives are structured, and, perhaps, even societies at large are governed.

The ability of taking a second-person stance thus links a person's ontology to their normative significance. It is both intrinsic to persons, and thus ontologically constitutive, as well as the normative foundation of socially ascribing personhood, and thus the source of personhoods' normativity. Such a socially-based approach

escapes the Normative Fallacy because it does not base personhood on pre-existing normative commitments from which ontological conditions are inferred, but it normatively accounts for persons' most fundamental ontological condition. The paradigmatic structure of person lives is inherently normative, since it is governed around organizing the way we deal with each other. This, in turn, is based on persons' innate capacity of taking a second-person stance. Persons are social by default, as it were, and therefore need to organize their collective lives accordingly. Moreover, taking a second-person stance is not reducible to intrinsic ontological conditions of persons alone, for it does not merely infer normative significance from ontological observations.

With the ability of taking a second-person stance comes a measure of social ascription of personhood that is in accordance with the corresponding behavior of relevant others in social settings. In line with persons' inherently social nature, entities are persons because they treat each other as such. Out of this social engagement arises a space of shared meaning. Personhood is, then, on the one hand, a normative concept defined by social conventions, and, on the other hand, an ontological taxonomy of social entities.

To avoid the ontology-normativity gap that is present in Cognitivism, I suggest that the innate capacity of taking a second-person stance can serve as both personhoods' ontologically and its normatively necessary condition. In keeping with persons' fundamental ontological nature, personhood's normativity arises out of the de facto human social organization, and must as such be responsive to constraining ontological facts about the world. Whereas the Cognitivist ontology of personhood presupposes the normative, a social ontology is constituted by it.

5.5 The Role of Moral Intuitions in Weighing Cognitivism and the Social Ontology of Persons

It might be argued that deciding between Cognitivist and social ontological views of personhood comes down to competing moral intuitions regarding the best way of justifying persons' unique moral status.[12] What I take to be a significant advantage of social ontological accounts is that they do not in the same way as Cognitivism crucially rely on moral intuitions; certainly no primacy is given to moral intuitions when determining ontological conditions of personhood.

There are substantial reasons to be wary of the reliability of moral intuitions, ranging from the frequent disagreement of "moral experts" and vast cultural differences to the susceptibility of moral intuitions to framing effects (Sinnott-Armstrong 2008; Andow 2016). I take this worry to be elevated when moral intuitions are not only invoked to figure out how to get things right in cases of applied ethics, but are invoked to justify the moral status of persons per se. This more demanding task calls for an inferential corroboration of such intuitions. But Cognitivists can hardly provide a further corroboration of their ad hoc intuition that persons have a unique, full moral status from the start. This would require Cognitivists to point to inherent normative features of persons that necessitate the

[12] I owe this point to an anonymous reviewer.

need of such beings to accord moral status to each other. Since Cognitivism is, as I have argued, based on Moral Intrinsicalism, such explanations are very hard to come by. This is so because there is a mistake in thinking that the normative orientation is *justified* by intrinsic ontological conditions. The real problem is in separating the normative and ontological conditions of persons and thinking one *justifies* the other; that we can determine what persons must be ontologically from the fact that we accord them a particular moral status or that we can determine what the moral status of persons should be from their ontological conditions.

On a social ontological view, it is rather that we are, by nature, beings who accord moral status to each other; this is not grounded in something more basic, but defines who we are, and in that way has an ontological as well as a normative significance build in. The kinds of beings we are, just *are* beings who engage in *normative* social relations. So the ontology does not justify the normativity, nor does the normativity tell us what ontological conditions to look for (the ones that would justify the normativity). The ontological condition just is the normative nature.

The fundamental ontological condition of sociality that is present from birth makes it normatively necessary to treat each other in accordance with that ontological condition, otherwise we would cease to exist. Now, granting that collective survival is *good* in a normatively relevant sense, it seems apparent that the condition allowing us to do so is normative in nature. The normative-ontological condition of personhood as manifested in taking a second-person stance is thus neither an 'inference from ontology' nor an 'ad hoc normative assumption', but an empirically substantiated hypothesis about what persons are most fundamentally.

5.6 The Normative Significance of Taking a Second-Person Stance

It is sensible to ask why taking a second-person stance has an intrinsic normative significance that higher-order cognition lacks. Answering this question illustrates how a social ontology is normatively constitutive of personhood (rather than justifying persons' normativity), less demanding in ascribing moral status, and more inclusive (thus better able to account for marginal cases) than Cognitivism.

It is not that taking a second-person stance has something more in terms of normative significance than higher-order cognition does; but rather that adapting a second-person stance is already enough of normative significance to sufficiently motivate a special moral status of beings that display that feature. Taking a second-person stance does not so much justify the ascription of moral worth as constitute it. This is so because to take the second-person stance is to ascribe moral worth to others, or to take a moral interest in them; so persons are by their very nature the kinds of beings that do this. When we see that persons come into being by pre-reflectively realizing that others share their normative nature, normativity is set as a constitutive condition of personhood. Not an ontological condition that justifies according a unique normative significance to its possessors, but an inherent feature of persons that constitutes (rather than justifies) their normativity.

Saying that self-consciousness has in-build normativity as well (as in being able to relate to others in a normatively significant way) doesn't do the trick for the Cognitivist. This is so since the normativity of self-consciousness originates from its

pre-reflective relational nature, and does not spring into existence once someone acquires self-reflective cognition. The burden of proof, then, shifts to Cognitivists to explain why the bar for according a special moral status is set so high.

A frequently adopted Cognitivist strategy to try justifying persons' right to life is to posit that in order to desire to continue living, we first must understand the concept of life and death, and think of ourselves as temporally extended agents. Only then do we appreciate our existence in a way that makes it wrong to take persons' lives. No doubt, such self-reflection requires higher-order cognition. But it appears too demanding to justify persons' right to life, for understanding a concept is neither necessary nor sufficient for being able to desire what that concept conveys. Here is an analogy. I do not understand the conceptual intricacies of happiness (occasionally it seems as though I don't even have the faintest idea of what it means). I thus lack the higher-order cognitive capacities required to conceptually sort out what happiness is. Yet, on a pre-reflective level, I nonetheless, and quite strongly, desire to be happy. So, too, does the desire to continue living not presuppose an understanding of the concept of life and death nor an awareness of one's diachronic identity. A desire to continue living is not per se a cognitively demanding endeavor. It is, rather, an integral, often tacit component of persons' lives.

So, persons' special moral status is neither based on mere sentience nor on higher-order cognition. But it comes into play once beings pre-reflectively see themselves in relation to others. Such a 'relational completion' initiates the existence of entities that are something else entirely: persons that are governed by a unique condition of inherently relating themselves to others. Since moral concern crops up only when our actions are in some way related to others, the capacity for so relating is a good place to start explaining the normative significance of persons.

5.7 Conventionalism and the Social Ontology of Persons

An insightful approach to dissolve the problem of how personhood's ontology and normativity are intertwined is Conventionalism about personhood. On this view, as seminally put forward by Braddon-Mitchell and Miller (2004), normative considerations play a central role in determining what persons are ontologically; but not necessarily so, since had these normative practices been different, persons' ontology would have been different as well. Persons are viewed as conventional constructs, "objects whose existence logically depends on conventions. … According to logical conventionalism, the existence of the relevant conventions is part of the truth conditions for claims about the existence of persons" (ibid., 458). In a way, then, there is no a priori ontological fact of the matter as to what persons are. But there are normative facts about persons, contingent on how the world happens to be, that determine what persons are ontologically.

On the face of it, Conventionalism looks like a paradigmatic case of the Normative Fallacy. But this is not so. Rather, Conventionalism is a promising route to evading the Normative Fallacy, and, as I shall sketch in what follows, compatible with a social ontology of personhood. Here's why. Since conventions about persons (some of which are normative in nature) are part of what determines their ontology,

it is sensible to assume that such conventions—particularly in the case of established conventions that evolved over a long period of time and are deeply entrenched in society—cohere with how beings of that sort de facto function. Now, were persons stripped of what Braddon-Mitchell and Miller call 'settled conventions' (conventions that go largely unnoticed because we have no need to think about them), they would cease to be persons altogether (ibid., 462). And so there is no pre-existing settlement of what determines persons normatively, but person-directed practices (the import of which I spell out below) that evolve conventionally, and in that way contingently define persons' ontology.

Even though the authors do not set out to conclusively entrench what specific conventions are most fundamental about persons, there is a case to be made for how Conventionalism is compatible with a social ontology of personhood. This is so since persons' ability of taking a second-person stance is arguably one of the most fundamental settled conventions about persons. A means of evading the Normative Fallacy is, then, by way of showing how some of what Braddon-Mitchell and Miller call 'person-directed practices'—attitudes persons have towards themselves and other persons—are hard-wired, not malleable and thus determine what persons are essentially. Now, were these person-directed practices different, so would personhood's ontology be different. But since these person-directed practices as settled conventions happen to be in a specific way, and perhaps some couldn't have been otherwise by a posteriori necessity, it is difficult to see how anything can be a person if it wasn't for their fundamental practice of taking a second-person stance. On this view, persons are a special case of conventional beings because their existence depends both on contingent de facto human practices, and persons themselves, in keeping with their nature, instantiate these practices that constitute them (ibid., 468). A social ontology of personhood based on taking a second-person stance thus seems perfectly compatible with the Conventionalist idea that "persons exist only if they exhibit person-directed practices" (ibid., 469). Adopting the Conventionalist vocabulary, taking a second-person stance is a 'hard-wired settled convention' about personhood, and therefore constitutive of what persons are both normatively and ontologically.

5.8 Animalism and the Social Ontology of Persons

It goes without saying that my aim is not to altogether dispatch the importance of higher-order cognition as a crucial feature of personhood; rather, to set personhood on a less demanding footing. This, however, is not to assert that we are persons just by virtue of being human organisms. We are persons because of our embeddedness in a complex web of social interactions, which is, on the one hand, predisposed by our basic biological and psychological make-up, and on the other hand, enabled by the social infrastructure around us that we have created in accordance with our nature.

There might be stages of human existence in which, according to the here advocated sketch of a social ontological approach, human beings no longer qualify as persons. Or, at least, their personhood might be severely diminished. For example, permanent vegetative state (PVS) patients are arguably no longer able to

engage in social interactions. However, unlike Cognitivism, a social ontology can give a more refined description of these marginal cases. Since personhood does not necessarily involve higher-order cognition on the basis of which a person's unique moral status is secured, it can be argued that even in PVS cases, there is still a residue of the socially ascribed status of personhood in place that can plausibly account for their unique moral status.

There are, it seems, normative concerns that reasonably persist toward those whose higher-order cognition has considerably waned. We supply them with continuous medical treatment, we visit them, we talk to them. In short, we keep treating them as members of our society. Suggesting that what we care about doesn't vanish altogether when higher-order cognition disappears. There are at least two related reasons for this: we think that the PVS patient in front of us still remains to be an entity worth caring about in a way very similar to how we have previously cared about her, and we think that there is a tight ontological connection between the entity now (without higher-order cognition) and before (with higher-order cognition). Adhering to a Cognitivist conception of personhood commits us to deny that PVS patients are persons ontologically; this, though, is at odds with our normative concerns about such beings as forceful parts of everyday life, for it seems as though we in fact *do* accord them moral status, and target them normatively. It is either ontologically implausible to say that such beings aren't persons, or it is normatively implausible to say that such beings differ significantly in moral status from persons, since this would contradict our forceful de facto normative practice. Both issues seem to derive from the misfit of ontological and normative conditions that Cognitivist claim to be constitutive of personhood.

Admittedly, PVS cases are more challenging to social ontological views of personhood than other marginal cases such as infancy or cognitive impairment, since these beings display pre-reflective social tendencies as required for taking a second-person stance. Whether PVS patients retain a measure of such tendencies is a yet to be determined empirical question.

6 Concluding Remarks

The upshot of this article was to demonstrate how Cognitivism allegedly defines personhood ontologically, but thereby disregards that this definition is based on pre-existing normative convictions. From this methodological problem arise difficulties in making the case for personhood as an ontological category that shall account for persons' unique moral status. Asserting a Cognitivist ontology of personhood on the basis of there being a *normative demand* for persons so defined falls short, since it lacks both an identification of, and more importantly, a justification for the inference from normative convictions to ontological conditions. I have described this argumentative move as a Normative Fallacy, and argued that this is a serious threat to the plausibility of Cognitivism.

My positive contribution drew on Marya Schechtman's social ontological view of personhood, emphasizing persons' pre-reflective ability to engage in social interactions. In a subsequent step, I suggested that a social ontology of personhood

generally fares better in avoiding the Normative Fallacy. I argued that some of our person-related practices and concerns apply to individuals who do not possess higher-order cognition (e.g., infants). Thus, the relation which constitutes a unit of these practices and concerns cannot be one that requires higher-order cognition. Accordingly, infants are persons at particular life stages, they are persons without some of the typical attributes of adult persons, but they are entirely persons. Just as someone with a heart disease is still a human animal, only without some of a human animal's typical functionality, so is a person with cognitive deficits still a person. This is so because persons are essentially relational beings, such that their conditions for individuation and identification cannot be given independently of how they stand in relation to others. For this reason, a plausible conception of personhood must take into account the social embeddedness of persons as their most salient condition. This sociality is based on the innate capacity of what I called taking a second-person stance, which takes place prior to, and independent of, higher-order cognition.

The need to drop a sharp distinction between personhood's ontology and normativity calls into question the idea of a purely metaphysical view of persons. It is unclear whether such a metaphysical view that posits some intrinsic person-constitutive conditions—independent of how persons stand in relation to others—could hold across all possible worlds. After all, persons are relational creatures to the innermost center of their being. I therefore venture to understand personhood as a key subject of empirically informed social ontology, rather than seeing persons as a metaphysical puzzle.

A social ontological view of personhood entails implications of an interdisciplinary nature, to be treated in a second moment of reflection from here. Important ethical questions may arise which, perhaps, have a more general impact on normative reasoning and practical concerns. After all, the constitution of personhood fundamentally underlies, theoretically and practically, the way we conduct our lives.

Acknowledgements I am grateful to Dieter Birnbacher, Pedro Chaves, Lucas Jurkovic, Luca Lavagnino, Heidi Maibom, Susana Monsó, Neil Roughley, Gregory Walters, Katherine Wayne, and the anonymous referees for Erkenntnis for their insightful comments that helped improving the paper significantly. I am particularly indebted to Marya Schechtman for her ingenious comments and constant encouragement. I've presented this work at the Carleton University Philosophy Colloquium in Fall 2014, and at the Boston Conference on Persons in Summer 2015. Many thanks to the audiences at these events—particularly to Gabriele Contessa and Andrew Brook at Carleton—for their valuable feedback.

References

Andow, J. (2016). Reliable but not home free? What framing effects mean for moral intuitions. *Philosophical Psychology, 29*(6), 904–911.

Andrews, K. (2012). *Do apes read minds? Toward a new folk psychology*. Cambridge: MIT Press.

Baker, L. (1999). What am i? *Philosophy and Phenomenological Research, 59*(1), 151–159.

Baker, L. (2000). *Persons and bodies: A constitution view*. Cambridge: Cambridge University Press.

Baker, L. (2007). *The metaphysics of everyday life: An essay in practical realism*. Cambridge: Cambridge University Press.

Baker, L. (2013). *Naturalism and the first-person perspective*. New York: Oxford University Press.

Baker, L. (2015). Human persons as social entities. *Journal of Social Ontology, 1*(1), 77–87.
Behne, T., et al. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41,* 328–337.
Braddon-Mitchell, D., & Miller, K. (2004). How to be a conventional person. *The Monist, 87*(4), 457–474.
Campbell, T. (1970). The normative fallacy. *Philosophical Quarterly, 20*(81), 368–377.
Carpenter, M. (2010). Social cognition and social motivations in infancy. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 106–128). Oxford: Wiley-Blackwell.
Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development 63*(4): Serial No. 255.
Chappell, T. (2011). On the very idea of criteria for personhood. *The Southern Journal of Philosophy, 49*(1), 1–27.
De Waal, F. (2014). Natural normativity: The 'is' and 'ought' of animal behavior. *Behaviour, 151,* 185–204.
DeGrazia, D. (1996). *Taking animals seriously: Mental life and moral status*. New York: Cambridge University Press.
English, J. (1975). Abortion and the concept of a person. *Canadian Journal of Philosophy, 5,* 233–243.
Evnine, S. (2008). *Epistemic dimensions of personhood.* New York: Oxford University Press.
Feinberg, J. (1980). Abortion. In T. Regan (Ed.), *Matters of life and death* (pp. 183–217). Philadelphia: Temple University Press.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*(1), 5–20.
Kemmerling, A. (2014). Why is personhood conceptually difficult? In M. Welker (Ed.), *The depth of the human person. A multidisciplinary approach* (pp. 15–44). Michigan: Grand Rapids.
Kittay, E. (2005). At the margins of moral personhood. *Ethics, 116*(1), 100–131.
Kugiumutzakis, G. (1998). Neonatal imitation in the intersubjective companion space. In S. Braten (Ed.), *Intersubjective communication and emotion in early ontogeny* (pp. 63–88). Cambridge: Cambridge University Press.
Kusch, M. (2014). The metaphysics and politics of corporate personhood. *Erkenntnis, 79*(9), 1587–1600.
Locke, J. (1975). *An essay concerning human understanding. Edited by P. H. Nidditch*. Oxford: Clarendon Press.
Meltzoff, A., & Moore, K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75–78.
Mittelstraß, J. (2003). Philosophy or the search for anthropological constants. In U. Staudinger & U. Lindenberger (Eds.), *Understanding human development Dialogues with lifespan psychology* (pp. 483–494). Boston: Kluwer Academic Publishers.
Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science, 7*(1), F1–F9.
Nagy, E., & Molnar, P. (2004). Homo imitans or homo provocans? The phenomenon of neonatal initiation. *Infant Behavior and Development, 27,* 57–63.
Schechtman, M. (2010). Personhood and the practical. *Theoretical Medicine and Bioethics, 31*(4), 271–283.
Schechtman, M. (2014). *Staying alive—Personal identity, practical concerns, and the unity of a life*. New York: Oxford University Press.
Schilbach, L., et al. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences, 36*(4), 393–414.
Shoemaker, D. (2007). Personal identity and practical concerns. *Mind, 116*(462), 317–357.
Shoemaker, D. (2016). The stony metaphysical heart of animalism. In S. Blatti & P. Snowdon (Eds.), *Animalism* (pp. 303–328). Oxford: Oxford University Press.
Singer, P. (1979). *Practical ethics*. Cambridge: Cambridge University Press.
Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 2: The cognitive science of morality* (pp. 47–76). Cambridge, MA: MIT Press.
Tannenbaum, J., & Jaworska, A. (2013). The grounds of moral status. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/sum2013/entries/grounds-moral-status/. Accessed 10 June 2017.
Tomasello, M., et al. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28,* 675–691.

