Nachdenken und Vordenken – Herausforderungen an die Philosophie

Herausgeber:
Oliver Petersen, Dagmar Borchers, Thomas Spitzley, Manfred Stöckler

gap
gesellschaft für
analytische
philosophie

Proceedings von GAP.7
Nachdenken und Vordenken –
Herausforderungen an die Philosophie

Herausgeber: Oliver Petersen, Dagmar Borchers,
Thomas Spitzley, Manfred Stöckler

Online-Veröffentlichung der
Universität Duisburg-Essen (DuEPublico)
2012

ISBN 978-3-00-036440-2

Vorwort

Vom 14.-17.9.2009 fand in Bremen der siebte Internationale Kongress "Nachdenken und Vordenken – Herausforderungen an die Philosophie" statt. Neben Hauptvorträgen und Kolloquiumsvorträgen gab es auch wieder Sektionsvorträge, in denen nicht nur, aber insbesondere auch viele Nachwuchskräfte ihre Forschungsarbeit präsentieren und im Anschluss diskutieren konnten. Von nicht allen aber einigen der deutlich über 200 Sektionsvorträge wurden seitens der Vortragenden Ausarbeitungen vorgenommen. Diese befinden sich in dieser Veröffentlichung.

Zwar wurden für die Ausarbeitungen Formatangaben gemacht, aber natürlich bedurften die Einreichungen doch noch der ein oder anderen Politur. Für die unglaubliche Unterstützung bei dieser sehr aufwendigen Arbeit danke ich Jannike Hensel, Johanna Krull, Anne Vogelgesang und vor allem Rebecca Hub.

Die Konferenz hätte ohne die organisatorischen Tätigkeiten des Bremer Philosophischen Instituts und insbesondere ohne die der Kongressausrichter Dagmar Borchers und Manfred Stöckler nicht so stattfinden können, wie sie das getan hat. Das Gleiche gilt für die Arbeiten der Kongressassistentinnen Nadja Niestädt und Kerstin Schnaars, des gesamten GAP-Vorstandes, der Stegmüller- und Ontos-Preis-Kommissionen, der Sektionsleiter und der Gutachter der Sektionseinreichungen. Allen sei hiermit für ihre großartige Mitarbeit gedankt. Ganz besonders gilt der Dank jedoch Thomas Spitzley, der wie kein Anderer zum Gelingen dieser Konferenz beigetragen hat.

Schließlich sei selbstverständlich auch allen Geldgebern gedankt, die diesen Kongress mitfinanziert haben: der DFG, der Universität Bremen (und besonders dem Fachbereich 9), der Sparkasse Bremen, den Unifreunden Bremen und der Philosophischen Gesellschaft Bremen.

Oliver Petersen

Inhalt

1 Logik und Wissenschaftstheorie

Nichtmonotone Reduktionsregeln für Dispositionen

Holger Andreas
holger.andreas@lrz.uni-muenchen.de
Seminar für Philosophie, Logik und Wissenschaftstheorie - LMU München

Abstract/Zusammenfassung

The analysis of dispositions has received much attention in the past two decades and was inspired by new thought experiments of what a disposition can be like. Arguably, there are three types of analysis that can be distinguished: 1) conditionals in the form of reduction sentences, 2) conditionals in the form of counterfactuals, 3) the view of dispositions as non-reducible modal properties or powers. The detailed description of so-called finkish dispositions by Martin (1994) had a profound impact on recent publications. Lewis (1997) and Malzkorn (2000) are attempts to refine the counterfactual analysis, whereas Molnar (1999) is an alleged refutation of Lewis (1997) and Mumford (2001) such a refutation of Malzkorn (2000). The latter two authors clearly favour an account of dispositions in terms of non-reducible powers and primitive modal facts being resistant to a conditional analysis.

In this paper, an attempt is made to refine the conditional analysis in terms of reduction sentences so as to account for finkish dispositions. The key idea of the refinement is the use of methods of defeasible reasoning. To my knowledge, such a strategy has not been pursued in the literature so far. What is the motivation to improve upon the old-fashioned reduction sentence analysis of dispositions by means of defeasible reasoning? Questioning the alleged victory of a realist analysis in terms of powers and primitive modal properties could be suspected to be the major motivation. My primary intention, however, is to make understood the inferential techniques on which the ascription of disposition concepts is based in ordinary and scientific linguistic practice. The non-conditional realist analysis has not much to say about this. It is doubtful, moreover, whether the counterfactual considerations coming with Lewis´s account are the ones governing the ascription of disposition concepts. Thus, it will be claimed that the present reduction sentence approach is a more promising candidate for a meaning analysis of disposition concepts than Lewis´s

counterfactual analysis as well as the non-conditional analysis in terms of powers.

Die Analyse von Dispositionen ist in den letzten beiden Jahrzehnten ein Gegenstand der intensiven Diskussion gewesen und dabei durch neue Gedankenexperimente inspiriert worden. Drei verschiedene Typen der Analyse lassen sich unterscheiden: 1) Konditionalsätze in der Form von Reduktionssätzen; 2) Konditionalsätze in der Form von kontrafaktischen Konditionalen; 3) die Auffassung, dass Dispositionen irreduzible modale Eigenschaften oder Kräfte sind. Die Beschreibung von Dispositionen, die Martin (1994) als „finkish" bezeichnet - wir sprechen im Folgenden von Fink-Dispositionen -, hatte einen nachhaltigen Einfluss auf die Publikationen der jüngeren Vergangenheit. Lewis (1997) und Malzkorn (2000) erarbeiten Vorschläge zur Verbesserung der kontrafaktischen Analyse. Molnar (1999) hingegen zielt darauf ab, den Ansatz von Lewis (1997) zu widerlegen, während Mumford (2001) behauptet, eine Widerlegung des Ansatzes von Malzkorn (2000) anzugeben. Mumford und Molnar glauben und argumentieren, dass Dispositionen über irreduzible Kräfte und einfache modale Fakten verstanden werden müssen und sich damit einer konditionalen Analyse entziehen.

In dem vorliegenden Aufsatz wird der Versuch unternommen, die konditionale Analyse in der Form von Reduktionssätzen zu verbessern durch Methoden des nichtmonotonen Schliessens. Im Ergebnis dieser Verbesserung können auch Fink-Dispositionen durch diese Form der Analyse erfasst werden. Welche Motivation steht hinter dem vorliegenden Versuch, die altmodische Methode der Reduktionssätze durch Methoden des nichtmonotenen Schließens zu verbessern? Man könnte darin vor allem den Versuch einer Kritik am scheinbaren Erfolg einer realistischen Analyse von Dispositionen erkennen. Meine primäre Intention hingegen ist die Analyse der inferentiellen Beziehungen, aufgrund welcher die Zuschreibung von Dispositionsprädikaten vorgenommen wird. Die nicht konditionale Analyse liefert keinen Beitrag hierfür. In Anbetracht der viel kritisierten epistemischen Probleme, mit denen der Lewissche Ansatz für Konditionalsätze verbunden ist, muss darüber hinaus in Frage gestellt werden, ob dieser Ansatz tatsächlich unsere Praxis der Zuschreibung von Dispositionsprädikaten angemessen erfasst. Daher möchte ich die These aufstellen, dass die vorliegende *Default-Theorie* für Dispositionen, die eine Verbesserung der Carnapschen Reduktionssätze darstellt, sowohl der Lewisschen, konditionalen Analyse als auch der nicht konditionalen „Analyse" über einfache und ursprüngliche Kräfte überlegen ist.

1. Die konditionale Analyse

Es gibt zwei Typen der konditionalen Analyse, zum einen in der Form von kontrafaktischen Konditionalsätzen, zum anderen in der Form von sogenannten Reduktionssätzen. Wesentlich für die Analyse ist die Unterscheidung zwischen einer Testbedingung T, einer charakteristischen Manifestation M und der Disposition D. Im Folgenden möchte ich die beiden Typen der konditionalen Analyse darlegen, um dann zur Kritik an dieser Form der Analyse überzugehen.

1.1. Die kontrafaktische konditionale Analyse

Unter Verwendung des von Lewis eingeführten kontrafaktischen Konditionals lässt sich die kontrafaktische konditionale Analyse folgendermaßen formalisieren:

$$(1) \quad \forall x(D(x) \leftrightarrow \forall t(T(x,t) \boxdot\!\!\rightarrow M(x,t)))$$

In Worten: Ein Gegenstand hat die Disposition D genau dann wenn gilt: Wenn x zu einem Zeitpunkt t einem Test unterworfen wird, dann zeigt x die für die Disposition charakteristische Manifestation zum Zeitpunkt t. D(x) steht entsprechend für die Disposition D, T(x,t) für die Testbedingung und M(x,t) für die charakteristische Manifestation.

An dieser Stelle verdient auch eine nicht formalisierte, kausale Variante der kausalen Analyse Beachtung. Danach hat ein Gegenstand x eine Disposition D genau dann wenn gilt: x hat die Eigenschaft G, die verursachen würde, dass x unter einer Testbedingung T die charakteristische Manifestation M zeigt (vgl. Molnar, 1999, S. 2).

1.2. Carnaps Ansatz der Reduktionssätze

Anders als im Logischen Aufbau der Welt (1928) vertritt Carnap in der Abhandlung Testability and Meaning (1936/37) eine schwächere Form des Reduktionismus, die eine ausdrückliche Verneinung der These der Übersetzbarkeit von theoretischen Aussagen in empirisch-basale Aussagen einschließt. Das basale Reduktionssatz-Schema hat die folgende Form:

$$(2) \quad \forall x \forall t(T(x,t) \rightarrow (M(x,t) \rightarrow D(x,t)))$$

Dieses Schema ist nur für sogenannte Augenblicksdispositionen gültig, d. h. für Dispositionen, die nur für den Moment der Manifestation präsent sind. Eine Modifikation des Schemas für Permanentdispositionen sieht folgendermaßen aus:

$$(3) \quad \forall x \forall t (T(x,t) \rightarrow (M(x,t) \rightarrow D(x)))$$

Das Schema (3) kann in der folgenden Weise variiert werden:

$$(4) \quad \forall x \forall t (T(x,t) \rightarrow (M(x,t) \leftrightarrow D(x)))$$

Diese Variante gestattet es, aus dem Nichtmanifestwerden der Disposition unter der Testbedingung T die Abwesenheit der Disposition zu schließen. Weitere Differenzierungen, die Carnap innerhalb des Reduktionssatz-Schemas einführte, werden für die nun folgenden Überlegungen nicht relevant sein.

2. Die Kritik an der konditionalen Analyse

2.1. Spekulative Fink-Dispositionen

Was ist eine Fink-Disposition? Ein Gegenstand hat eine solche Disposition, wenn er disponiert ist, in bestimmter Weise auf eine Testbedingung zu reagieren, aber dieses Disponiertsein verliert, sobald er tatsächlich einer Testbedingung unterworfen wird. Darüber hinaus spricht man von dem Vorliegen einer Fink-Disposition, wenn der Gegenstand x an sich keine Disposition hat, in bestimmter Weise auf eine Testbedingung zu reagieren, aber eine solche Disposition gewinnt, sobald er tatsächlich einer Testbedingung ausgesetzt wird. Die Beispiele, die für Fink-Dispositionen angegeben werden, sind eher spekulativer Natur und ohne wissenschaftliche Basis. So muss man sich zum Beispiel vorstellen, dass ein Eisblock seine Zerbrechlichkeit genau in dem Moment verliert, wenn er mit einem scharfen Gegenstand bearbeitet wird (Martin 1994). Es ist daher ein bedeutsames Verdienst von Lewis (1997) und Bird (1998), Beispiele für Fink-Dispositionen der realen Welt mit wissenschaftlicher Basis angegeben zu haben.

Warum stellt die behauptete Existenz von Fink-Dispositionen ein Problem für die kontrafaktische Analyse von Dispositionen dar? Betrachten wir den Fall, in dem ein Gegenstand x die Disposition D durch die Testbedingung T gewinnt. Nehmen wir des Weiteren an, dass zum Zeitpunkt t der Gegenstand nicht einer Testbedingung unterworfen wird. In diesem Fall ist das Definiens der Definition (1) erfüllt, so dass dem Gegenstand x die Disposition zugeschrieben wird,

obwohl x zum Zeitpunkt t die Disposition gar nicht aufweist, da - den Voraussetzungen nach - zum Zeitpunkt t gar keine Testbedingung vorliegt (Mumford 1998, Molnar 1996, Martin 1994). Dies scheint nicht akzeptabel zu sein.

Um den mit Fink-Dispositionen verbundenen Schwierigkeiten Rechnung zu tragen, ist die folgende Verbesserung der kontrafaktischen Analyse von Lewis (1997) entwickelt worden: Ein Gegenstand x ist disponiert, eine Manifestation M zu zeigen unter einer Testbedingung T genau dann, wenn x eine intrinsische Eigenschaft B hat derart, dass wenn T vorliegt und x unter dieser Testbedingung die Eigenschaft B behalten würde, dann würde der Gegenstand x die Manifestation M zeigen, und zwar aufgrund eines kausalen Zusammenhangs mit B. Es scheint offensichtlich zu sein, warum und in welcher Weise diese Verbesserung dem Problem der Fink-Dispositionen Rechnung trägt.

Bereiten Fink-Dispositionen irgendwelche Schwierigkeiten für den anderen Typ der konditionalen Analyse, Carnaps Ansatz der Reduktionssätze? Die ebenso überraschende wie erfreuliche Antwort auf diese Frage ist, dass der Ansatz der Reduktionssätze nicht vom Problem der Fink-Dispositionen betroffen ist. Die gerade beschriebene Situation, in welcher die Instanzierung des Definiens wahr und die Instanzierung des Definiendums falsch ist, konnte nur deshalb auftreten, weil das Definiens den Modus einer kontrafaktischen Aussage hat. Eine solche Form der Aussage wird in Carnaps Ansatz der Reduktionssätze nicht verwendet.

2.2. Fink-Dispositionen der realen Welt

Bird (1998) gibt das folgende Beispiel für eine Fink-Disposition der realen Welt an: Ein bestimmtes Gift hat die Disposition, einen Menschen zu töten. Doch selbst nach Einnahme des Giftes bleibt noch genügend Zeit, ein Gegengift zu nehmen, welches den Tod verhindert. Nichtsdestoweniger würden wir sagen, dass das Gift die Disposition hat, tödlich für den menschlichen Organismus zu sein. Charakteristisch für Fink-Dispositionen der realen Welt ist, dass intervenierende Bedingungen eintreten können, die das Manifestwerden der Disposition verhindern. Dieser Typ von Fink-Dispositionen stellt ein echtes Problem für das Reduktionssatz-Schema dar. Eine Anwendung dieses Schemas würde in der Tat zu dem kontraintuitiven Ergebnis führen, dass das Gift nicht die Disposition zu töten hat. Im folgenden Abschnitt werde ich zeigen, wie das Reduktionssatz-Schema durch Methoden des nichtmonotonen Schließens derart verbessert werden kann, dass auch Fink-Dispositionen der realen Welt angemessen analysiert werden.

3. Nichtmonotone Inferenzen

In Vorbereitung einer Default-Theorie für Dispositionen sollen in diesem Abschnitt in aller Kürze die Grundbegriffe der Default-Logik eingeführt werden. Dieses System zeichnet sich durch universale Verwendbarkeit und intuitive Plausibilität aus und wird entsprechend für viele Anwendungen gegenüber anderen nichtmonotonen Systemen bevorzugt (Antoniou 1997). Zunächst zur Unterscheidung zwischen monotonen und nichtmonotonen Begründungen. Die Konsequenzrelation einer Logik ist monoton genau dann, wenn die Erweiterung der Prämissenmenge niemals dazu führt, dass eine Konklusion aus dieser Prämissenmenge ungültig wird. Im Fall nichtmonotoner Logiken hingegen kann der Fall eintreten, dass die Gültigkeit einer Folgerung aus einer Menge von Prämissen aufgehoben wird durch eine Erweiterung der Prämissenmenge.

Default-Regeln repräsentieren allgemeine Zusammenhänge, die allerdings nicht ohne Ausnahmen gelten. Das bekannteste Beispiel für eine solche Regel ist die Aussage, dass alle Vögel die Fähigkeit haben zu fliegen. Default-Regeln haben in der Default-Logik die folgende syntaktische Gestalt:

(5) $A : B_1, \ldots B_n / C$

In Worten: Wenn A gilt und ¬B1, ..., ¬Bn nicht von der Wissensbasis - der jeweils aktuellen Überzeugungsmenge - abgeleitet werden können, dann C. Anders formuliert: Wenn A und es konsistent ist anzunehmen, dass auch B1, ..., Bn den Wert das Wahre haben, dann C. B1, ..., Bn werden auch als Konsistenz- oder Rechtfertigungsbedingungen bezeichnet. Eine besonders einfache und intuitiv leicht verständliche Form haben sogenannte normale Defaults:

(6) $A : C / C$

Eine Default-Theorie ist ein geordnetes Paar (D, W), bestehend aus einer Menge von Defaults D und einer Menge von Sätzen W in einer formalen Sprache der ersten Stufe. Die zuletzt genannte Menge repräsentiert die in strikter Weise gültigen Behauptungen der jeweiligen Default-Theorie. Neben atomaren und molekularen Sätzen kann W auch allquantifizierte Sätze enthalten. Inferenzbeziehungen werden in der Default-Logik über den Begriff der Erweiterung einer Default-Theorie (D, W) eingeführt:

Definition 1. Sei (D,W) eine Default-Theorie. Der Operator Γ ordnet jeder Menge S von Formeln die kleinste Menge U von Formeln zu mit den Eigenschaften i) $W \subset U$; ii) $Cn(U) = U$; iii) wenn $A : B_1 \ldots B_n / C \in D$, $U \models A$, $S \not\models \neg B_i, 1 \leq n$, dann $C \in U$.

Definition 2. Ein Satz ϕ ist eine *simple Konsequenz* einer Default-Theorie $T = (D,W)$ - in Symbolen: $T \mid\sim_c \phi$ g.d.w. ϕ Element mindestens einer Erweiterung von T ist.

Definition 3. Ein Satz ϕ ist eine *skeptische Konsequenz* einer Default-Theorie $T = (D,W)$ - in Symbolen: $T \mid\sim_s \phi$ g.d.w. ϕ Element aller Erweiterungen von T ist.

Im Englischen werden die simplen Konsequenzen auch „credulous“ (leichtgläubig) bezeichnet.

4. Eine Default-Theorie für Dispositionen

Versuchen wir nun, die neuen Ausdrucksmöglichkeiten der Default-Logik zu nutzen, um das etwaige Vorhandensein von intervenierenden Bedingungen bei dem Test von Dispositionen zu berücksichtigen. Die folgende Default-Regel besagt nichts anderes, als dass eine Disposition nur dann manifest wird, wenn die Behauptung des Manifestwerdens tatsächlich in konsistenter Weise mit unseren Überzeugungen vereinbar ist.

(7) $\quad T(x,t) \wedge D(x) : M(x,t) / M(x,t)$

In Worten: Wenn ein Gegenstand x der Testbedingung T zum Zeitpunkt t unterworfen wird, dann manifestiert sich die Disposition, es sei denn, das Manifestwerden M der Disposition lässt sich nicht in konsistenter Weise zu unseren Überzeugungen hinzufügen. Darüber hinaus gilt das folgende Axiom ohne Ausnahme:

(8) $\quad P(x,t) \rightarrow \neg M(x,t)$

Wenn für einen Gegenstand x eine intervenierende Bedingung P zum Zeitpunkt t vorliegt, dann manifestiert sich die Disposition nicht. Das folgende Axiom besagt, dass das Manifestwerden M zum Zeitpunkt t eindeutige Evidenz für das Vorliegen einer entsprechenden Disposition D ist:

(9) $\quad T(x,t) \rightarrow (M(x,t) \rightarrow D(x))$

Schließlich benötigen wir noch eine Regel, mit der wir das Nichtvorliegen einer Disposition folgern können daraus, dass diese Disposition unter der entsprechenden Testbedingung nicht manifest wird:

(10) $T(x,t) \wedge \neg M(x,t) : \neg P(x)/\neg D(x)$

In Worten: Wenn der Gegenstand x zum Zeitpunkt t der Testbedingung T unterworfen wird und dabei die Disposition D nicht manifest wird, dann liegt D nicht vor, sofern es konsistent ist anzunehmen, dass keine intervenierende Bedingung zum Zeitpunkt t vorgelegen hat.

Die soeben entwickelte Default-Theorie für Dispositionen repräsentiert nun in angemessener Weise unsere inferentiellen Übergänge, mit denen wir das Vorliegen von Dispositionen Gegenständen zuschreiben und deren Manifestwerden vorhersagen. Sei W_0 die Menge der Axiome unserer Default-Theorie, also diejenige Menge, die genau die Axiome (8) und (9) als Elemente enthält. Sei D_0 die Menge der Default-Regeln der Default-Theorie, also die Menge, bestehend aus den Defaults (7) und (10). Nehmen wir nun an, dass ein Gegenstand a die Disposition D hat und zum Zeitpunkt t_0 der Testbedingung T unterworfen wird. In Symbolen: $KB = (D_0, W_0 \cup \{D(a)\} \cup \{T(a,t_0)\})$. Dann gilt:

(11) $KB \mid\sim_s M(a,t_0)$

Dies bedeutet, aus dem Vorliegen der Disposition kann geschlussfolgert werden, dass sich dieselbe auch tatsächlich manifestiert, allerdings nur in fallibler Weise. Neu hinzukommende Informationen können die Gültigkeit dieser Schlussfolgerung wieder aufheben. Nehmen wir an, dass zum Zeitpunkt t_0 eine intervenierende Bedingung P vorliegt. Dann gilt:

(12) $KB \mid\not\sim_c M(a,t_0)$

Die Erweiterung der Prämissenmenge um die Annahme, dass eine intervenierende Bedingung vorliegt - $KB = (D_0, W_0 \cup \{D(a)\} \cup \{T(a,t_0)\} \cup \{P(a,t_0)\})$-, hat die Konsequenz, dass das Manifestwerden der Disposition nicht mehr gefolgert werden kann, was in korrekter Weise die Sachlage von Birds Beispiel (1998) wiedergibt.

Betrachten wir nun die Zuschreibung einer Disposition für den Fall, dass die Disposition D nicht manifestiert wird, obwohl der Gegenstand einer Testbedingungen unterworfen wurde. In Symbolen: $KB = (D_0, W_0 \cup \{T(a,t_0)\} \cup \{\neg M(a,t_0)\})$. Dann gilt:

(13) $KB \mid\sim_s \neg D(a)$

Wenn wir jedoch erfahren oder erschließen können, dass eine intervenierende Bedingung z um Zeitpunkt t_0 vorgelegen ha t, dann können wir nicht mehr ausschließen, dass die Disposition D vielleicht doch vorgelegen hat, obwohl sie gar nicht manifest wurde:

(14) $KB \not| J\sim_s \neg D(a)$

Hiermit ist gezeigt worden, dass es ein System von Axiomen und Inferenzregeln gibt, in dem Fink-Dispositionen der realen Welt in angemessener und, wie ich behaupten möchte, eleganter Weise rekonstruiert werden können. Dieses System stellt eine Verbesserung des Carnapschen Reduktionssatz-Schemas d ar. Mumford (1998) und Molnar (1996) haben zu Recht darauf hingewiesen, da ss Fink-Dispositionen im Rahmen des Carnapschen Ansatzes nicht erfasst werden können. Nach den Ergebnissen der vorliegenden Untersuchung s ollte daraus jedoch nicht die Schlussfolgerung gezogen werden, dass eine inferenzlogische Charakterisierung v on Dispositionen nicht möglich ist. Anders als Mumford (1998) behauptet, gibt es durchaus eine Möglichkeit der Rekonstruktion von Dispositionen im Rahmen der Wissenschaftslogik.

Literaturverzeichnis

Antoniou, G.: Nonmonotonic Reasoning, MIT-Press, Cambridge, 1997

Carnap, R.: Der Logische Aufbau der Welt, Meiner, Berlin, 1928

Carnap, R.: "Testability and Meaning". Philosophy of Science 3/4, 1936/37. S. 419-471/1-40

Dummett, M.: Truth and Other Enigmas, Duckworth, London, 1978

Lewis, D.: "Finkish Dispositions". The Philosophical Quarterly 47, (187), 1997. S. 143-158

Malzkorn, W.: "Realism, Functionalism and the Conditional Analysis of Dispositions". The Philosophical Quarterly, 50 (201), 2001. S. 452-469

Martin, C. P.: "Dispositions and Conditionals". The Philosophical Quarterly, 44 (174), 1994. S. 1-8

Mumford, S.: "Realism and the Dispositional Analysis of Dispositions: Reply to Malzkorn". The Philosophical Quarterly 51 (201), 2001. S. 375-378

Mumford, S.: Dispositions, Oxford, 1998

Molnar, G.: "Are Dispositions Reducible?". The Philosophical Quarterly 49 (194), 1996. S. 1-17

Daten, Phänomene und empirische Adäquatheit – van Fraassens unerledigte Hausaufgaben

Jochen Apel
jochen-apel@web.de
Philosophisches Seminar der Universität Heidelberg

Abstract/Zusammenfassung

In their paper *Saving the Phenomena* James Bogen and James Woodward introduce the distinction between data and phenomena and develop an objection to Bas van Fraassen's constructive empiricism which builds on this distinction. The aim of my paper is to rebut Bogen and Woodward's objection.