Tooley, M. (1972). Abortion and infanticide. *Philosophy & Public Affairs, 2*(1), 37–65.
Trevarthen, C. (1980). The foundations of intersubjectivity: Development of interpersonal and cooperative understanding in infants. In D. Olson (Ed.), *The social foundations of language and thought*. New York: W.W. Norton & Co.
Wittgenstein, L. (1953). *Philosophical investigations*. In G. Anscombe & R. Rhees (Eds.), G. Anscombe (Trans.). Oxford: Blackwell.

Medicine, Health Care and Philosophy
https://doi.org/10.1007/s11019-022-10089-w

SCIENTIFIC CONTRIBUTION

Personal identity, possible worlds, and medical ethics

Nils-Frederic Wagner[1]

Accepted: 23 April 2022

Abstract
Thought experiments that concoct bizarre possible world modalities are standard fare in debates on personal identity. Appealing to intuitions raised by such evocations is often taken to settle differences between conflicting theoretical views that, albeit, have practical implications for ethical controversies of personal identity in health care. Employing thought experiments that way is inadequate, I argue, since personhood is intrinsically linked to constraining facts about the actual world. I defend a moderate modal skepticism according to which intuiting across conceptually incongruent worlds constitutes 'invalid intuition-inferences'—i.e., carrying over intuitions gathered from facts about possible worlds that are at odds with facts about the actual world, for the purpose of making claims about real-life persons and their identity, leads to conceptual incongruences. Such a methodological fallout precludes accurate, informative judgments about personal identity in the actual world, calling into question the adequacy of thought experimental considerations for potential real-world applications in medical ethics.

Keywords Personal identity · Thought experiments · Intuitions · Psychological continuity · Animalism · Medical ethics

Introduction

Controversies about personal identity figure prominently in wide-ranging ethical issues in health care, such as abortion (McInerney 1990; Warren 1977; Oderberg 1997); advance directives, in particular with regard to neurodegeneration (Buchanan 1988; DeGrazia 1999; Vollmann 2001; Limbaugh 2016); and Deep Brain Stimulation (DBS) (Lipsman and Glannon 2013; Nyholm and O'Neill 2016; Müller et al. 2017). Employed as therapeutic interventions for neurodegenerative diseases, DBS bears the potential to significantly alter patients' psychological make-up. As such, DBS can have an impact on the ontological, moral, and legal status of patients undergoing such treatment. In the case of advance directives, patients suffering from neurodegenerative diseases such as Alzheimer's must make a call on behalf of their future selves, that, however, might no longer be identical to the 'original' self that has signed the advance directive. The moral permissibly of abortion is closely linked to the moral status of fetuses and the persons they potentially become: if a fetus is considered a person in the making, abortion would immorally deprive it of a 'future like ours', or so goes Marquis's (1989) seminal argument.

These and related discussions about potential identity disruptions in medical ethics are often implicitly (occasionally explicitly) based on philosophical theories of personal identity. Theories that are, in turn, frequently defended by appeal to counterfactual thought experiments that, despite being logically possible, are at odds with the set of facts of the actual world that affect real-persons and their identity. These kind of thought experiments, as I'll argue, have no bearing on real-world cases, and should be taken with a grain of salt. Before turning to argue why this is so, some conceptual tidying up is in order.

The focal point of theorizing about personal identity has been to sort out two related questions:

(1) what are synchronic conditions of personhood, and
(2) how do persons, so defined, persist through time?

Not all and only human beings are persons. The concept of personhood in principle allows for non-human persons, as well as artificial and alien persons. But human persons are the solely uncontested case to date, and thus inform our theorizing about personhood and personal identity. There is, then, reason to start with the constitution of human persons

✉ Nils-Frederic Wagner
n.wagner@uni-mainz.de

[1] Institute for the History, Theory, and Ethics of Medicine, University of Mainz Medical Center, Am Pulverturm 13, 55131 Mainz, Germany

Published online: 26 May 2022

as the paradigm. Accordingly, many recent attempts to answer (1) are based on what I shall call Orthodox Commitments about personhood.[1]

> Realism: 'Person' is a natural kind, picking out creatures that de facto exist in the actual world.
>
> Naturalism: Persons are biological beings whose existence is a matter of empirical facts.
>
> Cognitivism: Persons are equipped with higher-order cognition that enables diachronic self-consciousness.

In a nutshell, persons as we know them, are real, biological beings that, via higher-order cognition, can first-personally conceive of themselves as themselves persisting over time (hereafter, simply 'real-life persons').

Orthodox Commitments entail that personhood is intrinsically linked to, and constrained by, facts about the actual world. This is because these constraining facts have crucially informed the conceptual genesis of personhood, and continue to govern its practical application. Had the actual world been different, allowing, say, for people splitting into two equally adequate successors, the concept of personhood, too, would have evolved differently. Questions of, for example, divided identity that don't crop up as things stand (apart from thought-experimental worries), would have arisen as legitimate concerns in a world where people do split. Call the intimate relation between personhood and the actual world's constraining facts Intrinsic Linkage.

Even though Orthodox Commitments appear both independently plausible and are widely agreed upon, philosophers that share these commitments often disagree as to their entailment for answering (2). For brevity's sake, I focus on two main contenders:

> Friends of Psychological Continuity theories hold that, necessarily, person x at t_1 is identical to person y at t_2 if and only if x and y are psychologically continuous (e.g., connected via perpetual chains of psychological continuity).
>
> Friends of Animalism hold that, necessarily, person x at t_1 is identical to person y at t_2 if and only if x and y are biologically continuous (e.g., connected via perpetual chains of biological continuity).

How psychological and biological continuity is spelled out precisely differs between various proponents of these views. For my purposes, it shall suffice, though, to have a general overview in place.

Here's how I proceed: in section two, I discuss reasons why thought experiments loom large in settling differences between rival views of personal identity. In so doing, I distinguish between hypothetical thought experiments that are in keeping with facts about the actual world relevant to persons and their identity, and counterfactual thought experiments that are at odds with these facts. In section three, I survey cerebrum transplant thought experiments as a case in point. In section four, I argue that counterfactuals of such nature are inadequate to settle differences between rival views of personal identity. The reason for this is that intuiting from facts about possible worlds, where these facts violate those of the actual world, to then reapply these intuitions to the identity of real-life persons, constitutes invalid intuition-inferences. Because of Intrinsic Linkage, such intuition-inferences lead to conceptual incongruences across worlds and, thus, cannot generate accurate, informative judgments about personal identity in the actual world. In section five, I argue, furthermore, that such invalid intuition-inferences are at odds with Orthodox Commitments, as they conflate de re necessity and de re possibility about persons. In section six, I look at the impact that the rise of experimental methods in philosophical debates on personal identity has recently had on real-life cases in medical ethics. In so doing, I argue for a more balanced approach of both theory informing the practical approach clinical ethicists take towards real-life cases, as well as real-life cases inform theorizing about personal identity. In section seven, I take stock.

Thought-experimenting in personal identity

Differences between rival views of personal identity often appear most strikingly in thought experiments that frequently concoct bizarre counterfactual propositions. Derek Parfit's early work on personal identity has played a pivotal role in reigniting this way of thought-experimenting. Yet, Parfit tells us that "different views about personal identity make different claims about actual people, and ordinary lives" (Parfit 1984). So, thought experiments about possible worlds are not mere illustrations of theories and their implications. Nor ends in themselves. Rather, they are introduced as adequate, genuine attempts to sharpen our conceptual understanding of personhood and personal identity regarding real-life persons.

Along these lines, Parfit offers a reason why thought experiments about possible world modalities are evoked,

[1] Frankfurt (1971), Dennett (1976), Baker (2000), Wiggins (2001), McMahan (2003), Olson (2007), Shoemaker (2008), and Parfit (2012) are among the most prominent advocates of Orthodox Commitments.

claiming that "the difference between these views is clearer when we consider certain imaginary cases. Most of the arguments appeal, in part, to such cases. It may be impossible for some of these cases to occur, whatever progress may be made in science and technology" (ibid.).[2] Ever since, there has been little departure from such liberal application of thought experiments that are profoundly at odds with the facts about the actual world relevant to persons and their identity.[3] It is, for example, common currency for advocates of Psychological Continuity theories and Animalism alike to employ thought experiments to showcase how their views differ. And, more importantly, to elicit intuitions that allegedly pull the uninitiated towards their respective view. The primary reason for employing thought experiments is, then, to carve out what is conceptually essential about persons and their identity by isolating core conceptual features. According to the prevailing view, doing so requires imagining away ontologically insignificant contingencies about the world that real-life persons happen to inhabit.

What follows is not a critique of thought experiments that aim at imaging away ontologically insignificant contingencies about persons *per se*; let alone critiquing thought experiments in general. Rather, I take issue with the widespread tendency to employ thought experiments that evoke possible world modalities to make inferential claims across worlds, and thereby disregard Intrinsic Linkage. Accordingly, it is useful to distinguish between two different types of thought experiments that are commonly employed in debates on personal identity.

A family of thought experiments that I take to be methodologically adequate are what I shall call

> Hypotheticals: i.e., thought experiments that are in keeping with the set of facts of the actual world that affect real-life persons and their identity.

Hypotheticals are frequently used in both philosophy and science. By and large, they pose few problems in method, though, as stated by Coleman (2000), "they can certainly cause great disagreement over the results that they may suggest."

A well-known example of a Hypothetical in personal identity is Thomas Reid's (1785/1969) Brave Officer, where we are asked to imagine a small boy who once was flogged for having stolen an apple. When that small boy grew up to become a brave officer, he still remembered the flogging. And when the brave officer became an old general, he likewise remembers how he once was a brave officer. However, the old general no longer remembers having been flogged as a small boy. We are then invited to intuit whether the old general is identical to the small boy, despite no longer remembering the flogging. The results yielded by this Hypothetical might be controversial; but its method is not. Brave Officer itself cannot reveal whether the old general is identical to the small boy. It can only serve as a test case for theories of personal identity that have certain implications vis-à-vis the case. These implications are then compared to intuitions pumped by the Hypothetical, and squared with plausible conceptual commitments. Take the memory criterion that Brave Officer is directed against. If continuous first-personal memory is both necessary and sufficient for personal identity, the small boy and the brave officer are identical. As are the brave officer and the old general. But the small boy and the old general are not identical, since there is no first-personal memory relation between the two. This conclusion is not only counterintuitive to most, but also reveals that the memory criterion violates a plausible conceptual commitment about personal identity: if A is identical to B, and B is identical to C, then, by transitivity, A must also be identical to C.

Such analysis of Brave Officer has led many to conclude that the memory criterion of personal identity is implausible. This goes to show that Hypotheticals have a legitimate place in the debate and can help carving out core conceptual conditions of personal identity.

A family of thought experiments that I take to be methodologically inadequate are what I shall call

> Counterfactuals:[4] i.e., thought experiments that are logically possible, but at odds with the set of facts of the actual world that affect real-persons and their identity.

Thought experiments of such nature are particularly widespread in the personal identity literature. Teletransportation, fission and fusion, as well as brain/cerebrum transplants—to name but a few—figure prominently.

One of the most pertinent Counterfactuals is John Locke's (1698/2012) the Prince and the Cobbler. Locke, intending to establish consciousness as a necessary and sufficient condition of personal identity, asks us to concur that "should the soul of a prince, carrying with it the consciousness of the prince's past life, enter and inform the body of a cobbler, as

[2] Parfit understands that some thought experiments will remain forever impossible. Whereas others are merely technologically impossible or, in fact, possible. I return to these distinctions shortly.

[3] Wilkes (1988), Gendler (1998, 2002), and Kipper (2016) are wary of employing counterfactual thought experiments in personal identity. Coleman (2000), and Beck (2006; 2016), on the other hand, defend their adequacy.

[4] I use the terms Hypotheticals and Counterfactuals exclusively to denote factually possible and factually impossible thought experiments with regard to persons and their identity, respectively. So, when I say counterfactual, I do not talk about conditionals that make claims about circumstances that would have followed, had the actual sequence of events been different.

soon as deserted by his own soul, everyone sees he would be the same person with the prince, accountable only for the prince's actions."

In what follows, I look at a modern-day variant (the naturalization) of the Prince and the Cobbler, and argue why such Counterfactuals, despite appearances, do not yield insights into the ontology of real-life persons.[5]

Cerebrum transplant counterfactuals

It is widely acknowledged that real-life persons undergo continuous biological changes that are no threat to their diachronic personal identity. The human body's cells are constantly replaced, and the brain cell connections and chemistry are frequently changing without having an identity-compromising effect; neither on one's biological nor on one's psychological make-up. In all ordinary cases, when psychological continuity is in place, so is biological continuity; though not vice versa.[6]

To contrast Psychological Continuity theories with Animalism, it is tempting to imagine what were to happen if psychological continuity were present, but biological continuity were not. Since Locke's the Prince and the Cobbler with its Cartesian ring of an immaterial soul as the carrier of consciousness is no longer very popular with naturalistically minded philosophers, there has been a shift towards cerebrum transplant Counterfactuals. Accordingly, the soul has been replaced by the cerebrum as the seat of psychological continuity. Cerebrum transplants seem particularly pertinent since they appear to do justice to Naturalism about personhood. That way, cerebrum transplants strike most as less bizarre than, for example, teletransportation or fission Counterfactuals.

Here's a typical portrayal of a cerebrum transplant: Imagine *A*'s cerebrum is successfully transplanted into *B*'s head, while leaving *A*'s brainstem and midbrain regions intact such that *A*'s organism remains alive. Imagine further that this makes the resulting *B* psychologically continuous with *A* before the transplant had occurred by any standard: *A*'s mental states are physically realized throughout the process, and there are no troublesome rival candidates (Olson 2016). Now, who wakes up after the procedure? The seemingly natural intuition is that, were such things to happen, person *A* would be transferred with their cerebrum. Call this Transplant Intuition. Shoemaker (1963) presents, as do many others, such cerebrum transplant Counterfactuals as decisive evidence for Psychological Continuity theories against Animalism. Modern-day Animalists such as Snowdon (2014), however, disagree, denying the force of Transplant Intuition.[7]

When discussing intuitions gathered from possible world modalities, it is vital to keep in mind that the 'results', so yielded, are taken to settle differences between rival views of personal identity regarding real-life persons. It's not quite the claim that only people in some possible world where cerebrum transplants take place are transferred with their cerebrum. The point is, rather, that pondering these Counterfactuals is supposed to reveal that Psychological Continuity is the correct view of personal identity in real life. Granted, for argument's sake, that the transplant intuition offers enough of a compelling reason to drop Animalism. We are not, then, identical to the living organism left behind in a cerebrum transplant. Rather, we cease to exist once our psychology is gone; at least we are no longer inhabiting that cerebrum-robbed organism (Parfit 2012). Practically, this could mean that, given the serve deterioration of autobiographical memory in Alzheimer's that comes with a reported loss of sense of identity (El Haj et al. 2017), advance directives regarding someone in the late stages of Alzheimer's, with little to no psychological continuity linking them to the initial signee, should not be considered authoritative. By the same token, in the absence of advance directives, there is seemingly no point in interviewing close relatives to reconstruct the presumed patient's will since the patient currently undergoing treatment is no longer identical to the would-be signee.

There are, however, several constraining facts about the actual world that preclude cerebrum transplants from ever happening.[8] For one, the underlying assumption that the cerebrum *alone* maintains psychological continuity is called into question by evidence from cognitive science. Theories

[5] A recent approach in experimental philosophy has been to study the robustness of folk intuitions regarding thought experiments in personal identity (Blok et al. 2005; Nichols and Bruno 2010; Berniūnas and Dranseika 2016). These studies do not, however, aim at defending the role of intuitions in assessing theories of personal identity. Rather, the goal is to show "that if it is appropriate for philosophers to rely on intuitions in assessing theories of personal identity, then it will help to identify which intuitions are especially robust (Nichols and Bruno 2010)". I am not arguing against studying folk intuitions about personal identity, neither regarding counterfactual nor hypothetical cases. What I am concerned with is the counterfactual method in itself; i.e., employing intuitions pumped by thought experiments that are at odds with the relevant facts of the actual world about persons and their identity to inform the de facto ontological make-up of real people.

[6] Persistent vegetative state is an obvious example where biological continuity obtains but psychological continuity has vanished.

[7] For a recent analysis of transplant intuitions in the debate on the metaphysical soundness of Animalism see Skrzypek and Mangino (2021).

[8] An anonymous referee pointed out that the empirical constraints not so much preclude cerebrum transplants from happening, but call into question the claim that such transplants are sufficient (or, indeed, necessary) to preserve personal identity when psychological and biological continuity come apart.

Table 1 Constraining facts: possible world vs. actual world

Possible world P	Actual world Q
Personal identity obtains iff $\{X_1, X_2, \ldots, X_n\}$	Personal identity obtains iff $\{Y_1, Y_2, \ldots, Y_n\}$
Where $\{X_1, X_2, \ldots, X_n\}$ is the set of constraining facts about personal identity in P	Where $\{Y_1, Y_2, \ldots, Y_n\}$ is the set of constraining facts about personal identity in Q

of embodied cognition (Clark 1997, 1999; Lakoff and Johnson 1999) highlight the interdependence of brain and body. Roughly, the cognitive science of embodied cognition holds that a person's mind is deeply dependent upon their bodily features. That is, aspects of a person's body beyond the brain play a significant causal or physically constitutive role in cognitive processing (Wilson et al. 2021). Even if one were able to successfully transplant an entire functioning brain (let alone just the cerebrum), the psychological make-up of the resulting person would be shaped and informed by the constitution of an altogether different body. Granting that the old body and the new were much alike, they'd inevitably still be ever so slightly different, and so would be the resulting person's psychological make-up. Schechtman (1997) has called this the 'Brain–body Problem' and presented an alternative 'Distributed View' of the mind which coheres well with evidence from cognitive science. A further line of empirical research suggests that there is a strong 'brain–body historicity' based on immunological mechanisms observed in brain tissue transplantations. The immune system distinguishes the body's own tissue from foreign tissue only on the basis of the quality of the inserted material, whereas the quantity of inserted material is largely irrelevant. Even if the quantity of foreign inserted material is small, the immune system may still reject it. Thus, from an immunological perspective, there appear to be no principled differences between brain tissue transplantations and entire cerebrum transplantations: both are subject to the close interdependence between brain and body (Munzer 1994). We cannot expect, therefore, to transplant a cerebrum into someone else's head, assuming that this would result in the original person's distinct psychology having been transplanted. Rather, the entire body's vital functions, including, but not limited to, the functioning cerebrum, are necessary to sustain a person's distinct psychological make-up—suggesting that psychological continuity supervenes upon biological continuity.

Psychological continuity—qua being constrained by contingent empirical facts about the human body's nature—coincides via nomological necessity with biological continuity. Imagining apart psychological continuity from biological continuity, as we are asked to do in cerebrum transplant Counterfactuals, so as to isolate psychological continuity as the dependent variable, and to substitute biological continuity with independent variables (or different causes of psychological continuity), violates the nomologically necessary interdependence of psychological and biological continuity. If psychological continuity supervenes upon biological continuity in all actual cases, the mere conceptual possibility of them coming apart can't serve as a valid source of intuition when it comes to puzzles about real-life persons.

Despite appearances, Transplant Intuition is not just unreliable when employed to inform judgments about personal identity in the actual world, but largely irrelevant. For, in stepping into the counterfactual perspective, we are intuiting about beings whose envisioned physiological constitution is decisively different from real-life persons, such that reapplying these intuitions back to the actual world constitutes a change of subject.

In the succeeding section, I abstract from cerebrum transplant Counterfactuals to argue, more generally, that intuitions about personal identity gathered from possible worlds that differ in their facts from the actual world to the point of conceptual incongruence, are inadequate when reapplied to real-life persons.

Intuiting across conceptually incongruent worlds

Counterfactuals consult modalities about possible worlds where things are (often strikingly) different from how they are in the actual world. Typically, physical constraints that preclude, say, fission or teletransportation from actually happening, are imagined away. The implicit assumption seems to be, then, that the concept of personal identity is insensitive to constraining facts of the actual world such that personal identity can be isolated, transferred to some possible world, tested in those possible conditions to finally reapply the intuitions so gathered by transferring them back to the actual world. In so doing, the gathered intuitions from possible worlds are applied to hypothetical cases to see whether the theoretical implications that were drawn out by counterfactual thought-experimenting appear intuitively plausible. If these theoretical implications do not appear plausible in light of the counterfactual, conceptual engineering is undertaken to adjust the theory of personal identity accordingly.[9]

[9] I am grateful to an anonymous referee for having me flesh out the idea of 'intuition-transfer'.

Table 1 compares a possible world *P* where, by stipulation, the set of facts that constrain personal identity is different from the set of facts that constrain personal identity in the actual world *Q*. Say X_1 about *P* allows for *insert your favorite Counterfactual*; whereas Y_1 about *Q* precludes said Counterfactual. We have made it true, by stipulation, that people inhabiting *P* survive (or, as the case may be, do not survive) changes enabled by X_1 that people in *Q*, because of Y_1, never face. It is hard to see how intuiting about whether people in *Q* would survive a scenario that never occurs[10] can, in principle, have a sensible—let alone accurate—answer. Since such things never happen to people in *Q*, the conceptual apparatus that has evolved in conjunction with facts about *Q* is ill-equipped to deal with such Counterfactuals. Employing intuitions gathered from pondering what happens to people in *P* to then make inferences about personal identity in *Q* is invalid because of factual incongruences between worlds. Thus, it is unsurprising that pondering about what happens to people in *P* where, say, teletransportation is possible, evokes intuitions that are invalid when reapplied to real-life persons inhabiting *Q*. Such Counterfactuals throw a spanner in the works, then, by leading us to question whether the concept of personal identity applies to these sorts of Counterfactuals, where, in fact, it does not.

Recall that according to Intrinsic Linkage there is an intimate relation between personhood and relevant constraining facts about the actual world. Had *Q* been different, such that X_1 would not obtain but Y_1 would obtain, the concept of personhood, too, would have evolved differently. Intrinsic Linkage suggests, then, that factual incongruences between $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_n\}$ imply conceptual incongruences between personhood in *P* and personhood in *Q* that renders intuiting across worlds invalid. What decides the validity of a thought experiment about personal identity is thus the question as to whether the 'intuition-inference' from *P* to *Q* implicitly attempts to carry over facts about *P* to *Q* that are incongruent. If so, intuiting across worlds is invalid.

One might object that, since we have historically been mistaken about facts of the actual world, and have based our concept of personal identity on these alleged facts that turned out to be erroneous, Intrinsic Linkage might not be as tight after all. For example, during the heyday of Dualism, it seemed plausible that personal identity is to be analyzed in terms of the persistence of an immaterial soul. With growing knowledge about the actual world, though, we came to realize that the existence of an immaterial soul is rather unlikely. Accordingly, the concept of personal identity has been adjusted, and the soul theory has largely been abandoned.

Rather than viewing our epistemic limitations as an objection to Intrinsic Linkage, our conceptual responsiveness to relevant facts about the actual world indicates that there is an intimate relation between personhood and these constraining facts. We are, and ought to be, prepared to revise our concept of personhood, given newly acquired evidence. In this spirit, Bakhurst (2005) contends, "the marks of personhood issue from facts about what we are, so that there can be weighty truths, presently obscure to us, the discovery of which would dictate how we should think of ourselves." Furthermore, our *epistemic limitations* that suggest an epistemologically contingent relation between personal identity and what we currently know about the actual world do not rule out an *ontological dependency* between personal identity and relevant constraining facts about the actual world. The correct theory of personal identity cannot be divorced from those facts, but must be responsive to them. When imagining away the actual world's constraints on persons and their identity, the conceptual boundaries of personal identity dissolve with them. At the very least, the concept of personal identity becomes so blurry that it no longer evokes reliable intuitions—let alone informing sensible judgments about real-life persons.

I now turn to look at how invalid intuition-inferences might be based on conflating de re necessity and de re possibility about persons, and how such a conflation is at odds with Orthodox Commitments. That is, what might constitute personhood in some possible world carries no weight on the whereabouts of real-life persons.

Violating orthodox commitments

To see how the common practice of invalidly intuiting across worlds might be connected to conflating different de re modalities about persons, it is useful to draw a distinction between de re necessity and de re possibility.

If, via Counterfactuals, we are to isolate what is conceptually essential about persons and their identity per se, the features that constitute personhood must be steady across worlds. Call this de re necessity about personhood, according to which,

[10] By an event 'never occurring' I mean that the event is impossible to occur due to the actual world's constraining facts—not just that the event has not yet occurred. An anonymous referee has rightly pointed to the epistemic limitations that we face in figuring out whether an event is impossible to occur necessarily or whether it just so happens that it hasn't occurred yet. Given the epistemic uncertainty regarding the future state of science and technology, I do not mean to suggest that counterfactually generated philosophical intuitions should not be accorded any role at all in theorizing about personal identity. For example, counterfactual thought experiments can be useful to carve out differences between rival views that appear less strikingly so in ordinary cases.

> in every possible world containing persons, persons are F; whereby F contains every constitutive feature all and only persons possess necessarily.

As per Cognitivism, one such feature is higher-order cognition. However, there are logically possible worlds where persons are just like us; except, they have no psychological features at all (akin to Chalmers's Zombies). Such Zombie-like persons might still employ the de dicto practice of successfully ascribing personhood to each other in their respective possible world. Zombie-like persons might, for example, be held morally accountable for their actions; not based on any actual mental life though, but solely based on their behavior. Cognitivism, then, is not steady across worlds. It is, however, a widely agreed upon feature of real-life persons that few philosophers are willing to drop.

If, via Counterfactuals, we take the more moderate aim to isolate what is conceptually essential about persons and their identity per alia, the features that constitute personhood must be steady only within worlds. Call this de re possibility about personhood according to which,

> in at least one possible world containing persons, persons are G; whereby G contains every constitutive feature all and only persons possess necessarily in that particular world.

This might well be true; however, we cannot expect de re possibility to enable valid inferences across worlds. Zombie-like persons (lacking higher-order cognition, and any sort of consciousness, for that matter), for example, might very well count as persons in some possible world. But that does not yield any insights into real-life persons that are constituted differently (possessing higher-order cognition that enables them to track themselves over time).

If Cognitivism is true, persons are equipped with higher-order cognition that supervenes upon biological facts about their brains and bodies. Personal identity can, then, only be analyzed properly by being responsive to biological facts about higher-order cognition.

Furthermore, both de re necessity and de re possibility about personhood are in tension with Realism. If 'person' is a natural kind, persons cannot exist outside of their natural habitat. We cannot, it seems, have it both ways: holding on to Realism and conceptually removing persons from the actual world, placing them in some possible world. There is a related problem with holding on to Naturalism: if persons are biological beings whose existence is a matter of empirical facts, it is conceptually erroneous to disregard these empirical facts when considering Counterfactuals, and, simultaneously expect inferences drawn from pondering such scenarios to be informative regarding real-life persons.

Intuiting from Counterfactuals to make claims about real-life persons and their identity, both in terms of de re necessity and de re possibility, thus requires dropping any number of Orthodox Commitments. Biting the bullet, though, comes at a high price that, presumably, few philosophers are prepared to pay.

Having tidied up some of the conceptual muddle, I now turn to shed light on a few promising strategies that have recently been put forward to deal with troublesome implications of theoretical convictions in personal identity derived from counterfactuals when it comes to real-life cases.

Theoretical convictions, empirical studies, and real-life applications

The previously mentioned studies of folk intuitions regarding personal identity (footnote 5) point to the recent rise of experimental methods in philosophical discussions on personal identity. These empirically-informed approaches are an important step towards challenging the weight accorded to armchair theorizing in discussions of healthcare and health policy issues related to personal identity.[11]

In an online survey, Strohminger and Nichols (2014) investigated—employing case-study experiments (including a version of the brain transplant)—that moral traits, rather than other cognitive functions, are perceived to be the most integral part of personal identity. Contrary to theoretical convictions of psychological continuity views that do not properly account for the importance of differences in psychological traits in preserving personal identity, these findings suggest that folk notions of personal identity are largely informed by what the authors call the 'essential moral self', according to which the mental faculties affecting social relationships, with a particularly keen focus on moral traits, are most relevant to personal identity.

In a follow-up study, Strohminger and Nichols (2015) studied changes in personal identity in patients with various kinds of neurodegenerative diseases (dementia, Alzheimer's, and amyotrophic lateral sclerosis) as perceived by patients' relatives. Participants were told that the purpose of the research was to investigate how the neurodegenerative disease affected personal relationships. Accordingly, participants were asked questions indicative of how much the patient had changed since the onset of the disease. The study's results suggest that damage to the moral faculty is particularly threatening to the personal identity of patients as perceived by relatives. While other cognitive deficits did not show measurable impact on personal identity. Neurodegenerative diseases such as frontotemporal dementia that attack the brain's moral processing faculty

[11] I am grateful to an anonymous referee for urging me to discuss the important experimental work on personal identity.

have shown the greatest effect on perceived change in personal identity; whereas neurodegenerative diseases such as amyotrophic lateral sclerosis that affect mostly cognitive processing have shown the least effect on perceived change in personal identity. Needless to say, the presumed personal identity of patients as third-personally perceived by relatives cannot simply be converted into what patients themselves experience first-personally, and so there are limits to what we can learn from these results. But if these studies are at all indicative of what people want for themselves in terms of medical treatment, the results should be taken seriously into account when it comes to advance directives.

More generally, these kind of empirical studies have a potentially important impact on theorizing about personal identity since they call into question widespread views according to which personal identity consists mainly in unspecified psychological continuity. At the very least, these results indicate that a more fine-grained theoretical analysis of just which features of psychological continuity in detail are identity-preserving is required; suggesting that the coarse-grained concept of psychological continuity is not equipped with the necessary conceptual sophistication needed to inform real-life decisions about personal identity. Furthermore, if taken at face value, the 'essential moral self' view has serious ramifications for personal identity in light of moral enhancement. If moral traits are essential to personal identity, altering one's moral traits via, say, pharmacological intervention could, in principle, change a person to the point of becoming a different person altogether. Crutchfield (2018) goes so far as to suggest that moral enhancement can 'kill' the enhanced person.

With regard to the potential threat that DBS poses to personal identity, Bluhm et al. (2020) have recently made the case for utilizing empirical data gathered from patient reports both for improving patient care, and informing theories of personal identity. Their findings suggest the actual experiences of patients having undergone DBS cohere much more with a relational understanding of personal identity, according to which personal identity is formed within a web of social relations (Schechtman 2014), than with metaphysical reductionism about personal identity, where persons can be entirely reduced to the existence of certain psychological and/or biological states and their various relations (Parfit 1984). Bluhm et al. (2020) report, for example, that the majority of people that undergo DBS do not feel that they have changed in any fundamental way after DBS; at least not more so than the alterations they have experienced as a result of their illness, or of pharmacological treatments. Taking these patient narratives seriously, then, leads to a more nuanced reading of these reports that may have concrete, practical and theoretical implications. Contextualizing actual patients' experiences calls for asking 'what is it like to be a person being treated with DBS' rather than asking 'whether DBS is a threat to personal identity'. This is not just a semantic sleight of hand, but might contribute to a better understanding of what actually happens to DBS patients' personal identity, and thus help create tailor-made health care policies.

Concluding remarks

I have argued against the adequacy of employing counterfactual thought experiments that are at odds with relevant facts about the actual world to make claims about real-life persons and their identity. Personhood is deeply rooted in the actual world's constraining facts such that persons can't be conceptually isolated from the inner workings of the world they inhabit, without changing the concept fundamentally. In so doing, we are not talking about the identity of persons *as we know them* anymore, but about the identity of imaginary persons* instead. What makes persons* persist, however, carries no weight for real people.

If my arguments are on the right track, the onus lies with proponents of Counterfactuals to demonstrate the need and adequacy of reverting to such scenarios. Rather than pondering bizarre Counterfactuals, it might be worthwhile taking more seriously real-life puzzles that are far from solved. What happens to the identity of persons suffering from disorders of consciousness (such as persistent vegetative state) or neurodegenerative diseases (such as late stages of Alzheimer's), for example, is extensively discussed in the medical ethics literature of advance directives. Given the ontological dependency between personhood and the relevant actual world's constraining facts, these and other conditions might deserve more theoretical attention than they currently garner. Resulting empirically-informed theories of personal identity will be both ontologically more plausible, and better able to shed light on novel clinical applications that potentially alter real people's identities.

Acknowledgements I am grateful to two thoughtful referees at *Medicine, Health Care and Philosophy* for their constructive feedback that helped improving an earlier version of this article. Special thanks to Simon Beck, Niël Henk Conradie, Carl Friedrich Gethmann, Owen King, and Thomas Schirmer for insightful written input and fruitful discussions.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Bakhurst, D. 2005. Wiggins on persons and human nature. *Philosophy and Phenomenological Research* 71 (2): 462–469.

Baker, L. 2000. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.

Beck, S. 2006. These bizarre fictions: thought-experiments, our psychology and our selves. *Philosophical Papers* 35 (1): 29–54.

Beck, S. 2016. Technological fictions and personal identity: On Ricoeur, Schechtman and analytic thought experiments. *Journal of the British Society for Phenomenology* 47 (2): 117–132.

Coleman, S. 2000. Thought experiments and personal identity. *Philosophical Studies* 98 (1): 51–66.

Dennett, D. 1976. Conditions of personhood. In *The Identities of Persons*, ed. Amelie O. Rorty, 175–196. Berkeley: University of California Press.

Gendler, T. 1998. Exceptional persons: on the limits of imaginary cases. *Journal of Consciousness Studies* 5 (5–6): 592–610.

Gendler, T. 2002. Personal identity and thought-experiments. *Philosophical Quarterly* 52 (206): 34–54.

El Haj, M., J. Roche, K. Gallouj, and M.C. Gandolphe. 2017. Autobiographical memory compromise in Alzheimer's disease: A cognitive and clinical overview. *Geriatr Psychol Neuropsychiatr Vieil* 15 (4): 443–451.

Kipper. 2016. Substance and the concept of personal identity. *Ergo* 3 (1): 1–26.

Marquis, D. 1989. Why abortion is immoral. *Journal of Philosophy* 86 (4): 183–202.

Nyholm, S., and E. O'Neill. 2016. Deep brain stimulation, continuity over time, and the true self. *Cambridge Quarterly of Healthcare Ethics* 25 (4): 647–658.

Olson, E. 2007. *What Are We? A Study in Personal Ontology*. New York: Oxford University Press.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon.

Reid, T. 1785/1969. *Essays on the Intellectual Powers of Man*. Cambridge: MIT Press.

Shoemaker, S. 2008. Persons, animals, and identity. *Synthese* 163 (3): 313–324.

Skrzypek, J.W., and D. Mangino. 2021. Should animalists be '"Transplanimalists"'? *Axiomathes* 31: 105–124.

Parfit, D. 2012. We are not human beings. *Philosophy* 87 (1): 5–28.

Strohminger, N., and S. Nichols. 2014. The essential moral self. *Cognition* 131 (1): 159–171.

Nichols, S., and M. Bruno. 2010. Intuitions about personal identity: An empirical study. *Philosophical Psychology* 23 (3): 293–312.

Oderberg, D. 1997. Modal properties, moral status and identity. *Philosophy and Public Affairs* 26 (3): 259–298.

Buchanan, A. 1988. Advance directives and the personal identity problem. *Philosophy and Public Affairs* 17 (4): 277–302.

DeGrazia, D. 1999. Advance directives, euthanasia, and the someone else problem. *Bioethics* 13 (5): 373–391.

McMahan, J. 2003. *The Ethics of Killing*. New York: Oxford University Press.

McInerney, P. 1990. Does a fetus already have a future-like-ours? *Journal of Philosophy* 87 (5): 264–268.

Lakoff, G., and M. Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Snowdon, P. 2014. *Persons, Animals, Ourselves*. New York: Oxford University Press.

Vollmann, J. 2001. Advance directives in patients with Alzheimer's disease; ethical and clinical considerations. *Medicine, Health Care and Philosophy* 4 (2): 161–167.

Warren, M. 1977. Do potential people have moral rights? *Canadian Journal of Philosophy* 7 (2): 275–289.

Wiggins, D. 2001. *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.

Olson, E.T. 2016. The role of the brainstem in personal identity. In *Animals: New Essays*, ed. A. Blank. Philosophia: Verlag.

Wilkes, K. 1988. *Real People*. Oxford: Clarendon.

Müller, S., M. Bittlinger, and H. Walter. 2017. Threats to neurosurgical patients posed by the personal identity debate. *Neuroethics* 10 (2): 299–310.

Munzer, S. 1994. Transplantation, chemical inheritance, and the identity of organs. *British Journal for the Philosophy of Science* 45 (2): 555–570.

Schechtman, M. 1997. The brain-body problem. *Philosophical Psychology* 10 (2): 149–164.

Clark, A. 1999. Embodied, situated, and distributed cognition. In *A Companion to Cognitive Science*, eds. W. Betchel, and G. Graham, 506–517. Malden: Blackwell Publishing.

Clark, A. 1997. *Being There: Putting Brain Body and World Together Again*. Cambridge, Massachusetts: MIT Press.

Limbaugh, D. 2016. Animals, advance directives, and prudence: Should we let the cheerfully demented die? *Ethics, Medicine and Public Health* 2 (4): 481–489.

Locke, J. 2012. *An Essay Concerning Human Understanding*. Oxford: Clarendon.

Blok, S., G. Newman, and L.J. Rips. 2005. Individuals and their concepts. In *Categorization Inside and Outside theLaboratory*, ed. W.-K. Ahn, R.L. Goldstone, B.C. Love, A.B. Markman, and P. Wolff, 127–149. Washington, DC: American Psychological Association.