The constructive empiricist's demand for empirical adequacy is, according to Bogen and Woodward, identical with a demand for theories which explain and predict observed data. However, in scientific practice theories do not explain and predict observed data, but rather phenomena, which have to be inferred from the data. This leads Bogen and Woodward to the conclusion that constructive empiricism is incompatible with actual scientific practice. I will present two objections to this argument. First, I will show that there are good reasons to believe that van Fraassen is not committed to the narrow notion of empirical adequacy which Bogen and Woodward suggest. Rather empirical adequacy can be explicated in a way such that it includes a certain subset of phenomena, namely phenomena which are represented by "patterns in data sets". Second, I will argue that, even if one accepts Bogen and Woodward's narrow notion of empirical adequacy, it is not the case that focusing on inferred phenomena instead of observed data in scientific theorizing and model-building is incompatible with the aim of science van Fraassen proposes. Both arguments allow a defence of constructive empiricism, but each builds on a different explication of empirical adequacy. Thus, although Bogen and Woodward's considerations do not provide a convincing objection, they nevertheless force the constructive empiricist to give a more precise explication of empirical adequacy by settling for one or the other line of argument.

In ihrem Aufsatz *Saving the Phenomena* stellen James Bogen und James Woodward die Unterscheidung zwischen Daten und Phänomenen vor und entwickeln anhand dieser ein Argument gegen Bas van Fraassens Konstruktiven Empirismus. Ziel meines Aufsatzes ist es, diesen Einwand zurückzuweisen.

Bogen und Woodward zufolge ist die Forderung des Konstruktiven Empiristen nach empirisch adäquaten Theorien identisch mit der Forderung danach, dass Theorien beobachtete Daten erklären und vorhersagen sollen. In der wissenschaftlichen Praxis würden jedoch nicht beobachtete Daten, sondern auf Grundlage der Daten erschlossene Phänomene durch Theorien erklärt und vorhergesagt. Deshalb, so der Vorwurf, sei der Konstruktive Empirismus mit der tatsächlichen wissenschaftlichen Praxis unvereinbar. Gegen diese Argumentation werde ich zwei Einwände vorbringen. Erstens werde ich aufzeigen, dass es gute Gründe dafür gibt, dass van Fraassen nicht auf den engen Begriff von empirischer Adäquatheit festgelegt ist, den Bogen und Woodward ihm unterstellen. Der Begriff der empirischen Adäquatheit kann viel-

mehr so verstanden werden, dass er eine bestimmte Teilklasse der Phänomene umfasst, nämlich solche, die durch „Muster in Datensätzen“ repräsentiert werden. Zweitens argumentiere ich dafür, dass es, selbst wenn man Bogens und Woodwards enges Verständnis von empirischer Adäquatheit akzeptiert, nicht der Fall ist, dass die Fokussierung auf erschlossene Phänomene statt auf beobachtete Daten in der wissenschaftlichen Modell- und Theoriebildung mit van Fraassens Position unvereinbar ist. Beide Argumentationslinien erlauben die Verteidigung des Konstruktiven Empirismus, sie setzen jedoch jeweils eine unterschiedliche Auffassung von empirischer Adäquatheit voraus. Wenngleich aus Bogens und Woodwards Überlegungen somit kein überzeugender Einwand erwächst, zwingen sie den Konstruktiven Empiristen dennoch dazu, den Begriff der empirischen Adäquatheit genauer zu spezifizieren, indem er sich für eine der beiden Argumentationslinien entscheiden muss.

1. Einleitung

Bas van Fraassens Konstruktiver Empirismus ist die in den vergangenen 30 Jahren wahrscheinlich meistdiskutierte Alternativposition zum Wissenschaftlichen Realismus. Allein aufgrund des Umfangs der Debatte, die seit dem Erscheinen von *The Scientific Image* zwischen Wissenschaftlichen Realisten und Konstruktiven Empiristen geführt wird, ist es kein Wunder, dass bisher nicht alle vorgebrachten Argumente von der jeweiligen Gegenseite aufgenommen und diskutiert wurden. Auf beiden Seiten gibt es noch unerledigte Hausaufgaben. Überraschend ist allerdings, dass gerade diejenige Problematik nicht eingehender untersucht wurde, die ich in diesem Aufsatz behandeln werde.

Bereits 1988 führen James Bogen und James Woodward in ihrem Aufsatz *Saving the Phenomena* die Unterscheidung zwischen Daten und Phänomenen ein und präsentieren einen auf dieser Unterscheidung beruhenden Einwand gegen den Konstruktiven Empirismus, demzufolge van Fraassens Position mit der wissenschaftlichen Praxis unvereinbar ist. Überraschend ist van Fraassens Nichtbeachtung dieses Einwandes deshalb, weil Bogens und Woodwards Daten-Phänomen-Unterscheidung in der zeitgenössischen Wissenschaftstheorie weithin anerkannt wird. Sollte sich aus solch einer allgemein akzeptierten Unterscheidung tatsächlich ein überzeugender Einwand gegen den Konstruktiven Empirismus ergeben, würde dies van Fraassen vor ernsthafte Schwierigkeiten stellen. Van Fraassen, so könnte man sagen, hat deshalb nicht irgendeine, sondern eine besonders wichtige Hausaufgabe bisher nicht erledigt. In diesem Aufsatz möchte ich versuchen diese Lücke zu schließen und dafür argumentieren, dass auf Grundlage der Daten-Phänomen-Unterscheidung kein überzeugendes Argument gegen den Konstruktiven Empirismus entwickelt werden kann, indem ich zwei Verteidigungsmöglichkeiten für den Konstruktiven Empiristen vorstelle. Allerdings muss der Konstruktive Empirist, je nachdem für welche der beiden Möglichkeiten er sich entscheidet, den Begriff der empirischen Adäquatheit unterschiedlich ausbuchstabieren. Insofern weist Bogens und Woodwards Ein-

wand, obwohl er letztlich nicht überzeugt, darauf hin, dass die Position des Konstruktiven Empirismus weiterer Klärung bedarf.

Ich werde folgendermaßen vorgehen: In Abschnitt 2 stelle ich die Auffassung des Konstruktiven Empirismus vor und führe aus, wie sie von van Fraassen begründet wird. Abschnitt 3 behandelt die Daten-Phänomen-Unterscheidung und den auf dieser Unterscheidung beruhenden Einwand Bogens und Woodwards gegen den Konstruktiven Empirismus. In Abschnitt 4 präsentiere ich zwei unterschiedliche Verteidigungsstrategien für den Konstruktiven Empiristen und argumentiere dafür, dass jede dieser Strategien hinreichend ist, um Bogens und Woodwards Einwand zurückzuweisen.

2. Die Position des Konstruktiven Empirismus

Das Anliegen des Konstruktiven Empirismus ist es, die Frage zu beantworten, was das Ziel der Wissenschaft ist. Van Fraassen zufolge lautet die Antwort auf diese Frage, dass Wissenschaft darauf abzielt, empirisch adäquate Theorien zu finden, d.h. Theorien, die wahre Aussagen über die beobachtbaren Teile der Wirklichkeit machen.[1] Bei wissenschaftlichen Aussagen über Unbeobachtbares hingegen wird der Konstruktive Empirist zum Agnostiker: Weder können noch müssen wir wissen, ob solche Aussagen wahr sind. Damit grenzt van Fraassen seine Auffassung insbesondere vom Wissenschaftlichen Realismus ab, demzufolge es das Ziel der Wissenschaft ist, wahre Theorien zu finden, d.h. Theorien, die auch diejenigen Bereiche, die uns nicht direkt in der Erfahrung zugänglich sind, korrekt beschreiben.

Doch warum sollte man sich eine solche Auffassung zu eigen machen? In aktuelleren Aufsätzen stellt van Fraassen zwei Kriterien vor, die zur Bewertung wissenschaftstheoretischer Positionen herangezogen werden sollen. Diese Kriterien sind:

a) epistemische Bescheidenheit
b) Angemessenheit gegenüber der wissenschaftlichen Praxis[2]

Das Kriterium der epistemischen Bescheidenheit besagt, dass man (*ceteris paribus*) solche Positionen akzeptieren sollte, die ein möglichst geringes Risiko, irrtümlich eine falsche Schlussfolgerung als wahr anzuerkennen, auf sich nehmen. Die Hochschätzung epistemischer Bescheidenheit ist ein zentrales Kennzeichen des van Fraassen'schen Empirismus, also seiner übergeordneten philosophi-

1 Vgl. van Fraassen (1980), S. 12.

2 Vgl. Monton und van Fraassen (2003). Der Ausdruck „epistemische Bescheidenheit" wird innerhalb der Debatte um den Konstruktiven Empirismus meines Wissen zum ersten Mal von Alspector-Kelly (2001) verwendet, den Ausdruck „Angemessenheit gegenüber der wissenschaftlichen Praxis" übernehme ich von Berg-Hildebrand und Suhm (2006).

schen Haltung, die sich insbesondere durch ihre metaphysikkritische Attitüde auszeichnet. Gute Empiristen sollten, laut van Fraassen, Wissensansprüche nur in Bereichen anmelden, die zumindest im Prinzip Gegenstand der Erfahrung sein können.

Das Angemessenheitskriterium wiederum besagt, dass eine philosophische Zielbestimmung der Wissenschaft mit der wissenschaftlichen Praxis in Einklang stehen muss. Das, was Wissenschaftler *de facto* tun, muss in ihrem Licht sinnvoll erscheinen. Dieses Kriterium ergibt sich ebenfalls aus van Fraassens empiristischer Haltung: Der Empirist ist, van Fraassen zufolge, ein Bewunderer der modernen Naturwissenschaft. Sie ist für ihn das Paradigma rationaler Erkenntnisgewinnung.[3]

Die Aufgabe der Wissenschaftstheorie ist es, diejenige Position zu identifizieren, die im Hinblick auf beide Kriterien am besten abschneidet. Diesen Theoriewahlprozess beschreibt van Fraassen folgendermaßen:

> Consider a range of possibilities with 'science aims to give us true theories' on the far right side, and 'science aims to give us theories which are true in what they say about what is being observed right now' on the far left side. Realists submit that attention to the practice of good science, where bold conjectures and audacious theorizing have been rewarded with much predictive success, moves us towards the right. Empiricists, who would wish for epistemic modesty in their paradigm of rational inquiry, would tend towards the left. Constructive empiricism finds an equilibrium point between the two extremes, thus respecting both desiderata.[4]

Der Vergleich anhand der beiden oben genannten Kriterien zeichnet den Konstruktiven Empirismus gegenüber dem Wissenschaftlichen Realismus aus. Dass der Konstruktive Empirismus epistemisch bescheidener als der Wissenschaftliche Realismus ist, ist offensichtlich. Wenn der Wissenschaftliche Realist eine Theorie akzeptiert, hält er auch die Aussagen der Theorie über Unbeobachtbares für (annäherungsweise) wahr. Der Konstruktive Empirist geht dieses epistemische Wagnis hingegen nicht ein. Aufwiegen könnte der Wissenschaftliche Realismus das geringere Maß an epistemischer Bescheidenheit, wenn er hinsichtlich des Angemessenheitskriteriums besser abschneiden würde. Van Fraassen zufolge trifft dies jedoch nicht zu, denn ein Wissenschaftler, der überprüfen möchte, ob eine Theorie wahr ist, würde sich nicht anders verhalten als einer, der ihre empirische Adäquatheit feststellen möchte. Schließlich stehen uns, um uns von der Wahrheit einer Theorie zu überzeugen, nur die in der Erfahrung gegebenen Belege zur Verfügung.[5] Deshalb lässt sich hinsichtlich des Angemessenheitskriteriums kein Unterschied zwischen beiden Positionen ausmachen. Insgesamt

3 Vgl. van Fraassen (2002), S. 63
4 Monton und van Fraassen (2003), S. 407.
5 Vgl. auch van Fraassen (1985), S. 255.

schneidet somit der Konstruktive Empirismus besser ab als der Wissenschaftliche Realismus. Mithin sollten wir Konstruktive Empiristen sein.[6]

3. Unerledigte Hausaufgaben: Bogens und Woodwards Einwand gegen den Konstruktiven Empirismus

Durch die vorhergehenden Überlegungen sind wir nun in der Position, die Scharnierstelle identifizieren zu können, an der Bogens und Woodwards Einwand ansetzt. Dazu werde ich zunächst in Abschnitt 3.1 die Daten-Phänomen-Unterscheidung vorstellen, um anschließend in Abschnitt 3.2 herauszuarbeiten, wie Bogen und Woodward diese Unterscheidung gegen van Fraassen zur Anwendung bringen.

3.1 Die Unterscheidung zwischen Daten und Phänomenen

In ihrem berühmten Aufsatz *Saving the Phenomena* und einer Reihe daran anschließender Texte führen Bogen und Woodward die Daten-Phänomen-Unterscheidung in die wissenschaftstheoretische Debatte ein.[7] Ziel ihrer Überlegungen ist es, ein unter Wissenschaftstheoretikern weit verbreitetes, aber zu stark vereinfachendes Wissenschaftsbild zu korrigieren. Diesem Bild zufolge liefern wissenschaftliche Theorien Erklärungen und Vorhersagen für Daten, d.h. für die beobachtbaren Ergebnisse experimenteller Messungen. Bogen und Woodward wenden gegen diese Auffassung ein, dass ein genauerer Blick auf die wissenschaftliche Praxis zeige, dass dies schlicht und einfach falsch sei. Theorien machen keine Vorhersagen und liefern keine Erklärungen für Daten, sondern auf Grundlage der Daten wird „bloß“ auf das Vorliegen bestimmter Phänomene geschlossen. Erst diese erschlossenen Phänomene sind es dann, die durch Theorien erklärt und vorhergesagt werden und die als Belege für Theorien herangezogen werden können.[8]

---

6 Allerdings folgt dies noch nicht aus der bisherigen Betrachtung. Van Fraassen muss noch ausschließen, dass es Auffassungen gibt, die epistemisch bescheidener sind als der Konstruktive Empirismus und dem Angemessenheitskriterium dennoch hinreichend gut gerecht werden. Dies tut er unter Rückgriff auf eine Überlegung von Gideon Rosen (1994), S. 162. Vgl. Monton und van Fraassen (2003), S. 407.

7 Vgl. Bogen und Woodward (1988; 1992; 2003), Bogen (2009a; 2009b) und Woodward (1989; 2000; 2009).

8 Es ist offensichtlich, dass Bogen und Woodward den Ausdruck „Phänomen“ bzw. „phenomenon“ anders verwenden als van Fraassen. Bei van Fraassen ist der Ausdruck für Beobachtbares reserviert, für Bogen und Woodward gilt dies klarerweise nicht. Ich verwende den Ausdruck in diesem Aufsatz so, wie er von Bogen und Woodward gebraucht wird. Zum Verhältnis zwischen van Fraassens Phänomenbegriff und dem Bogens und Woodwards vgl. auch Apel (2009).

Daten und Phänomene unterscheiden sich dabei in folgenden Hinsichten: Daten sind wahrnehmbare und öffentlich zugängliche Ergebnisse von Messungen. Sie sind an bestimmte Kontexte und insb. an das Vorhandensein bestimmter experimenteller Apparaturen gebunden. Typischerweise können sie nicht außerhalb dieser Kontexte auftreten. Darüber hinaus sind Daten das Ergebnis komplexer kausaler Wechselwirkung derart, dass es typischerweise nicht möglich ist, eine Theorie aufzustellen, die einzelne Datenpunkte erklärt oder vorhersagt. Was erklärt und vorhergesagt wird, sind vielmehr, wie bereits erwähnt, Phänomene. Diese zeichnen sich durch stabile und regelmäßige Eigenschaften aus. Sie sind im Gegensatz zu Daten nicht an spezifische experimentelle Kontexte gebunden, sondern können mittels verschiedener experimenteller Prozeduren nachgewiesen werden. Phänomene sind uns nicht direkt in der Wahrnehmung zugänglich, sondern sie müssen aus den Daten erschlossen werden. Bogen und Woodward gehen darüber hinaus davon aus, dass Phänomene in der Regel unbeobachtbar sind.[9]

Die Unterscheidung zwischen Daten und Phänomenen lässt sich leicht anhand eines einfachen Beispiels illustrieren: der Bestimmung des Schmelzpunktes von Blei.[10] Der Schmelzpunkt von Blei ist ein Phänomen im Sinne Bogens und Woodwards. Er wird bestimmt, indem man eine Vielzahl von Messungen durchführt, bei denen eine Bleiprobe zum Schmelzen gebracht wird und die entsprechende Temperatur mittels eines geeigneten Thermometers gemessen wird. Auch wenn die verwendete Messapparatur gut funktioniert und systematische Fehler ausgeschlossen werden können, erhält man durch dieses Vorgehen eine Vielzahl von Messpunkten (die Daten), die sich alle (leicht) voneinander unterscheiden. Unter bestimmten Bedingungen betrachten wir es als gerechtfertigt, den Mittelwert der Verteilung als gute Abschätzung für den „wahren Wert“ des Schmelzpunktes von Blei anzusehen. Wir schließen somit auf Grundlage der Daten unter Rückgriff auf mathematisch-statistische Verfahren und weitere Hintergrundannahmen auf das Phänomen, dass der Schmelzpunkt von Blei bei 327°C liegt. Das Phänomen wird somit nicht direkt beobachtet, sondern es muss auf Grundlage der vorliegenden Daten erschlossen werden, indem wir ein regelmäßiges „Muster“ in den uns vorliegenden Datensätzen identifizieren. Erst dieses „Muster“ ist es dann, das als Beleg für oder gegen die Geltung physikalischer Theorien herangezogen werden kann und das durch solche Theorien erklärt wird.[11]

---

9 Vgl. z.B. Bogen und Woodward (1988), S. 305, 314, 351.

10 Vgl. Bogen und Woodward (1988), S. 307-310.

11 Die Rede von Mustern orientiert sich an einer Formulierung von Woodward (1989), S. 396-397. Allerdings werden nicht alle Phänomene in dieser Weise durch Muster in Datensätzen repräsentiert. Vielmehr muss bei vielen Phänomenen darüber hinaus ein Kausalschluss auf die Ursache solcher Muster erfolgen. So wird beispielsweise das Phänomen der schwachen neutralen Ströme, ein weiteres Beispiel Bogens und Woodwards, als Ursa-

Ich möchte es an dieser Stelle bei dieser groben Skizze der Daten-Phänomen-Unterscheidung belassen. Zum einen, weil ich der Auffassung bin, dass Bogen und Woodward selbst dann, wenn man ihre weiteren Ausführungen in Betracht zieht, letztlich keine vollständig überzeugende Klärung des Daten- und des Phänomenbegriffs anbieten können.[12] Zum anderen, weil die bisherigen Erläuterungen ausreichen, um den hier im Zentrum stehenden Einwand gegen den Konstruktiven Empirismus angemessen bewerten zu können.

3.2 Der Einwand: Deskriptive Angemessenheit und die Unterscheidung zwischen Daten und Phänomenen

Bogens und Woodwards Kritik am Konstruktiven Empirismus beruht im Wesentlichen auf folgender Überlegung: Dass eine Theorie empirisch adäquat im Sinne van Fraassens ist, bedeutet nichts anderes, als dass diese Theorie wahre Aussagen über Daten macht. Schließlich werden nur Daten, nicht aber Phänomene beobachtet. Die Theorien, über die wir *de facto* verfügen, behandeln jedoch erschlossene Phänomene; sie liefern keine Erklärungen und Vorhersagen für einzelne Datenpunkte.[13] Deshalb, so der Vorwurf, erweise sich der Konstruktive Empirismus letztlich als inkonsistent.[14]

Doch wo genau liegt die vermeintliche Inkonsistenz? Um diese Frage zu beantworten, nehme man eine der zentralen Behauptungen des Konstruktiven Empirismus als Ausgangspunkt:

> KE: Das Ziel der Wissenschaft ist es, Theorien aufzustellen, die wahre Aussagen über Beobachtbares machen.

Zudem betrachte man folgendes Argument, das sich auf Grundlage der von Bogen und Woodward angegebenen Charakteristika von Daten und Phänomenen formulieren lässt:

P1:	Unsere gegenwärtigen wissenschaftlichen Theorien machen Aussagen über erschlossene Phänomene, nicht über beobachtete Daten.
P2:	Phänomene sind unbeobachtbar, Daten sind beobachtbar.
K1:	Unsere gegenwärtigen wissenschaftlichen Theorien machen keine wahren Aussagen über Beobachtbares.

che für bestimmte charakteristische Spuren in der Blasenkammer erschlossen. Vgl. Bogen und Woodward 1988, S. 327-331 und Woodward (1989), S. 404-410.

12 Einige Unklarheiten und Probleme des Ansatzes von Bogen und Woodward sowie Überlegungen dazu, worin die wesentlichen Merkmale beider Begriffe im Einzelnen bestehen könnten, stelle ich in Apel (2009) vor.

13 Vgl. Bogen und Woodward (1988), S. 351, Woodward (1989), S. 450-452 und Bogen (2009a).

14 Vgl. Bogen und Woodward, S. 337.

KE und K1sind offensichtlich miteinander verträglich, denn KE ist eine These über das Ziel der Wissenschaft und dass es das Ziel der Wissenschaft ist, wahre Aussagen über Beobachtbares zu machen, ist ohne weiteres damit vereinbar, dass unsere gegenwärtigen Theorien diesem Ziel nicht gerecht werden.
Zieht man jedoch zusätzlich die von mir im zweiten Abschnitt skizzierte Art und Weise in Betracht, auf die van Fraassen seine Position begründet, ergibt sich ein Problem für den Konstruktiven Empiristen. Der Konstruktive Empirist fordert deskriptive Angemessenheit gegenüber der wissenschaftlichen Praxis, weil er die moderne Naturwissenschaft als Paradigma rationaler Erkenntnisgewinnung bewundert. Aus diesem Grund sollte van Fraassen meines Erachtens die These vertreten, dass unsere gegenwärtigen naturwissenschaftlichen Theorien ihrem Ziel zumindest zu einem gewissen Grad gerecht werden, denn eine epistemische Praxis, die ihr Ziel völlig verfehlt, als Paradigma rationaler Erkenntnisgewinnung zu betrachten, erscheint absurd. Dann allerdings ist der Konstruktive Empirist auf die Behauptung festgelegt, dass unsere derzeitigen Theorien zumindest ein gewisses Maß an empirischer Adäquatheit erreichen, denn empirische Adäquatheit ist gemäß des Konstruktiven Empirismus das Ziel der Wissenschaft.[15] Diese Überlegung lässt sich folgendermaßen zusammenfassen:

P3:	KE
P4:	Wissenschaft ist das Paradigma rationaler Erkenntnisgewinnung.
P5:	Wenn Wissenschaft das Paradigma rationaler Erkenntnisgewinnung ist, dann wird sie zum gegenwärtigen Zeitpunkt ihrem Ziel (zumindest zu einem gewissen Grad) gerecht.
K2:	Gegenwärtige wissenschaftliche Theorien erlauben es uns, (zumindest zu einem gewissen Grad) wahre Aussagen über Beobachtbares zu machen.

K1 und K2 widersprechen sich. Dies könnte der Punkt sein, an dem die Inkonsistenz liegt, die Bogen und Woodward van Fraassen vorwerfen. Wenn K1 wahr wäre, würde der Konstruktive Empirismus das von ihm selbst ins Feld geführte Kriterium der deskriptiven Angemessenheit verfehlen und damit droht die Gefahr, dass sich letztlich doch der Wissenschaftliche Realismus als diejenige Position erweist, die den Gleichgewichtspunkt zwischen den Desideraten der epistemischen Bescheidenheit und der deskriptiven Angemessenheit findet.

15 Weder van Fraassen noch ich haben eine ausgearbeitete Theorie dazu anzubieten, wie man unterschiedliche Grade empirischer Adäquatheit feststellen, vergleichen und evtl. sogar quantifizieren kann. Aber die intuitive Idee hinter der Rede von Graden empirischer Adäquatheit besteht darin, dass eine Theorie empirisch adäquater als eine andere ist, wenn sich aus ihr mehr korrekte Vorhersagen ableiten lassen und diese Vorhersagen unterschiedliche Bereiche betreffen sowie ein höheres Maß an quantitativer Präzision erreichen.

4. Wie man die Hausaufgaben erledigen könnte: Zwei Verteidigungsmöglichkeiten für den Konstruktiven Empiristen

Im Folgenden möchte ich zwei unterschiedliche Strategien skizzieren, die es dem Konstruktiven Empiristen ermöglichen, Bogens und Woodwards Einwand zurückzuweisen. Ich halte beide Strategien für durchführbar, sie setzen jedoch jeweils eine andere Explikation des Begriffs der empirischen Adäquatheit voraus. Während es Bogen und Woodward also nicht gelingt, ein überzeugendes Argument gegen den Konstruktiven Empirismus zu entwickeln, zwingen ihre Überlegungen den Konstruktiven Empiristen immerhin dazu, seine Auffassung genauer zu spezifizieren.

Um Bogens und Woodwards Einwand zu parieren, muss der Konstruktive Empirist K1 oder K2 zurückweisen. Da ich K2 für plausibel halte, sollte der Konstruktive Empirist sich meines Erachtens auf K1 konzentrieren. Dementsprechend greift eine der Argumentationsstrategien, die ich hier vorstellen werde, P1 an und die andere P2. Beginnen werde ich mit letzterer.

4.1 Sind alle wissenschaftlichen Phänomene unbeobachtbar?

Ich beginne mit der Diskussion von P2, da vieles dafür spricht, dass van Fraassen diese Voraussetzung bestreiten würde. So beschreibt er in *Empiricism and the Philosophy of Science* die Aufgabe theoretischer Modelle, die unbeobachtbare Entitäten postulieren, in der folgenden Weise:

> The whole point of having theoretical models is that they should fit the phenomena, that is, fit models of the data.[16]

Äußerungen wie diese legen nahe, dass empirische Adäquatheit ein Begriff ist, der sich auf Datenmodelle bezieht. Ohne hier in die Details zu gehen, kann der Begriff des Datenmodells so verstanden werden, dass er das bezeichnet, was ich zu Beginn dieses Textes als „Muster in Datensätzen“ bezeichnet habe.[17] Man denke hier an das Phänomen des Schmelzpunktes von Blei, das erschlossen wird, indem ein regelmäßiges Muster in einem Datensatz identifiziert wird. Die-

16 Van Fraassen (1985), S. 271. Vgl. auch van Fraassen (2002), S. 163-164 und van Fraassen (2006), S. 31. Zu beachten ist im Hinblick auf die zitierte Stelle natürlich, dass van Fraassen den Ausdruck „phenomena“ anders als Bogen und Woodward verwendet.