Lipsman, N., and W. Glannon. 2013. Brain, mind and machine: what are the implications of deep brain stimulation for perceptions of personal identity, agency and free will? *Bioethics* 27 (9): 465–470.

Berniūnas, R., and V. Dranseika. 2016. Folk concepts of person and identity: A response to Nichols and Bruno. *Philosophical Psychology* 29 (1): 96–122.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68 (1): 5–20.

Wilson, Robert A., Lucia Foglia, Lawrence Shapiro, and Shannon Spaulding. 2021. "Embodied Cognition", The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.).

Bluhm, R., L. Cabrera, and R. McKenzie. 2020. What we (should) talk about when we talk about deep brain stimulation and personal identity. *Neuroethics* 13: 289–301.

Crutchfield, P. 2018. Moral enhancement can kill. *Journal of Medicine and Philosophy* 43 (5): 568–584.

Strohminger, N., and S. Nichols. 2015. Neurodegeneration and identity. *Psychological Science* 26 (9): 1469–1479.

Shoemaker, S. 1963. *Self-knowledge and Self-identity*. Ithaca: Cornell University Press.

Schechtman, M. 2014. *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford, UK: Oxford University Press.

Philos. Technol. (2019) 32:135–154
https://doi.org/10.1007/s13347-018-0313-7

RESEARCH ARTICLE

Doing Away with the Agential Bias: Agency and Patiency in Health Monitoring Applications

Nils-Frederic Wagner[1]

Received: 14 July 2017 / Accepted: 9 April 2018 / Published online: 23 April 2018

Abstract Mobile health devices pose novel questions at the intersection of philosophy and technology. Many such applications not only collect sensitive data, but also aim at persuading users to change their lifestyle for the better. A major concern is that persuasion is paternalistic as it intentionally aims at changing the agent's actions, chipping away at their autonomy. This worry roots in the philosophical conviction that perhaps the most salient feature of living autonomous lives is displayed via agency as opposed to patiency—our lives go well in virtue of what we do, rather than what happens to us. Being persuaded by a device telling us how to conduct our lives seemingly renders the agent passive, an inert recipient of technological commands. This agential bias, however, has led to a marginalization of patiential characteristics that are just as much part of our lives as are agential characteristics. To appreciate the inherent interlocking of acting and being acted upon, it is vital to acknowledge that agency and patiency are correlates, not mutually exclusive opposites. Furthermore, it is unclear whether an action can only count as agential so long as its causes are internal. Drawing on the extended mind and extended will framework, I argue that mHealth applications merely serve as volitional aids to the agent's internal cognition. Autonomously set goals can be achieved more effectively via technology. To be persuaded by an mHealth device does not mainly—let alone exclusively—emphasize patiency; on the contrary, it can be an effective tool for technologically enhancing agency.

Keywords Mobile health monitoring · Smartphone applications · Persuasion · Autonomy · Paternalism · Agency and patiency

✉ Nils-Frederic Wagner
nils-frederic.wagner@uni-due.de

[1] Department of Philosophy, University of Duisburg-Essen, Room: LE 329, Forsthausweg 2, 47057 Duisburg, Germany

1 Introduction

It is estimated that by 2020, roughly 70% of the world's population will be using smartphones; currently, the worldwide dissemination is at around 50%.[1] While such devices continue to be used mostly recreationally, the burgeoning tendency to *smarten up* our lives has long reached the medical sphere. New technological means enable an ever-increasing number of users to track and analyze a vast amount of sensitive, personal health-related data. Mobile health applications (commonly referred to as "mHealth apps") that typically run on smartphones are the most pervasive form of such devises. A recent study across 16 countries found that, as of today, around one third of the online population uses mobile health devices.[2] Core features of most such apps already at hand are the tracking of movements, nutrition, and sports activities. Physiological parameters like heart rate and blood pressure that are known to highly correlate with emotional states, and wellness factors such as quality of sleep, and social interaction are also monitored but may require gadgets like chest straps or heart rate watches. These smart devices facilitate a previously unheard-of efficacy of self-monitoring. With the rise of such novel, intimate technologies, a variety of philosophical issues crop up concerning, most pertinently, data security, responsibility, paternalism, autonomy (Krieger 2013; Owens and Cribb, forthcoming), as well as conflicts of interest between different stakeholders.

Concerns about users' autonomy become ever more pressing since a growing number of such applications do not merely collect data, but also aim at persuading users to change their lifestyle for the better, i.e., living a healthier, more active life. To achieve this goal, app designers utilize a variety of persuasive strategies that potentially erode users' autonomy and threaten their agency.

Here's a case that shall help illustrating the concern. Suppose you are going for lunch with your colleague. When you are about to order, a message from your recently acquired mHealth app pops up telling you to have the healthy salad option (needless to say, you would much rather have fish and chips). Having salad, the device says, will lower your cholesterol level which in turn will make you feel better in the long run. From automatically checking the weekly cafeteria menu online, the device knows the food options and comes up with the healthiest choice for everyday. This is done according to a complex algorithm, taking into account both your physical parameters as well as all previously chosen meals ever since you have been using the device to ensure that your diet is more balanced. You are in a clear epistemic disadvantage here; matters that are considered by the app to generate choices are just too complex for you to fully comprehend. That way, the app has an expertise you lack. Let us suppose further that if you go for salad, the device rewards you by allocating health points to your account, say in the form of green leaves. Since this app is quite popular among your friends, you have entered into a competition. Whoever has collected most green leaves by the end of the year will be declared winner and can look forward to being invited to a fancy getaway by the other participants. Lots of reasons to go for salad, it seems. But what if you cannot resist your cravings and go for fish and chips anyway? In

[1] https://techcrunch.com/2015/06/02/6-1b-smartphone-users-globally-by-2020-overtaking-basic-fixed-phone-subscriptions/ (retrieved May 2017).
[2] http://www.gfk.com/global-studies/global-studies-fitness-tracking/ (retrieved June 2017).

that case, the app will deduct the green leaves you earned yesterday and will also send a message to your friends letting them know that you have indulged in some tasty but unhealthy meal.

When imagining such cases, most of us, I take it, feel a certain unease regarding our autonomy. Is it really you deciding to have salad for lunch? Or is it the app deciding for you? Since you are at an epistemic disadvantage, the app knows better what is good for you anyway. So, on the face of it, it looks as though you have been paternalized by some technological device, perhaps a non-human agent, using more or less subtle elements of persuasion (maybe even manipulation) to impose its will on you. Also, it employs motivational triggers by introducing elements of reward and competition to nudge you into complying.

In what follows, I focus on whether persuasive mHealth apps do indeed dislodge autonomy by constituting a paternalistic intervention into people's lives. I argue that, despite appearances, there are good reasons to believe that these systems do not per se pose a threat to agency. Under certain conditions, mHealth apps even bear the potential to technologically ameliorate agency.

I proceed as follows: Firstly, I present both paradigmatic views that argue for mHealth apps' potential to enhance users' autonomy and paradigmatic views that argue for mHealth apps' potential to threaten users' autonomy. I then sort out the conceptual interrelation of autonomy, persuasion, and paternalism in the context of mHealth apps. Subsequently, I make some remarks on how these concepts figure in the common understanding of agency and patiency in philosophy of mind and action. I argue that a widespread agential bias has led to an underappreciation of patiential concerns that make up significant proportions of our lives. Finally, I make the case for understanding some persuasive elements of mHealth apps as what I shall call "volitional aids" which are, considering the theories of extended mind and extended will, part of the agent's own, albeit extended cognitive architecture as opposed to external interferences, suggesting that, understood in this way, some of these apps are effectively outsourced parts of our minds and wills that potentially enhance agency.

2 Autonomy, Persuasion, Paternalism

Owens and Cribb are among those who have recently argued for the autonomy enhancement potential of mHealth apps. In particular, they think that such apps can foster users' deliberation and decision-making capacities: "By providing access to biomedical data and generating awareness of habits, behaviours and performances, there is good reason to think these technologies can support processes of deliberation about health that enhance their users' procedural autonomy. For example, information about one's heart rate, sleeping patterns, mobility or calorific intake might help people make important decisions that directly affect their health" (Owens and Cribb, forthcoming, 5).

Recent empirical studies indicate that some users feel autonomous and motivated when employing mHealth apps in their daily routine. It remains elusive how much theoretical weight should be given to a sample of users' assessment of such apps. Nonetheless, as the following summary of user reports illustrates, there is reason to at least surmise some autonomy enhancement potential: "Users reflected positively on the

use of the apps, with one user felt that the autonomy-supportive style was evident in terminology used. Users felt motivational value from seeing steps, styles and advice. User attitudes reinforced autonomy stating it made use of the device more engaging and positively influenced sustained or repeat use. Generally, users enjoyed the level of autonomy they were granted by the apps. However, some stated a need for apps to balance autonomy with more self-directed goal creation to support their engagement" (Asimakopoulos et al. 2017, 7).

In contrast to the views such sketched, several scholars have argued that technology in health care in general, and mHealth apps in particular, pose threats to users' autonomy. In what follows, I highlight some of the most pressing worries such views articulate before relating these views to an analysis of persuasion and paternalism in this context.

Timmer et al. argue that due to new technological means of persuasion it "might be harder for the individual to make an autonomous choice about the goals he is being persuaded to, or whether he consent[s] to the use [of] persuasive technologies" (Timmer et al. 2015, 196). They argue that safeguarding autonomy in these new means of persuasive technology is even more important when the setting of their application takes place in sensitive contexts like health care (ibid., 197). When persuasive technologies appear in what the authors call "collective applications," "for instance in healthcare and insurance—research is needed on the role of these third parties as providers of persuasion and how they impact the users' autonomy" (ibid., 201).

Several authors that critically engage with new mHealth technologies such as Lanzing point to a tension between disclosing sensitive personal information and safeguarding one's autonomy: "self-tracking breaks down informational privacy boundaries that otherwise enable autonomous self-presentation within different social contexts" (Lanzing 2016, 10). Lanzing further thinks that users' autonomy is comprised because of a potential breach of information privacy in mHealth apps, since, one the one hand, users are encouraged to collect and share as much data as possible, both to increase functionality and persuasion. But on the other hand, privacy of information is a hallmark of living autonomously. Success stories about empowerment, self-control, and self-improvement camouflage the reality of decontextualization, thinks Lanzing, where we expose too much to an undefined (future) audience, which limits our capacity to run our lives for ourselves. Altogether, this constitutes a violation of users' privacy that can undermine their autonomy on a more fundamental level (ibid., 15).

Nordgren also places particular importance on privacy in personal health monitoring, submitting that frequently "the user has no autonomy regarding which information is to be collected, transmitted, processed and used" (Nordgren 2015, 155). To avert this issue, Nordgren suggests a context-sensitive balancing of automated privacy protection that might be feasible in some circumstances and autonomously chosen privacy protection that might be called for in other circumstances (ibid., 163).

Sharon holds against the idea of mHealth apps as empowering users that, "self-tracking for health is disempowering, insofar as it invites an increased control of others—health promoters, friends and followers, and even the internalized health promoter of one's own super ego—over oneself" (Sharon 2017, 99). She further posits that "discourses of empowerment and healthy citizenship are seen as concealing economic realities that are often detached from the interests of citizens and patients

and of creating new forms of discipline, subjection, and social control—of imposing limits on the autonomy of individuals" (ibid., 106f.).

Even though these views focus on different aspects of autonomy, they commonly see threats to users' autonomy as a problematic interference with their agency that ought to be circumvented since agency is something worth aspiring to when it comes to living well. For this reason, I describe such views as having an "agential bias." In Section 3, I say more about how this agential bias might be rooted in the common tendency in philosophy to see our lives as going well, first and foremost, in virtue of agential features.

Although most of the literature leans towards either of the two sides just sketched, some authors see both negative and positive aspects of personal health monitoring regarding users' autonomy. Here is one such view: "However, PHM [personal health monitoring] can also restrict the lifeworld, impinging the system's economic and power concerns on the individual lifeworld such that restrictions are placed or information demanded in order to maintain institutional structures. Hence PHM has the potential to act both as the repressive father, dictating behaviour and routine and demanding information for his own purposes, or the supportive mother offering both reassurance but also an environment which supports the autonomy of the patient" (Mittelstadt et al. 2014, 50).

2.1 Key Aspects of Persuasive mHealth Devices

The semantics of persuasion plays a central role in assessing the issue as to whether its application in mHealth apps poses a threat to users' autonomy, possibly eroding their agency. To a first approximation, persuading someone to doing (or omitting)[3] something is to intentionally try changing their actions via their convictions and intentions that lead to action. Whereas persuasion has a largely positive connotation in social psychology and health science (Cialdini et al. 2005), its reputation in the philosophy of action is rather seedy; sometimes, persuasion is seen as akin to (or at best in between) manipulation and convincing (O'Keefe 2012).

Persuasive technologies are generally designed such that they provide technological means to intentionally, and often permanently, change users' behavior via their convictions and intentional states by constantly providing feedback on what is understood as inadequate behavior and by incentivising in various ways what is deemed desired behavior (cf. Fogg 2003).[4] Persuasion is thus inherently normative as its main rationale is not to merely describe something or inform someone, but to persuade, or as some argue, to manipulate people into doing something.

Key characteristics of persuasive technologies in the context of mHealth applications are that they work with body sensors, implemented on smart devices such as smartphones and smart watches that continuously collect and display the recorded data, that they provide real-time persuasive feedback, that they function widely automated without the need for human control, that they are customized so as to

[3] Whenever I talk about "doing" in the context of persuasion and autonomy, I take "omitting" to be implied. For brevity's sake, I hereafter omit "omitting."

[4] Davis et al. ' (2015) scoping review surveys health-related behavior change theories as they are put forward in the social sciences.

accommodate users' specific needs, and that they are context-sensitive, taking into account users' current condition (Koelle et al. 2014). Furthermore, the design of such applications is supposed to carefully consider technological, social, and interactive components of human-computer interfaces that enable a smooth human-computer interaction.

Even though persuasion is generally taken to be a sustained effort to change someone's behavior, with paternalism and patronization lurking in the shadows, ideally, there are various measures in place that prevent technological persuasion from being unethical: (1) the desired behavior change should be achieved without deception (i.e., neither in terms of concealing the striven-for outcome nor in terms of disguising the measures taken to achieve that goal), (2) users should voluntarily decide to use such technologies, and (3) the intended goals of persuasion should be kept transparent (Chatterjee and Price 2009), as much as this is possible without running into problems of *persuasive backfiring* (i.e., the triggering of unintended outcomes of behavior change).

Before discussing principles of ethically sound persuasion in more detail, I turn to motivate potential threats of paternalistic interventions to people's autonomy. Firstly, I discuss defining conditions and standard cases of paternalism and then relate main elements of these scenarios to technologically aided paternalistic interventions.

2.2 Paternalistic Interventions and the Principle of Respecting Autonomy

The widely held principle of respecting autonomy suggests that persuasion chips away at the agent's autonomy since it figuratively (or, as the case may be, literally) talks the agent into doing something they would not have done otherwise, out of their own free will. It is frequently argued that such persuasive interventions violate the principle of respecting autonomy since they constitute a form of paternalism, interfering with the agent's own volition; albeit motivated by the conviction that the agent will be better off when persuaded into changing their behavior accordingly (Enoch 2016). Some hold that the basis for a behavioral adaptation so achieved is not that the agent was rationally convinced to do so, but rather deceptively manipulated into doing so (Spahn 2012). This line of thought stems from persuasion's rather shady reputation in philosophy, where it is often seen as inherently paternalistic and thus at odds with respecting the agent's autonomy.

In numerous writings, perhaps most succinctly in the *Stanford Encyclopedia of Philosophy*, Gerald Dworkin (1972, 2005, 2015, 2017) essays a definition of paternalistic interventions that helps understanding why persuasion is often held in disesteem in philosophy, and why it might be seen as posing a threat to agency.

Dworkin suggests the following conditions as an analysis of *X acts paternalistically towards Y by doing* (*omitting*) *Z*:

1. *Z* (or its omission) interferes with the liberty or autonomy of *Y*.
2. *X* does so without the consent of *Y*.
3. *X* does so only because *X* believes *Z* will improve the welfare of *Y* (where this includes preventing his welfare from diminishing) or in some way promote the interests, values, or good of *Y*.

Even though Dworkin does not explicitly say that *X* and *Y* are two distinct agents, each having their own set of intentions, his analysis implicitly suggests that that is what he has in mind. Paternalism, then, occurs when one agent imposes their will on another agent with the benevolent, albeit patronizing intent of promoting (or at least preserving) the other's well-being or interests more generally.[5]

Given the involvement of two distinct agents in paternalistic interventions à la Dworkin and others, (2) in conjunction with (1) constitutes a violation of *Y*'s autonomy *eo ipso*. One cannot both respect someone's autonomy and, at the same time, interfere with their autonomy by acting on them without having previously obtained informed consent. However, the conjunction of (1) and (2) leaves open whether interfering with someone's autonomy were morally permissible if the interfered-with, acted-upon, or paternalized agent had given their consent.[6] (3) subverts *Y*'s autonomy inasmuch as it implies that *X* is in a better epistemic position to judge (or is in some other way more competent) than *Y* themselves with regard to figuring out what actions and intentions that lead to action are conducive to *Y*'s well-being; thus questioning their decision-making capacity, rationality or whichever agential feature might be required for the task at hand. (1) most obviously relies on the assumption that the paternalistic action *Z* is performed or initiated by *X*, whereby *X* is an agent in their own right, having beliefs and intentional states they want to impose on *Y*. Given that (2) supposes that *X* performs *Z* on *Y* without their consent, evidently, *X* and *Y* are taken to be two distinct agents.[7] Now, while this might hold true in standard cases of paternalism (for example, a parent taking away their drunken child's car keys to prevent them from causing an accident or from getting pulled over for DUI), when mHealth apps are concerned, it is far from clear whether there actually are two distinct agents at play. The question, then, becomes: are such apps "covert agents" pushing someone else's agenda by proxy with the means of technological persuasion, or are these apps, perhaps, just an extension of the agent's own volition? It goes without saying that the "covert-agent-concern" largely depends on the app-creator's intentions and on its design. For example, an app commissioned by an health insurance company with the aim of reducing costs by persuading their policyholder to, say, quit smoking, might well be paternalistic in that another agent (a team of app-designers on behalf of the company's CEO) tries imposing their will on users by means of technological persuasion. But importantly, this need not always be the case, nor is this necessarily so in mHealth apps.

Dworkin's three conditions of paternalism rely on the prima facie tenable assumption that *X* and *Y* are distinct agents, each representing their own set of intentions. Paternalism, then, seems to occur just in case *X* acts upon *Y* by meeting at least one of Dworkin's three conditions. While this might be reasonable in (2), since this condition requires another agent *ipso facto*, it is not necessarily true in (1) and (3). I venture that

[5] This reading is suggested by many other accounts of paternalism. Michael Cholbi (2017), to name just one example, has recently put forward a version of rational will that ranks the wrongfulness of paternalistic interventions in terms of the extent to which such acts replace the paternalizee's practical rationality with the paternalizer's and the degree of mistrust in the paternalizee's rational agency displayed by the paternalistic intervention.