17 Suppes (1962) führt den Terminus „Datenmodell“ in die wissenschaftstheoretische Debatte ein. Frigg und Hartmann (2006) geben eine instruktive Erläuterung dessen, was Datenmodelle sind: „A model of data is a corrected, rectified, regimented, and in many instances idealized version of the data we gain from immediate observation, the so-called raw data. Characteristically, one first eliminates errors (e.g. removes points from the record that are due to faulty observation) and then presents the data in a ‚neat' way, for instance by drawing a smooth curve through a set of points.“

ses und ähnliche Phänomene werde ich im Folgenden als *Musterphänomene* bezeichnen. Sie sind dadurch charakterisiert, dass in Aussagen über Musterphänomene beobachtbaren Entitäten (Blei) beobachtbare Eigenschaften (das Haben einer bestimmten Schmelztemperatur) zugeschrieben werden, die exakte Ausprägung dieser Eigenschaften aber erschlossen werden müssen.

Die gerade zitierte Textstelle kann man so lesen, dass van Fraassen die Position vertritt, dass auch Aussagen über Musterphänomene vom Konstruktiven Empiristen für wahr gehalten werden sollten. Sollte dies tatsächlich seine Auffassung sein, so könnte er dem Einwand Bogens und Woodwards begegnen. Um zu sehen, warum dies der Fall ist, muss man zunächst in Betracht ziehen, dass es van Fraassen als das Ziel der Wissenschaft ansieht, wahre Aussagen über das Beobacht*bare*, nicht bloß über das Beobacht*ete*, zu machen. Diese Erhöhung des epistemischen Risikos ist erforderlich, um dem Kriterium der deskriptiven Angemessenheit gerecht zu werden. Analog zu dieser Überlegung könnte der Konstruktive Empirist im Lichte des Einwands von Bogen und Woodward einräumen, dass es erforderlich ist, auch Schlussfolgerungen als wahr zu akzeptieren, bei denen man auf der Grundlage vorliegender Datensätze auf die Existenz von Musterphänomenen schließt. Nichtsdestotrotz kann der Konstruktive Empirist weiterhin daran festhalten, dass eine solche Position, die sich nur darauf verpflichtet, dass es das Ziel der Wissenschaft ist, wahre Aussagen über Musterphänomene (und natürlich über die verhältnismäßig wenigen direkt beobachtbaren Phänomene) zu machen, epistemisch bescheidener ist, als ein Wissenschaftlicher Realismus, der sich darüber hinaus darauf festlegt, dass auch Aussagen über unbeobachtbare Gegenstände wie Elektronen und Atome wahr sein müssen. Insbesondere gehen mit der Etablierung von Musterphänomenen keine ontologischen Verpflichtungen auf neue Entitäten einher; ein Schritt den van Fraassen für besonders problematisch hält.[18]

Es stellt sich allerdings die Frage, ob solch ein weiter Begriff der empirischen Adäquatheit, der auch Musterphänomene umfasst, für van Fraassen akzeptabel wäre. Da der Empirist van Fraassen an verschiedenen Stellen darauf hinweist, dass man seine Überzeugungen auf solche Sachverhalte beschränken sollte, die zumindest potenziell Gegenstand unserer Erfahrung sein können,[19] wäre dies nur dann der Fall, wenn Musterphänomene ebensolche potenziellen Gegenstände der Erfahrung sind, d.h., wenn sie beobachtbar sind. Bogen und Woodward bestreiten dies, allerdings ohne eine ausdrückliche Begründung für diese These zu liefern. Sie führen zwar aus, dass Musterphänomene nicht beobachtet, sondern

18 Vgl. Ladyman et al. (1997), S. 316.

19 Zum Beispiel: „[…] how could anyone who does not say *credo ut intelligam* be baffled by a desire to limit belief to what can at least be in principle be disclosed in experience? Or, more to the point, by the idea that acceptance in science does not require belief in truth beyond those limits?“ Van Fraassen (1985), S. 258, Hervorhebung im Original.

erschlossen werden, aber daraus folgt selbstverständlich nicht, dass sie unbeobachtbar sind. Hierfür müssten weitere Gründe angeführt werden.

Betrachten wir zur Klärung der Frage, ob Musterphänomene beobachtbar sind, folgende Aussage: „Der Schmelzpunkt von Blei liegt bei 327°C“. Diese Aussage ist semantisch äquivalent zu der Aussage „Wenn die Bedingungen $B_1 \ldots B_n$ erfüllt sind, dann schmilzt eine Bleiprobe bei 327°C“. Deshalb handelt es sich beim Phänomen des Schmelzpunktes von Blei um den Sachverhalt, dass Blei die dispositionale Eigenschaft hat unter geeigneten experimentellen Bedingungen, d.h. beim Vorliegen einer reinen Bleiprobe und unter Abschirmung aller externen Störfaktoren, bei 327°C zu schmelzen. Erinnern wir uns nun daran, wie van Fraassen den Begriff der Beobachtbarkeit erläutert: X ist genau dann beobachtbar, wenn es *mögliche* Beobachtungsbedingungen für X gibt, d.h. Bedingungen, unter denen wir X beobachten würden, wenn sie realisiert würden.[20] Die entscheidende Frage ist, welche Art von Möglichkeit bei dieser Erläuterung im Spiel ist und ob die gerade erwähnten geeigneten experimentellen Bedingungen gemäß dieser Auffassung als mögliche Beobachtungsbedingungen klassifiziert werden können.

Da van Fraassen bei der Bestimmung des Beobachtbarkeitsbegriffs explizit auf unsere sinnesphysiologische Ausstattung und die Gesetze der finalen Physik und Biologie verweist, scheint der plausibelste Kandidat für die relevante Modalität naturgesetzliche Möglichkeit unter Konstanthaltung unserer sinnesphysiologischen Fähigkeiten zu sein.[21] Für viele der sog. Musterphänomene gilt dann Folgendes: Es handelt sich bei ihnen um Sachverhalte über Entitäten und ihre dispositionalen Eigenschaften. Die Manifestationsbedingungen dieser Eigenschaften sind die oben angesprochenen geeigneten experimentellen Bedingungen. Wenn solche Bedingungen realisiert würden, wenn wir also beispielsweise reine Bleiproben herstellen und sämtliche Störfaktoren abschirmen könnten, dann würden wir die entsprechenden Phänomene beobachten.[22] Zwar sind solche Bedingungen in unserer Welt häufig nicht realisierbar, aber dennoch kann man sagen, dass man in einer Welt, die nichts außer der entsprechenden experimentellen Apparatur enthalten würde, d.h. in einer Welt mit reinen Bleiproben, aber ohne Störfaktoren, Vorkommnisse des entsprechenden Phänomens beobachten könnte.

Bei genauerer Betrachtung erscheint es jedoch fraglich, ob ein solches Beobachtbarkeitsverständnis tatsächlich mit van Fraassens Empirismus vereinbar ist. Wir wissen, wie schon erwähnt, dass die geeigneten experimentellen Bedin-

20 Vgl. van Fraassen (1980), S. 16.

21 Vgl. auch van Fraassen (1980), S. 17.

22 Dies ist im Falle, dass es sich bei dem jeweiligen Phänomen um einen allgemeinen Sachverhalt handelt, natürlich so zu verstehen, dass wir nicht Generalisierungen selbst, sondern nur Instanzen dieser beobachten würden. Wir können nicht beobachten, dass alle Raben schwarz sind, sondern nur einzelne Vorkommnisse von schwarzen Raben.

gungen in unserer Welt in der Regel nicht realisiert werden können. Zwar darf man die Frage, ob es naturgesetzlich möglich ist, ideale experimentelle Bedingungen zu realisieren, nicht mit der Frage verwechseln, ob es unter der Voraussetzung idealer experimenteller Bedingungen naturgesetzlich möglich ist, ein bestimmtes Phänomen zu beobachten,[23] aber dennoch ist nicht klar, auf welcher Grundlage *der Empirist* das Verhalten bestimmter physikalischer Systeme unter idealen Bedingungen als potenziellen Gegenstand der Erfahrung klassifizieren kann. Man könnte van Fraassen deshalb vorwerfen, dass er einen für ihn als Empiristen problematischen Möglichkeitsbegriff in Anschlag bringen muss, um seinen Beobachtbarkeitsbegriff so auszubuchstabieren, wie es im Lichte der Daten-Phänomen-Unterscheidung erforderlich ist.

Jedoch sehe ich eine Möglichkeit, wie van Fraassen ebendiesem Einwand begegnen könnte. Diese mögliche Argumentationslinie beruht auf einer von Andreas Hüttemann eingeführten Unterscheidung zwischen zwei Typen von Dispositionen: sog. kontinuierlich manifestierbaren und diskontinuierlich manifestierbaren Dispositionen.[24] Erstere sind Hüttemann zufolge auch im Rahmen einer empiristischen Epistemologie unproblematisch. Ich werde im weiteren Verlauf versuchen, Gründe dafür anzuführen, dass Aussagen über Musterphänomene als Zuschreibungen kontinuierlich manifestierbarer Dispositionen zu physikalischen Systemen aufzufassen sind und als solche, Hüttemann folgend, auch für den Empiristen epistemologisch unproblematisch sind. Ich werde damit letztlich die These verteidigen, dass van Fraassen gerade wegen der epistemologischen Harmlosigkeit dieses Dispositionstyps guten Grund dazu hat, Musterphänomene als beobachtbar anzusehen.

Der Unterschied zwischen beiden Dispositionstypen lässt sich anhand der Dispositionen der Zerbrechlichkeit und der Löslichkeit erläutern. Die Zerbrechlichkeit eines Glases ist eine diskontinuierlich manifestierbare Disposition. Diese Disposition wird manifest, wenn eine entsprechende Manifestationsbedingung realisiert wird, also z.B., wenn das Glas aus hinreichender Höhe auf einen Marmorfußboden fällt. Solange das Glas noch nicht auf dem Boden aufgeschlagen ist, ist es nicht zerbrochen, in dem Moment, wo es aufschlägt, ändert sich schlagartig (d.h. diskontinuierlich) sein Zustand. Bei der Löslichkeit von einer Portion Salz in Wasser verhält sich dies anders. Die Manifestationsbedingung dieser Disposition liegt vor, wenn genügend Wasser auf eine gegebene Menge Salz gegossen wurde. Dann löst sich das Salz vollständig im Wasser auf. Das Verhalten des kombinierten Systems aus Salz und Wasser ist dabei jedoch eine kontinuierliche Funktion des Grades, zu dem die Dispositionsbedingung realisiert wurde, d.h., es findet hier eine kontinuierliche Zustandsveränderung des Salzes statt. Je mehr Wasser auf das Salz gegossen wird, desto mehr Salz löst

23 Diesem nahe liegenden Missverständnis sitzt beispielsweise Philip Kitcher auf. Vgl. Kitcher (1993), S. 152.

24 Vgl. Hüttemann (1997), S. 145-151 und Hüttemann (1998), S. 130-133.

sich auf. Dispositionen dieser Art bezeichnet Hüttemann als kontinuierlich manifestierbare Dispositionen.

Der für unsere Diskussion entscheidende Punkt ist nun, dass wir, Hüttemann zufolge, dafür, dass ein bestimmtes Objekt oder ein bestimmter Objekttyp eine kontinuierlich manifestierbare Disposition besitzt, auch dann Belege haben können, wenn die Manifestationsbedingungen niemals vollständig realisiert wurden. Auch wenn unsere Portion Salz niemals mit so viel Wasser in Berührung kommt, dass sie sich vollständig in Wasser auflöst, haben wir allen Grund zu der Annahme, dass sie dies tun würde, wenn sie mit einer entsprechenden Wassermenge in Berührung käme, da sich das Salz bereits dann teilweise aufgelöst haben wird, wenn die Manifestationsbedingungen annähernd realisiert wurden.[25] Hüttemann zieht hieraus folgende Schlussfolgerung:

> The lesson is that CMDs [continuously manifestable Dispositions; J.A.] are epistemologically as innocuous as any ordinary property. Empiricists therefore have no reason to recoil from employing the concept of a CMD.[26]

Werden also Beschreibungen von Musterphänomenen als Aussagen aufgefasst, in denen beobachtbaren Objekten kontinuierlich manifestierbare Dispositionen zugeschrieben werden, dann sind sie auch für den Konstruktiven Empiristen epistemologisch unproblematisch.

Aufgrund der Tatsache, dass wir die in Aussagen über Musterphänomene zugeschriebenen dispositionalen Eigenschaften bei der sukzessiven Annäherung an die Manifestationsbedingungen auch graduell immer mehr realisieren (dies tun wir, wenn wir kontrollierte Experimentalsituationen herstellen und sukzessive optimieren), ist es äußerst plausibel, Aussagen über Musterphänomene in ebendieser Weise aufzufassen. Deshalb gibt es keinen Grund für den Empiristen das kontrafaktische Konditional „Wenn geeignete experimentelle Bedingungen realisiert würden, dann würden wir X beobachten“ nicht für wahr zu halten.

4.2 Machen Theorien tatsächlich keine Aussagen über Daten?

Für jemanden, der nicht bereit ist, der Argumentation im oberen Abschnitt zu folgen, werde ich an dieser Stelle einen weiteren Argumentationsgang vorstellen, der es dem Konstruktiven Empiristen ebenfalls ermöglicht, seine Position im Lichte der Daten-Phänomen-Unterscheidung aufrecht zu erhalten. Dabei setze ich Bogens und Woodwards P2 als korrekt voraus und diskutiere die Plausibilität von P1. P1 besagt, dass gegenwärtige wissenschaftliche Theorien Aussagen über erschlossene Phänomene, nicht über beobachtete Daten machen.

---

25 Vgl. Hüttemann (1998), S. 131. Bei diskontinuierlich manifestierbaren Dispositionen ist dies klarerweise anders: Der Zustand des fallenden Glases kurz vor dem Aufprall gibt uns keinerlei Information darüber, wie sich sein Zustand nach dem Aufprall ändern wird.

26 Hüttemann (1998), S. 132.

Wieder bildet das Schmelzpunktbeispiel den Ausgangspunkt meiner Überlegungen. Bogen und Woodward haben darauf hingewiesen, dass wir aus Theorien über die atomaren Bindungen in Festkörpern die Aussage ableiten können, dass Blei unter idealen experimentellen Bedingungen bei 327°C schmilzt. Dieses Phänomen wird von der atomaren Festkörpertheorie behandelt. Sie erklärt uns, auf welche Weise die einzelnen Bleiatome in einer idealen Bleiprobe miteinander gebunden sind und welche Energie erforderlich ist, um diese Bindungen aufzubrechen. Nehmen wir an, van Fraassen würde P2 akzeptieren und somit u.a. davon ausgehen, dass die atomare Festkörpertheorie mit dem Schmelzpunkt von Blei ein unbeobachtbares Phänomen erklärt. In diesem Fall müsste van Fraassen in den sauren Apfel beißen und auch bezüglich des Wahrheitswertes wissenschaftlicher Aussagen über dieses Phänomen agnostisch bleiben. Aber wäre dies für den Konstruktiven Empirismus tatsächlich problematisch? Dazu müsste es der Fall sein, dass die Beschäftigung mit solchen Phänomenen, also zum Beispiel das Bilden von Modellen, die ideale Fälle beschreiben, nichts zum Erreichen des Ziels der empirischen Adäquatheit beiträgt (oder diesem sogar abträglich ist). Dass dies offenkundig nicht der Fall ist, lässt sich aber anhand einer einfachen Überlegung plausibel machen:

Die Annahme, dass Bei unter idealen experimentellen Bedingungen bei 327°C schmilzt, ist überaus nützlich, um wahre Aussagen über Beobachtbares zu machen. Diese Annahme ermöglicht es uns, beispielsweise korrekte Vorhersagen über Messdaten im Rahmen bestimmter Konfidenzintervalle zu machen. Wenn wir den Schmelzpunkt durch die Erhebung geeigneter Daten und anschließender Mittelwertbildung etabliert haben, dann können wir diesen Wert dazu benutzen, vorherzusagen, dass der nächste Messwert, der an der entsprechenden Apparatur gemessen wird, mit hoher Wahrscheinlichkeit in einem bestimmten Intervall um 327°C liegen wird. D.h., die Phänomenbehauptung über das erschlossene Phänomen des Schmelzpunktes von Blei kann mit Hilfe weiterer Annahmen durchaus in Beziehung zu Sachverhalten gesetzt werden, die wir beobachten. Diese weiteren Annahmen sind in etwa folgender Art: „Wenn die experimentellen Bedingungen nicht weit von idealen experimentellen Bedingungen abweichen und unser Messinstrument im Rahmen einer gewissen Messgenauigkeit zuverlässig ist, dann weicht auch der Messwert nicht weit vom wahren Schmelzpunkt ab.“ Zwar erhalten wir auf diese Weise keine exakten Vorhersagen, sondern nur solche mit einer gewissen Unschärfe und zudem kann im Einzelfall auch eine größere Abweichung als erwartet auftreten, aber wir werden dennoch bessere Vorhersagen machen, als wenn wir uns nicht an der erschlossenen Aussage über den Schmelzpunkt unter idealen experimentellen Bedingungen orientieren. Wenn wir nichts über diesen Schmelzpunkt wüssten, könnten wir nicht sagen, in welchem Temperaturbereich die nächste Bleiprobe schmelzen wird. Die Phänomenbehauptung über den unbeobachtbaren Schmelzpunkt

von Blei ist demzufolge Teil einer Theorie, die einen gewissen Grad von empirischer Adäquatheit erreicht.

Gleiches gilt auch für weiterführende Theorien, die den Schmelzpunkt in Beziehung zur atomaren Struktur des Bleis setzen. Mit Hilfe dieser können wir z.B. Aussagen darüber machen, wie sich der Schmelzpunkt bei einer Erhöhung des äußeren Drucks oder bei einer Verunreinigung der Probe ändern wird, indem wir überlegen, wie diese Faktoren die atomaren Bindungen, die unser theoretisches Modell beschreibt, beeinflussen. Im Anschluss können wir ein Experiment konstruieren, in dem sich diese Überlegungen testen lassen, und das Modell auf diese Weise überprüfen. Theorien über unbeobachtbare Mechanismen auf der atomaren Ebene sind besonders dann nützlich, wenn es darum geht, Aussagen über das Verhalten physikalischer Systeme in Situationen zu machen, die sich von bisher realisierten Situationen stark unterscheiden, d.h. in Situationen, in denen wir nicht bloß induktiv darauf schließen, dass ein System in einer Situation, die den uns bereits bekannten Situationen ähnelt, wieder ein ähnliches Verhalten zeigen wird. Durch solche Mikrotheorien sind wir in der Lage erfolgreich neuartige Vorhersagen (die berühmt-berüchtigten „novel predictions“) zu machen und so den Grad der empirischen Adäquatheit unserer Theorien wesentlich zu steigern.

Bogens und Woodwards Behauptung, dass wir die beobachteten Datenpunkte weder exakt vorhersagen noch erklären, mag in einem gewissen Sinne richtig sein: In der Regel bilden Wissenschaftler keine theoretischen Modelle, die einzelne Datenpunkte vorhersagen und erklären sollen. In diesem Sinne gilt ihr Interesse eher den Musterphänomenen, d.h. idealen, prototypischen Fällen. Dies bedeutet jedoch keineswegs, dass die theoretische Beschäftigung mit solchen Fällen völlig von der empirischen Adäquatheit und somit von der von uns erfahrbaren Wirklichkeit abgekoppelt wäre. Im Gegenteil: Eine solche Behauptung wäre absurd. Dann bliebe beispielsweise die erfolgreiche Anwendung physikalischer Theorien in Ingenieurstätigkeiten vollkommen unverständlich. Ingenieure greifen auf naturwissenschaftliche Theorien zurück, um Instrumente zu entwickeln, die verlässlich beobachtbare Wirkungen erzeugen. Infrarotfernbedienungen sollen zuverlässig vom ersten auf das zweite Programm umschalten, Laser in Supermarktkassen sollen die richtigen Preise auslesen etc. Dies sind beobachtbare Sachverhalte und die entsprechenden Geräte wurden unter Rückgriff auf wissenschaftliche Theorien konstruiert. Die Quantenmechanik ist in diesem Sinne eine empirisch adäquatere Theorie als die klassische Elektrodynamik, da sie es uns ermöglicht, wenn auch in äußerst komplexen Zwischenschritten, die Aussage abzuleiten, dass ein Laserscanner in einer Supermarktkasse, wenn ein bestimmter Barcode eingelesen wird, einen bestimmten Preis anzeigen wird. Ganz offensichtlich sind wir demnach dazu in der Lage, unter Anwendung der entsprechenden Theorien wahre Aussagen über Beobachtbares zu machen.

P1 in der Rekonstruktion von Bogens und Woodwards Argument ist somit zurückzuweisen. Selbst wenn wir zugestehen, dass die Phänomene, die unsere Theorien erklären und die wir letztlich als Belege heranziehen, erschlossen und unbeobachtbar sind, so lassen sich dennoch mit Hilfe dieser Theorien wahre Aussagen über Beobachtbares machen. Dies geschieht zwar nicht mit beliebiger Präzision, aber immerhin im Rahmen bestimmter Konfidenzintervalle. Die Präzision solcher Aussagen kann im Verlauf der Theoriebildung erhöht werden und zudem werden neuartige Vorhersagen ermöglicht. Folglich kann der Konstruktive Empirist K1 und damit den Einwand Bogens und Woodwards gegen den Konstruktiven Empirismus zurückweisen.

5. Ergebnisse der Untersuchung

Zusammenfassend kann man festhalten, dass es Bogen und Woodward nicht gelungen ist, ein überzeugendes Argument gegen den Konstruktiven Empirismus zu formulieren. Ich habe im Vorhergehenden zwei Möglichkeiten zur Verteidigung der van Fraassen'schen Position ausgeführt:

Der Konstruktive Empirist kann zum einen dafür argumentieren, dass Bogens und Woodwards Einwand fehlgeht, weil der Begriff der empirischen Adäquatheit durchaus so expliziert werden kann, dass er eine Teilklasse der Phänomene, die Musterphänomene, umfasst. Entscheidet er sich für diesen Weg, muss der Konstruktive Empirist allerdings plausibel machen, dass Musterphänomene beobachtbar (im Sinne seines Beobachtbarkeitsbegriffs) sind. Dies kann geschehen, indem man Aussagen über Musterphänomene als Zuschreibungen kontinuierlich manifestierbarer Dispositionen auffasst und dafür argumentiert, dass diese auch für einen Empiristen epistemologisch unproblematisch sind.

Sollte er diesen Weg nicht gehen wollen, so steht dem Konstruktiven Empiristen zum anderen die Option offen, dafür zu argumentieren, dass das Bilden wissenschaftlicher Theorien und Modelle, die Musterphänomene behandeln, dem Ziel der empirischen Adäquatheit auch dann zuträglich ist, wenn man hinsichtlich des Wahrheitswertes der entsprechenden Aussagen agnostisch bleibt. Dies ist der Fall, da diese Theorien und Modelle die Vorhersage von beobachtbaren Daten zumindest mit einer gewissen quantitativen Genauigkeit erlauben. Damit ist Bogens und Woodwards Behauptung widerlegt, dass wissenschaftliche Theorien ausschließlich Aussagen über Phänomene, nicht aber über Daten machen.

Letztlich erweist sich Bogens und Woodwards Einwand deshalb als nicht mehr als eine bisher unerledigte Hausaufgabe, die sich mit einigem Fleiß lösen lässt. Gleichzeitig ist er jedoch auch ein Musterbeispiel dafür, dass die Bearbeitung von Hausaufgaben mehr ist, als eine lästige Pflicht. Gute Hausaufgaben sind derart, dass man bei ihrer Bearbeitung etwas dazulernt. Dies gilt auch für

die Aufgabe, die Bogen und Woodward dem Konstruktiven Empiristen vorgesetzt haben, denn bei der Auseinandersetzung mit ihrem Einwand kann er lernen, dass der Begriff der empirischen Adäquatheit der Präzisierung bedarf: Den beiden in diesem Aufsatz ausgeführten Argumentationsgängen entsprechend, kann der Begriff entweder so weit gefasst werden, dass er auch Musterphänomene umfasst, oder so eng, dass er dies nicht tut. Der Konstruktive Empirist wird durch Bogens und Woodwards Überlegungen deshalb auf die Notwendigkeit hingewiesen, genauer zu klären, wie strikt oder liberal die Erkenntnisgrenzen einer empiristischen Wissenschaftsphilosophie gezogen werden sollen.

Literaturverzeichnis

Alspector-Kelly, M.: "Should the Empiricist be a Constructive Empiricist". Philosophy of Science, 68, 2001. S.413-431

Apel, J.: "On the meaning and the epistemological relevance of the notion of a scientific phenomenon". Synthese (online first version), 2009. DOI 10.1007/s11229-009-9620-y

Berg-Hildebrand, A. und Ch. Suhm: "The Hardships of an Empiricist". In: *Berg-Hildebrand, A. und Ch. Suhm (Hrsg.)*: Bas C. van Fraassen - The Fortunes of Empiricism. Ontos, Frankfurt/Main, 2006. S. 57-67

Bogen, J.: "'Saving the phenomena' and saving the phenomena". Synthese (online first version), 2009a. DOI 10.1007/s11229-009-9619-4

Bogen, J.: "Observation in Science". In: The Stanford Encyclopedia of Philosophy, Spring 2009, 2009b. http://plato.stanford.edu/entries/science-theory-observation/ (24.02.2009)

Bogen, J. und J. Woodward: "Saving the Phenomena". The Philosophical Review, 97, 1988. S. 303-352

Bogen, J. und J. Woodward: "Observations, Theories and the Evolution of the Human Spirit". Philosophy of Science, 59, 1992. S. 590-611

Bogen, J. und J. Woodward : "Evading the IRS". Poznan Studies in the Philosophy of Science and the Humanities, 20, 2003. S. 223-245

*Frigg, R., and S. Hartmann: "*Models in Science". In: The Stanford Encyclopedia of Philosophy, Spring 2006. http://plato.stanford.edu/archives/spr2006/entries/models-science/ (12.10.2007)

Hüttemann, A.: Idealisierungen und das Ziel der Physik. Eine Untersuchung zum Realismus, Empirismus und Konstruktivismus in der Wissenschaftstheorie. De Gruyter, Berlin, 1997

Hüttemann, A.: "Laws and Dispositions". Philosophy of Science, 65, 1998. S. 121-135

Kitcher, P.: The Advancement of Science. Oxford University Press, Oxford, 1993

Ladyman, J., I. Douven, L. Horsten, and B. C. van Fraassen: "A Defence of van Fraassen's Critique of Abductive Inference: Reply to Psillos". The Philosophical Quarterly, 47, 1997. S. 305-321

Monton, B., and B. C. van Fraassen: "Constructive Empiricism and Modal Nominalism". British Journal for Philosophy of Science, 54, 2003. S. 405-422

Rosen, G.: "What is Constructive Empiricism?". Philosophical Studies, 74, 1994. S. 143-178

Suppes, P.: "Models of Data". In: *Ernest Nagel, Patrick Suppes and Alfred Tarski (Hrsg.)*: Logic, Methodology, and Philosophy of Science - Proceedings of the 1960 International Congress. Stanford University Press, Stanford, 1962. S. 252-261

van Fraassen, B.: The Scientific Image. Clarendon Press, Oxford, 1980

van Fraassen, B.: "Empiricism in the Philosophy of Science". In: *P.M. Churchland and C.A. Hooker (Hrsg.):* Images of Science. University of Chicago Press, Chicago, 1985. S. 245-368

van Fraassen, B.: The Empirical Stance. Yale University Press, Yale, 2002

Woodward, J.: "Data and Phenomena". Synthese, 79, 1989. S.393-472

Woodward, J.: "Data, Phenomena, and Reliability". Philosophy of Science, 67, 2000. S. 63-179

Woodward, J.:"Data and phenomena: a restatement and defense". Synthese (online first versi*on),* 2009. DOI 10.2007/s11229-009-9618-5

Wissenschaftlicher Empirismus, Robustheit und der Schluss auf die beste Erklärung

Tobias Breidenmoser
breidenmoser@googlemail.com
Zentrum für Logik, Wissenschaftstheorie und Wissenschaftsgeschichte,
Universität Rostock

Abstract/Zusammenfassung

In this paper I will sketch a kind of Selective Realism which is similar to Constructive Empiricism of Bas van Fraassen yet emphasize the importance of the concept of robustness. I will analyse some objections against Constructive Empiricism to develop my own version of scientific empiricism which can avoid these objections by using the concept of robustness. Hereafter I will show that even scientific empiricists have not to be worry about every inference to the best explanation.

In diesem Essay möchte ich eine Version des selektiven Wissenschaftlichen Realismus skizzieren, die an den Konstruktiven Empirismus von Bas van Fraassen angelehnt ist und den Begriff der Robustheit als zentrales Konzept nutzt. Zunächst werde ich Einwände gegen den Konstruktiven Empirismus analysieren, um anschließend mit dem Konzept der Robustheit eine eigene Version des Wissenschaftlichen Empirismus zu formulieren, die diesen Einwänden nicht ausgesetzt ist. Abschließend werde ich zeigen, dass für wissenschaftliche Empiristen nicht alle Schlüsse auf die beste Erklärung ungerechtfertigt sein müssen.

1. Wissenschaftlicher Empirismus

Wissenschaftliche Realisten, die an die absolute Wahrheit von wissenschaftlichen Theorien glauben, können sich nur schwer gegen den Einwand der pessimistischen Metainduktion behaupten. Moderne Realisten beschränken ihren Realismus daher auf ausgereifte und erfolgreiche Theorien. Erfolgreich ist eine Theorie dann, wenn sie neuartige Voraussagen macht und diese empirisch bestätigt wurden. Außerdem glauben Realisten nur an die angenäherte Wahrheit der besten Theorien, d.h. nur an die Wahrheit ihrer wichtigen Teile. Dies ist aus Sicht von P. Kyle Stanford (2006) jedoch nur ein Phyrrus-Sieg, denn es ist nur in der Retrospektive erkennbar, welche Teile einer Theorie wichtig und beständig sind. So hielt James Clerk Maxwell beispielsweise den mechanischen Äther für einen essentiellen Bestandteil seiner Theorie. Auch der moderne Wissenschaftliche Realist ist daher vor das Problem gestellt, dass er die wahren und beständigen Teile einer gegenwärtigen Theorie nicht eindeutig auszeichnen kann.

Eine solche Möglichkeit bietet der Konstruktive Empirismus von Bas van Fraassen (1980). Während Realisten glauben, dass unsere besten wissenschaftlichen Theorien zumindest angenähert wahr sind und sich ihre theoretischen Terme auf reale Entitäten beziehen, ist seine Meinung über Glaubwürdigkeit und Ziel der Wissenschaften bescheidener: Van Fraassen zufolge müssen wir nur müssen nur an die empirische Adäquatheit wissenschaftlicher Theorien glauben, d.h. nur an die Wahrheit der Aussagen einer Theorie über Beobachtbares. Aussagen über Unbeobachtbares gegenüber können wir uns dagegen agnostisch verhalten.

Die Existenz einer klaren Unterscheidung von beobachtbaren und unbeobachtbaren Entitäten ist eine Grundvoraussetzung des Konstruktiven Empirismus, allerdings keine unkontroverse. Grover Maxwell (1971) nahm einen kontinuierlichen Übergang von leicht beobachtbaren zu schwer beobachtbaren Entitäten an, zwischen denen eine Grenze nicht scharf und nur willkürlich gezogen werden könnte. Außerdem könnten Drogen entwickelt werden, die unsere Sinnesorgane erweitern oder ein menschlicher Mutant mit erweiterten Sinnesorganen geboren werden könnte.

Darüber hinaus kann Paul Churchland (1985) zufolge die Grenze unserer Beobachtung durch technische Hilfsmittel immer weiter verschoben werden. Beobachtungen können mit einem Elektronenmikroskop genauso gemacht werden wie mit bloßem Auge. Dies macht er mit einem Gedankenexperiment deutlich, bei dem ein Humanoid mit einem Elektronenmikroskop über dem linken Auge geboren wird. Da das Elektronenmikroskop ein Teil seines Organismus ist, könnte dieser Humanoid Viren und DNA-Stränge sprichwörtlich sehen. Es wäre unplausibel zu behaupten, dass der Humanoid diese Dinge sehen könnte und Menschen nicht, da der einzige Unterschied zwischen den Beobachtungen des Humanoiden und der Menschen im kausalen Ursprung des Instruments liegt.

Es gibt somit keine klare Grenze zwischen Beobachtung und Theorie. Allerdings kann auch eine vage Grenze eine Grenze sein, sofern man klare Fälle von beobachtbaren und unbeobachtbaren Entitäten auf beiden Seiten angeben kann. Van Fraassen illustriert dies mit den Beispielen der Jupitermonde als klar beobachtbare und Elektronen als klar unbeobachtbare Entitäten. Diese Beispiele sind jedoch weder unkontrovers noch intuitiv einleuchtend. Unkontroverse beobachtbare Entitäten wären vor allem mittelgroße physikalische Objekte um uns herum, wie Tische, Stühle und Kaffeemaschinen. Ein Beispiel für eine *für uns* nicht beobachtbare Entität ist Immanuel Kant, da es keine Umstände gibt, unter denen wir Kant jemals beobachten könnten. Egal wie viele Mikroskope oder Drogen wir entwickeln, wir werden Kant nie direkt beobachten können, sondern versuchen aus den vorhandenen Phänomenen, nämlich der Existenz von historischen Dokumenten zu schließen, dass es Kant tatsächlich gab. Ebenso wenig können wir Dinosaurier beobachten, da diese längst ausgestorben sind. Um eine Millionen von Lichtjahren entfernte Supernovae zu beobachten, sind wir dagegen zu früh geboren. Selbst wenn man unter Unbeobachtbarkeit also prinzipielle

Unbeobachtbarkeit versteht, lassen sich klare Beispiele für beobachtbare und unbeobachtbare Entitäten finden.

Der Konstruktive Empirismus ist somit nicht deshalb problematisch, weil er zwischen beobachtbaren und unbeobachtbaren Entitäten unterscheidet. Es könnte allerdings der Fall sein, dass Van Fraassen diese Unterscheidung für viel zu bedeutsam hält, da Überzeugungen über Beobachtbares für ihn stets besser gerechtfertigt sind als Überzeugungen über Unbeobachtbares. Doch Churchland (1985, 40) zufolge können wir nicht einfach zwischen beobachtbaren und unbeobachtbaren Dingen unterscheiden, sondern es gibt eine Dreiteilung aus (1) beobachtbaren und beobachtete Dingen, (2) beobachtbaren, aber nicht beobachteten Dingen und (3) unbeobachtbaren Dingen. Van Fraassen glaubt nicht an Aussagen über (3), doch zwischen (2) und (3) besteht nur ein marginaler epistemischer Unterschied. Elliott Sober (1993) betont, dass weder die Jupitermonde noch das AIDS-Virus beobachtet, sondern beide lediglich mittels Instrumenten aufgespürt wurden. Für unser gegenwärtiges Wissen ist es epistemisch irrelevant, dass die Jupitermonde prinzipiell auch direkt beobachtet werden können, das AIDS-Virus jedoch nicht. Aussagen über (2) können in einigen Fällen sogar besser gerechtfertigt sein als Aussagen über (3). Wenn es einen Yeti gibt, ist er beobachtbar, doch Alan Musgrave (1985) betont zu Recht, dass wir bessere Gründe haben, an die Existenz von unbeobachtbaren Elektronen zu glauben als an die Existenz des Yetis. Wenn es jedoch keinen signifikanten Unterschied zwischen Aussagen über (2) und (3) gibt und Van Fraassen sich gegenüber Aussagen über (3) agnostisch verhält, muss er das auch gegenüber Aussagen über (2) tun. Dadurch gerät er jedoch an den Rand des Skeptizismus.

Nancy Cartwright (2007) hat in jüngster Zeit versucht, diesen Einwand zu entkräften. Sie stellte fest, dass man auch als Antirealist auf die guten Voraussagen von Theorien vertrauen kann, wenn deren gute Voraussagefähigkeit bereits durch unsere Beobachtungen bestätigt wurde. Mithilfe dieser Theorien kann man auch als Konstruktiver Empirist Aussagen über Beobachtbares glauben, das noch nicht beobachtet wurde. Doch diese Verteidigung ist nur begrenzt gültig. Um zu wissen, dass unserer Theorien verlässliche Voraussagen machen, muss sich ihre gute Voraussagekraft an gleichartigen Entitäten bewährt haben. Sie können nicht bestätigen, dass es zwischen Karbon und Jura einen Urkontinent namens Pangaea gab, Menschen und Schimpansen einen gemeinsamen Vorfahren haben, die französische Revolution 1789 begann und Kant fast sein ganzes Leben in Königsberg verbracht hat. All diese Behauptungen lassen sich durch Beobachtungen nicht bestätigen. Van Fraassen muss sie daher konsequenterweise aus der rationalen Ontologie der Wissenschaft verbannen.

Doch dieser Schritt ist sehr unplausibel. Fast jeder Wissenschaftler würde alle diese Aussagen für gerechtfertigt halten und es wäre sehr merkwürdig, wenn kaum eine historische Aussage gerechtfertigt werden kann, weil historische Entitäten für uns nicht mehr beobachtbar sind. Der Konstruktive Empirismus

scheint somit defekt zu sein, da empirische Adäquatheit nicht korrekt zwischen gerechtfertigten und ungerechtfertigten wissenschaftlichen Behauptungen unterscheiden kann.

2. Robustheit

Um die Vorteile des Konstruktiven Empirismus zu retten, möchte ich daher nicht zwischen beobachtbaren und unbeobachtbaren Entitäten unterscheiden, sondern zwischen robusten und fragilen wissenschaftlichen Behauptungen. Robuste Behauptungen sind außerordentlich stabil, da sie weder von neuen empirischen Tatsachen noch von neuen Theorien noch von plötzlich entdeckten Fehlern leicht widerlegt werden können. Aus diesem Grund bilden einen Indikator für Wahrheit, selbst wenn sie fallibel sind und sich in seltenen Einzelfällen als falsch herausstellen können. Anstelle der empirischen Adäquatheit von Van Fraassen tritt somit Robustheit als epistemischer Leitwert von Wissenschaft: Wissenschaft hat das Ziel, Theorien mit robusten Behauptungen aufzustellen und die Akzeptanz einer Theorie impliziert nur den Glauben an die Wahrheit ihre robusten Behauptungen.

Doch wieso sind robuste Behauptungen so schwer zu widerlegen? Der Begriff der Robustheit wurde entscheidend durch William Wimsatt (1981) geprägt und hängt eng mit dem Konzept der multiplen Determiniertheit zusammen. Etwas ist multipel determiniert, wenn es auf verschiedene unabhängige Weisen hergeleitet, identifiziert oder gemessen werden kann. Dadurch ist eine Annahme über die Existenz einer Entität, ein Gesetz oder ein Prozess nicht von bestimmten Hintergrundtheorien abhängig. Selbst wenn sich herausstellt, dass in einem Experiment oder einer Herleitung ein Fehler aufgetreten ist, kann das Ergebnis beibehalten werden, da es weitere Experimente und Herleitungen gibt, die von dem Fehler nicht beeinflusst werden. Wenn verschiedenartige Lösungswege zum gleichen Ergebnis kommen, hat man es daher mit einem robusten Ergebnis zu tun.

Der Begriff der Robustheit wurde lange Zeit kaum beachtet und erst in jüngster Zeit als einer von vielen epistemischen Werten wiederentdeckt. Doch für eine Modifikation des Konstruktiven Empirismus kann Robustheit außerordentlich fruchtbar sein. Während es zweifelhaft ist, dass nur Aussagen über Beobachtbares gerechtfertigt sind, ist eine Beschränkung auf robuste wissenschaftliche Behauptungen selbstevident und beinahe analytisch. Bereits William Wimsatt hat betont, dass Robustheit die verlässlichen Elemente der Wissenschaft von den spekulativen Elementen abgrenzt. Einige Philosophen, vor allem aus der philosophischen Schule des Neuen Experimentalismus, haben bereits die Wichtigkeit von Robustheit für sicheres und stabiles wissenschaftliches Wissen be-

tont. Bisher hat allerdings niemand systematisch dafür argumentiert, dass Robustheit wichtiger als andere epistemische Werte sein könnte.

Dagegen möchte ich dafür plädieren, Robustheit als einziges und entscheidendes Kriterium zu nutzen, um gerechtfertigte wissenschaftliche Überzeugungen von Postulaten, Heuristiken und pragmatisch nützlichen Hilfsmitteln zu unterscheiden. Ebenso wie im Konstruktiven Empirismus gibt es einen zentralen epistemischen Leitwert, während alle anderen epistemischen Werte wie Einfachheit, Einheitlichkeit oder große Erklärungskraft nur eine pragmatische Rolle spielen. Robustheit anstelle von empirischer Adäquatheit führt ebenso zu einem selektiven Skeptizismus und zur Unterscheidung zwischen Akzeptanz einer Theorie und Glauben an alle ihre Behauptungen. Allerdings spielt empirische Adäquatheit selbst in dieser Position keine große Rolle mehr, so dass es irreführend wäre, sie lediglich als eine Modifikation des Konstruktiven Empirismus zu bezeichnen.

3. Robustheit und der Schluss auf die beste Erklärung

Ein praktisches Beispiel, wie man durch multiple Determiniertheit zu sicheren wissenschaftlichen Resultaten kommt, findet man bei Ian Hacking (1983), selbst wenn dieser die Vokabel „Robustheit“ nicht verwendet. Hacking nutzt multiple Determinierbarkeit, um zwischen realen Entitäten und Artefakten unterscheiden zu können. Die Zellstruktur von roten Blutkörperchen kann sowohl mit einem starken Lichtmikroskop als auch mit einem schwachen Elektronenmikroskop beobachten. Für Hacking wäre es ein Wunder, wenn zwei vollkommen verschiedene physikalische Vorgänge immer wieder dasselbe Ergebnis hervorrufen würden. Man müsste einen Mikroskoptäuscherdämon annehmen, um dieses Phänomen anders zu erklären, als dass man tatsächlich die Strukturen von realen Entitäten sieht.

Dieses Argument wird als *Argument of Coincidence* bezeichnet. Doch ist es wirklich so gut, wie es klingt? Richard Reiner und Robert Pierson (1995) weise darauf hin, dass dieses Argument entgegen der Beteuerung von Hacking ein Schluss auf die beste Erklärung ist, da man die identische visuelle Konstellation bei den verschiedenartigen Messungsverfahren durch die Realität der roten Blutkörperchen erklärt. Eine alternative Erklärung wäre, dass wir nur zufällig dieselben Ergebnisse erhalten, diese Erklärung ist allerdings weniger plausibel. Doch da die Gültigkeit des Schlusses auf die beste Erklärung umstritten ist, kann man auch Hackings Argument nicht trauen. Damit wird auch die Möglichkeit infrage gestellt, wissenschaftliche Behauptungen mit dem Konzept der Robustheit zu rechtfertigen.

Um das Argument of Coincidence und damit auch Robustheit als epistemischen Leitwert zu rechtfertigen, muss sichergestellt sein, dass die Einwände ge-

gen den Schluss auf die beste Erklärung zumindest für diesen Fall nicht gerechtfertigt sind. In seinem Buch *Laws and Symmetry* formuliert Bas Van Fraassen (1989) zwei Einwände gegen den Schluss auf die beste Erklärung. Sein erstes Argument ist das *Argument from Indifference*. Es besagt, dass unsere empirischen Evidenzen oft durch sehr viele unterschiedliche und sich widersprechende Hypothesen erklärt werden können. Daher ist jede Erklärung für sich genommen sehr unwahrscheinlich, inklusive der besten Erklärung. Wenn aber die beste aller Erklärungen für ein bestimmtes Phänomen sehr unwahrscheinlich ist, kann sie auch nicht gerechtfertigt sein. Stathis Psillos (1996) bestreitet allerdings, dass man für jede Erklärung *gute* Alternativen finden kann. Denn die Alternativen dürfen keine wissenschaftsfernen Trivialfälle sein und müssen annähernd genauso gute Erklärungen liefern wie die beste Erklärung. Dies kontern James Ladyman, Van Fraassen und andere (1997) wiederum damit, dass allein die Möglichkeit, gute Erklärungen finden zu können, gegen den Schluss auf die beste Erklärung spricht. Psillos (1999) hält die generelle Möglichkeit der Existenz von *guten* Alternativen aber nicht für bewiesen und schiebt die Beweislast wieder den Antirealsten zu. Somit scheint es hier ein Patt zu geben, da beide Parteien der jeweils anderen die Beweislast für ihre Thesen aufdrücken und niemand einen Beweis seiner eigenen These erbringt.

Der Diskussion um Van Fraassens zweites Argument geht es zunächst ähnlich. Um ein Phänomen zu erklären, reicht es dem *Argument from the bad lot* zufolge nicht aus, die Wahrscheinlichkeiten der gegebenen Hypothesen gegeneinander abzuwägen. Man benötigt die Zusatzprämisse, dass sich die richtige Hypothese bereits unter den gegebenen Alternativen befindet, da anderenfalls die beste Erklärung lediglich die beste falsche Erklärung wäre. Wir können diesem Argument zufolge allerdings nicht wissen, ob sich die richtige Hypothese bereits unter den von uns bekannten möglichen Erklärungen befindet. Ein Schluss auf die beste *verfügbare* Erklärung kann aber keinesfalls die Wahrheit dieser Erklärung garantieren. Psillos betrachtet dies zwar als theoretische Möglichkeit, doch unser Hintergrundwissen rüstet uns mit epistemischen Privilegien aus, mit denen wir die wahre Hypothese in den meisten Fällen finden. Van Fraassen sieht wiederum bereits die Möglichkeit der Nichtberücksichtigung der wahren Hypothese als ernsthafte Bedrohung für den realistischen Standpunkt an.

Auch hier haben wir es zunächst mit einer Pattsituation zu tun. Doch durch das *Argument of unconceived Alternatives* von Kyle Stanford (2006) hat Van Fraassen vor einiger Zeit beträchtliche Schützenhilfe bekommen. Stanford macht deutlich, dass sich die meisten Hypothesen, die ein Phänomen erklären könnten, außerhalb unseres epistemisch zugänglichen Horizontes befinden. Wir haben schlichtweg nicht die Zeit, das Geld, die technischen, mathematischen oder kognitiven Möglichkeiten, alle Alternativen überblicken zu können. Dies erscheint plausibel: Einsteins Relativitätstheorie ist eine gute Alternative zur Mechanik Newtons, doch im 17. Jahrhundert war die Mathematik noch nicht

weit genug entwickelt, um sie überhaupt als Alternative berücksichtigen zu können. Ebenso konnte Newtons Mechanik nicht als Alternative zur Physik des Aristoteles berücksichtigt werden, bevor Newton und Leibniz die Infinitesimalrechnung entwickelt hatten. Stanford zufolge kam es darüber hinaus in der Geschichte der Wissenschaft sogar häufig Alternativen zu einer etablierten Theorie vor, die aber von den Anhängern des herrschenden Paradigmas nicht zur Kenntnis genommen wurden, bevor dieses Paradigma in eine Krise geriet. Daher ist es nicht nur möglich, sondern sehr wahrscheinlich, dass sich auch die wahre Erklärung unter den unzugänglichen Hypothesen befindet. Auch unser Hintergrundwissen kann dies nicht verhindern, sondern muss sich Stanford zufolge in vielen Fällen sogar ändern, damit wir gute alternative Erklärungen überhaupt akzeptieren.

Stanford lehnt den Schluss auf die beste Erklärung allerdings nicht grundsätzlich ab, da dies einem generellen Skeptizismus sehr nahe käme. Erinnern wir uns daran, dass Van Fraassens Konstruktiver Empirismus auch deshalb als unplausibel zurückgewiesen wurde, da unser historisches Wissen vor allem auf abduktiven Schlüssen beruht. Kant ist in der Gegenwart keine beobachtbare Entität mehr, doch ein Gesamtwerk von Kants Schriften und ein umfassender Nachlass neben vielen offiziellen und persönlichen Schriften ist kaum anders zu erklären als durch die reale Existenz Kants als historische Person. Um den Radius gerechtfertigter Schlüsse auf die beste Erklärung abstecken zu können ist es entscheidend, den Unterschied zwischen der Existenz Kants und der Newton'schen Mechanik als Erklärungen für gewisse Phänomene zu betonen. Denn im Gegensatz zur Newton'schen Mechanik scheint es zur Existenz Kants weder eine verfügbare Alternative zu geben, noch ist eine unberücksichtigte Alternative zu erwarten. Zwar folgt die Existenz Kants nicht aus den vorhandenen Daten, denn es könnte ja auch ein Philosophenteam gegeben haben, die alle unter dem Pseudonym „Kant“ veröffentlicht haben oder Gott könnte die Welt erst vor fünf Minuten mit all unseren scheinbaren Erinnerungen erschaffen haben. Doch diese Alternativen ähneln Verschwörungstheorien und können als Rivalen einer ernsthaften wissenschaftlichen Annahme ignoriert werden. Schließt man alle irrelevanten Alternativen aus, ist die Existenz Kants die *einzige* gute Erklärung für unsere vorhandenen Daten.

Das legitime Ignorieren von irrelevanten Alternativen findet sich in erkenntnistheoretischen Positionen der Theorie der Relevanten Alternativen, bei der es um Kriterien von Wissen geht. In dieser von Alvin Goldman, Fred Dretske, David Lewis und anderen Philosophen entwickelten Position weiß ein Subjekt S genau dann, dass p, wenn S alle relevanten Alternativen zu p ausräumen kann. Welche Alternativen relevant sind, ist hierbei kontextabhängig. Diese Ideen sind auch auf den Schluss auf die beste Erklärung anwendbar. Es scheint keine guten Gründe dafür zu geben, warum in der Wissenschaftsphilosophie Überzeugungen mit guten Alternativen gerechtfertigt sein sollen, wenn sie es in der Erkenntnis-

theorie nicht sind. Daher sollte eine wissenschaftliche Erklärung ebenfalls nur dann gerechtfertigt sein, wenn man alle relevanten alternativen Erklärungen ausräumen kann. Im wissenschaftlichen Kontext sind skeptische Hypothesen wie böse Täuscherdämonen selbstverständlich niemals relevant, wohl aber die Gefahr unentdeckter Alternativen.

Der Schluss auf die beste Erklärung ist somit nur dann gerechtfertigt, wenn es genau eine gute Erklärung gibt, andere Erklärungen nicht relevant sind und relevante, aber unberücksichtige Erklärungen nicht auftauchen können. In diesem Fall spricht Alexander Bird (2007) von einem Schluss auf die *einzige* Erklärung. Diese Schlussart ist weitaus besser gerechtfertigt als ein Schluss auf die beste Erklärung mit mehreren plausiblen, rivalisierenden Erklärungen.