[6] This question requires a separate treatment, but is not of central importance for my purposes.

[7] The issue as to whether one can, in principle, interfere with one's own autonomy or whether this necessarily requires another agent is interesting, but a detailed analysis thereof must remain a task for another day. At first blush, it looks as though actively deciding to renounce one's autonomy is itself an act of autonomy.

(1) and (3) need not always involve another agent. Why is that? I can, in principle, act upon myself paternalistically for example by forming distal intentions and putting measures in place that will make my future self comply. Think of, for example, Ulysses tying himself to the mast to resist the Sirens' song, or Parfit's Russian nobleman requesting his wife to hold him to the promise to distribute large portions of his wealth once he reaches a certain age even though his older self might have a change of heart; although, this case is less clear, since the anticipated change in attitude does not need to involve a decline in rationality. Nevertheless, once the implicit claim that being acted upon inevitably requires two distinct agents is dropped, it is much more contentious whether paternalism always interferes with agents' autonomy. In fact, a situation where *X* acts upon *Y*, where *X* is not a distinct agent but a technological device acting *in the service* of *Y* (i.e., as an extension of, and not in opposition to, *Y*'s will), might not constitute a case of paternalism at all—surely, it does not seem to pose an obvious threat to the agent's autonomy. It might even be a genuine expression of the agent's autonomy.

Before returning to the idea of expressing one's autonomy through being acted upon more thoroughly, I shall address the following question: What is it with this apparent contrast of paternalism and agency? It appears that paternalistic interventions address us primarily as patients and thereby chip away at our autonomy. In what follows, I want to change what I take to be a misguided focus by suggesting that patiency need not be agency-eroding but can, under certain circumstances, be a display of agency. Admittedly, expressing one's autonomy by being acted upon is unusual and difficult to grasp since autonomy is ordinarily displayed by one's actions, not by what happens to us (for lack of a proper noun; "inaction" does not seem to do the trick since we are not necessarily inactive when something happens to us). The underlying dichotomy between actions and things that happen to us plays a crucial part in agency's reputation as displaying inherently active features of agents' lives. So, it might be worth taking a closer look at the allegedly opposing concepts of agency (as in acting) and patiency (as in being acted upon). Such an analysis shall help revealing that, perhaps, there is not such a sharp divide between these two aspects of people's lives after all.

I now turn to make some remarks on the common tendency in philosophy of mind and action to underappreciate patiential traits of our lives and to spell out how this agential bias might have given rise to the misconception of technological persuasion as agency-eroding. In a subsequent step, I hope to show that exercising patiential characteristics of people's lives can, under certain conditions, enhance agency—if not paradigmatically, then at least more commonly than initially thought.

3 The Agential Bias

The wide notion of "patients" and "patiency" as technical terms in philosophy has a different connotation than the narrow notion of patients in ordinary language, particularly in healthcare settings. Philosophically, being a patient describes, broadly, the passivity of someone who undergoes some action or to whom something is done. This passivity of patients is chiefly contrasted with the activity displayed by agents. Roughly, the philosophical contrast between agents and patients is captured by the slogan: "agents do things, whereas things are done to patients." This philosophical dichotomy

is initially independent of the settings in which agents act and patients are acted on. The medical notion of patients, which is commonplace in ordinary language, locates patients in the vicinity of health care. Patients are thus people that suffer from a medical condition for which they receive medical treatment, either in outpatient or inpatient care. In what follows, I am mainly concerned with the philosophical notion of patiency. In the context of mHealth apps, however, "philosophical patients" are in some sense also "medical patients," but only contingently so. My arguments do not, therefore, rely on the medical notion of patients. I say more about this in Section 3.1.

Agency enjoys a considerable privilege in philosophy of mind and action, as does moral agency in ethics. Philosophers often emphasize that our lives go well in virtue of what we *do*, rather than in virtue of what *happens to* us (Lott 2016).[8] A paradigmatic way of phrasing the agential bias is put forward by Mark LeBar when he says that his view of living a good human life "is agentist, not patientist … we are first of all agents, who live by acting on their world" (LeBar 2013, 69 f.). To be an agent is, by and large, to actively partake in life; agents mold the world around themselves. Patients, on the other hand, are people to whom things happen; passive sufferers, molded by life's happenings. On that view, when our lives begin, we start out as depended patients, and, if everything goes well, in the course of our adult lives, we evolve into fairly independent agents. It goes without saying that this is but an ideal we aspire to, never to be fully achieved.

To appreciate the conjunction of acting and being acted upon, it is important to acknowledge that agency and patiency are *correlates*, not mutually exclusive opposites. Soran Reader (2007) neatly describes how the active features of agents' lives that are taken to be exclusive to agency (on her account action, capability, choice, and independence) all have a corresponding "other side" to them (on her account, passion, liability, necessity, and dependency, respectively). Reader characterizes this other side of agency as "a complementary aspect which necessarily accompanies the aspect valorised as 'positive' and assumed to furnish the essence" (ibid., 588) of what makes an agent.

Focusing on action, Reader carves out two aspects in which agency is necessarily accompanied by its complementary other side: Firstly, she claims that agents themselves suffer from their action, and thus always are, inevitably, in some relevant sense, patients as well as agents. For example, when I ride my bike, pedaling hard, I do not just move my bike forward, but I also suffer the pedals' resistance. Another example Reader cites is that when I hit you, it is not just you that suffers from my punch, but it is also I that suffers from your resistance to the blow. In the second sense, according to Reader, every action requires a patient at the receiving end of that action. The person being hit in the previous example is a most obvious case of a patient. So, every action requires both an agent initiating the action (whereby the agent is to some extent also a patient with respect to that very action) and a patient being passively affected by the action. As we have seen, in some cases, agent and patient involved in a particular

[8] There is much to be said about the interplay between acting and being acted upon. Mikael M. Karlsson (2002), for example, cashes out the distinction between things we do and things that happen to us in an Aristotelian attempt of "self-movement." Richard Taylor (1982) questions the metaphysical distinctiveness of action and suggests a more practical approach, claiming that we decide whether something is an action or someone an agent when we encounter them; as we go along, so to speak.

situation can be one and the same person. When I lift my cup of tea, I am both an agent sipping from my cup and a patient suffering the cup's touch at my lips.[9]

Since agency and patiency are complementary parts of a person's life, Reader submits that ascribing agency metaphysical primacy in the constitution of personhood is unfounded. It is up for debate whether one must follow Reader all the way to this metaphysical conclusion. Certainly, as an anonymous referee rightly pointed out, additional arguments are needed to confute the metaphysical claim Reader prematurely rejects; perhaps agency does deserve priority in the metaphysical constitution of personhood. All the same, a decisive decision on this matter is not necessary for my purposes, and so I remain agnostic about the matter here. From a more practical point of view, drawing attention to the tight linkage between agency and patiency is a valuable insight that should help alleviate the agential bias by emphasizing the complementary nature of these two integral aspects of people's lives.

Another reason why it is difficult to do away with the agential bias is the common misconception to view patients on a par with mere *objects*, forfeiting or lacking agential features altogether. Along these lines, Krakauer issues a Heideggerian and Foucault inspired worry regarding the technologically induced exposure of agents as mere objects in healthcare technology: "This challenging or provoking of beings to expose themselves as objects which thereby also poses or establishes beings as objects is precisely what Heidegger calls 'the essence of technology.' Foucault took as his task to follow the path indicated by this Heideggerian thought through the language of medicine. Foucault's labor of listening to medical language hears that even the autonomous individual, the subject itself, has acquired, and been reduced to, 'the status of an object'" (Krakauer 1998, 533). But this is not so. When I am being acted on, even in the crudest sense, I thereby do not cease to be an agent altogether. It is just that I might not currently exercise most of my agential features. Now, the idea is to remedy the agential bias that marginalizes patiential parts of people's lives by acknowledging that agency is manifested not only when we are acting as agents but also when we are being acted upon as patients, and furthermore, to see that we are, inevitably, patients all the time. Since, as mentioned previously, every action has both agential and patiential characteristics. And so, these patiential parts are no failure, nor are they of lesser value than agential characteristics in attaining agency. Patiency dialectically completes agency. In seeing that agency presupposes patiency, we might find reason to drop the idealization of agential characteristics as the main components for valuing our lives in favor of a more balanced view that can help appreciating patiential characteristics just as much. What happens to us as patients, in acting and in being acted on, may define us

[9] An anonymous referee pressed the point that the interpretation of predicates like "suffering" seems very generous here. What is meant to be shown by emphasizing things like suffering (and similar patiential expressions) is the idea that in every instance of human action, there is always simultaneously an aspect in which things passively happen either to someone else involved in that action or to that very agent initiating the action (this is why I have earlier described agency and patiency as correlates). The reviewer further says, and rightly so, that it would be more intuitive to say that agents sometimes also "overcome," for example, some "resistance" but that this does not make them patients. This might be so, but I think there is still reason to resist this intuitive appeal and instead hold on to the conceptual conviction that agents are in some relevant sense always also passive with regard to their actions. "Overcoming" some resistance, in this example, suggests that the agent actively does so overcome, when in fact they might just passively "endure" something that happens to them, something that is an integral part of the action but that the agent has nonetheless no influence or control over and is thus patiential with regard to the action.

as much as what we do as agents. And so, being a patient all the time is not as such a reduced or unpleasant condition but an unavoidable fact about us. In order to appreciate the complementary nature of agency and patiency, we ought to explore what passively happens to agents, what constraints them, the contingencies they are subjected to, as well as their display of active, agential characteristics.

3.1 Patiency and Paternalism in Health Care

When it comes to medical settings, the notion of patiency in relation to paternalism requires an additional treatment from what I have said so far concerning the broader philosophical notion of patients and paternalism. The special relation between medical patients and healthcare professionals is constituted by two assumptions. For one, medical patients rely and trust on the expertise of healthcare professionals and base their decisions on what clinicians recommend. Healthcare professionals, on the other hand, are supposed to act in ways that are both beneficial to the patients in their charge and at the same time respect patients' autonomy with regard to their medical care. When healthcare professionals act on the basis of what is good for their patients alone, and thereby ignore patients' autonomy, they act paternalistically. The asymmetry of expertise between patients and healthcare professionals allows for asking when, if ever, medical paternalism is called for. I cannot attempt to give a decisive answer to this question here, but I think that Groll (2014) has a point when he argues that the burden of proof should lie on those who think that medical paternalism is sometimes justified.

Since mHealth apps are not only used by healthy individuals to track their fitness level but also prescribed by physicians to monitor medical conditions, the distinction between the philosophical and the medical notion of patients becomes blurry and so does the question as to whether a medical form of paternalism, which potentially involves (technological) persuasion, might be acceptable in such cases. A further complication that makes a decision as to if and when medical paternalism in the context of mHealth can be appropriate is the fact that users differ vastly in their health literacy which impacts users' autonomy. What could be an enhancement of someone's autonomy, might constitute a threat to someone else's autonomy. Mantovani et al. (2014) put it as follows:

> In addition, it must be pointed out that apps mobile devices are not used by abstract individuals, but by people with flesh and bones, different levels of understanding and even different capacities for the exercise of individual autonomy. The ability of an individual to be able to gauge truly the exact nature of his/her situation in an mHealth environment will vary enormously between people such as a teenager or an elderly patient. In a real-life environment (in a hospital, for example) a healthcare provider would be able to guide users/patients through the process of consent, explain the consent form that needs to be signed and to answer possible questions. Current medical apps often leave the user alone and even require him/her to open up additional links to find information on external sites" (ibid., 57).

Furthermore, as pointed out by Mittelstadt et al., "While autonomy is increased by the release of the lifeworld from the confines of hospitalisation, PHM still allows the

system to invade the lifeworld and exert control through the quantisation and regulation of behaviour in the personal environment. … The visibility of the PHM may also affect the user's identity, as PHM use becomes part of who they are, and affect behavioural patterns derived from the lifeworld. Behavioural patterns must be adapted to meet the requirements of PHM, whether that is in routines of monitoring by recording and transmitting physiological and behaviour data, or by routine of intervention, where therapies are conducted in response to the output of the PHM" (Mittlestadt et al. 2014, 50).

Recent attempts to increase health literacy, particularly in chronic conditions, have led to a shift from patients that once were inactive sufferers to active, competent partakers in their own health care. As such, patients become "knowledgeable about their condition, health services and their rights as a patient; skilled and organised in self-managing it; actively involved in information seeking and use; communicative with health professionals in an assertive manner; able to seek and negotiate treatment options" (Edwards et al. 2012, 6). Users with such a level of health literacy that use mHealth apps will be very much in control of the way they use the app and will have a clear idea of what to expect from the app.[10]

Having cleared some conceptual ground as to how agency and patiency are deeply intertwined, I now turn to put into perspective the widely held conviction that persuasive mHealth devices primarily emphasize patiential characteristics of people's lives. Removing the hurdle of the agential bias, my goal is to show that this patiential emphasis needs not to have a negative connotation. Rather, as I argue in what follows, patiency can be an extension of the agent's own mind and volition—perhaps even widening the scope of their autonomy. In making the case for complementing agency with patiency in mHealth apps, I suggest some strategies that help preserving and ultimately widening agents' autonomy in this context.

4 Volitional Aids: Enhancing Agency by Design

MHealth apps that are aimed at persuading users to change their behavior have philosophically been challenged mainly on two related grounds:

(a) Based on eroding users' autonomy due to their persuasive character. Rossi and Yudell, for example, claim that "persuasive (as opposed to manipulative) health communication infringes upon autonomy if and when it exerts a controlling influence, and persuasion may infringe upon autonomy if risk or health messages fail to provide message recipients with the information they are due" (Rossi and Yudell 2012, 201).
(b) For addressing users primarily as patients. Sharon (2017) as previously mentioned, sees the potential control over users enabled by mHealth apps as a threat to their autonomy, degrading their level of agency towards inactive patients.

Sharon (2017), as previously mentioned, sees the potential control over users enabled by mHealth apps as a threat to their autonomy, degrading their level of agency towards inactive patients.

Now, at first glance, it seems evident that being persuaded by a device that tells us how to conduct significant portions of our lives renders the agent passive, an inert

[10] Thanks to an anonymous referee for drawing my attention to this.

recipient of technological commands. If these intrusive commands are, for good measure, disguised as persuasive suggestions (rather than straightforward imperatives), and thus chip away at users' autonomy, employing such devices should be avoided at all costs. Or so it seems, given the agential bias.

Heretofore, I have tried to show that persuasion is neither necessarily paternalistic nor inherently at odds with autonomy and that being acted upon is an inevitable, ever present part of agents' lives—not something to be evaded, as those who hold agency dear might think. I am now in a position to look at the philosophically sometimes underappreciated positive side of mHealth apps. In so doing, I bring together the extended mind and extended will theses with a more balanced account of agency that encompasses both agential and patiential characteristics, contending that there is a way of technologically harnessing patiency to enhance agency.

4.1 Extended Minds and Extended Wills

One contentious claim that, perhaps, sparks the critique of mHealth apps as agency-eroding is the juxtaposition of users on the one hand and technological devices on the other. This sharp division between agents as bearers of mental states and technology has not been made explicit in the context of mHealth apps (to my knowledge) but might, perhaps, be an implicit reason why apps are seen as mere tools that are not part of the agent's cognitive or volitional apparatus. This follows both straightforwardly from traditional conceptions of the mind such as physicalism and from recent criticisms of the extended mind thesis that see the mind as staying "safely within the boundaries of the body and brain" (Weiskopf 2008, 275). mHealth apps as tools of external influence could then, if they were to impose someone else's interest on users, present a threat to their autonomy. This traditional boundary between agents and their environment has been called into question ever since Andy Clark's and David Chalmers's extended mind thesis (Clark and Chalmers 1998). On this view, certain technological devices are literally extensions of people's minds, enabling agents to extend their minds beyond the physical boundaries of their bodies. What kind of cognitive processes qualify as being realized extendedly depends on a sensible conditional:

> If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head (ibid., 8; italics in original)!

Clark and Chalmers distinguish between what they call "extended cognition" which involves various physical and computational artifacts employed by agents, such as calculators and notebooks. These are enactive systems that are not confined to the physical boundaries of one's body, but nevertheless an extension of the agent's cognitive apparatus into the environment. Functionally, extended cognition plays the same role as does internal cognition, or as in the calculator example, extended cognition complements, even enhances internal cognition. For example, most of us are reasonably decent at mental arithmetic but nowhere near the performance of a calculator.

A stronger and more controversial thesis is what Clark and Chalmers describe as "extended mind"—the claim that some mental states, particularly beliefs and desires, can literally be stored and manifested on devices outside of one's own body without thereby ceasing to be one's own beliefs and desires. Proponents of this view take smartphones to be apt examples of mind extenders. Not only do these devices complement the agent's internal cognition, as say in their functions as calculators, but they also do store information such as pictures, call logs, and directions that constitute the agent's own mental states. Take, for example, finding one's way around in a familiar but not so frequently visited area. When recalling the way to get from here to there, it makes no difference, so says the extended mind thesis, whether we reach the destination by accessing our internal memory or by consulting our smartphone's memory that has that information stored on our behalf. Either way, we are recollecting an existing belief.

Some authors have expanded the extended mind theory to an outsourcing of decision-making capacities, which they call "extended will" (Heath and Anderson 2010). On this view, people make use of the ability to offload various motivational and cognitive processes to their environment, broadly construed. Due to such outsourcing, the environment can provide the necessary "scaffolding" (Sterelny 2010) that enables agents to successfully solve various problems of self-regulation whose accomplishment by traditional, internal means might have been impossible due to a temporary or permanent scarcity of the corresponding internal resources. From the extended will perspective, persuasive technologies do not appear as a threat to agents' autonomy but rather as a form of what I call volitional aids, assisting agents with the accomplishment of difficult tasks. Such technologies, seen in this light, do not undermine agents' autonomy because they are genuinely parts of our own decision-making processes. Just as outsourced beliefs are genuine parts of the extended mind, so are outsourced volitions genuine parts of the extended will.

Granted that both the extended mind and the extended will theses are hotly debated, they nevertheless present a sensible platform for suggesting that mHealth apps and an agent's internal cognitive and volitional apparatus need not be mutually exclusive opposites. Particularly, the extended mind/will theses call into question the claim that mHealth apps are agents of their own, either in virtue of imposing the app designer's will on users or by representing some other agent's vested interest by proxy.

One might argue that there is no tight conceptual connection between the agent/patient distinction and the extended mind/will discussion.[11] This might be so, but realizing that patiency is an integral part of every agent's life—even though it is mainly constituted by things that happen to us—helps understanding why external devices such as mHealth apps that occasionally render the agent passive in a similar way by, for example, nudging users, can nevertheless be genuine parts of the agent's cognitive or volitional apparatus. The very fact that things happen to us that are beyond our direct initiation does not necessarily render these happenings foreign.