Hackings Beispiel der roten Blutkörperchen, dessen Struktur wir sowohl mit einem Lichtmikroskop als auch mit einem Elektronenmikroskop sehen können, ist eine Instanz des Schlusses auf die einzige Erklärung. Es wäre ein Wunder, wenn die identischen visuellen Ergebnisse nur zufällig immer wieder auftauchen würden. Wunder sind im normalen wissenschaftlichen Kontext aber niemals relevant. Somit wird das *Argument of Coincidence* nicht von den Einwänden gegen den Schluss auf die beste Erklärung bedroht, selbst wenn es ein abduktiver Schluss ist.

Der Schluss auf die einzige Erklärung dient als Basis zur Rechtfertigung von robusten Behauptungen durch multiple Determinierbarkeit. Eine einzelne Herleitung, Beobachtung oder Messung ist theoriegeladen, fehleranfällig und nicht vertrauenswürdig. Wenn allerdings viele unabhängige Herleitungsweisen zu identischen Ergebnissen kommen, kann man nur schwer annehmen, dass alle Herleitungen fehlerhaft sind und trotzdem dasselbe Ergebnis liefern. Eine wissenschaftliche Methodologie, die sich dieser Begründung bedient, kann man als Robustheits-Analyse bezeichnen. Während Robustheit ein genereller, abstrakter epistemischer Wert ist, sind Robustheits-Analysen präzise und für Einzeluntersuchungen angepasste wissenschaftliche Methodologien. Die These, dass sich Wissenschaft dynamisch verhält und sich wissenschaftliche Methoden mit der Zeit verändern ist kompatibel mit der Annahme, dass es jederzeit in den fortgeschrittenen Wissenschaften Formen von Robustheits-Analysen geben kann, durch die stabile wissenschaftliche Ergebnisse sichergestellt werden. Hiermit soll allerdings nicht behauptet werden, dass allein durch Robustheits-Analysen sichere wissenschaftliche Behauptungen erzielt werden können. Allerdings sind Robustheits-Analysen hinreichend für dieses Ziel und können in vielen wissenschaftlichen Disziplinen angewandt werden.

Durch Robustheits-Analyse können mittels multipler Determiniertheit aus theorieverseuchten Daten verlässliches Wissen gewonnen werden. Es spricht daher einiges dafür, diese Ergebnisse als gerechtfertigt anzusehen. Abstrakte theoretische Behauptungen basieren dagegen auf unsicheren Schlüssen auf die besten Erklärungen, mit denen die Gefahr relevanter aktueller oder unentdeckter

Alternativen nicht vermieden werden kann. Eine Position, die Robustheit als epistemischen Leitwert akzeptiert, zeichnet daher besonders das wissenschaftliche Wissen aus, das durch Experimente gewonnen wird und ist theoretischen Hypothesen gegenüber skeptisch eingestellt.

Literaturverzeichnis

Bird, Alexander: "Inference to the Only Explanation". Philosophy and Phenomenological Research, 74, 2007. S. 424–32

Cartwright, Nancy: Why be hanged for even a lamb? In: *Monton, B.* (*Hrsg.):* Images of Empiricism: Essays on Science and Stances, with a Reply from Bas C. van Fraassen. Oxford University Press, Oxford, 2007. S. 33-45.

Churchland, Paul: The Ontological Status of Observables: In Praise of the Superempirical Virtues. In: *Churchland, Paul und Hooker, Clifford (Hrsg.):* Images of Science: Essays on Realism and Empiricism (with a reply from Bas C. van Fraassen). University of Chicago Press, Chicago, 1985. S. 35-47

Hacking, Ian: Representing and Intervening. Cambridge University Press, Cambridge, 1983

Ladyman, James, Douven, Igor, Horsten, Leon und van Fraassen, Bas: "A Defense of Van Fraassen's Critique of Abductive Inference": Reply to Psillos. The Philosophical Quarterly, 47, 1997. S. 305-321

Maxwell, Grover: "The ontological status of theoretical entities". In: *Feigl, Herbert und Maxwell, Grover (Hrsg.)*: Minnesota Studies in the Philosophy of Science 3. University of Minnesota Press, Minneapolis, 1962. S. 3-14

Musgrave, Alan: "Realism versus Constructive Empiricism." In: *Churchland, Paul und Hooker, Clifford (Hrsg.)*: Images of Science: Essays on Realism and Empiricism (with a reply from Bas C. van Fraassen). University of Chicago Press, Chicago, 1985. S. 197-221

Psillos, Stathis: "On Van Fraassen's Critique of Abductive Reasoning." The Philosophical Quarterly, 46, 1996. S. 31-47

Psillos, Stathis: Scientific Realism: How Science Tracks Truth. Routledge, London, 1999

Reiner, Richard und Pierson, Robert: "Hacking's Experimental Realism: An Untenable Middle Ground". Philosophy of Science, 62, 1995. S. 60-69

Sober, Elliott: "Epistemology for Empiricists." In: *Wettstein, H. (Hrsg.):* Midwest Studies in Philosophy 18. Philosophy of Science. University of Notre Dame Press, Notre Dame, 1993. S. 39-61

Stanford, P. Kyle: Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford University Press, Oxford, 2006

Van Fraassen, Bas: The Scientific Image. Oxford University Press, Oxford, 1980

Van Fraassen, Bas: Laws and Symmetry. Oxford University Press, Oxford, 1989

Wimsatt, William: "Robustness, Reliability, and Overdetermination". In: *Brewer, M. und Collins, B*: Scientific Inquiry and the Social Sciences. Jossey-Bass, San Francisco, 1981. S. 24-163

Ist Wahrheit positiv?
Ein Paradox der Gödelschen Positivität

Gregor Damschen
gregor.damschen@unilu.ch
Universität Luzern

Abstract/Zusammenfassung

Is Truth Positive? A Paradox of Gödel's Positiveness
In his "Ontological Proof", Kurt Gödel introduces the notion of a second-order value property, the positive property *P* (K. Gödel, Ontological proof, in: Collected Works, vol. III, ed. S. Feferman et al., Oxford U.P., 1995, 403-404). $P(\varphi)$ is true if φ is a first-order property and φ is positive, and 'positive' means 'purely good'. Moreover, the second axiom of the proof states that for any property φ: $P(\neg\varphi) \leftrightarrow \neg P(\varphi)$. That is, if any property φ is positive, its negation is not positive, and vice versa. In this paper, I put forward that Gödel's concept of positiveness, especially the second axiom, leads into a paradox when we apply it to the following self-reflexive sentences which are not trivial nor paradoxical nor meaningless:
(A) The truth-value of A is not positive;
(B) The truth-value of B is positive.
Given axiom 2, sentences A and B paradoxically cannot be both true or both false, and it also follows that it is impossible that one of the sentences is true whereas the other is false.

In diesem Aufsatz wird gezeigt, dass die Frage, ob Wahrheit positiv sei, ein Paradox ergibt, wenn man das Axiom 2 des Konzepts der Positivität zugrundelegt, das Kurt Gödel in seinem ontologischem Beweis der Existenz Gottes entwickelt hat (K. Gödel, Ontological proof, in: Collected Works, vol. III, ed. S. Feferman et al., Oxford U.P. 1995, 403-404). Gödel versteht unter der ‚positiven Eigenschaft' *P* eine nicht-relative Werteigenschaft zweiter Stufe, die auf Eigenschaften erster Stufe φ gemäß Axiom 2 folgendermaßen angewendet werden muss: $P(\neg\varphi) \leftrightarrow \neg P(\varphi)$. Das bedeutet, dass bei jedem Paar einer Eigenschaft erster Stufe und ihrer Negation genau eine Eigenschaft positiv ist. Gegeben seien zudem die beiden nicht-trivialen, nicht-paradoxen und nicht-sinnlosen selbstbezüglichen Sätze A und B:
(A) Der Wahrheitswert von A ist nicht positiv.
(B) Der Wahrheitswert von B ist positiv.
Unter der Annahme der Sätze A und B, des Axioms 2 sowie der Invarianz der positiven Eigenschaft *P* gegenüber den Trägern der Eigenschaften erster Stufe φ, auf die *P* angewendet wird, ergibt sich das paradoxe Resultat, dass A und B we-

der zusammen wahr noch zusammen falsch sein können, und auch keiner der beiden Sätze wahr sein kann, wenn der andere falsch ist.

The way of paradoxes is the way of truth.
(Oscar Wilde)

Wenn Oscar Wilde Recht hat und der Weg der Paradoxien der Weg zur Wahrheit ist, dann wäre paradoxerweise auch der Weg der *Paradoxien der Wahrheit* der Weg zur Wahrheit. Paradoxien der Wahrheit wie der Lügner und der Verstärkte Lügner sind tatsächlich immer wieder Anlass für gehaltvolle Reflexionen über den Begriff der Wahrheit gewesen. Jede Neuentdeckung im Bereich der Paradoxien der Wahrheit könnte also unser Verständnis der Wahrheit erhöhen. Aus diesem Grunde möchte ich ein neues Problem, vielleicht sogar eine neue Paradoxie der Wahrheit vorstellen, die sich aus der Frage ergibt, ob Wahrheit positiv sei. Es wird sich zeigen, dass Kurt Gödels Positivitätseigenschaft (insbesondere das Axiom 2 des Gödelschen Beweises der Existenz Gottes) zusammen mit zwei selbstbezüglichen Sätzen, die Aussagen über die Positivität von Wahrheitswerten machen, zu einem Paradox der positiven Wahrheit führt.

I. Positivität

Was ist Positivität? Ein formales System der Positivität, das hier zugrunde gelegt werden soll, hat Kurt Gödel in seinem „ontologischen Beweis“ der Existenz Gottes skizziert (Gödel 1995, 403-404). Dieses modallogische Argument kursierte seit den 70er Jahren des letzten Jahrhunderts unter Logikern und Philosophen, wurde aber erst 1987 publiziert. Kernstück des Beweises ist die Annahme der sogenannten ‚positiven Eigenschaft‘.

Gödel erläutert den in seinem Beweis verwendeten Grundbegriff der Positivität *P* folgendermaßen: Die Positivität *P* ist eine Eigenschaft zweiter Stufe, die auf Eigenschaften φ erster Stufe angewendet wird. In der ersten Zeile des ontologischen Beweises wird Positivität deshalb als die Eigenschaft $P(\varphi)$ eingeführt: „φ is positive (or $\varphi \in P$)“ (Gödel 1995, 403). Zweitens ist ‚*P*‘ ein Wertprädikat, die Positivität *P* eine Werteigenschaft: „Positive means positive in the moral aesthetic sense (independently of the accidental structure of the world).“ (Gödel 1995, 404). Darüber, was unter dem „moral aesthetic sense“ der Positivität zu verstehen ist, gibt ein Tagebucheintrag Gödels weitere Auskunft: „It is possible to interpret the positive as perfective; that is, ‘purely good’, that is, such as implies no negation of ‘purely good’.“ (Gödel 1995, 434-35 = „Phil XIV“, S. 105). Positivität ist also drittens keine Werteigenschaft, die erst relativ zu etwas anderem ihren Wert erhält, sondern sie besitzt ihren Wert absolut. Mit dem Hinweis darauf, dass Positivität als „*reines* Gutsein“ verstanden werden muss, grenzt sich Gödel explizit von der Annahme ab, positiv sei bereits das, was in

irgendeiner Hinsicht gut ist: „The interpretation of ‘positive property’ as ‘good’ (that is, as one with positive value) is impossible, because the greatest advantage + the smallest disadvantage is negative“ („Phil XIV“, S. 105). Gödels Positivität ist viertens am Perfektionenbegriff orientiert – wobei Perfektionen Eigenschaften sind, die sich nicht selbst und keiner anderen Perfektion widersprechen: „It [sc. positive, G.D.] may also mean pure “attribution” as opposed to privation (or containing privation)“ (Gödel 1995, 404). Mit diesen inhaltlichen Erläuterungen der Positivität als einer perfekten Werteigenschaft zweiter Stufe ist schließlich eine fünfte Annahme verbunden: Wenn eine Eigenschaft erster Stufe positiv ist, dann ist ihr Positivsein *invariant* demgegenüber, wer oder was Träger der Eigenschaft erster Stufe ist („independently of the accidental structure of the world“; Gödel 1995, 404).

II. Ein Axiom der Positivität und zwei Sätze über Wahrheit

Wenn Wahrheit eine Eigenschaft erster Stufe ist, stellt sich die Frage, ob Wahrheit positiv ist oder nicht. Folgende Gödelsche Axiome (Gödel 1995, 403-404) tragen nun zum Paradox der positiven Wahrheit bei:

1.) Positivität ist eine Eigenschaft zweiter Stufe: $P(\varphi)$.

2.) Das Gödelsche Axiom 2 besagt, dass für jede beliebige Eigenschaft φ gilt: entweder sie oder ihr Komplement bzw. ihre Negation ist positiv.

Axiom 2: $P(\neg\varphi) \leftrightarrow \neg P(\varphi)$ *Dichotomie; Ultrafilter-Axiom*
Bei jedem Paar einer Eigenschaft und ihrer Negation ist *genau eine* Eigenschaft positiv.

Dieses Axiom lässt sich leicht in zwei Axiome zerlegen, die beide gesondert behandelt werden sollten:

Axiom 2.1: $\neg P(\varphi) \rightarrow P(\neg\varphi)$
Wenn eine Eigenschaft erster Stufe nicht positiv ist, dann ist ihre Negation positiv.
Mindestens ein Glied des Paares ist positiv.

Axiom 2.2: $P(\neg\varphi) \rightarrow \neg P(\varphi)$
Wenn eine Eigenschaft erster Stufe positiv ist, dann ist ihre Negation nicht positiv.
Höchstens ein Glied des Paares ist positiv.

Nehmen wir noch eine weitverbreitete Annahme über die Wahrheit und zwei selbstbezügliche Sätze über Wahrheitswerte hinzu. Die Annahme über die Wahrheit lautet:

Annahme 1: Wahrheit ist eine Eigenschaft von Wahrheitswertträgern.

Die beiden selbstbezüglichen Sätze über Wahrheitswerte lauten:

(A) Der Wahrheitswert von A ist nicht positiv.
(B) Der Wahrheitswert von B ist positiv.

Aus dem Axiom 2, der Annahme 1 und den beiden Sätzen A und B sowie der Annahme, dass Positivität invariant ist, ergibt sich das Paradox der positiven Wahrheit. Wie genau, werde ich im Folgenden zeigen.

III. Sind Wahrheitswerte positiv?

Die selbstbezüglichen Sätze A und B sind nicht trivial, nicht paradox und nicht sinnlos. Trivial sind sie nicht, da sie weder analytisch sind, noch aus analytischen Sätzen gewonnen werden können, noch ihr Wahrheitswert sofort zu bestimmen wäre. Paradox sind sie nicht, da sich aus ihnen nicht ergibt, dass sie wahr sind genau dann, wenn sie falsch sind. Sinnlos sind sie ebenfalls nicht, da sie syntaktisch wohlgeformt sind und die in ihnen vorkommenden Begriffe zusammen einen Gedanken wiedergeben, den man für wahr oder falsch halten kann. Die Selbstbezüglichkeit der beiden Sätze allein ist auch kein Grund, sie für sinnlos zu halten, denn selbstbezügliche Sätze wie „Dieser Satz enthält fünf Wörter“ scheinen wahrheitsfähig zu sein. Dies vorausgesetzt, kann man sich dann fragen, welchen Wahrheitswert die Sätze A und B haben. Für A ergeben sich zwei Möglichkeiten.

(i) Wenn A wahr ist, ist der Wahrheitswert von A nicht positiv. Wenn der Wahrheitswert von A nicht positiv ist, stimmt es, was A zum Ausdruck bringt, und A ist wahr. A ist also genau dann wahr, wenn der Wahrheitswert von A nicht positiv ist.
(ii) Wenn A falsch ist, ist der Wahrheitswert von A positiv. Wenn der Wahrheitswert von A positiv ist, stimmt es nicht, was A zum Ausdruck bringt, und A ist falsch. A ist also genau dann falsch, wenn der Wahrheitswert von A positiv ist.

Wenn A wahr ist, kann der Wahrheitswert von A nicht zugleich falsch sein, da A sonst zugleich wahr und falsch wäre, was zum Widerspruch führt. Dasselbe gilt für den Fall, dass A falsch ist. Auch hier gilt, wenn A falsch ist, ist auch sein Wahrheitswert falsch. Es ergibt sich deshalb, dass die Wahrheit von A keine positive Eigenschaft, die Falschheit von A eine positive Eigenschaft ist:

1. Die Falschheit von A ist positiv.
2. Die Wahrheit von A ist nicht positiv.

Nimmt man das Axiom 2.1 hinzu, ergibt sich für A:

3. Wenn die Wahrheit von A nicht positiv ist, ist die Nicht-Wahrheit von A positiv.

Aus den Sätzen 2 und 3 ergibt sich per Modus ponens:

4. Die Nicht-Wahrheit von A ist positiv.

Unter der Anname, dass Falschheit die Negation der Wahrheit ist, ergibt sich aus Satz 1 dasselbe:

5. Die Nicht-Wahrheit von A ist positiv.

Im Falle des Satzes B erhält man dementsprechend andere Ergebnisse:

6. Die Falschheit von B ist nicht positiv.
7. Die Wahrheit von B ist positiv.

Unter Hinzunahme von Axiom 2.1 ergibt sich aus Satz 6:

8. Die Wahrheit von B ist positiv.

Setzt man das Axiom 2.1 voraus, ergibt sich also, dass sowohl unter der Annahme der Wahrheit als auch unter der Annahme der Falschheit von Satz A die Nicht-Wahrheit von A positiv ist. Umgekehrt verhält es sich im Falle von B: Sowohl unter der Annahme der Wahrheit als auch der Falschheit von B ist die Wahrheit von B positiv.

IV. Positivität ist invariant

Wenn positive Eigenschaften Perfektionen sind und sich die Positivität auf die Perfektion der Eigenschaft bezieht, dann ändert sich die Perfektion der Eigenschaft nicht dadurch, dass sie von (möglicherweise) verschiedenen Individuen exemplifiziert wird. Wenn beispielsweise reine Schönheit positiv ist, dann ist es mit Blick auf die Positivität der reinen Schönheit egal, wer die reine Schönheit exemplifiziert. Nehmen wir also folgendes Axiom der Positivitätsinvarianz (PosInv) hinzu:

Axiom PosInv: Wenn es etwas gibt, das eine Eigenschaft φ hat und φ positiv ist, dann ist die Positivität von φ invariant gegenüber dem Träger von φ.

Wenn wir das Axiom der Positivitätsinvarianz auf Satz 5 anwenden, erhalten wir:

9. Wenn die Nicht-Wahrheit von A positiv ist, dann ist Nicht-Wahrheit positiv.

Aus den Sätzen 5 und 9 ergibt sich per Modus ponens:

10. Nicht-Wahrheit ist positiv. $P(\neg W)$

Wenn wir das Axiom der Positivitätsinvarianz auf Satz 8 anwenden, erhalten wir:

11. Wenn die Wahrheit von B positiv ist, dann ist Wahrheit positiv.

Aus den Sätzen 8 und 11 ergibt sich dann abschließend:

12. Wahrheit ist positiv. $P(W)$

V. Das Paradox der positiven Wahrheit

Es gibt nur genau vier Fälle, wie sich die sinnvollen Sätze A und B logisch zueinander verhalten können: Beide Sätze sind wahr, beide falsch oder jeweils einer ist wahr, während der andere falsch ist. Ein einfacher formaler Beweis zeigt, dass jeder der vier Fälle zu einem formalen Widerspruch führt.

1. $W(A) \wedge W(B)$: $P(\neg W) \wedge P(W)$

$P(\neg W) \wedge P(W)$ + Axiom 2.2 — *Widerspruch:* $\neg P(\neg W) \wedge P(\neg W)$

Beweis:

1.	$P(\neg W) \wedge P(W)$	
2.	$\forall\varphi(P(\neg\varphi) \rightarrow \neg P(\varphi))$	Axiom 2.2
3.	$\forall\varphi(P(\varphi) \rightarrow \neg P(\neg\varphi))$	aus 2, Kontraposition von Axiom 2.2
4.	$P(W) \rightarrow \neg P(\neg W)$	$\forall$-Elimination, W/φ in 3
5.	$P(W)$	aus 1, Simplifikation
6.	$\neg P(\neg W)$	aus 4 und 5, Modus ponens
7.	$P(\neg W)$	aus 1, Simplifikation
Also: 8.	$\neg P(\neg W) \wedge P(\neg W)$	aus 6 und 7, Konjunktionseinführung

Q.E.D. — *Widerspruch*

2. $F(A) \wedge F(B)$: $P(\neg W) \wedge P(W)$

$P(\neg W) \wedge P(W)$ + Axiom 2.2 — *Widerspruch:* $\neg P(\neg W) \wedge P(\neg W)$

Beweis wie in 1.

3. $W(A) \wedge F(B)$: $P(\neg W) \wedge P(W)$

$P(\neg W) \wedge P(W)$ + Axiom 2.2 — *Widerspruch:* $\neg P(\neg W) \wedge P(\neg W)$

Beweis wie in 1.

4. $F(A) \wedge W(B)$: $P(\neg W) \wedge P(W)$

$P(\neg W) \wedge P(W)$ + Axiom 2.2 *Widerspruch:*
$\neg P(\neg W) \wedge P(\neg W)$

Beweis wie in 1.

VI. Ergebnis

Die Aussagen A und B über die Positivität ihrer Wahrheitswerte ergeben zusammen mit Gödels Axiomen 2.1, 2.2 und dem Axiom der Positivitätsinvarianz ein Paradox: Die singulären Sätze A und B können weder zusammen wahr noch zusammen falsch noch jeweils einer wahr und einer falsch sein. Denn jede der vier möglichen Alternativen erzeugt einen formalen Widerspruch. Wir hatten jedoch angenommen, dass die Sätze A und B nicht trivial, nicht paradox und nicht sinnlos sind.

Dieses Ergebnis gilt nicht nur für Gödels positive Eigenschaft *P*, sondern für *jede* beliebige Eigenschaft zweiter Stufe Ψ, die in ein formales System integriert ist, das zumindest strukturell analoge Axiome zu den Axiomen 2.1, 2.2 und dem Axiom der Positivitätsinvarianz enthält. Das Puzzle, das es zu lösen gilt, ergibt sich, wenn man selbstreflexive Aussagen wie A und B über die Positivität der Wahrheitswerte von A und B hinzunimmt. Das Problem scheint also nicht allein durch Gödels Positivitätseigenschaft erzeugt zu werden, sondern – wie bei den anderen bekannten Paradoxien der Wahrheit – auch durch das Prädikat der Wahrheit und durch den Selbstbezug der Sätze. Wie das Problem durch das Positivitäts- und das Wahrheitsprädikat gemeinsam erzeugt wird und wie das Paradox der positiven Wahrheit gelöst werden könnte, bleibt noch zu klären.

Literaturverzeichnis

Gödel, Kurt: Ontological proof. In: Collected Works, vol. III, ed. S. Feferman et al.. Oxford U.P., Oxford, 1995. S. 403-404 (plus Appendix B: Texts relating to the ontological proof. S. 429 ff.)

Das Ende der Paradigmenwechsel

Ludwig Fahrbach
Ludwig.Fahrbach@googlemail.com
Heinrich-Heine-Universität, Düsseldorf

Abstract/Zusammenfassung

In this paper I defend scientific realism against the argument from pessimistic meta-induction. I take scientific realism to be the position that licences an inference from success of a scientific theory to its approximate truth. The argument from pessimistic meta-induction maintains that this kind of inference is undermined by numerous counterexamples, i.e., by theories from the history of science that were successful, but false. In order to defend scientific realism against the pessimistic meta-induction, I adopt a notion of success that admits of degrees. I aim to show that our current best theories enjoy far higher degrees of success than any of the successful, but refuted theories of the past.

In diesem Aufsatz verteidige ich den wissenschaftlichen Realismus gegen die pessimistische Metainduktion. Unter wissenschaftlichem Realismus verstehe ich die Position, dass wir vom Erfolg von wissenschaftlichen Theorien auf ihre annähernde Wahrheit schließen dürfen. Die pessimistische Metainduktion ist ein Argument gegen diesen Schluss: Die Wissenschaftsgeschichte ist voll von Theorien, die ein Zeitlang erfolgreich waren, dann aber widerlegt wurden. Für meine Verteidigung des wissenschaftlichen Realismus verwende ich einen Begriff des Erfolgs von wissenschaftlichen Theorien, der Grade zulässt. Ich möchte zeigen, dass unsere gegenwärtig besten Theorien viel höhere Erfolgsgrade genießen als alle widerlegten Theorien.

Der wissenschaftliche Realismus

Unter wissenschaftlichem Realismus verstehe ich in diesem Aufsatz die Position, die das folgende Schlussprinzip gutheißt: Wenn eine wissenschaftliche Theorie erfolgreich ist, dann ist sie annährend wahr. Nennen wir dieses Schlussprinzip das EIW-Prinzip, für **E**rfolg **i**mpliziert **W**ahrheit. Wissenschaftliche Realisten wenden das EIW-Prinzip auf unsere gegenwärtig besten Theorien an, etwa auf die Evolutionstheorie, die Plattentektonik und die Atomtheorie der Materie, und schließen, dass diese annährend wahr sind.

Im EIW-Prinzip tauchen drei Begriffe auf, der Begriff des Erfolges, der Begriff der Wahrheit und der Begriff der annähernden Wahrheit. Der Begriff des Erfolges wird später ausführlich besprochen und präzisiert. Im Augenblick nur soviel: Eine wissenschaftliche Theorie nenne ich „erfolgreich zu einem gegebenen Zeitpunkt“, wenn, soweit den Wissenschaftern zu diesem Zeitpunkt be-

kannt, die Theorie hinreichend viele richtige und keine falschen empirischen Konsequenzen hat.

Zum Begriff der Wahrheit ließe sich viel sagen, aber ich möchte mich mit zwei Hinweisen begnügen. Wissenschaftliche Realisten vertreten oft eine Korrespondenztheorie der Wahrheit. Das ist völlig in Ordnung. Doch braucht man sich nicht auf eine solche festzulegen, denn für unsere Belange würde auch eine deflationäre Auffassung des Wahrheitsbegriffes ausreichen (Paul Horwich 1998). Der Begriff der *annähernden* Wahrheit wirft ebenfalls eine Menge Probleme auf. Eine allgemeine Analyse dieses Begriffes konnte bisher nicht geliefert werden. Jedoch können Wissenschaftler diesen Begriff in den meisten konkreten Situationen ohne Probleme anwenden, und das reicht, um das EIW-Prinzip brauchbar zu machen.

Im Folgenden werde ich die Qualifizierung „annährend“ meist weglassen und einfach von “Wahrheit“ von Theorien sprechen; es ist aber immer annährende Wahrheit gemeint. Auch werde ich meist „wissenschaftlicher Realismus“ mit “Realismus” abkürzen. Außerdem verwende ich die Bezeichnung „Theorie“ in einem sehr weiten Sinne. Ich bezeichne damit auch Naturgesetze, theoretische Aussagen und sogar Klassifikationssysteme wie das Periodensystem der Elemente. Der Grund hierfür ist, dass es viele Aussagen dieser Art gibt (etwa die Aussagen, die im Periodensystem der Elemente enthalten sind), auf die der Realist ebenfalls das EIW-Prinzip anwenden möchte, d.h. ebenfalls von ihrem Erfolg auf ihre Wahrheit schließen möchte, und die ebenfalls durch die pessimistische Metainduktion angegriffen sind.

Das EIW-Prinzip ist ein induktives Prinzip, im weiten Sinn von Induktion, der alle nicht-deduktiven gültigen Schlüsse umfasst. Allgemein gesprochen geht es in der Debatte um den wissenschaftlichen Realismus um die Frage, wie weit unser induktives Schließen uns über die Beobachtung hinaustragen kann, d.h. welche Formen induktiven Schließens wahrheitszuträglich sind und welche nicht. Wissenschaftliche Realisten sind eher optimistisch und meinen, dass Schlussprinzipien wie das EIW-Prinzip wahrheitszuträglich sind, wohingegen Antirealisten auf verschiedene Weise und aus unterschiedlichen Gründen eher pessimistisch sind und Schlussprinzipien dieser Art eher ablehnen.

Die Gegenposition zum Realismus ist also der Antirealismus. Auch hier gibt es eine ganze Reihe verschiedener Versionen, aber ich werde in diesem Aufsatz keine konkrete Version definieren und besprechen, denn es geht mir ausschließlich um die Verteidigung des Realismus und nicht die Diskussion der verschiedenen Formen von Antirealismus.[1] Unter „Antirealismus“ können wir daher einfach die Negation des Realismus verstehen, also die Ablehnung des EIW-

1 Für einen Vergleich verschiedener Formen von Antirealismus speziell in Bezug auf die pessimistische Metainduktion siehe Fahrbach (2009a, 2009b).

Prinzips. Antirealisten sind dann Leute, die den Realismus anzugreifen versuchen, indem sie Argumente gegen das EIW-Prinzip vorbringen.

Typischerweise haben Definitionen des wissenschaftlichen Realismus neben epistemischen auch semantische, metaphysische und pragmatische Komponenten, aber ich beschränke mich in meiner Diskussion auf epistemische Fragen und werde daher die anderen Komponenten beiseite lassen.[2]

Eine Weise, den Realismus zu definieren, ist nicht sinnvoll. Sie besteht darin, dass eine Menge von wissenschaftlichen Theorien, etwa unsere gegenwärtig besten Theorien, eingegrenzt wird und dann deren Wahrheit behauptet wird.[3] Ein solches Vorgehen ist nicht geeignet, um den Realismus zu definieren. Denn wenn die Wissenschaft weiter fortschreitet, kann sich die Menge der besten Theorien ändern, zum Beispiel größer werden. Dann ändern sich auch die Überzeugungen des Realisten. Daran ist nichts auszusetzen, doch wollen wir nicht sagen, dass sich dann sein *Realismus* ändert. Vielmehr bleibt sein Realismus der gleiche. Der Realist akzeptiert die gegenwärtig besten Theorien, *weil sie erfolgreich sind*. Das ist sein Grund, und dieser Grund sollte in die Definition des Realismus aufgenommen werden. Dies wird durch das EIW-Prinzip geleistet. Wenn sich dann die Menge der erfolgreichen Theorien ändert, hat das weder Auswirkungen auf das EIW-Prinzip noch auf den Realismus.

Das Wunderargument

Das wichtigste Argument zur Stützung des Realismus ist das Wunderargument (Putnam 1978, Smart 1960). In seiner einfachsten Form ist es ein direkter Appell an unsere Intuitionen: „Wäre es nicht ein Wunder, wenn unsere besten Theorien trotz ihres hohen Erfolges falsch wären? Wäre es beispielsweise nicht ein Wunder, wenn sich chemische Substanzen immer wieder so verhielten, wie wenn sie aus den 92 chemischen Elementen zusammengesetzt wären, es aber in Wirklichkeit nicht sind?“ Neben dieser einfachsten Form des Wunderargumentes gibt es verschiedene Fortentwicklungen, wie z.B. den Schluss auf die beste Erklärung, doch will ich hier nicht weiter auf diese eingehen.[4]

Es scheint mir, dass sich feststellen lässt, dass alle Versionen des Wunderargumentes am Ende an „Intuitionen“ appellieren müssen.[5] Diese Intuitionen wer-

2 Darstellungen der Realismusdebatte finden sich in Jarrett Leplin (1997), André Kukla (1998), Stathis Psillos (1999), Jaako Niiniluoto (1999), Alan Musgrave (1999), Kyle Stanford (2006), Derek Turner (2007), Anjan Chakravartty (2007) und Ladyman/Ross (2008).

3 So definiert zum Beispiel Devitt (2005, p. 769) den Realismus.

4 Siehe Richard Boyd (1983), Psillos (1999, Kapitel 4), Philip Kitchers “Galilean strategy” (2001) und Gerhard Schurz (2008).

5 Darüber, was unter „Intuitionen“ genau zu verstehen ist, kann mehr sehr unterschiedlicher Meinung sein. Realisten mögen für Intuitionen eine apriorische Gültigkeit beanspruchen,

den von Antirealisten nicht geteilt. Sie weisen das Wunderargument daher zurück. Wir können also beobachten, dass sich die beiden Parteien in der Realismusdebatte – Realisten und Antirealisten – über die Haltbarkeit des EIW-Prinzips und ähnlicher Prinzipien wie den Schluss auf die beste Erklärung streiten, wobei Realisten sie für begründbar halten, während Antirealisten dies verneinen. In dieser Debatte kann, wie so oft in der Philosophie, keine Partei die andere überzeugen, und eine Entscheidung ist nicht in Sicht. Die Meinungsverschiedenheit scheint auf einem Aufeinanderprallen unterschiedlicher Intuitionen zu gründen, die nicht miteinander vereinbar sind.

Die pessimistische Metainduktion

Doch nun, so können wir die Realismusdebatte fortsetzen, präsentiert der Antirealist ein Argument gegen das EIW-Prinzip, das unabhängig vom bisherigen Verlauf der Debatte ist und nichts mit dem Zusammenprallen von unterschiedlichen Intuitionen zu tun hat. Dieses Argument ist die pessimistische Metainduktion (kurz PMI). Die PMI hat die Prämisse, dass die Wissenschaftsgeschichte voll von Theorien ist, die eine Zeitlang erfolgreich waren und von Wissenschaftlern akzeptiert wurden, jedoch später widerlegt wurden.[6] Nehmen wir zunächst einmal an, dass diese Prämisse korrekt ist. Dann stellen die erfolgreichen, aber widerlegten Theorien *Gegenbeispiele* gegen den Schluss vom Erfolg auf die Wahrheit dar. Sie unterminieren das EIW-Prinzip.[7]

Antirealisten müssen die Prämisse der PMI belegen. Das haben sie gemacht. Berühmt ist Larry Laudans Liste (1981) von solchen Theorien. Er erwähnt beispielsweise das geozentrische Weltbild der frühen Astronomie, die kalorische Theorie der Wärme, die Vitalkrafttheorie der Physiologie, die Phlogistontheorie, verschiedene Äthertheorien des Lichts, usw.

Der Antirealist argumentiert dann gegen den Realisten wie folgt. Selbst wenn man zunächst die Perspektive des Realisten, also die Intuitionen des Realisten, auf denen das Wunderargument und die Begründung des EIW-Prinzips beruht, akzeptiert, muss man einsehen, dass die PMI die Aufgabe all dieser Dinge erzwingt. Aus der Perspektive des Realisten muss man nämlich zwei Argumente gegeneinander abwägen, das Wunderargument und die PMI. Das Wunderargument stützt das EIW-Prinzip, und die PMI unterminiert dieses Prinzip. Die bei-

---

oder sie mögen meinen, dass sie irgendwie in unserer allgemeinen empirischen Erfahrung über die Welt gründen, oder sie mögen zugeben, dass sie nichts weiter sind als Meinungen (David Lewis, 1983, S. 10).

6 So bemerkt Jim Holt „Scientific progress … takes place by funerals." (2005)

7 Diese Form der PMI erwähnen unter anderem Psillos (1999, ch. 5), Peter Lewis (2001), Devitt (2005) und Kitcher (2001).

den Argumente müssen gegeneinander abgewogen werden.[8] Das Ergebnis des Abwägens ist, so meint der Antirealist, dass die PMI das Wunderargument aussticht. Denn während das Wunderargument letztlich auf Intuitionen basiert, deren Natur unklar und deren epistemischer Status umstritten ist, geht die PMI von harten Fakten der Wissenschaftsgeschichte aus, von konkreten Gegenbeispielen gegen den Schluss vom Erfolg auf die Wahrheit. Wie kann man ein Schlussprinzip besser untergraben als durch konkrete Gegenbeispiele? Folglich ist die PMI stärker als das Wunderargument. Selbst wenn man also anfänglich die Intuitionen des Realisten teilt, muss man, so der Antirealist, seine Meinung aufgrund der PMI ändern und das EIW-Prinzip aufgeben. Wenn man das tut, hat das natürlich Konsequenzen für unsere Einstellung zu unseren gegenwärtig besten Theorien: An ihre Wahrheit zu glauben ist dann nicht mehr rational rechtfertigbar.

Ich kann nun das Ziel dieses Aufsatzes genauer formulieren. Ich möchte den Realismus gegen den Angriff durch die PMI verteidigen. Dieses Ziel ist bescheiden, denn ich werde nicht in die Realismusdebatte, wie sie oben präsentiert wurde, eingreifen, sondern einfach die Perspektive des Realisten voraussetzen, d.h. den Realismus nur relativ zu seinen eigenen Intuitionen und bestätigungstheoretischen Ansichten verteidigen und die Intuitionen und bestätigungstheoretischen Ansichten von Antirealisten außen vorlassen. Ich werde insbesondere nicht versuchen, ein Argument gegen den Antirealisten zu konstruieren (obgleich die Aussichten hierfür meines Erachtens sehr gut sind, siehe Fahrbach (2010b) für erste Schritte in diese Richtung). Mein Ziel ist also rein defensiv. Ich möchte zeigen, dass das EIW-Prinzip nur geringfügig modifiziert werden muss, um es vor den Gegenbeispielen zu schützen und mit der Wissenschaftsgeschichte kompatibel zu machen.

Der Begriff des Erfolgs

Für meine Verteidigung des Realismus muss ich den Begriff des Erfolges genauer definieren. Meine Definition soll zwei Desiderata erfüllen. Erstens bleibt dieser Begriff, wie allseitig beklagt wird, in der Literatur der Realismusdebatte meistens ziemlich unbestimmt. Deswegen möchte ich den Begriff zumindest ein Stück weit präzisier machen. Das zweite Desiderat ist, dass meine Definition für die meisten Realisten akzeptabel sein soll.[9] Das erreiche ich, indem ich zum einen die Präzisierung der Definition nicht besonders weit treibe, sondern immer

8 Um sich in der Realismusdebatte ein Urteil zu bilden, muss man natürlich am Ende *alle* Argumente gegeneinander abwägen, aber ich denke, es ist lehrreich, die Abwägung an dieser Stelle der Debatte mit diesen beiden zentralen Argumenten zu machen.

9 Ob sie hingegen für Antirealisten akzeptabel ist, ist unwichtig, denn für meine Verteidigung des Realismus setze ich ja die Perspektive des Realisten voraus.

noch recht allgemein halte, und zum anderen in meiner Definition wohlbekannten Ideen des Testens und Bestätigens von Theorien verwende, die, wie ich hoffe, unter Realisten weitgehend unkontrovers sind.

Betrachten wir also ganz allgemein, wie Theorien durch die Beobachtung getestet werden. Um einen Test einer Theorie durchzuführen, leiten Wissenschaftler aus der Theorie eine beobachtbare Konsequenz her. Nennen wir jede beobachtbare Konsequenz einer Theorie, die Wissenschaftler aus ihr herleiten, eine „Vorhersage der Theorie". Wissenschaftler sammeln zudem Beobachtungen. Ein Test der Theorie besteht dann in einem Vergleich zwischen einer Vorhersage und einer Beobachtung. Wenn Vorhersage und Beobachtung übereinstimmen, hat die Theorie den Test bestanden und gewinnt ein gewisses Maß an Erfolg. Wenn die Theorie hinreichend viele Tests besteht, dann gilt sie als *erfolgreich* (wobei wir offen lassen können, was „hinreichend viele" genau bedeutet).

Wenn Vorhersage und Beobachtung nicht übereinstimmen, hat die Theorie den Test nicht bestanden. Wir können dann sagen, dass der fehlgeschlagene Test für die Theorie eine Anomalie darstellt. Wenn die Anomalie „wesentlich" ist, oder die Anomalien sich akkumulieren, ist die Theorie widerlegt.[10] Sie ist dann natürlich nicht erfolgreich. Die Wissenschaftler müssen dann nach anderen Theorien Ausschau halten, und ein Theoriewechsel mag vonstatten gehen. Solange eine Theorie nicht an „wesentlichen" Anomalien leidet und die Anomalien sich nicht akkumulieren, gilt sie nicht als widerlegt.

Betrachten wir nun den Begriff des *Erfolgsgrades*. Der Erfolgsgrad einer Theorie zu einer bestimmten Zeit wird durch die Zahl, Vielfalt und Strenge der Tests bestimmt, die die Theorie bis zu diesem Zeitpunkt bestanden hat. Wenn eine Theorie mit der Zeit mehr Tests, vielfältigere Tests oder strengere Tests besteht, dann steigt ihr Erfolgsgrad. Das ist alles, was wir zur Definition des Begriffes des Erfolgsgrades benötigen. Aus der Definition folgt, dass die gleiche Theorie zu verschiedenen Zeiten verschiedene Erfolgsgrade genießen kann. Ferner können sich zum gleichen Zeitpunkt verschiedene Theorien hinsichtlich des Erfolgsgrades unterscheiden. Später werden wir sehen, dass diese Unterschiede sehr groß sein können.

Ich verwende die Begriffe „Test" und „Vorhersage" in einem sehr weiten Sinn. Ein Test einer Theorie liegt immer dann vor, wenn die Theorie irgendwie mit der Erfahrung in Kontakt kommt und bestätigt oder entkräftet wird. Insbesondere liegt auch dann ein Test vor, wenn die Bestätigung oder Entkräftung durch den Test nur schwach ist. Dann ist auch die Änderung im Erfolgsgrad der Theorie nur gering. Ebenso verwende ich die Bezeichnung „Vorhersage" in einem weiten Sinn. Sie bezeichnet jede beobachtbare Konsequenz einer Theorie, die Wissenschaftler aus der Theorie hergeleitet haben und zum Testen der Theorie verwenden können. So zählen Konsequenzen einer Theorie auch dann als

10 Vergleiche Paul Hoyningen-Huene (1993, Kap. 7)

Vorhersagen, wenn sie keine *neuen Arten* von Phänomenen vorhersagen, d.h. nicht „neuartige Vorhersagen“ („novel predictions“) sind, sondern lediglich schon bekannte Phänomene vorhersagen oder sogar bei der Konstruktion der Theorie verwendet wurden. Die Gründe für diese weiten Verwendungsweisen der Begriffe „Test“ und „Vorhersage“ wird später deutlich werden.

Es versteht sich von selbst, dass die hier vorgestellten Erläuterungen und Definitionen zum den Begriffen des Testens und des Erfolgs von Theorien immer noch recht allgemein sind. Aber, wie gesagt, ein gewisses Maß an Allgemeinheit ist erwünscht, denn der Begriff des Erfolgs soll mit möglichst vielen Versionen des Realismus vereinbar sein.

Das modifizierte EIW-Prinzip

Wenn wir die Wissenschaftsgeschichte betrachten, dann ist die Behauptung sehr plausibel, dass im Laufe der Wissenschaftsgeschichte die Erfolgsgrade der jeweils besten Theorien im Großen und Ganzen kontinuierlich wuchsen. Sie wuchsen sowohl *bei* Theoriewechseln wie auch *zwischen* Theoriewechseln. *Bei* Theoriewechseln wuchsen sie, weil die jeweiligen Nachfolgetheorien die Erfolge der jeweiligen Vorgängertheorie im Allgemeinen übernommen haben und zusätzlich die Anomalien der Vorgängertheorie in Erfolge umgewandelt haben, denn eigens zur Lösung der Anomalien wurden sie im Allgemeinen konstruiert. *Zwischen* Theoriewechseln wuchsen die Erfolgsgrade, weil in der Wissenschaftsgeschichte die Menge und Qualität der Beobachtungen, die Präzision der Messapparate, die Genauigkeit der Vorhersagen, usw. die ganze Zeit über stetig wuchs.

Für den Realisten liegt dann die folgende Erwiderung auf die PMI nahe. Wie gerade festgestellt, wuchsen im Laufe der Wissenschaftsgeschichte die Erfolgsgrade der jeweils besten Theorien kontinuierlich an. Daraus folgt, dass unsere gegenwärtig besten Theorien höhere Erfolgsgerade genießen als alle widerlegten Theorien der Vergangenheit, etwa denen auf Laudans Liste. Wir können daher das EIW-Prinzip so modifizieren, dass es nur den Schluss von Erfolg auf die Wahrheit *für gegenwärtige Erfolgsgrade* erlaubt. Dieser Schluss wird dann nicht durch Gegenbeispiele aus der Wissenschaftsgeschichte unterminiert, und der Realismus ist vor dem Angriff durch die PMI geschützt.[11] Um das modifizierte EIW-Prinzip zu untermauern, kann der Realist wieder das Wunderargument heranziehen. Das Wunderargument seinerseits beruht am Ende nach wie vor auf den Intuitionen des Realisten, doch haben wir diese vorausgesetzt, denn es geht

11 Varianten dieses Arguments werden von Leplin (1997, p. 141), Stanford (2006), Psillos (1999) und anderen erwähnt oder andiskutiert.

uns ja nur um eine Verteidigung des Realismus aus einer realistischen Perspektive.

Diese Antwort auf die PMI ist zwar noch sehr unterentwickelt, aber sie weist in die richtige Richtung. Ich möchte sie nun weiter ausbauen.

Die zentrale Behauptung

Zur Ausarbeitung der eben präsentierten Idee für die Verteidigung des Realismus beginne ich mit der folgenden Beobachtung. Wenn wir Laudans Liste von erfolgreichen, aber widerlegten Theorien untersuchen, stellen wir fest, dass alle Einträge älter sind als 100 Jahre. Das gleiche gilt für praktisch alle Beispiele von widerlegten Theorien, die in der philosophischen Literatur diskutiert werden. Wie so oft in der Philosophie beschränken sich die meisten Diskussionen auf einige wenige Beispiele, in diesem Fall auf drei, nämlich die Phlogistontheorie, die kalorische Theorie der Wärme und die Äthertheorien des Lichts. Kyle Stanford (2006) hat sich die Mühe gemacht, drei weitere Fallbeispiele ausführlich darzustellen und zu diskutieren, doch stammen diese ebenfalls aus dem 19. Jahrhundert. Halten wir also fest, dass praktisch alle Beispiele von erfolgreichen, widerlegten Theorien älter sind als 80 Jahre.

Nennen wir Theorien, die vor mehr als 80 Jahren erfolgreich waren, „alte Theorien". Genau genommen ist diese Eigenschaft nicht eine Eigenschaft von Theorien selbst, sondern von *Theoriestadien*, d.h. Theorien *zu bestimmten Zeiten*. So ist die Evolutionstheorie im Jahre 1900 eine alte Theorie, heute dagegen nicht. Der Einfachheit halber wende ich diese Eigenschaft aber direkt auf Theorien an.

Ich möchte nun zeigen, dass der Altersunterschied zwischen widerlegten Theorien und gegenwärtig besten Theorien auch einen Unterschied in den Erfolgsgraden zwischen diesen beiden Arten von Theorien zur Konsequenz hat. Ich möchte die folgende Behauptung verteidigen: Die gegenwärtig besten Theorien genießen einen viel höheren Erfolgsgrad als alle alten Theorien. Dies ist die zentrale Behauptung dieses Aufsatzes. Nennen wir dabei den Erfolgsgrad der alten Theorien "moderat" und den Erfolgsgrad der gegenwärtig besten Theorien „hoch". Man beachte dann, dass die zentrale Behauptung gar nicht von den widerlegten Theorien der Vergangenheit handelt. Solange ich also Belege für sie präsentiere, geht es nur indirekt um die widerlegten Theorien.

Bevor wir fortfahren, möchte ich einige weitere Beispiele von Theorien, die zu unseren gegenwärtig besten gehören, vorlegen. Hier ist eine Liste von solchen Theorien. (Man beachte, dass der Realist immer nur die *annähernde* Wahrheit dieser Theorien behauptet.)

- das Periodensystem der Elemente[12]
- die Evolutionstheorie
- "Sterne sind wie unsere Sonne."
- die Erhaltung von Masse/Energie
- die kinetische Gastheorie
- die Keimtheorie von ansteckenden Krankheiten
- „Alle lebenden Organismen der Erde bestehen aus Zellen"
- $E = mc^2$
- usw.[13]

Das Ziel ist also nun, die Erfolgsgrade dieser Theorien mit den Erfolgsgraden von „alten" Theorien zu vergleichen. Um den Vergleich anzustellen, verwende ich eine Größe, von der ich zeigen möchte, dass sie mit Erfolgsgraden korreliert ist. Diese Größe ist die wissenschaftliche Arbeit, die Wissenschaftler in einer gewissen Zeit leisten.

Das exponentielle Wachstum der wissenschaftlichen Arbeit

Definieren wir den Begriff der wissenschaftlichen Arbeit. Unter wissenschaftlicher Arbeit sollen alle Tätigkeiten und Aktivitäten verstanden werden, denen Wissenschaftler den lieben langen Tag nachgehen, wenn sie Wissenschaft treiben. Dazu gehören das Sammeln von Beobachtungen, die Konstruktion neuer Theorien, das Durchführen von Experimenten, das Testen von Theorien, usw. Natürlich verbringen Wissenschaftler auch viel Zeit mit anderen Dingen, etwa mit der Lehre und administrativem Tätigkeiten, doch wird sich gleich zeigen, dass wir nicht zu entscheiden brauchen, ob wir diese Dinge als wissenschaftliche Arbeit zählen oder nicht.

Wie lässt sich die wissenschaftliche Arbeit messen? Die Annahme ist plausibel, dass die wissenschaftliche Arbeit, die Wissenschaftler in einem gewissen Zeitraum leisten, ganz grob proportional zu zwei Größen ist, der Zahl der Wissenschaftler, die in diesem Zeitraum lebt und arbeitet, und der Zahl der wissenschaftlichen Veröffentlichungen, die Wissenschaftler in diesem Zeitraum herausbringen.[14] Beide Größen haben im Verlauf der Wissenschaftsgeschichte ein

12 Wie eingangs bemerkt, verwende ich den Begriff einer Theorie in einem sehr weiten Sinn.

13 Meine Liste enthält keine Theorien aus der Grundlagenphysik (wie die Quantenmechanik), denn ich glaube, dass solche Theorien in der Realismusdebatte einen Spezialfall darstellen, der gesondert behandelt werden sollte.

14 Es gibt noch einige andere Methoden, die Menge der wissenschaftlichen Arbeit zu messen. Dazu gehören die finanziellen Aufwendungen von Staat und Industrie, die Zahl der Universitäten, die Zahl der Doktoranden, und andere mehr. Soweit Daten vorhanden, er-

exponentielles Wachstum durchlaufen. So wuchs die Zahl der wissenschaftlichen Veröffentlichungen im Laufe der letzten Jahrhunderte mit einer Verdopplungsrate von ungefähr 15 bis 20 Jahre. Die Zahl der Wissenschaftler erfuhr ein ähnliches Wachstum.[15] Das bedeutet, dass, wenn wir beispielsweise eine Verdopplungszeit von 20 Jahren annehmen, in den letzten 20 Jahren die Hälfte aller wissenschaftlichen Arbeit geleistet wurde und in der ganzen Zeit davor die andere Hälfte, und in den letzten 40 Jahren drei Viertel aller wissenschaftlichen Arbeit geleistet wurde, und in der Zeit davor ein Viertel (siehe Bild 1). Für unsere Zwecke bedeutet es, dass in den letzten 80 Jahren mindestens 90% aller wissenschaftlichen Arbeit geleistet wurde.