In what follows, I take the thesis that mHealth apps are serious candidates for both mind and will extenders as a tenable way of understanding the relation between agents and these kinds of systems. This leaves me with the question of what features such

[11] I owe this point to an anonymous referee.

devices must have in order to initially at least retain users' autonomy and to potentially even enhance it.

I now turn to discuss three principles of persuasion that shall help rendering mHealth devices ethically sound and spell out how an agency enhancement via persuasive mHealth apps can work by way of looking at some real-life examples.

4.2 Ethically Sound Principles of Persuasion

Spahn (2012) who sees merit in persuasion suggests three useful principles that ensure the preservation of as much agency as possible while at the same time taking advantage of the effectiveness of technological persuasion that enables users to reach goals more efficiently.

(1) Persuasion should be based on prior (real or counterfactual) consent.

Any persuasive device that even initially bears the potential of preserving users' autonomy must meet the gold standard of obtaining informed consent before the device is put to use. Once it is ensured that users apprehend and consent to the device's persuasive character, Dworkin's second criterion of paternalistic interventions (*X* does something to *Y* without their consent) is circumvented. Practically, this could work for example by presenting users' educational videos that sincerely display the device's workings and persuasive goals. If this is implemented as a prerequisite for being able to operate the device, informed consent is warranted.

(2) Ideally, the aim of persuasion should be to end persuasion.

Persuasive technologies involve a specific kind of human-technology interaction that differs from regular human-to-human communication in at least two important ways: both parties have limited resources to influence each other, and there is no mutual communication possible in the sense that the device does not understood users' feelings and thus cannot adequately respond to them. The interaction between users and persuasive technologies can thus be characterized as an asymmetrical relation with the goal of changing users' behavior for the better. When it comes to the means of persuasion, it is important to distinguish between what might be called *manipulative* persuasion and *educational* persuasion. Manipulative persuasion aims at creating a dependent person that is in permanent need of guidance. The aim of educational persuasion, on the other hand, is to empower users and thus to promote their independency and autonomy. The autonomous user is, then, able to end the asymmetrical relation and to educate themselves. For example, if an mHealth app educates users to think about their nutrition behavior and helps implementing a healthier diet, eventually the persuasive technology is no longer needed, and users will be able to stick to their newly acquired routine by themselves.

(3) Persuasion should grant as much autonomy as possible to the user.

Persuasive technologies preserve users' autonomy just in case that these devices do not take over users' large-scale decision-making capacity. This issue is particularly tricky since one major asset of persuasive technologies is precisely to take over some of

users' choices, prompting them to follow behavioral recommendations generated by the device. Now, the key to autonomy preservation and, ultimately, to autonomy enhancement is to ensure that large-scale goals of behavior change are set autonomously.

4.3 Volitional Aids and Second-Order Autonomy

Preserving and ultimately enhancing users' autonomy might be achieved by what I have earlier described as volitional aids. If the overall goals of behavior change are set by users themselves (i.e., autonomously), there is no other agent involved and thus no threat to autonomy. The device merely aids users' initial intentions to achieve their goals via technological persuasion. To further embellish this idea, I differentiate between first- and second-order autonomy:

> First-order autonomy can, in this context, be described as an agent's exerted capacity to autonomously make decisions on a small-scale level. For example, deciding at whim what to have for lunch, how to get to work, whether to hit the gym today.
>
> Second-order autonomy can, in this context, be described as an agent's exerted capacity to autonomously make decisions on a large-scale level. For example, deciding to improve one's diet, or to live a more active life by increasing one's sports activities.

In the spirit of Harry Frankfurt's hierarchical model of autonomy, agents are autonomous with respect to their actions if and only if their first-order autonomous small-scale decision-making capacity is approved of (or sanctioned by) their second-order autonomous large-scale decision-making capacity. Second-order autonomy oversees first-order autonomy, as it were.

The suggested view of second-order autonomy as an agent's exerted capacity to autonomously make decisions on a large-scale level that might include the occasional forfeit of first-order autonomous choices combined with the hypothesis that mHealth apps can be seen as extended minds/wills potentially sheds new light on Spahn's second principle of ethically sound persuasion as ultimately aiming to end persuasion.[12] If I am correct in conceptualizing mHealth apps as volitional aids, there is no need to aim for an end of persuasion. By using such volitional aids, we simply outsource internal willpower to the environment that nevertheless remains part of the agent's own volitional apparatus which is based on a second-order autonomous choice to employ such technologies.

It is important to keep in mind that mHealth apps are not imposed on users but autonomously employed (except, perhaps, in a few medical settings). That said, mHealth apps do not only target people with health conditions but are also frequently used by healthy people and created for prevention purposes. Lifestyle-based interventions such as motivational goal setting, action planning, or self-monitoring are becoming more feasible for individuals due to personalized mobile technologies (Orrell and Brayne 2015). Recent reviews suggest that self-monitoring applications have great potential to aid and modify people's lifestyle (Burke et al. 2015) and to encourage self-management in chronic conditions and patient autonomy (Boulos et al. 2011; Landry 2015). First encouraging effects with respect to the use of such apps have been demonstrated with respect to lifestyle issues such as physical activity, diet, and weight

[12] I am grateful to an anonymous referee for pointing me in that direction.

control (Carter et al. 2013; Glynn et al. 2014; Lubans et al. 2014). Both recreationally and medically, mHealth apps are mainly used for tracking and monitoring purposes only or additionally for helping users to change their behavior. When employed with the explicit goal of being assisted to change one's behavior, users expect such devices to persuade them towards a wanted outcome.[13] In these circumstances, persuasion can hardly count as posing a threat to autonomy, but rather as an expression of users' second-order autonomous choice to use such devices.

Now, in some cases, acting truly autonomously might mean to voluntarily relinquish or outsource one's first-order autonomy for the purpose of enhancing one's second-order autonomy. Let us return to the initial lunch example that can now be redescribed in the following way:

If I have autonomously decided that I want to improve my diet as an exercise of second-order autonomy, I will have salad for lunch even though this might be at odds with satisfying my cravings for fries as an exercise of first-order autonomy. Persuasive mHealth apps can serve as volitional aids in such scenarios since they have the potential to increase one's second-order autonomy by incentivizing a first-order autonomous behavior that is in accordance with the previously set second-order autonomous goal. In a way, then, by increasing the level of patiency regarding small-scale decisions ("I'll go with what the device tells me"), the overall level of agency is enhanced, trading patiency in the fine print for agency in the heading, as it were. Since the behavior change is self-initiated, and thus based on the agent's own intentions and motivations, there is no threat to second-order autonomy but an enhancement thereof. Importantly, autonomy-enhancing persuasive apps treat users as reason-responsive agents by way of presenting reasons and motivational incentives for self-initiated behavior changes rather than simple imperatives. Weintraub and Barilan (2001) go as far as to suggest that the value of autonomy traces back to persons' right to be respected as agents who can argue, persuade, and be persuaded in matters of utmost personal significance such as decisions about medical care. These authors suggest that autonomy should and could be respected only after such an attempt of persuasion has been made.

How can this work in practice? One established technique that helps changing one's behavior by effectively translating previously set goals into action are so-called implementation intentions (Gollwitzer 1999; Roughley 2016). Such "if-then plans" are psychological constructs for establishing new routines that aim at long-term behavioral changes. The basic structure of implementation intentions looks as follows:

If situation *x* arises, I will initiate the goal-directed response *y*.

Whereby *x* constitutes the if-component, representing a critical situation containing behavioral cues. *Y* constitutes the then-component, representing the goal-directed behavior. For implementation intentions to work most effectively, the striven-for plans must be both viable and precise.

Coming back to the previous example, an implementation intention that promotes the agent's second-order autonomous goal to improve their diet could be the following conditional: "If there is an healthy option at the cafeteria for lunch, I will go for it." If

[13] Thanks to an anonymous referee for pointing this out.

mHealth apps are used to remind users of that intention and the corresponding cue by, say, popping up a message at noon, so much the better for their effective goal achievement. Implementation intentions facilitate second-order autonomy through harnessing patiential features, namely following previously set goals by hewing to persuasive suggestions. "The device tells me to do so, so I'll comply"—thereby making the achievement of the previously set second-order autonomous goals more efficacious. Setting such large-scale goals preserves autonomy, and technological persuasion makes their achievement more effective.

It is, however, crucial to keep some caveats in mind when looking for practical solutions to make persuasive mHealth apps work. Importantly, the goals set by second-order autonomous decision making capacities must be self-motivated in order to have their intended effect. According to the "self-determination theory" (Ryan and Deci 2000), "autonomous motivation" is much more effective by way of sustainably changing one's behavior compared to what Ryan and Deci call "controlled motivation." When agents are autonomously motivated, they gain self-support and reinforcement through their own actions; the motivation emanates from the self, and the behavior is thus self-determined (Hager et al. 2014, 567). On the other hand, controlled motivation is an external, introjected regulation of one's behavior (e.g., avoiding punishment or feelings of guilt). There is ample evidence suggesting that autonomous motivation has the most pervasive effects on behavior change, particularly on health-related behavior (ibid., 578). Pavey and Sparks (2010) further show that autonomous motivation increases an healthier lifestyle by promoting intentions to reduce behavior that is harmful to one's health.

5 Concluding Remarks

In this paper, I have tried to show that persuasive mHealth applications are, despite appearances, not necessarily at odds with users' autonomy. This is so, I have argued, for two main reasons. (1) Once the misguided assumption of a sharp divide between agency and patiency is mitigated, it becomes clear that displaying agency can be extended to patiential characteristics of our lives. For example, complying with an apps' suggestion to bike to work instead of taking the car, notwithstanding one's current lazy preference for driving to work, might initially appear to render the agent a patient, a passive recipient of technological commands. However, harnessing this patiential feature can be fully compatible with one's autonomy if it is an exercise of adhering to a previously set large-scale autonomous goal such as wanting to increase one's physical activity. (2) Drawing on the extended mind and extended will theories, I have argued that ethically sound persuasive technologies do not constitute intrusive external interventions into people's lives but are rather what I have called volitional aids, assisting agents with the accomplishment of difficult tasks that might have been impossible to achieve otherwise due to a temporary or permanent scarcity of the corresponding internal resources. Ethically sound persuasive apps, thus, need not be paternalistic and can even bear the potential to enhance agents' autonomy when applied with caution.

Acknowledgements Special thanks to the audience and the referees of the 20th Conference of the Society for Philosophy and Technology (SPT) in Darmstadt for their helpful comments on earlier versions of this paper. Thanks also to the referees of *Philosophy and Technology* for their valuable feedback.

Funding Information I am grateful to the German Federal Ministry of Education and Research (BMBF) for their generous financial support.

References

Asimakopoulos, S., Asimakopoulos, G., & Spillers, F. (2017). Motivation and user engagement in fitness tracking: heuristics for mobile healthcare wearables. *Informatics, 2017*, 1–16.

Boulos, M. N. K., Wheeler, S., Tavares, C., & Jones, R. (2011). How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX. *Biomedical Engineering Online, 10*, 24.

Burke, L. E., Ma, J., Azar, K. M. J., Bennett, G. G., Peterson, E. D., Zheng, Y., & Quinn, C. C. (2015). Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. *Circulation, 132*(12), 1157–1213.

Carter, M. C., Burley, V. J., Nykjaer, C., & Cade, J. E. (2013). Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. *Journal of Medical Internet Research, 15*(4), e32.

Chatterjee, S., & Price, A. (2009). Healthy living with persuasive technologies: framework, issues and challenges. *Journal of the American Medical Informatics Association, 16*(2), 171–178.

Cholbi, M. (2017). Paternalism and our rational powers. *Mind, 126*(501), 123–153.

Cialdini, R., et al. (2005). Persuasion and health: creating positive behaviour change. In J. Kerr, R. Weitkunat, & M. Moretti (Eds.), *ABC of behavior change: a guide to successful disease prevention and health promotion* (pp. 247–258). Edinburgh: Elsevier Science.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Davis, R., Campbell, R., Hildon, Z., Hobbs, L., & Michie, S. (2015). Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychology Review, 9*(3), 323–344.

Dworkin, G. (1972). Paternalism. *The Monist, 56*, 64–84.

Dworkin, G. (2005). Moral paternalism. *Law and Philosophy, 24*(3), 305–319.

Dworkin, G. (2015). Defining paternalism. In Thomas Schramme (ed.), *New perspectives on paternalism and health care*, Springer, 17–29.

Dworkin, Gerald, "Paternalism", The Stanford encyclopedia of philosophy (Spring 2017 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/spr2017/entries/paternalism/>.

Edwards, M., Wood, F., Davies, M., & Edwards, A. (2012). The development of health literacy in patients with a long-term health condition: the health literacy pathway model. *BMC Public Health, 12*, 1–15.

Enoch, D. (2016). What's wrong with paternalism: autonomy, belief and action. *Proceedings of the Aristotelian Society, 116*(1), 21–48.

Fogg, B. J. (2003). *Persuasive technology: using computers to change what we think and do. The Morgan Kaufmann series in interactive technologies*. Amsterdam, Boston: Morgan Kaufmann Publishers.

Glynn, L. G., Hayes, P. S., Casey, M., Glynn, F., Alvarez-Iglesias, A., Newell, J., & Murphy, A. W. (2014). Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial. *The British Journal of General Practice, 64*(624), e384–e391.

Gollwitzer, P. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist, 54*(7), 493–503.

Groll, D. (2014). Medical paternalism—part 2. *Philosophy Compass, 9*(3), 194–203.

Heath, J., & Anderson, J. (2010). Procrastination and the extended will. In M. White & C. Andreou (Eds.), *The thief of time* (pp. 233–252). New York: Oxford University Press.

Karlsson, M. (2002). Agency and patiency: back to nature? *Philosophical Explorations, 5*(1), 59–81.

Koelle, M., Kranz, M., & Lindemann, P. (2014). Persuasive technologies and applications. *Advances in Embedded Interactive Systems, 3*(2).

Krakauer, E. (1998). Prescriptions: autonomy, humanism and the purpose of health technology. *Theoretical Medicine and Bioethics, 19*, 525–545.

Krieger, W. (2013). Medical apps: public and academic perspectives. *Perspectives in Biology and Medicine, 56*(2), 259–273.

Landry, K. (2015). Using eHealth to improve health literacy among the patient population. *Creative Nursing, 21*(1), 53–57.
Lanzing, M. (2016). The transparent self. *Ethics and Information Technology, 18*(1), 9–16.
LeBar, M. (2013). *The value of living well*. Oxford: Oxford University Press.
Lott, M. (2016). Agency, patiency, and the good life: the passivities objection to eudaimonism. *Ethical Theory and Moral Practice, 19*(3), 773–786.
Lubans, D. R., Smith, J. J., Skinner, G., & Morgan, P. J. (2014). Development and implementation of a smartphone application to promote physical activity and reduce screen-time in adolescent boys. *Frontiers in Public Health, 2*, 42.
Mantovani, E., et al. (2014). eHealth to mHealth—a journey precariously dependent upon apps? *European Journal of ePractice, 20*, 48–66.
Mittlestadt, B., Fairweather, B., Shaw, M., & McBride, N. (2014). The ethical implications of personal healthcare monitoring. *International Journal of Technoethics, 5*(2), 37–60.
Nordgren, A. (2015). Privacy by design in personal health monitoring. *Health Care Analysis, 23*, 148–164.
O'Keefe, D. J. (2012). Conviction, persuasion, and argumentation: untangling the ends and means of influence. *Argumentation, 26*(1), 19–32.
Orrell, M., & Brayne, C. (2015). Dementia prevention: call to action. *The Lancet, 386*(10004), 1625.
Owens, J. & Cribb, Alan (forthcoming). 'My Fitbit thinks I can do better!' Do health promoting wearable technologies support personal autonomy? *Philosophy and Technology*: 1–16.
Pavey, L. J., & Sparks, P. (2010). Autonomy and reactions to health-risk information. *Psychology and Health, 25*(7), 855–872.
Reader, S. (2007). The other side of agency. *Philosophy, 82*(4), 579–604.
Rossi, J., & Yudell, M. (2012). The use of persuasion in public health communication: an ethical critique. *Public Health Ethics, 5*(2), 192–205.
Roughley, N. (2016). *Wanting and intending: elements of a philosophy of practical mind*. Dordrecht: Springer.
Sharon, T. (2017). Self-tracking for health and the quantified self: re-articulating autonomy, solitary and authenticity in an age of personalized healthcare. *Philosophy & Technology, 30*(1), 93–121.
Spahn, A. (2012). And lead us (not) into persuasion…? Persuasive technology and the ethics of communication. *Science and Engineering Ethics, 18*(4), 633–650.
Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*(4), 465–481.
Taylor, R. (1982). Agent and patient: is there a distinction? *Erkenntnis, 18*(2), 223–232.
Timmer J., Kool L., van Est R. (2015) Ethical challenges in emerging applications of persuasive technology. In: MacTavish T., Basapur S. (eds) Persuasive technology. PERSUASIVE 2015. Lecture Notes in Computer Science, vol 9072. Springer, 196–201.
Weintraub, M., & Barilan, M. (2001). Persuasion as respect for persons: an alternative view of autonomy and of the limits of discourse. *Journal of Medicine and Philosophy, 26*(1), 13–34.
Weiskopf, D. A. (2008). Patrolling the mind's boundaries. *Erkenntnis, 68*(2), 265–276.

Literatur

Aharoni, E., Antonenko, O. & Kiehl, K.A. (2011). Disparities in the moral intuitions of criminal offenders: The role of psychopathy. Journal of Research in Personality, 45(3), 322–327.

Andow, J. (2016). Reliable but not home free? What framing effects mean for moral intuitions.Philosophical Psychology, 29(6), 904–911.

Andrews, K. (2012). Do apes read minds? Toward a new folk psychology. Cambridge: MIT Press.

Asimakopoulos, S., Asimakopoulos, G. & Spillers, F. (2017). Motivation and user engagement in fitness tracking: heuristics for mobile healthcare wearables. Informatics, 2017, 1–16.

Baars, B., Franklin, S. & Ramsoy, T. (2013). Global Workspace Dynamics: Cortical "Binding and Propagation" Enables Conscious Contents. Frontiers in Psychology, 4, 200.

Baker, L. (1999). What am i? Philosophy and Phenomenological Research, 59(1), 151–159.

Baker, L. (2000). Persons and bodies: A constitution view. Cambridge: Cambridge University Press.

Baker, L. (2007). The metaphysics of everyday life: An essay in practical realism. Cambridge: Cambridge University Press.

Baker, L. (2013). Naturalism and the first-person perspective. New York: Oxford University Press.

Baker, L. (2015). Human persons as social entities. Journal of Social Ontology, 1(1), 77–87.

Bakhurst, D. (2005). Wiggins on persons and human nature. Philosophy and Phenomenological Research, 71 (2), 462–469.

Barello, S., Palamenghi, L. & Graffigna, G. (2020). The Mediating Role of the Patient Health Engagement Model on the Relationship Between Patient Perceived Autonomy Supportive Healthcare Climate and Health Literacy Skills. International Journal of Environmental Research and Public Health (IJERPH), 17, 1741.