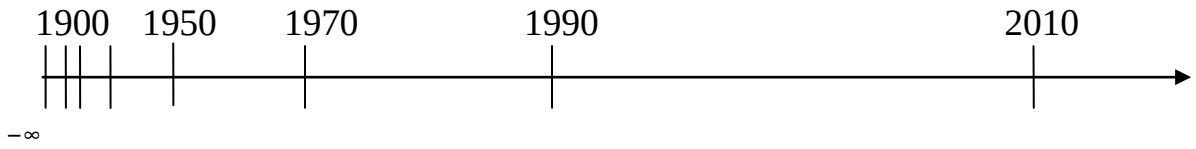

Bild 1. Eine Darstellung der Zeitachse für eine Verdopplungszeit der wissenschaftlichen Arbeit von 20 Jahren. Die Länge der Intervalle auf der x-Achse ist proportional zur Menge der in der entsprechenden Zeit geleisteten wissenschaftlichen Arbeit.

Weil das Wachstum so stark war, können wir alle Unbestimmtheiten in der Definition des Begriffes der wissenschaftlichen Arbeit vernachlässigen. Ebenso macht für unsere Zwecke der ständig wachsende Druck auf Wissenschaftler, immer mehr zu publizieren („publish or perish"), praktisch keinen Unterschied.

Wissenschaftliche Arbeit ist in folgender Weise mit Erfolgsgraden von Theorien verknüpft. Wir haben den Erfolgsgrad einer Theorie zu einem bestimmten Zeitpunkt als etwas definiert, dass von der Zahl, Vielfalt und Strenge der Tests abhängt, die die Theorie bis zu diesem Zeitpunkt bestanden hat. Dabei war der Begriff des Tests sehr weit gefasst: jede Gelegenheit, bei der die Theorie durch die Beobachtung bestätigt oder entkräftet werden kann, ist ein Test der Theorie. Um nun eine Theorie zu testen, müssen Wissenschaftlicher ganz allgemein zwei Arten von wissenschaftlicher Arbeit leisten, sie müssen erstens Beobachtungen sammeln und zweitens Vorhersagen aus den jeweiligen Theorien herleiten. Die erste Art von Aktivität ist offenkundig eine wichtige Art von wissenschaftlicher Arbeit. Für sie müssen Wissenschaftler oft ein erhebliches Maß an Zeit und Anstrengung – also von wissenschaftlicher Arbeit – aufbringen.[16] Aber auch die

---

geben diese Methoden die gleichen Resultate wie die beiden im Haupttext diskutierten Größen.

15 Für Daten, Details und Literaturhinweise siehe Fahrbach (2009a).

16 Beobachtungen und Daten sind nicht nur nötig, um sie mit Vorhersagen der jeweiligen Theorie zu vergleichen, sondern auch, um Vorhersagen aus der Theorie zu gewinnen. Aus der Theorie alleine lassen sich meistens keine empirischen Konsequenzen herleiten, viel-

Herleitung von Vorhersagen aus der Theorie kann, je nach Theorie, großen Aufwand seitens der Wissenschaftler verlangen. Insbesondere in den physikalischen Wissenschaften, aber zunehmend auch in anderen Wissenschaften, liegen die Theorien in mathematischer Form vor, d.h. sind mit Hilfe von Gleichungen, etwa Differentialgleichungen, formuliert. Diese müssen gelöst werden, um Vorhersagen aus der Theorie zu gewinnen. Sie zu lösen, kann jedoch einen erheblichen Aufwand für den Wissenschaftler bedeuten. Zum einen müssen Methoden entwickelt werden, mit denen die jeweilige Art von Gleichung gelöst werden kann, zum anderen müssen in den konkreten Anwendungen der Methoden konkrete numerische Werte berechnet werden (vgl. Humphreys 2004). All diese Tätigkeiten sind somit wichtige Arten von wissenschaftlicher Arbeit.

Es sind dann die folgenden Korrelationen plausibel (siehe Bild 2): (1) Ein Anstieg der wissenschaftlichen Arbeit führt zu einem Anstieg der Zahl und Qualität von Beobachtungen und Vorhersagen. (2) Dieser Anstieg führt zu einem Anstieg der Zahl und Qualität von Tests von Theorien. (3) Dieser Anstieg wiederum führt für diejenigen Theorien, die all diese Tests bestehen, zu einem Anstieg ihres Erfolgsgrades.

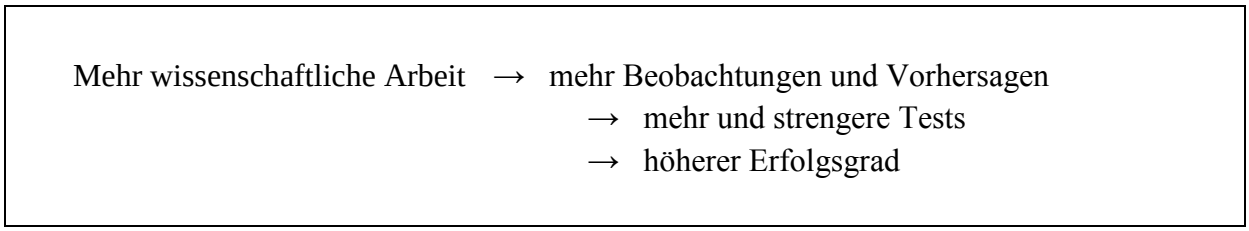

Bild 2

Das Argument für die zentrale Behauptung

Mit diesen Korrelationen können wir nun ein Argument formulieren, dass die zentrale Behauptung, wonach unsere gegenwärtig besten Theorien viel höhere Erfolgsgrade genießen als alle alten Theorien, stützen soll. Wie wir gerade festgestellt haben, ist bei weitem der größte Teil der wissenschaftlichen Arbeit in den letzten Jahrzehnten geleistet worden. Unsere gegenwärtig besten Theorien haben von dieser Arbeit profitiert und zwar über die drei eben erwähnten Korrelationen: All die wissenschaftliche Arbeit resultierte in einem starken Anwachsen der Qualität und Vielfalt von Beobachtungen und Vorhersagen; diese resultierte in einem starken Anwachsen der Qualität und Vielfalt von Tests von Theorien; dass unsere gegenwärtig besten Theorien in den letzten Jahrzehnten stabil waren bedeutet, dass sie all diese Tests bestanden haben; folglich erlebten sie

---

mehr benötigt man dazu (neben Hilfshypothesen) empirische Prämissen über die konkreten Testbedingungen, über Anfangs- und Randbedingungen, etc., und diese muss man aus Daten und Beobachtungen gewinnen.

einen riesigen Erfolgsschub; daher genießen sie heute einen „hohen“ Erfolgsgrad. Für alte Theorien gilt das alles nicht. Sie sind älter als 80 Jahre, konnten vom Anstieg der wissenschaftlichen Arbeit nicht profitieren und haben daher keinen Erfolgsschub erfahren. Folglich war ihr Erfolgsgrad nur „moderat“. Dies ist das Argument für die zentrale Behauptung.

Gegen dieses Argument kann man viele Einwände vorbringen, doch möchte ich aus Platzgründen nur einen diskutieren. Dieser Einwand richtet sich gegen die beiden Proportionalität (1) und (2). In vielen wissenschaftlichen Gebieten hat der Anstieg der wissenschaftlichen Arbeit ganz offensichtlich nicht zu einem entsprechenden Anstieg der Daten und Vorhersagen geführt. Zum Beispiel haben wir sehr wenige Daten über den Ursprung der Menschheit oder den Ursprung des Lebens auf der Erde, und aufgrund dieses Mangels an Daten haben wir auch keine guten Theorien über diese Dinge. Dies sind Gegenbeispiele gegen die Proportionalität (1).

Aber auch gegen die Proportionalität (2) gibt es viele Gegenbeispiele. So gibt es in den Wirtschaftswissenschaften heute eine Unmenge von empirischen Daten und von Computermodellen, die all die Rechenkapazität nutzen, die unsere besten Rechner heute zur Verfügung stellen; beides hat aber nicht zu besonders erfolgreichen ökonomischen Theorien geführt. Ähnliches gilt für eine ganze Menge anderer wissenschaftlicher Gebiete, etwa für große Teile der Krebsforschung oder große Teile der Erforschung des Gehirns. Ganz allgemein floriert die Wissenschaft heute wie nie zuvor, aber das zeigt gerade, dass es in all den wissenschaftlichen Disziplinen noch sehr viele offene Fragen gibt, für die wir entweder höchstens mehr oder wenige plausible Antworten haben, oder oft genug überhaupt keine Ahnung, wie Antworten aussehen könnten, und all das trotz des riesigen Anstiegs der wissenschaftlichen Arbeit. Es ist also nicht zu sehen, so der Einwand, dass die Schritte (1) bis (3) echte Korrelationen sind. Daraus folgt, dass nicht zu sehen ist, dass es einen systematischen Zusammenhang zwischen der wissenschaftlichen Arbeit und den Erfolgsgraden gibt; eher scheinen die beiden Größen ziemlich unabhängig voneinander zu sein.

In Reaktionen auf diesen Einwand gibt es eine kurze und eine lange Erwiderung. Ich stelle hier nur die kurze Erwiderung vor und verschiebe die lange auf eine andere Gelegenheit. Zunächst muss man zugeben, dass es viele Fälle gibt, in denen die Korrelationen (1) bis (3) nicht bestehen. Es ist sicher nicht plausibel, dass ein Anstieg der wissenschaftlichen Arbeit in einem wissenschaftlichen Gebiet *automatisch* einen Anstieg der Erfolgsgrade der Theorien in diesem Gebiet zur Folge hat. Auch ist ganz offensichtlich, dass es ganz allgemein gesprochen trotz des exponentiellen Anstiegs der wissenschaftlichen Arbeit sehr viele Lücken in unserem Wissen gibt. Der Zusammenhang zwischen wissenschaftlicher Arbeit und Erfolgsgraden der besten Theorien ist nicht so einfach und direkt, wie es vorderhand vielleicht erschien. Vielmehr ist er kompliziert, kontingent und abhängig von den Gegebenheiten des jeweiligen wissenschaftlichen

Gebietes. Indes wäre es – und das ist jetzt die Erwiderung – sehr unplausibel, wenn *keine* wissenschaftliche Theorie in *keinem* wissenschaftlichen Gebiet vom Anstieg der wissenschaftlichen Arbeit profitiert hätte. So groß wie das Wachstum war, wäre das extrem überraschend. Und es sind genau diejenigen Theorien, die am meisten von diesem Anstieg profitiert haben, die der Realist im Sinn hat, wenn er von unseren gegenwärtig besten Theorien spricht. Genau für diese Theorien soll die obige Liste Beispiele liefern. Dies war die kurze Erwiderung auf den Einwand. Weil die lange Erwiderung, wie gesagt, zu lange für diesen Artikel ist, werde ich jetzt nur noch ein Beispiel für die Argumentation vorstellen.

Das Periodensystem der Elemente

Den Zusammenhang zwischen wissenschaftlicher Arbeit und Erfolgsgraden möchte ich am Beispiel der wichtigsten Theorie der Chemie, des Periodensystems der Elemente, verdeutlichen. Wie die restliche Wissenschaft hat auch die Chemie ein exponentielles Wachstum durchlaufen. Im Laufe der letzten zwei Jahrhunderte wuchsen sowohl die Zahl der Chemiker als auch die Zahl ihrer Veröffentlichungen im großen und ganzen auf ähnlich exponentielle Weise wie in der restlichen Wissenschaft. Wie Jochen Schummer (1999, 92) bemerkt, hat es (bezogen auf das Jahr 1999) „während der letzten 15 Jahre mehr Veröffentlichungen in der Chemie gegeben als jemals zuvor. … Dieses Jahr werden 100 mal so viele Artikel veröffentlich wie im Jahre 1901, als van't Hoff den ersten Nobelpreis für Chemie erhielt." Nun lässt sich aber im Falle der Chemie das Wachstum noch an einer anderen Größe festmachen, nämlich an der Zahl der neu synthetisierte chemischen Substanzen. Für diese beobachten wir eine sehr stabiles exponentielles Wachstum mit einer Wachstumsrate von 5.5%, was einer Verdopplungszeit von 12,9 Jahren entspricht (1997a, p. 111).[17]

Wie führt nun das Wachstum der Zahl der neu synthetisierte chemischen Substanzen zu einem Wachstum des Erfolgsgrades des Periodensystems der Elemente? Für jede chemische Substanz, die von Chemikern neu synthetisiert wird, macht das Periodensystem Vorhersagen. Es sagt beispielsweise voraus, dass die neue Substanz aus den chemischen Elementen des Periodensystems bestehen muss und eine mit Hilfe der chemischen Elemente beschreibbare Molekularstruktur aufweisen muss. Wenn eine solche Vorhersage eintrifft, dann steigt der Erfolgsgrades des Periodensystems. Dabei brauchen die Vorhersagen nicht besonders spezifisch zu sein, sondern können relativ allgemein sein, wie

17 Die Verdopplungszeit für neue chemische Substanzen von 12,9 Jahren ist kürzer als die Verdopplungszeit für die Zahl der Chemiker und ihrer Veröffentlichungen. Hierfür bietet Schummer eine Erklärung. Er zeigt, dass die Zahl der neuen chemischen Substanzen *pro Artikel* spürbar wuchs, das heißt, die Produktivität der Chemiker wuchs im Laufe der Jahrzehnte. (Schummer 1997a, p. 118)

die beiden Beispiele eben, in welchem Falle sie nur zu einem kleinen Anstieg des Erfolgsgrades des Periodensystems führen. Jedoch haben wir gerade gesehen, dass die Zahl der neuen chemischen Substanzen enorm gewachsen ist. Sie liegt heute bei 50 Millionen.[18] Folglich ist auch die Gesamtzahl der entsprechenden Vorhersagen enorm gewachsen. Praktisch alle diese Vorhersagen haben sich als korrekt erwiesen, denn sonst hätte das Periodensystem irgendwann aufgegeben werden müssen. Zusammengenommen haben sie also zu einem riesigen Erfolgsschub für das Periodensystem geführt. Folglich genießt das Periodensystem als eine unserer gegenwärtig besten Theorien einen viel höheren Erfolgsgrad als zum Beispiel das Periodensystem als alte Theorie vor mehr als 80 Jahren, und das ist gerade die zentrale Behauptung für den Fall des Periodensystems.

Verteidigung des Realismus

Zwar werden in der zentralen Behauptung unserer gegenwärtig besten Theorien nur *mit den „alten" Theorien* verglichen, aber wir können aus ihr eine ähnliche Behauptung gewinnen, in der unsere gegenwärtig besten Theorien *mit den widerlegten Theorien* vergleichen werden. Auf der einen Seite haben wir oben gesehen, dass praktisch alle widerlegten Theorien, die in der philosophischen Literatur diskutiert werden, vor mehr als 80 Jahren widerlegt wurden und somit zu den „alten" Theorien gehören. Also können wir die zentrale Behauptung auf diese Theorien anwenden und schließen, dass ihre Erfolgsgrade viel geringer sind als diejenigen unserer gegenwärtig besten Theorien. Wir haben einen solchen Erfolgsgrad „moderat" genannt.

Es stellt sich dann die Frage, ob es widerlegte Theorien gibt, die einen „hohen" Erfolgsgrad irgendwann in den letzten 80 Jahren genossen haben. Ich habe Ausschau nach solchen Theorien gehalten und mögliche Kandidaten gesammelt. Das Ergebnis ist, dass keiner meiner Kandidaten ein überzeugendes Beispiel für eine widerlegte Theorie mit "hohem" Erfolgsgrad ist. Vielmehr genossen sie alle höchstens „moderaten" Erfolg. Ich könnte nun meine Kandidaten präsentieren und nacheinander abhandeln, doch wäre das ein zweifelhaftes Unterfangen: Eine Liste von möglichen Beispielen zu präsentieren, nur um dann zu zeigen, dass sie doch keine Beispiele sind. Stattdessen überlasse ich es dem Antirealisten, Beispiele von widerlegten Theorien mit hohem Erfolgsgrad zu präsentieren. Zum Beispiel mag er Lehrbücher von wissenschaftlichen Disziplinen wie der Chemie, Biologie oder Astronomie der letzten Jahrzehnte, z.B. der 60er, 70er oder 80er Jahre, durchsuchen, um nach solchen Beispielen zu fanden. Solange er keine vorweisen kann, werde ich behaupten, dass es keine gibt.

18 http://www.cas.org/newsevents/connections/heterocycle.html

Wir gelangen mithin zu dem Schluss, dass es weder in der weiter zurückliegenden noch in der jüngsten Vergangenheit Fälle von widerlegten Theorien gibt, die „hohen“ Erfolg genossen haben. Daraus folgt zusammen mit der zentralen Behauptung, dass alle unsere gegenwärtig besten Theorien einen viel höheren Erfolgsgrad (nämlich „hohen“ Erfolgsgrad) genießen als alle widerlegten Theorien (welche nur „moderate“ Erfolgsgrade genossen).

Wie nun der Realismus gerettet werden kann, haben wir bereits gesehen: Der Realist modifiziert das EIW-Prinzip zu einem Schluss von „hohem“ Erfolgsgrad auf Wahrheit. Dieses modifizierte EIW-Prinzip ist nicht mehr durch Gegenbeispiele bedroht. Der Realist kann dann den Realismus modifizieren, sodass sie im Gutheißen des modifizierten EIW-Prinzips besteht. Der modifizierte Realismus ist dann vor dem Angriff durch die PMI sicher und somit eine Version des Realismus, die mit der Wissenschaftsgeschichte verträglich ist. Der modifizierte Realismus kann am Ende wieder mit dem Wunderargument gestützt werden.

Es versteht sich von selbst, dass die hier vorgestellt Argumentation noch große Lücken aufweist. Sie ist nicht mehr als eine Skizze gemeint, die durch weitere Daten und Argumente verbessert werden muss. Dennoch meine ich, dass sie schon eine gewisse argumentative Kraft hat. Es ist ja der Antirealist, der mit der pessimistischen Metainduktion ein Argument gegen den Realismus präsentieren möchte und es daher zum Laufen bringen muss. Die Beweislast liegt ganz bei ihm. Der Realist kann schon dadurch etwas gegen die pessimistischen Metainduktion ausrichten, dass er ernsthafte Zweifel an ihr weckt. Genau das zu erreichen, war das Ziel meiner Argumentation.

Literaturverzeichnis

Boyd, R.: "On the Current Status of the Issue of Scientific Realism." Erkenntnis 19, 1983. S. 45-90

Chakravartty, Anjan: A Metaphysics for Scientific Realism: Knowing the Unobservable. Cambridge University Press, Cambridge, 2007

Devitt, M.: “Scientific Realism”. In: *Frank Jackson and Michael Smith (Hrsg.):* The Oxford Handbook of Contemporary Analytic Philosophy. Oxford University Press, Oxford, 2005. S. 767-91

Fahrbach, L.: “The Pessimistic Meta-Induction and the Exponential Growth of Science”. In: *Alexander Hieke and Hannes Leitgeb* (*Hrsg.):* Reduction and Elimination in Philosophy and the Sciences. Proceedings of the 31th International Wittgenstein Symposium, 2009a

Fahrbach, Ludwig: "How the Growth of Science Ended Theory Change". Synthese, 2009b. DOI 10.1007/s11229-009-9602-0

Holt, Jim: "Madness About a Method." New York Times, December 11, 2005
Horwich, Paul: Truth, Oxford University Press, Oxford, 1998
Hoyningen-Huene, P.: Reconstructing Scientific Revolutions: Thomas S. Kuhn's Philosophy of Science. University of Chicago Press, Chicago, 1993
Humphreys, P.: Extending Ourselves: Computational Science, Empiricism, and Scientific Method. Oxford University Press, Oxford, 2004

Kitcher, P.: "Real realism: The Galilean Strategy". The Philosophical Review, 110 (2), 2001. S. 151–197
Kukla, A.: Studies in Scientific Realism. Oxford University Press, Oxford, 1998

Ladyman, James and Don Ross: Every Thing Must Go: Metaphysics Naturalized. Oxford University Press, Oxford, 2008
Laudan, L.: "A Refutation of Convergent Realism", Philosophy of Science, 48 March, 1981. S. 19-49
Leplin, J.: A Novel Defence of Scientific Realism. Oxford University Press, Oxford, 1997
Lewis, David: Philosophical Papers Volume I. Oxford University Press, New York, 1983
Lewis, P.: "Why The Pessimistic Induction Is A Fallacy", Synthese, 129, 2001. S. 371-380

Musgrave, A.: Essays on realism and rationalism. Rodopi, Amsterdam & Atlanta, 1999

Niiniluoto, Ilkka: Critical Scientific Realism, Oxford University Press, Oxford, 1999

Psillos, P.: Scientific Realism: How Science Tracks Truth. Routledge, New York and London, 1999
Putnam, Hilary: Philosophical Papers I. Cambridge University Press, Cambridge, 1975

Saatsi, J.. "On the Pessimistic Induction and Two Fallacies". In: Proceedings Philosophy of Science Assoc. 19th Biennial Meeting, PSA2004: PSA 2004 Contributed Papers, 2004
Schummer, Joachim: "Scientometric Studies on Chemistry I: The Exponential Growth of Chemical Substances, 1800-1995". Scientometrics, 39, 1997a. S. 107-123

Schummer, Joachim: "Scientometric Studies on Chemistry II: Aims and Methods of Producing New Chemical Substances". Scientometrics, 39, 1997b. S. 125-140

Schummer, Joachim: "Coping with the Growth of Chemical Knowledge". Educación Química, Vol. 10 No. 2, 1999. S. 92-101

Schurz, Gerhard: „Patterns of Abduction". *Synthese*, 164, 2008. S. 201-234

Smart, J. J. C.: Philosophy and Scientific Realism. Routledge, London, 1963

Stanford, P. Kyle: Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford University Press, 2006

Turner, Derek: Making Prehistory: Historical Science and the Scientific Realism Debate, Cambridge University Press, 2007

Persistenz in Minkowskis Raum-Zeit

Cord Friebe
cfriebe@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Bonn

Abstract/Zusammenfassung

Under the eternalist hypothesis that objects or events exist temporally but independently of being present two different views of persistence are on the market: Persisting objects *endure* if they are multiply located in (space-)time, and persisting objects *perdure* if they are singly located by having numerically different temporal parts. Recently several authors have argued that the special theory of relativity (SR) favours perdurantism over its endurantist rival. In this paper, I intend to show that the purported arguments are in fact only those against endurantism and that with a similar strategy we should argue against perdurantism as well: Enduring *and* perduring entities are *both* in conflict with SR, which undermines the eternalist hypothesis.

Auf Basis der statischen B-Theorie temporaler Existenz unterscheidet man zwei Weisen des Zeitüberdauerns: Persistierende Objekte *endurieren*, wenn sie in der (Raum-)Zeit mehrfach instantiiert sind, hingegen *perdurieren* sie, wenn sie nur einfach vorkommen und (instantane) zeitliche Teile haben. In jüngster Zeit ist dafür argumentiert worden, dass die spezielle Relativitätstheorie (SRT) Gründe *a posteriori* zugunsten der perdurantistischen Variante liefere. Ich werde dagegen zeigen, dass die vorgebrachten Argumente zunächst nur gegen den Endurantismus sprechen und dass man eigentlich mit derselben Strategie auch gegen den Perdurantismus argumentieren könnte. Beide Weisen des Zeitüberdauerns scheinen sonach im Konflikt mit der SRT zu stehen, was die eternalistische Voraussetzung in Frage stellt.

In der analytischen Philosophie der Zeit unterscheidet man zwei Typen von Theorien – die A-Theorien, wonach der Bezug zu einer transitorischen Gegenwart in irgendeiner Weise objektiv ist, und die B-Theorien, wonach lediglich die Früher-Später-Relation als objektiv gilt. Des Weiteren lassen sich zwei Wirklichkeitsauffassungen unterscheiden – eine dynamische, wonach das, was (*simpliciter*) existiert, mit der Zeit variiert, und eine statische, derzufolge dies nicht der Fall ist. Von den vier möglichen Zeit-Wirklichkeits-Auffassungen ist anscheinend nur die *statische B-Theorie*, der Eternalismus, mit der Raum-Zeit der speziellen Relativitätstheorie vereinbar. Sie lässt sich mit Blick auf Minkowskis Raum-Zeit wie folgt reformulieren:

Def.: Tempuslose Existenz (*simpliciter*): Ein Objekt x oder ein Ereignis e existiert in einem *nicht-perspektivischen* Sinn von „Existenz", wenn es an diesem oder jenem, also an irgendeinem Raumzeit-Punkt p lokalisiert ist – tempuslos, d.h. unabhängig davon, ob p als gegenwärtig gelten kann.[1]

Def.: Tempuslose Existenz (zu einer bestimmten Zeit) x (bzw. e) existiert *perspektivisch*, nämlich tempuslos, aber zeitpunktrelativ, wenn es an einer bestimmten Koordinatenzeit (*frame-time*) t^F lokalisiert ist.[2]

Auf Basis dieser Theorie temporaler Existenz unterscheidet man dann zwei Auffassungen des *Zeitüberdauerns*: Persistierende Objekte *endurieren*, wenn sie in der Raum-Zeit universalienanalog mehrfach instantiiert sind, wenn sie also zu mehr als einer Koordinatenzeit „wholly present" sind. Persistierende Objekte *perdurieren*, wenn sie hingegen nur einfach vorkommen und (instantane) zeitliche Teile haben. Vor-relativistisch analysiert man den tempuslosen Satz (bzw. die Schar tempusloser Aussagen) „a ist F zur Zeit t"[3] auf drei verschiedene Weisen – nämlich entweder perdurantistisch als „a(t) ist F", wonach die Eigenschaft F einem *zeitlichen Teil* von a zugesprochen wird, oder aber endurantistisch, wobei es dann wiederum zwei Möglichkeiten gibt, nämlich: *indexikalistisch* als „a ist F(t)", wonach F eine zeitpunktindizierte Eigenschaft ist, oder *adverbialistisch* als „a ist t-weise F", wonach a F auf eine *spezifische* Weise hat. Diese drei Möglichkeiten überträgt man auf den relativistischen Kontext, indem man die klassisch-absoluten Zeitpunkte durch relative Zeiten, also durch den 2-Parameter-Index t_i^F, ersetzt. Das Zeitpunktindizieren der Teile, der Eigenschaften oder des Habens von Eigenschaften geschieht nun also bezugssystemabhängig. Doch es geschieht *gleichermaßen* bezüglich der Teile, der Eigenschaften oder des Habens von Eigenschaften, so dass man annehmen sollte, dass, wenn Perdurantismus, Indexikalismus und Adverbialismus vor-relativistisch gleichberechtigte Analyse des Zeitüberdauerns und der zeitlichen Veränderung persistierender Objekte bilden, sie dies auch im relativistischen Kontext bleiben.