Bartels, D. & Pizzaro, D. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. Cognition, 121(1), 154–161.

Baumrin, B.H. (1968). Is there a naturalistic fallacy? American Philosophical Quarterly, 5(2), 79–89.

Beauchamp, T.L. & Childress, J.F. (2019). Principles of Biomedical Ethics, Eighth Edition. New York: Oxford University Press.

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. Brain Cognition, 55(1), 30–40.

Beck, S. (2006). These bizarre fictions: thought-experiments, our psychology and our selves. Philosophical Papers, 35(1), 29–54.

Beck, S. (2016). Technological fictions and personal identity: On Ricoeur, Schechtman and analytic thought experiments. Journal of the British Society for Phenomenology, 47(2), 117–132.

Behne, T. et al. (2005). Unwilling versus unable: Infants' understanding of intentional action. Developmental Psychology, 41, 328–337.

Berker, S. (2009). The normative insignificance of neuroscience. Philosophy and Public Affairs, 37(4), 293–329.

Berniūnas, R. & Dranseika, V. (2016). Folk concepts of person and identity: A response to Nichols and Bruno. Philosophical Psychology, 29(1), 96–122.
Beukema, S., Gonzalez-Lara, L.E., Finoia, P. et al. (2016). A hierarchy of event-related potential markers of auditory processing in disorders of consciousness. NeuroImage: Clinical, 12: 359–371.
Blok, S., Newman, G. & Rips, L.J. (2005). Individuals and their concepts, in: Ahn, W.-K., Goldstone, R.L., Love, B.C., Markman, A.B & Wolff, P.: Categorization Inside and Outside the Laboratory, 127–149. Washington, DC: American Psychological Association.
Bluhm, R., Cabrera, L. & McKenzie, R. (2020). What we (should) talk about when we talk about deep brain stimulation and personal identity. Neuroethics, 13, 289–301.
Boddington, P., Northcott, A. & Featherstone, K. (2024). Personhood as projection: the value of multiple conceptions of personhood for understanding the dehumanisation of people living with dementia. Medicine, Health Care and Philosophy, 27(1), 93-106.
Boksem, M.A. & De Cremer, D. (2010). Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. Social Neuroscience, 5(1), 118–128.
Botvinick, M., Cohen, J.D. & Carter, C.S. (2000). Conflict monitoring and anterior cingulate cortex: An update. Trends in Cognitive Sciences, 8(12), 539–546.
Boulos, M. N. K., Wheeler, S., Tavares, C. & Jones, R. (2011). How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX. Biomedical Engineering Online, 10, 24.
Braddon-Mitchell, D. & Miller, K. (2004). How to be a conventional person. The Monist, 87(4), 457–474.
Buchanan, A. (1988). Advance directives and the personal identity problem. Philosophy and Public Affairs, 17(4), 277–302.
Burke, L. E.,Ma, J., Azar, K.M.J., Bennett, G.G., Peterson, E.D., Zheng, Y. & Quinn, C.C. (2015). Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association. Circulation, 132(12), 1157–1213.
Cacioppo, J., Crites, S.L. & Gardner, W. (1994). Attitudes to the right: Evaluative processing is associated with lateralized late positive event-related brain potentials. Personality and Social Psychology Bulletin, 22(12), 1205–1219.
Campbell, T.D. (1970). The normative fallacy. Philosophical Quarterly, 20(81), 368–377.
Carpenter, M., Nagell, K. & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development, 63(4), Serial No. 255, 1-142.
Carpenter, M. (2010). Social cognition and social motivations in infancy, in: Goswami, U. (Ed.): The Wiley-Blackwell handbook of childhood cognitive development (2nd ed.), 106–128. Oxford: Wiley-Blackwell.
Carter, M.C., Burley, V.J., Nykjaer, C. & Cade, J.E. (2013). Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. Journal of Medical Internet Research, 15(4), e32.
Casebeer, W.D. (2003). Moral cognition and its neural constituents. Nature Reviews Neuroscience, 4(10), 840–847.
Casebeer, W.D. & Churchland, P. (2003). The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. Biology and Philosophy, 18(1), 169–194.

Chang, L.J. & Sanfey, A.G. (2008). Emotion, decision-making and the brain, in: Houser, D. & McCabe, K. (Ed.): Neuroeconomics (Advances in Health Economics and Health Services Research, Volume 20), 31–53. Emerald Group Publishing.
Chappell, T. (2011). On the very idea of criteria for personhood. The Southern Journal of Philosophy, 49(1), 1–27.
Chatterjee, S. & Price, A. (2009). Healthy living with persuasive technologies: framework, issues and challenges. Journal of the American Medical Informatics Association, 16(2), 171–178.
Cholbi, M. (2017). Paternalism and our rational powers. Mind, 126(501), 123–153.
Churchland, P.S. (2012). Braintrust: What neuroscience tells us about morality. Princeton: Princeton University Press.
Churchman, C.W. (1956). Science and decisionmaking. Philosophy of Science, 23(3), 247–249.
Cialdini, R.B., Maner, J. & Gerend, M. (2005). Persuasion and health: creating positive behaviour change, in: Kerr, J., Weitkunat, R. & Moretti, M. (Eds.): ABC of behavior change: a guide to successful disease prevention and health promotion, 247–258. Edinburgh: Elsevier Science.
Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. Cambridge, Massachusetts: MIT Press.
Clark, A. & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19.
Clark, A. (1999). Embodied, situated, and distributed cognition, in: Betchel, W. & Graham, G. (Eds.): A Companion to Cognitive Science, 506–517. Malden: Blackwell Publishing.
Coleman, S. (2000). Thought experiments and personal identity. Philosophical Studies, 98(1), 51–66.
Contreras-Rodriguez, O., Pujol, J., Batalla, I. et al. (2014). Functional connectivity bias in the prefrontal cortex of psychopaths. Biological Psychiatry, 78(9), 647–655.
Crutchfield, P. (2018). Moral enhancement can kill. Journal of Medicine and Philosophy, 43(5), 568–584.
Cummings, M.A. (2015). The neurobiology of psychopathy: Recent developments and new directions in research and treatment. CNS Spectrums, 20(3), 200–206.
Davies, B. (2021). 'Personal Health Surveillance': The Use of mHealth in Healthcare Responsibilisation. Public Health Ethics, 14(3), 268-280.
Davis, R., Campbell, R., Hildon, Z., Hobbs, L. & Michie, S. (2015). Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychology Review, 9(3), 323–344.
Decety, J. & Cowell, J.M. (2014). Friends or foes: Is empathy necessary for moral behavior? Perspectives on Psychological Science, 9(4), 525–537.
DeCicco, J., Solomon, B. & Dennis, T. (2012). Neural correlates of cognitive reappraisal in children: An ERP study. Developmental Cognitive Neuroscience, 2(1), 79–80.
DeGrazia, D. (1996). Taking animals seriously: Mental life and moral status. New York: Cambridge University Press.
DeGrazia, D. (1999). Advance directives, euthanasia, and the someone else problem. Bioethics, 13(5), 373–391.
Dehaene, S. & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. Neuron, 70(2), 200–227.
Dennett, D. (1976). Conditions of personhood, in: Rorty, Amelie O. (Ed.): The Identities of Persons, 175–196. Berkeley: University of California Press.

De Waal, F. (2014). Natural normativity: The ‘is’ and ‘ought’ of animal behavior. Behaviour, 151, 185–204.
Dodd, J. & Stern-Gillet, S. (1995). The is/ought gap, the fact/value distinction and the naturalistic fallacy. Dialogue, 34(4), 727–746.
Douglas, H. (2007). Rejecting the ideal of value-free science, in: Kincaid, H., Dupr’E, J. & Wylie, A. (Eds.): Value-free science? ideals and illusions, 120–141. Oxford: Oxford University Press.
Dubljević, V. (2021). The Normative Implications of Recent Empirical Neuroethics Research on Moral Intuitions. Neuroethics, 14(3), 449-457.
Duke, A. & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. Cognition, 134, 121–127.
Dworkin, G. (1972). Paternalism. The Monist, 56, 64–84.
Dworkin, G. (2005). Moral paternalism. Law and Philosophy, 24(3), 305–319.
Dworkin, G. (2015). Defining paternalism, in: Schramme, T. (Ed.): New perspectives on paternalism and health care, 17–29. Cham a.o.: Springer.
Dworkin, G. (2017). Paternalism, in: Zalta, E.N. (Ed.): *The Stanford Encyclopedia of Philosophy (Spring 2017 Edition).* https://plato.stanford.edu/archives/spr2017/entries/paternalism/. Accessed 08 April 2024.
Edelman, G.M. & Tononi, G. (2000). A universe of consciousness: How matter becomes imagination. New York: Basic Books Inc.
Edwards, M., Wood, F., Davies, M. & Edwards, A. (2012). The development of health literacy in patients with a long-term health condition: the health literacy pathway model. BMC Public Health, 12, 1–15.
El Haj, M., Roche, J., Gallouj, K. & Gandolphe, M.C. (2017). Autobiographical memory compromise in Alzheimer’s disease: A cognitive and clinical overview. Geriatr Psychol Neuropsychiatr Vieil, 15(4), 443–451.
English, J. (1975). Abortion and the concept of a person. Canadian Journal of Philosophy, 5, 233–243.
Enoch, D. (2016). What’s wrong with paternalism: autonomy, belief and action. Proceedings of the Aristotelian Society, 116(1), 21–48.
Everett, J. & Kahane, G. (2020). Switching Tracks? Towards a Multi-Dimensional Model of Utilitarian Psychology. Trends in Cognitive Sciences, 23(2), 124-134.
Evnine, S. (2008). Epistemic dimensions of personhood. New York: Oxford University Press.
Feinberg, J. (1980). Abortion, in: Regan, T. (Ed.): Matters of life and death, 183–217. Philadelphia: Temple University Press.
Fogg, B.J. (2003). Persuasive technology: using computers to change what we think and do. The Morgan Kaufmann series in interactive technologies. Amsterdam, Boston: Morgan Kaufmann Publishers.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. Oxford Review, 5, 5–15.
Frankena, W.K. (1939). The naturalistic fallacy. Mind, 48(192), 464–477.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. Journal of Philosophy, 68(1), 5–20.
Franklin, S., Strain, S., Snaider, J., McCall, R. & Faghihi, U. (2012). Global workspace theory, its LIDAmodel and the underlying neuroscience. Biologically Inspired Cognitive Architectures, 1, 32–43.

Fronda, G., Angioletti, L. & Balconi, M. (2024). EEG Correlates of Moral Decision-Making: Effect of Choices and Offers Types. American Journal of Bioethics Neuroscience, 15(3), 191-205.
Gazzaniga, M.S. (2005). Facts, fictions and the future of neuroethics, in: Illes, J. (Ed.): Neuroethics: defining the issues in theory, practice, and policy, 141–148. Oxford: Oxford University Press.
Gendler, T. (1998). Exceptional persons: on the limits of imaginary cases. Journal of Consciousness Studies, 5(5–6), 592–610.
Gendler, T. (2002). Personal identity and thought-experiments. Philosophical Quarterly, 52(206), 34–54.
Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics?, in: Sinnott-Armstrong (Ed.): The cognitive science of morality: intuition and diversity, vol. 2 of Moral psychology, 1–26. Cambridge, MA: MIT Press.
Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. Topics in Cognitive Science, 2(3), 528–554.
Gillett, G., & Franz, E. (2014). Evolutionary neurology, responsive equilibrium, and the moral brain. Consciousness and Cognition, 45, 245–250.
Glynn, L. G., Hayes, P. S., Casey, M., Glynn, F., Alvarez-Iglesias, A., Newell, J. & Murphy, A. W. (2014). Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial. The British Journal of General Practice, 64(624), e384–e391.
Gollwitzer, P. (1999). Implementation intentions: strong effects of simple plans. American Psychologist, 54(7), 493–503.
Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M. & Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105–2108.
Greene, J.D. & Haidt, J. (2002). How (and where) does moral judgment work? Trends in Cognitive Sciences, 6(12), 517–523.
Greene, J.D. (2003). From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? Nature Reviews Neuroscience, 4(10), 847–850.
Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M. & Cohen, J.D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44(2), 389–400.
Greene, J.D. (2008). The secret joke of Kant's soul, in: Sinnott-Armstrong, W. (Ed.): The neuroscience of morality: emotion, brain disorders, and development, vol. 3 of Moral psychology, 35–79. Cambridge, MA: MIT Press.
Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E. & Cohen, J.D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107(3), 1144–1154.
Greene, J.D., Cushman, F.A., Stewart, L.E., Lowenberg, K., Nystrom, L.E. & Cohen, J.D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition, 111(3), 364–371.
Greene, J.D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. Ethics, 124(4), 695–726.
Groll, D. (2014). Medical paternalism—part 2. Philosophy Compass, 9(3), 194–203.
Hagger, M.S., Hardcastle, S.J., Chater, A., Mallett, C., Pal, S., Chatzisarantis, N.L.D. (2014). Autonomous and controlled motivational regulations for multiple health-related behaviors: between- and within-participants analyses. Health Psychology and Behavioral Medicine, 2(1), 565-601. https://doi.org/10.1080/21642850.2014.912945.

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. (2007). The new synthesis in moral psychology. Science, 316(5827), 998–1002.
Hare, R.M. (1991). The Language of Morals. Oxford: Oxford Clarendon Press.
Hare, R.D. (2003). The psychopathy checklist–Revised, 2nd ed.. Toronto: Multi-Health Systems.
Harman, O. (2012). Is the naturalistic fallacy dead (and if so, ought it be?). Journal of the History of Biology, 45(3), 557–572.
Harsay, H.A., Spaan, M., Wijnen, J.G. & Ridderinkhof, K.R. (2012). Error awareness and salience processing in the oddball task: shared neural mechanisms. Frontiers in Human Neuroscience, 6, 246.
Hart, E. (2024). Advance directives need full legal status in persons with dementia. Nursing Ethics, 31(7), 1247-1257.
Hauser, M.D. (2008). When your moral organ is right! Think, 7(19), 17–21.
Heath, J. & Anderson, J. (2010). Procrastination and the extended will, in: White, M. & Andreou, C. (Eds.): The thief of time, 233–252. New York: Oxford University Press.
Heilman, R.M., Crişan, L.G., Houser, D., Miclea, M. & Miu, A.C. (2010). Emotion regulation and decision making under risk and uncertainty. Emotion, 10(2), 257–265.
Hirstein, W. (2022). Neuroscience and Normativity: How Knowledge of the Brain Offers a Deeper Understanding of Moral and Legal Responsibility. Criminal Law and Philosophy, 16 (2), 327-351.
Hudson, W.D. (Ed.) (1969). The is-ought question: a collection of papers on the central problem in moral philosophy. London: Macmillan.
Huebner, B., Dwyer, S. & Hauser, M.D. (2009). The role of emotion in moral psychology. Trends in Cognitive Science, 13(1), 1–6.
Huettel, S., Song, A. & McCarthy, G. (2008). Functional magnetic resonance imaging. Sunderland, MA: Sinauer Associates, Inc.
Hughes, V. (2010). Science in court: Head case. Nature, 464(7287), 340–342.
Hume, D. (1978). A treatise of human nature. Oxford: Oxford University Press.
Huth, A., Nishimoto, A. & Gallant, J. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron, 76(6), 1210–1224.
Jiang, Q., Zhuo, L., Wang, Q. & Lin, W. (2022). The Neural Basis of Moral Judgement for Self and for Others: Evidence From Event-Related Potentials. Frontiers in Human Neuroscience, 16, 919499.
Kagan, S. (1998). Normative Ethics. Oxford: Westview Press.
Kahane, G. (2013). The armchair and the trolley: An argument for experimental ethics. Philosophical Studies, 162(2), 421–445.
Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. Social Neuroscience, 10(5), 1–10.
Karlsson, M. (2002). Agency and patiency: back to nature? Philosophical Explorations, 5(1), 59–81.
Kemmerling, A. (2014). Why is personhood conceptually difficult?, in: M. Welker (Ed.): The depth of the human person. A multidisciplinary approach, 15–44. Michigan: Grand Rapids.
Kendler, K.S. (2005). Toward a philosophical structure for psychiatry. American Journal of Psychiatry, 162(3), 433–440.
Kipper, J. (2016). Substance and the concept of personal identity. Ergo, 3(1): 1–26.
Kitcher, P. (2014). Is a naturalized ethics possible? Behaviour, 151(2–3), 245–260.

Kittay, E. (2005). At the margins of moral personhood. Ethics, 116(1), 100–131.
Knobe, J. & Nichols, S. (2007). An experimental philosophy manifesto, in: Knobe, J. & Nichols, S. (Ed.): Experimental philosophy, 3–14. Oxford: Oxford University Press.
Koelle, M., Kranz, M. & Lindemann, P. (2014). Persuasive technologies and applications. Advances in Embedded Interactive Systems, 3(2).
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. (2007). Damage to prefrontal cortex increases utilitarian moral judgments. Nature, 446(7138), 908–911.
Koenigs, M., Kruepke, M., Zeier, J. & Newman, J.P. (2012). Utilitarian moral judgment in psychopathy. Social Cognitive and Affective Neuroscience, 7(6), 708–714.
Krakauer, E. (1998). Prescriptions: autonomy, humanism and the purpose of health technology. Theoretical Medicine and Bioethics, 19, 525–545.
Krieger, W. (2013). Medical apps: public and academic perspectives. Perspectives in Biology and Medicine, 56(2), 259–273.
Kugiumutzakis, G. (1998). Neonatal imitation in the intersubjective companion space, in: Braten, S. (Ed.): Intersubjective communication and emotion in early ontogeny, 63–88. Cambridge: Cambridge University Press.
Kuhlmann, J., Wensing, G. (2006). The applications of biomarkers in early clinical drug development to improve decision-making processes. Current Clinical Pharmacology, 1(2), 185–191.
Kusch, M. (2014). The metaphysics and politics of corporate personhood. Erkenntnis, 79(9), 1587–1600.
Kutas, M. & Federmeier, K. (2011). Thirty years and counting: Findingmeaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621–647.
Lahat, A., Helwig, C. & Zelazo, P. (2012). An event-related potential study of adolescents' and young adults' judgments of moral and social conventional violations. Child Development, 84(3), 955–969.
Lakoff, G. & Johnson, M. (1999). Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought. New York: Basic Books.
Landeweerd, L. (2004). Normative-descriptive and the naturalistic fallacy. Global Bioethics, 17(1), 17–23.
Landry, K. (2015). Using eHealth to improve health literacy among the patient population. Creative Nursing, 21(1), 53–57.
Lanzing, M. (2016). The transparent self. Ethics and Information Technology, 18(1), 9–16.
LeBar, M. (2013). The value of living well. Oxford: Oxford University Press.
Lenartowicz, A., Lu, S., Rodriguez, C. et al. (2016). Alpha desynchronization and frontoparietal connectivity during spatial working memory encoding deficits in ADHD: A simultaneous EEGfMRI study. NeuroImage, Clinical 11, 210–223.
Leiser, S., Dunlop, J., Bowlby, M. & Devilbiss, D. (2011). Aligning strategies for using EEG as a surrogate biomarker: A review of preclinical and clinical research. Biochemical Pharmacology, 81(12), 1408–1421.
Leuthold H., Filik, R., Murphy, K. & Mackenzie, I.G. (2012). The on-line processing of socio-emotional information in prototypical scenarios: inferences from brain potentials. Social Cognitive and Affective Neuroscience, 7(4), 457–466.
Leuthold, L., Kunkel, A., Mackenzie, I.G. & Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. Social Cognitive and Affective Neuroscience, 10(8), 1021–1029.

Levy, N. (2007). Neuroethics: Challenges for the 21st century. Cambridge, MA: Cambridge University Press.
Li, W., Mai, X. & Liu, C. (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us? Frontiers in Human Neuroscience, 8, 74.
Libet, B.W. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences, 8(4), 529–566.
Limbaugh, D. (2016). Animals, advance directives, and prudence: Should we let the cheerfully demented die? Ethics, Medicine and Public Health, 2(4), 481–489.
Lindenberger, U. (Ed.): Understanding human development Dialogues with lifespan psychology,
483–494. Boston: Kluwer Academic Publishers.
Lipsman, N. & Glannon, W. (2013). Brain, mind and machine: what are the implications of deep brain stimulation for perceptions of personal identity, agency and free will? Bioethics, 27(9), 465–470.
Locke, J. (1975). An essay concerning human understanding. Oxford: Clarendon Press.
Locke, J. (2012). An Essay Concerning Human Understanding. Oxford: Clarendon.
Lott, M. (2016). Agency, patiency, and the good life: the passivities objection to eudaimonism. Ethical Theory and Moral Practice, 19(3), 773–786.
Lubans, D.R., Smith, J.J., Skinner, G. & Morgan, P.J. (2014). Development and implementation of a smartphone application to promote physical activity and reduce screen-time in adolescent boys. Frontiers in Public Health, 2, 42.
Luck, S. (2014). An introduction to the event-related potential technique. Cambridge, MA: MIT Press.
Macnamara, A., Foti, D. & Hajcak, G. (2009). Tell me about it: Neural activity elicited by emotional pictures and preceding descriptions. Emotion, 9(4), 531–543.
Maibom, H.L. (2014). To treat a psychopath. Theoretical Medicine and Bioethics, 35(1), 31–42.
Malatesti, L. & McMillan, J. (2024). The methods of Neuroethics. Cambridge University Press.
Mantovani, E. et al. (2014). eHealth to mHealth—a journey precariously dependent upon apps? European Journal of ePractice, 20, 48–66.
Marquis, D. (1989). Why abortion is immoral. Journal of Philosophy, 86(4), 183–202.
Martell, D.A. (2009). Neuroscience and the law: Philosophical differences and practical constraints. Behavioral Sciences & the Law, 27(2), 123–136.
McInerney, P. (1990). Does a fetus already have a future-like-ours?. Journal of Philosophy, 87(5), 264–268.
McLoughlin, G., Makeig, S. & Tsuang, M.T. (2014). In Search of biomarkers in psychiatry: EEG-based measures of brain function. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 165(2), 111–121.
McMahan, J. (2003). The Ethics of Killing. New York: Oxford University Press.
McMullin, E. (1982). Values in science. PSA: Proceedings of the biennial meeting of the Philosophy of Science Association, 1982(2), 3–28.
Meltzoff, A. & Moore, K. (1977). Imitation of facial and manual gestures by human neonates. Science, 198, 75–78.
Mendez, M. (2009). The neurobiology of moral behavior: Review and neuropsychiatric implications. CNS Spectrums, 14(11), 608–620.

Mittlestadt, B., Fairweather, B., Shaw, M. & McBride, N. (2014). The ethical implications of personal healthcare monitoring. International Journal of Technoethics, 5(2), 37–60.
Mittelstraß, J. (2003). Philosophy or the search for anthropological constants, in: Staudinger, U. & Lindenberger, U. (Eds.): Understanding human development Dialogues with lifespan psychology, 483–494. Boston: Kluwer Academic Publishers.
Mogilner, A., Grossman, J.A., Ribary, U. et al. (1993). Somatosensory cortical plasticity in adult humans revealed by magnetoencephalography. Proceedings of the National Academy of Sciences, 90(8), 3593–3597.
Moll, H. & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. Developmental Science, 7(1), F1–F9.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F. & Grafman, J. (2005). The neural basis of human moral cognition. Nature Reviews Neuroscience, 6(10), 799–809.
Moore, G.E. (1993). Principia ethica. Cambridge: Cambridge University Press.
Morse, S. (2005). Brain overclaim syndrome and criminal responsibility: A diagnostic note. Ohio State Journal of Criminal Law, 3, 397–412.
Motzkin, J.C., Newman, J.P., Kiehl, K.A. & Koenigs, M. (2011). Reduced prefrontal connectivity in psychopathy. Journal of Neuroscience, 31(48), 17348–17357.
Müller, S., Bittlinger, M. & Walter, H. (2017). Threats to neurosurgical patients posed by the personal identity debate. Neuroethics, 10(2), 299–310.
Munzer, S. (1994). Transplantation, chemical inheritance, and the identity of organs. British Journal for the Philosophy of Science, 45(2), 555–570.
Näätänen, R. & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. Psychophysiology, 24(4), 375–425.
Nagy, E. & Molnar, P. (2004). Homo imitans or homo provocans? The phenomenon of neonatal initiation. Infant Behavior and Development, 27, 57–63.
Nathan, Z., Katz, D. & Zafonte, R. (2007). Brain injury medicine: Principles and practice. New York, NY: Demos Medical Publishing.
Nichols, S. & Bruno, M. (2010). Intuitions about personal identity: An empirical study. Philosophical Psychology, 23(3), 293–312.
Nordgren, A. (2015). Privacy by design in personal health monitoring. Health Care Analysis, 23, 148–164.
Northoff, G. (2006). Neuroscience of decision making and informed consent: An investigation in neuroethics. Journal of Medical Ethics, 32(2), 70–73.
Northoff, G. (2009). What is neuroethics? Empirical and theoretical neuroethics. Current Opinion in
Psychiatry, 22(6), 565–569.
Nozick, R. (1981). Philosophical explanations. Cambridge, Mass.: Harvard University Press.
Nyholm, S. & O'Neill, E. (2016). Deep brain stimulation, continuity over time, and the true self. Cambridge Quarterly of Healthcare Ethics, 25(4), 647–658.
Oderberg, D. (1997). Modal properties, moral status and identity. Philosophy and Public Affairs, 26(3), 259–298.
O'Keefe, D. J. (2012). Conviction, persuasion, and argumentation: untangling the ends and means of influence. Argumentation, 26(1), 19–32.
Olson, E. T. (2007). What Are We? A Study in Personal Ontology. New York: Oxford University Press.
Olson, E.T. (2016). The role of the brainstem in personal identity, in: Blank, A. (Ed.): Animals: New Essays, 291-302. München: Philosophia Verlag.

Orrell, M. & Brayne, C. (2015). Dementia prevention: call to action. The Lancet, 386(10004), 1625.
Owens, J. & Cribb, A. (2019). 'My Fitbit Thinks I Can Do Better!' Do Health Promoting Wearable Technologies Support Personal Autonomy?. Philosophy and Technology, 32, 23–38. https://doi.org/10.1007/s13347-017-0266-2.
Parfit, D. (1984). Reasons and Persons. Oxford: Clarendon.
Parfit, D. (2012). We are not human beings. Philosophy, 87(1), 5–28.
Pascual, L., Rodrigues, P. & Gallardo-Pujo, D. (2013). How does morality work in the brain? A functional and structural perspective of moral behavior. Frontiers in Integrative Neuroscience, 7, 65.
Pavey, L.J. & Sparks, P. (2010). Autonomy and reactions to health-risk information. Psychology and Health, 25(7), 855–872.
Pletti, C., Sarlo, M., Palomba, D., Rumiati, R. & Lotto, L. (2015). Evaluation of the legal consequences of action affects neural activity and emotional experience during the resolution of moral dilemmas. Brain Cognition, 94, 24–31.
Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? Trends in Cognitive Sciences, 10(2), 59–63.
Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology, 118(10), 2128–2148.
Raine, A. & Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. Social Cognitive and Affective Neuroscience, 1(3), 203–213.
Reader, S. (2007). The other side of agency. Philosophy, 82(4), 579–604.
Reid, T. (1969). Essays on the Intellectual Powers of Man. Cambridge: MIT Press.
Reiner, P. (2011). The rise of neuroessentialism, in: Sahakian, J.I.B. (Ed.): Oxford handbook of neuroethics, 161–177. New York: Oxford University Press.
Rescher, N. (1990). How wide is the gap between facts and values? Philosophy and Phenomenological Research, 50, 297–319.
Roskies, A. (2002). Neuroethics for the new millenium. Neuron, 35(1), 21–23.
Rossi, J. & Yudell, M. (2012). The use of persuasion in public health communication: an ethical critique. Public Health Ethics, 5(2), 192–205.
Roughley, N. (2016). Wanting and intending: elements of a philosophy of practical mind. Dordrecht: Springer.
Rueda, J. (2021). Socrates in the fMRI Scanner: The Neurofoundations of Morality and the Challenge to Ethics. Cambridge Quarterly of Healthcare Ethics, 30(4), 606-612.
Rudner, R. (1953). The scientist qua scientist makes value judgments. Philosophy of Science, 20(1), 1–6.
Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68.
Sackris, D. & Larsen, R. (2022). The disunity of moral judgment: Evidence and implications. Philosophical Psychology, 1, 1-20.
Sarlo, M., Lotto, L., Rumiati, R. & Palomba, D. (2014). If it makes you feel bad, don't do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas. Physiology & Behavior, 130, 127–134.
Schechtman, M. (1997). The brain-body problem. Philosophical Psychology, 10(2), 149–164.
Schechtman, M. (2010). Personhood and the practical. Theoretical Medicine and Bioethics, 31(4), 271–283.

Schechtman, M. (2014). Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life. Oxford, UK: Oxford University Press.
Schilbach, L. et al. (2013). Toward a second-person neuroscience. Behavioral and Brain Sciences, 36(4), 393–414.
Schleidgen, S., Jungert, M.C. & Bauer, R.H. (2010). Mission: Impossible? On empirical-normative collaboration in ethical reasoning. Ethical Theory and Moral Practice, 13(1), 59–73.
Schlosser, M.E. (2014). The neuroscientific study of free will: A diagnosis of the controversy. Synthese, 191(2), 245–262.
Schomer, D. & Lopes da Silva, F. (2012). Niedermeyer's Electroencephalography. Philadelphia, PA: Lippincott Williams & Wilkins.
Searle, J.R. (1964). How to derive ''ought'' from ''is''. Philosophical Review, 73(1), 43–58.
Shafir, R., Schwartz, N., Blechert, J. & Sheppes, G. (2015). Emotional intensity influences pre-implementation and implementation of distraction and reappraisal. Social Cognitive and Affective Neuroscience, 10(10), 1329.
Sharon, T. (2017). Self-tracking for health and the quantified self: re-articulating autonomy, solitary and authenticity in an age of personalized healthcare. Philosophy & Technology, 30(1), 93–121.
Shelton, W. & Geppert, C. (2024). Limits of advance directives in decision-making around food and nutrition in patients with dementia. Journal of Medical Ethics, 50(11), 762-765.
Shoemaker, D. (2007). Personal identity and practical concerns. Mind, 116(462), 317–357.
Shoemaker, D. (2016). The stony metaphysical heart of animalism, in: Blatti, S. & Snowdon, P. (Ed.): Animalism, 303–328. Oxford: Oxford University Press.
Shoemaker, S. (1963). Self-knowledge and Self-identity. Ithaca: Cornell University Press.
Shoemaker, S. (2008). Persons, animals, and identity. Synthese 163(3), 313–324.
Singer, P. (1979). Practical ethics. Cambridge: Cambridge University Press.
Singer, P. (2005). Ethics and intuitions. Journal of Ethics, 9(3–4), 331–352.
Sinhababu, N. (2024). The Reliable Route from Nonmoral Evidence to Moral Conclusions. Erkenntnis, 89(6), 2321-2341.
Sinnott-Armstrong, W. (2008). Framing moral intuitions, in: Sinnott-Armstrong, W. (Ed.): Moral psychology, vol. 2: The cognitive science of morality, pp. 47–76. Cambridge, MA: MIT Press.
Sinnott-Armstrong, W. (2008). The neuroscience of morality: emotion, brain disorders, and development. Vol. 3, Moral philosophy. Cambridge, MA: MIT Press.
Skrzypek, J.W. & Mangino, D. (2021). Should animalists be '"Transplanimalists"'?. Axiomathes, 31, 105–124.
Snowdon, P. (2014). Persons, Animals, Ourselves. New York: Oxford University Press.
Song, C.Y., Wu, Q. & Zhuang, T.G. (2006). Hybrid weighted minimum norm method a new method based LORETA to solve EEG inverse problem. Conference Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society 2005, 1079–1082.
Spahn, A. (2012). And lead us (not) into persuasion…? Persuasive technology and the ethics of communication. Science and Engineering Ethics, 18(4), 633–650.
Sterelny, K. (2010). Minds: extended or scaffolded?. Phenomenology and the Cognitive Sciences, 9(4), 465–481.

Steriade, M. (2006). Grouping of brain rhythms in corticothalamic systems. Neuroscience, 137(4), 1087–1106.
Stevenson, C.L. (1944). Ethics and language. New Haven: Yale University Press.
Strohminger, N. & Nichols, S. (2014). The essential moral self. Cognition, 131(1), 159–171.
Strohminger, N. & Nichols, S. (2015). Neurodegeneration and identity. Psychological Science, 26(9), 1469–1479.
Stevenson, R.A., Zemtsov, R.K. & Wallace, M.T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. Journal of Experimental Psychology: Human Perception and Performance, 38(6), 1517–1529.
Sunstein, C. & Vermeule, A. (2005). Is capital punishment morally required? The relevance of life-life tradeoffs. University of Chicago Law & Economics Olin Working Paper No. 239.
Tannenbaum, J. & Jaworska, A. (2013). The grounds of moral status. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2013/entries/grounds-moral-status/. Accessed 10 June 2017.
Tanner, J. (2006). The naturalistic fallacy. Richmond Journal of Philosophy, 13. http://www.richmondphilosophy.net/rjp/rjp13_tanner.php. Accessed 11 May 2015.
Taylor, R. (1982). Agent and patient: is there a distinction? Erkenntnis, 18(2), 223–232.
Thompson, D.F., Ramos, C.L. & Willett, R.K. (2014). Psychopathy: clinical features, developmental basis and therapeutic challenges. Journal of Clinical Pharmacy and Therapeutics, 39(5), 485–495.
Thomson, J.J. (1985). The trolley problem. Yale Law Journal, 94, 1395–1415.
Timmer, J., Kool, L. & van Est, R. (2015). Ethical challenges in emerging applications of persuasive technology, in: MacTavish, T. & Basapur, S. (Ed.): Persuasive technology. PERSUASIVE 2015. Lecture Notes in Computer Science, vol 9072, 196–201. Cham: Springer.
Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28, 675–691.
Tooley, M. (1972). Abortion and infanticide. Philosophy & Public Affairs, 2(1), 37–65.
Trevarthen, C. (1980). The foundations of intersubjectivity: Development of interpersonal and cooperative understanding in infants, in: Olson, D. (Ed.): The social foundations of language and thought, 316-342. New York: W.W. Norton & Co.
Van Berkum, J., Holleman, B., Nieuwland, M., Otten, M. & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. Psychological Science, 20(9), 1092–1099.
Vollmann, J. (2001). Advance directives in patients with Alzheimer's disease; ethical and clinical considerations. Medicine, Health Care and Philosophy, 4(2), 161–167.
Wagner, N.-F. & Northoff, G. (2015). A Fallacious Jar? The Peculiar Relation between Descriptive Premises and Normative Conclusions in Neuroethics. Theoretical Medicine and Bioethics, 36(3), 215-235.
Wagner, N.-F., Chaves, P. & Wolff, A. (2017). Discovering the Neural Nature of Moral Cognition? Empirical, Theoretical, and Practical Challenges in Bioethical Research with Electroencephalography (EEG). Journal of Bioethical Inquiry, 14(2), 299-313.
Wagner, N.-F. (2019). Against Cognitivism About Personhood. Erkenntnis, 84(3), 657-686.

Wagner, N.-F. (2019). Doing Away with the Agential Bias: Agency and Patiency in Health Monitoring Applications. Philosophy and Technology, 32(1), 135-154.
Wagner, N.-F. (2022): Personal Identity, Possible Worlds, and Medical Ethics. Medicine, Health Care and Philosophy, 25, 429-437.
Walsh, F.M. (2008). The return of the naturalistic fallacy: A dialogue on human flourishing. Heythrop Journal, 49(3), 370–387.
Walter, W.G., Cooper, R., Aldridge, V.J., McCallum, W.C. & Winter, A.L. (1964). Contingent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. Nature, 203(4943), 380–384.
Warren, M. (1977). Do potential people have moral rights? Canadian Journal of Philosophy, 7(2), 275–289.
Weinberg, A. & Hajcak, G. (2010). Beyond good and evil: The time-course of neural activity elicited by specific picture content. Emotion, 10(6), 767–782.
Weintraub, M. & Barilan, M. (2001). Persuasion as respect for persons: an alternative view of autonomy and of the limits of discourse. Journal of Medicine and Philosophy, 26(1), 13–34.
Weiskopf, D. A. (2008). Patrolling the mind's boundaries. Erkenntnis, 68(2), 265–276.
Wieczorek, Michał & Rossmaier, L. (2023). Healthiness as a Virtue: The Healthism of mHealth and the Challenges to Public Health. Public Health Ethics, 16(3), 219-231.
Wiggins, D. (2001). Sameness and Substance Renewed. Cambridge: Cambridge University Press.
Wilkes, K. (1988). Real People. Oxford: Clarendon.
Wilson, D.S., Dietrich, E. & Clark, A.B. (2003). On the inappropriate use of the naturalistic fallacy in evolutionary psychology. Biology and Philosophy, 18(5), 669–681.
Wilson, R.A., Foglia, L., Shapiro, L., Spaulding, S. (2021). Embodied Cognition, in: Zalta, E.N. (Ed.): *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*. https://plato.stanford.edu/archives/sum2021/entries/embodied-cognition/. Accessed 08 April 2024.
Wittgenstein, L. (1953). Philosophical investigations. Oxford: Blackwell.
Wittgenstein, L. (2009). Philosophical Investigations. Oxford: Wiley.
Yoder, K. & Decety, J. (2014). The good, the bad, and the just: Justice sensitivity predicts neural response during moral evaluation of actions performed by others. Journal of Neuroscience, 34(12), 4161–4166.