In jüngster Zeit aber ist – auf Basis dieser Voraussetzungen[4] – von einigen Autoren (Balashov, vgl. Abschnitt 1; Gilmore, vgl. Abschnitt 2) dafür argumentiert worden, dass die SRT Argumente *a posteriori* zugunsten der *perdurantistischen* Auffassung liefern würde. Ich hingegen werde zu zeigen

1 Gegenüber der traditionellen eternalistischen *Simpliciter*-Existenz (‚at some time or other') wird relativistisch der Zeitpunkt durch einen Raumzeit-Punkt ersetzt.

2 Der klassische Zeitpunkt-Bezug wird relativistisch durch den Bezug auf eine bestimmte Koordinatenzeit ersetzt.

3 Es ist immer mitzudenken, dass a zu einer anderen Zeit G sein soll, also „a ist F zur Zeit t1" im Kontrast zu „a ist G zur Zeit t2".

4 Also erstens auf Basis der eternalistischen Auffassung temporaler Existenz und zweitens auf Basis der Annahme, dass klassische Zeitpunkte problemlos durch relative Koordinatenzeiten ersetzt werden können (was ja nicht selbstverständlich ist).

versuchen, dass die vorgebrachten Argumente sich eigentlich nur *gegen* den Endurantismus richten und dass sie, recht verstanden, ebenso gegen den Perdurantismus sprechen. Endurierende und perdurierende Objekte sind, so die These, *gleichermaßen unverträglich* mit der geometrischen Struktur von Minkowskis Raum-Zeit – was vielleicht gegen die eternalistische Voraussetzung spricht.

Zur Veranschaulichung des Problems betrachten wir ein ausgedehntes Gebiet der Raum-Zeit, einen vierdimensionalen Raumzeit-Wurm: Längs einer parallelen Menge von Gleichzeitigkeits(hyper)ebenen gebe es einen Zeitpunkt, zu dem das Objekt sich von räumlich ganz rot zu räumlich ganz grün instantan verändere. Dann gibt es eine weitere Menge ebenso paralleler, aber gekippter Gleichzeitigkeitsebenen, in Bezug auf welche die zeitliche Veränderung von rot nach grün *kontinuierlich* verläuft und bei der es bestimmte Elemente gibt, auf denen das Objekt räumlich *teilweise* rot und *teilweise* grün ist. Unstrittig ist daher, dass das, was in einem Bezugssystem ein *rein zeitliches Nacheinander* ist, in einem anderen, gleichberechtigten, (zum Teil) ein *räumliches Nebeneinander* ist. Zeitliche Veränderung und räumliche Variation scheinen sich einander anzunähern.

Schon allein diese (unstrittige) Angleichung von zeitlicher Veränderung und räumlicher Variation scheint den Perdurantismus nahezulegen, da diese Auffassung des Zeittüberdauerns zumindest insofern *raumanalog* ist, als Objekte den Raum auch derart erfüllen, dass sie dort (als Partikularien) nur einfach vorkommen und Teile haben. Darauf aber stützen sich die vorgebrachten Argumente eigentlich nicht, wäre man doch sonst an dieser Stelle schon fertig. Vielmehr legt dieses Bild den Perdurantismus nur intuitiv nahe, während mit den zu diskutierenden Argumenten der Anspruch verbunden ist, über diese Intuition hinauszugelangen.

1. Das endurantistische Erklärungsdefizit (Balashov, 2000)

Das erste Argument stammt von Yuri Balashov, der den Perdurantismus explanatorisch im Vorteil sieht: Im Gegensatz zu den Endurantisten könne ein Perdurantist nämlich erklären, *wie* die dreidimensionalen bezugssystemabhängigen räumlichen Gestalten der Objekte das invariante vierdimensionale Gebilde „erfüllen“. Da ein Endurantist bestreite, dass es ein invariantes vierdimensionales *Objekt* gibt[5], und vielmehr annehme, dass es bloß ein mehrfach lokalisiertes dreidimensionales Objekt mit einer Vielheit von räumlichen Formen gibt, sei es aus seiner Perspektive unerklärlich, *wieso* diese Formen „arrange themselves in-

5 Das Vierdimensionale ist danach vielmehr nur die *Geschichte* eines dreidimensionalen Objekts.

to a ‚nice' volume […] without ‚corrugation' and ‚dents'" (Balashov, 2000, 334). Nach perdurantistischer Auffassung hingegen gebe es eine „pre-existing ontological entity, the 4D perduring pole" (Balashov, 2000, 333)[6], die „objectively *stand[s] behind* all [its] 3D parts" (Balashov, 2000, 334). Im Gegensatz zum Endurantisten sei der Perdurantist nicht konfrontiert mit „'separate and loose' 3D shapes" (ebd.), die sich, mysteriöserweise, zu einer vierdimensionalen Einheit zusammenfügen. Für die Perdurantisten sei es „the invariant 4D shape of this volume that generates the whole multitude of 3D shapes" (Balashov, 2000, 333), während: „In the end, the endurantist must regard the infinite variety of perspectival relations as brute facts, with no unifying ground behind them" (Balashov, 2000, 338). Diese Behauptungen sind jedoch gleich in mehrerer Hinsicht fragwürdig.

Bevor sie kritisiert werden sollen, sei Balashov jedoch zunächst gegen die diesbezüglichen Einwände von Miller (2004) und Gibson/Pooley (2006) verteidigt: Sie argumentieren, dass ein Endurantist tatsächlich eine Erklärung für dieses Faktum habe, wenn man berücksichtige, dass es „various causal facts about an enduring object O at time t" gebe, „that make it the case that O will exist at t*" (Miller, 2004, 367). Die SRT erlaube es, zusammen mit anderen physikalischen Gesetzen, *vorherzusagen*, wo und wann Objekte, die in einem gegebenen Raumzeit-Gebiet „wholly present" sind, in anderen solchen Gebieten existieren (werden). Nach Gibson/Pooley (vgl. 2006, Abschnitt 6) kann man jedes räumlich ausgedehnte Objekt als aus punktuellen Teilchen zusammengesetzt denken, die exakt berechenbaren Weltlinien folgen, welche sich zu einem zusammenhängenden Ganzen fügten. Doch bestreitet Balashov meines Erachtens nicht, dass eine solche *physikalische* Erklärung gegeben werden könne: Empirische Erklärungen sind natürlich unabhängig von der ontologischen Weise des Zeitüberdauerns. Was Balashov intendiert, ist vielmehr, dass von einem *ontologischen* Standpunkt eben diese physikalische Erklärung für die Endurantisten ein mysteriöses nacktes Faktum sei. Denn ihnen zufolge gebe es eben keine „invariant 4D shape", keine „pre-existing entity", sondern lediglich dreidimensionale Formen, die, aus ontologischer Perspektive, „separate and loose" seien, obwohl verbunden durch physikalische Gesetze. Doch ist gerade dieses ontologische Bild irreführend.

Denn zunächst ist nicht zu sehen, warum aus perdurantistischer Sicht der vierdimensionale Raumzeit-Wurm gegenüber seinen dreidimensionalen zeitlichen Teilen ontologisch primär sein soll. Mark Heller etwa betont mehrfach, dass die Teile und das aus ihnen zusammengesetzte Ganze *gleichberechtigt* seien.[7] Ferner gibt es neben dem sogenannten *worm view* als perdurantistische Variante noch den von Sider sogar preferierten *stage view* (vgl. Sider, 2001, Ab-

6 „Pre-existing" ist hier sicherlich nicht zeitlich gemeint. Wie aber dann?

7 Beispielsweise: „[The temporal parts] are ontologically no more or less basic than the wholes that they compose" (Heller, 1992, 696).

schnitt 5.8), wonach die ursprünglichen Referenz-Gegenstände, auf die sich Namen und Begriffe beziehen, die dreidimensionalen Objekte der Wahrnehmung, also die instantanen zeitlichen Teile sind. Laut Sider, und insbesondere auch gemäß dem „worm-view", hat jedes vierdimensionale Objekt nur Eigenschaften *at some time or other*. So ist es beispielsweise bis zu einer Zeit t ständig und (räumlich) vollständig rot und ab dann ständig und (räumlich) vollständig blau, indem seine entsprechenden zeitlichen Teile *simpliciter* rot (bzw. blau) sind. Daraus folgt nicht, dass das perdurierende Objekt etwa noch Eigenschaften in dem Sinne *simpliciter* hätte, dass es im Ganzen ‚bunt' wäre. Zur Unterscheidung von Perdurantismus und Endurantismus ist es nämlich gänzlich irrelevant, wenn man dem perdurierenden Objekt eine „invariant 4D shape" zuschriebe, wie Balashov dies tut. Denn seine *auf Zeitpunkte bezogenen* Eigenschaften sind ja, weil seinen zeitlichen Teilen inhärierend, weder zeitpunktindiziert, noch hätte es sie auf zeitpunktindizierte Weise. Umgekehrt will nicht einleuchten, warum aus endurantistischer Sicht die räumlichen Gestalten „separate and loose" erscheinen. Da es eine „invariant 4D shape" in keinem Falle zu geben braucht, könnte man stattdessen mit gutem Recht behaupten, dass gerade den *Perdurantisten* die räumlichen Gestalten als „separate and loose" erscheinen, repräsentieren sie ihnen zufolge doch nicht nur qualitativ Verschiedenes, sondern auch *numerisch* (nämlich numerisch verschiedene zeitliche Teile). Ein Endurantist braucht dagegen keine Vielheit zu reduzieren, wie Balashov (vgl. 2000, 338) vermeint, da die ihm gemäß *bloß qualitativ* verschiedenen dreidimensionalen Objekte von vornherein *numerisch identisch* sind, weil mehrfach instantiiert (*multiply located*). Vom ontologischen Standpunkt droht die *Identität* vielmehr dem Perdurantisten verloren zu gehen, nicht dem Endurantisten.

Balashov glaubt also zum einen, d.h. in Bezug auf den Perdurantisten, dass der vierdimensionale Wurm seinen dreidimensionalen Teilen *auf dieselbe Weise* zugrunde läge, wie dreidimensionale Objekte im Raum ihren perspektivisch-zweidimensionalen Erscheinungen zugrunde liegen (vgl. Balashov, 2000, 334). Ganz ernsthaft führt er an entscheidender Stelle (vgl. Balashov, 2000, 332) die altbekannte Analogie von der Straße an, die man vertikal oder schräg schneiden kann, so dass folglich die Weglänge ihrer Überschreitung mit dem Winkel variiert. Es ist klar, dass kein Problem bestünde, wenn erlaubt wäre, den räumlichen Vergleich derart wörtlich zu nehmen.[8] Das aber ist eben nicht erlaubt: Der Raumzeit-Wurm ist kein *Raum*-Wurm, weil *allein* aus der ‚Perspektive' vom Nirgendwo und Nirgendwann ein wesentliches Merkmal der Zeit verloren geht, so dass *immer auch* die Perspektive von innerhalb der Raum-Zeit eingenommen werden muss.[9] Wie in jeder Theorie, die noch eine *über Zeit* sein will, gibt es

8 „One should simply take the diagramm in Figure 7 [eine räumliche Darstellung] literally" (Balashov, 2000, 336).

9 Die zugrunde gelegte eternalistische Auffassung hat ja sowohl einen nicht-perspektivischen als auch einen perspektivischen Sinn von „Existenz" (siehe oben).

auch im Eternalismus und folglich auch im eternalistischen Perdurantismus zumindest dem Anspruch nach zeitliche Sukzession – wenn auch kein ‚bewegtes Jetzt' – und also ein zeitliches Werden, das von räumlicher Variation unterschieden ist. Jedenfalls ist die *no-change-objection* in den Augen *aller* Autoren ein ernstzunehmender Einwand, gegen den zu argumentieren ist, was gerade nicht geschieht, wenn man „resort to the spatial analogy once again" (Balashov, 2000, 337). Ebenso also, wie schon im klassischen Kontext auch ein Perdurantist vor der Frage stand, wie denn bei sukzessive verschiedenen räumlichen Gestalten[10] noch die Identität des sich verändernden Objekts zu sichern sei, so steht er vor dieser Frage auch im Kontext der SRT. Und hier erschwert sich die Antwort – für ihn wie für einen Endurantisten –, weil die räumliche Gestalt nicht nur von Zeit zu Zeit wechseln kann, sondern im Allgemeinen auch von Bezugssystem zu Bezugssystem verschieden ist. Zwei Mengen von jeweils parallelen Gleichzeitigkeitsschnitten, die zueinander gekippt sind und daher sich schneiden, repräsentieren gleichberechtigt zwei Weisen des Nacheinanders zeitlicher Teile. Wie aber hat man sich denn diese verschiedenen Sukzessionen vorzustellen, die gewissermaßen ‚zugleich' verschiedene Richtungen haben? Diese Frage muss auch ein Perdurantist beantworten. In der *Darstellung* – also in der räumlichen Analogie – ist das alles schön anschaulich; doch ist diese eben nicht zu verwechseln mit dem *Dargestellten*.

Offenbar unbemerkt bedient sich Balashov der falschen Raum-Analogie aber auch bei der Charakterisierung des Endurierens. Indem er in den verschiedenen räumlichen Gestalten bloß eine Vielheit sieht, deren Einheit dem Endurantisten verloren gehe, übersieht er offenkundig die längst implizite Antwort der Endurantisten auf die Frage nach der Identität der verschiedenen räumlichen Gestalten. Ein eternalistischer Endurantist behauptet doch, dass ein und dasselbe dreidimensionale, persistierende Objekt in der Zeit *mehrfach instantiiert* sei, folglich als numerisch Identisches an diesem *und* an jenem Zeitpunkt lokalisiert sei. Dies wendet er nun auf den Kontext der SRT derart an, dass er sagt, ein und dasselbe dreidimensionale, rein raumartige Objekt sei *in der Raum-Zeit* mehrfach instantiiert. Zwar könne nicht ein und dasselbe Objekt zur selben Zeit mehr als ein *Raum*gebiet einnehmen – Objekte wären sonst absurderweise Universalien –, wohl aber könne ein und dasselbe Objekt am selben Orte so *wie* Universalien mehr als einen *Zeit*punkt einnehmen, ob dieser Zeitpunkt nun absolut ist wie klassisch oder bloß relativ wie relativistisch. Im relativistischen Kontext könnten räumliche Gestalten nun zwar auch von Bezugssystem zu Bezugssystem verschieden sein – ein numerisch identisches Objekt kann sowohl ganz rot als auch bloß teilweise rot sein. Verschiedene Raumzeit-Gebiete, an denen dem Endurantismus gemäß numerisch identische Objekte instantiiert sein sollen,

10 Sie sind dem Perdurantisten zufolge numerisch *und* qualitativ verschiedene zeitliche Teile, was problematisch macht, dass sie es von Demselben sind.

können sich folglich gar schneiden: Doch das mache die Sache der Mehrfach-Lokalisation lediglich unanschaulicher. Balashov hingegen scheint sie gar nicht einmal zu erwägen. Doch aus dem wahren Prinzip, dass ein Objekt zu einer bestimmten Zeit nur exakt ein *Raum*-Gebiet einnehmen könne, folgt eben nur mit Hilfe einer „illegitimate analogy“ (van Inwagen, 1990, 248), dass ein Objekt nur exakt ein *Raumzeit*-Gebiet einnehmen könnte.

Ohne diese falsche Raum-Analogie bricht Balashovs Argument von dem endurantistischen Erklärungsdefizit daher vorerst zusammen.

2. Das Problem sich schneidender Hyperebenen (Gilmore, 2006)

Inzwischen gibt es aber noch ein zweites Argument zugunsten der perdurantistischen Auffassung. Nach Cody Gilmore gibt es nämlich doch einen erheblichen Unterschied zwischen multipler Instantiierung in einer klassisch-absoluten Zeit und einer solchen in Minkowskis Raum-Zeit. Balashov macht laut Gilmore auf eine Schwierigkeit aufmerksam, die ein Endurantist tatsächlich habe, wenn er sein Prinzip der Mehrfach-Instantiierung auf die Raum-Zeit der SRT anwenden will.

Nehmen wir an, dass ein Objekt in einem bestimmten Raum-Zeit-Gebiet „wholly present“ ist. Dann müsse ein Endurantist bestimmte Bedingungen fordern, die bei der Zuweisung weiterer Raumzeit-Gebiete erfüllt sein müssen. Ein Endurantist könne nicht einfach sagen, dass bei der Zuweisung der Raumzeit-Gebiete keine Regeln herrschten. Schon klassisch wird man beispielsweise fordern wollen, dass das Zeitüberdauern eines Objekts *lückenlos* erfolgt. In der SRT sollte ein dreidimensionales Objekt *flach* sein. Gleichgültig, welche Kriterien man noch aufstellt: Die kontinuierliche Abfolge dreidimensionaler flacher Gebiete entlang einer *klassischen* Zeitachse, sollte sie erfüllen, will man den Endurantismus nicht einfach per Dekret ausschließen. Was aber, so fragt Gilmore, ist mit zwei sich schneidenden Raumzeit-Schnitten; mit dreidimensionalen Hyperebenen entlang zweier verschiedener Bezugssystem-Zeiten? Können auch solche Gebiete von *ein und demselben* Dreidimensionalen belegt sein, obwohl zwischen ihnen keine Kausalrelation bestehen kann, wie auch Balashov betont?

Bevor dieses Argument begutachtet werden soll, eine wichtige Bemerkung bezüglich des Kriteriums der *Flachheit*: Tatsächlich nämlich fordern *weder* Gilmore (vgl. 2006, Abschnitt 4.1) *noch* Gibson/Pooley (2006, mehrfach), dass das endurierende Objekt durch *flache* raumartige Hyperebenen dargestellt werden müsse. Diese Forderung sei zu eng, heißt es, da sie in allgemeinrelativistischen Raum-Zeiten ohnehin nicht erfüllt werden könne. Doch sind zumindest in solchen, möglicherweise realistischen, allgemeinrelativistischen Raum-Zeiten mit einer kosmologischen Zeit solche nicht-flachen Hyperflächen *auf Zeitpunkte bezogen*. In Minkowskis Raum-Zeit dagegen sind sie es nicht: Eine raumartige

Hyperfläche, die keine Gleichzeitigkeitshyperebene darstellen kann, repräsentiert etwas, das nicht auf Zeitpunkte bezogen werden kann – weder auf absolute Eigenzeiten noch auf relative Koordinatenzeiten. Nun könnte man zwar sagen, dass auch eine Menge nicht-flacher raumartiger Hyperflächen eine Abfolge von *zeitartig* getrennten Entitäten bilden könnte, doch zeigt dies eben nur, dass eine zeitartige Abfolge gerade nicht automatisch eine *zeitliche, temporale Sukzession* darstellt. Es ist nämlich nur schwer verständlich zu machen, dass ein Element dieser Menge *früher* ist als ein anderes, wenn weder das eine noch das andere relativ zu irgendeinem Zeitpunkt sein kann. Weder also kann meines Erachtens ein endurierendes Objekt in einem solchen nicht-flachen raumartigen Gebiet existieren noch ein perdurierendes Objekt dort einen zeitlichen Teil haben. Man würde den vierdimensionalen Raumzeit-Wurm *verräumlichen*, wenn man meint, er könnte von persistierenden Objekten derart ‚erfüllt' werden, dass sie – unmittelbar oder vermittelt durch Teile – auch in solchen (nicht-flachen) Teilmengen existierten. Vielmehr existiert das persistierende Objekt „zu mehr als einer Zeit", sonach in diesem Falle räumlich ausgedehnter Objekte zu den verschiedensten Koordinatenzeiten, sukzessive längs t^F und ebenso sukzessive längs t^{F*}, ob es nun enduriert oder perduriert. Zu keiner Zeit aber existiert es in nicht-flachen Hyperflächen. Die Tatsache, dass Gilmore und Gibson/Pooley Derartiges erlauben, zeigt, dass sie das Block-Universum gewissermaßen als *zeitlos* – und nicht als bloß *tempuslos* – auffassen, in einem Sinne nämlich, da Existenz (und Teilhabe) *non-relative to times* ist.[11]

Doch folgen wir weiter Gilmore: Ein notwendiges Kriterium dafür, dass ein Objekt in zwei verschiedenen zeitartig getrennten Gebieten der Raum-Zeit „wholly present" ist, sei, so Gilmore, eben das Bestehen einer *kausalen Relation* zwischen den Gehalten dieser Gebiete. Dann aber bestehe folgendes Problem: Zwei sich schneidende Gebiete könnten *nicht* von einem und demselben Objekt *exakt belegt* werden, weil zwischen ihnen keine Kausalbeziehung bestehen könne. Betrachte man nämlich eine Gleichzeitigkeitsebene von F, die in den Grenzen von P-Q ein räumlich homogen rotes Objekt darstellt, und eine sie schneidende Ebene von F*, die in den Grenzen von P'-Q' ein teilweise rotes und teilweise grünes Objekt repräsentiert. Der rote Teil von P'-Q' liege unterhalb von P-Q und der grüne Teil oberhalb. Dann gilt nicht nur, dass kein Punkt des grünen Teils von P'-Q' im Vergangenheitslichtkegel von irgendeinem Punkt von P-Q liegt, sondern auch umgekehrt, dass kein Punkt von diesem Teil von P-Q im Vergangenheitslichtkegel von irgendeinem Punkt von P'-Q' liegt. Überlichtgeschwindigkeiten und *backward causation* außer Betracht gelassen, folgt daraus offenbar, dass weder der Zustand von P-Q durch den von P'-Q' verursacht sein kann noch umgekehrt der von P'-Q' durch den von P-Q – im Widerspruch zu Gilmores Forderung. Die Bezugssystem-Abhängigkeit der räumlichen

11 In dem es also den temporal-perspektivischen Sinn von „Existenz" nicht gibt.

Gestalten, die es im vor-relativistischen Kontext *nicht* gibt, stellt den Endurantisten somit vor anscheinend unlösbaren Schwierigkeiten. Daher gebe es doch Gründe *a posteriori* zugunsten des Perdurantismus, wie es Balashov (mit schlechteren Argumenten) schon immer vertreten habe, so Gilmore.

Meines Erachtens aber überzeugt auch diese Argumentation nicht. Erneut ein Ausdruck der Gefangenschaft in der falschen Raum-Analogie wäre es nämlich, wenn man der Meinung wäre, dass einfach *jedes* dreidimensionale Teilgebiet des vierdimensionalen Raumzeit-Wurms, *weil ja „matter-filled"*, folglich auch durch ein Teil-Objekt im *perdurantistischen* Sinne belegt sein müsse. In der Anwendung des Perdurantismus auf die SRT stellt sich folglich die (auch) von Gilmore gar nicht erörterte Frage, wie der eternalistisch existierende vierdimensionale Raumzeit-Wurm sinnvoll *in Teile* geteilt werden kann, so dass sich *zeitliche* von *räumlichen* Teilen unterscheiden lassen. Zwar mag, wie gesagt, jeder beliebige zweidimensionale Schnitt durch ein dreidimensionales räumliches Objekt als ein *räumlicher* Teil desselben gelten, *zeitartige* Abschnitte von Weltlinien innerhalb des Raumzeit-Wurms, beispielsweise, oder *gekrümmte* dreidimensionale Teilgebiete desselben repräsentieren aber *weder* einen zeitlichen *noch* einen räumlichen Teil des persistierenden Objekts.[12] Folglich steht der Perdurantist vor einer analogen Schwierigkeit wie der Endurantist: Ein notwendiges Kriterium dafür, dass zwei verschiedene raumartige Gebiete der Raum-Zeit zwei *zeitliche* Teile *desselben* Objekts darstellen können, scheint zu sein, dass zwischen den Gehalten dieser Gebiete eine kausale Relation besteht.

Schon bei David Lewis (1976) war es ja keineswegs so, dass irgendwelche Kandidaten ganz ohne Kriterien als zwei zeitliche Teile desselben Objekts hätten gelten können. Zeitliche Teile eines Objekts sind nämlich *per definitionem* numerisch distinkt und können im Fall sich verändernder Objekte auch qualitativ verschieden sein. Was hätten sie dann gemeinsam, so dass sie Teile *Desselben* sein könnten? Nach Lewis muss zwischen diesen Kandidaten eine bestimmte Relation bestehen, zwar nicht die der Identität wie bei den Endurantisten, aber doch eine vergleichbare, die er „I-Relation" nennt. Auch zwischen zeitlichen Teilen muss eine Kausal-Relation gelten. So auch Sider (2001):

> [A] sequence of temporal parts counts as a continuant only if that sequence falls under a causal law. (Sider, 2001, 227)

Wenn nun aber zwischen den Gehalten der sich schneidenden, flachen raumartigen Gebieten *kein* Kausalverhältnis bestehen kann, wie Gilmore argumentiert, dann können sie ebensowenig zwei numerisch distinkte zeitliche Teile desselben Vierdimensionalen darstellen, wie sie von einem numerisch identischen Dreidimensionalen exakt belegt werden konnten. Wo keine Mehrfach-Instantiierung

12 Keineswegs ist also der relativistische Perdurantismus einfach die „doctrine of arbitrary spatiotemporal parts" (Gibson/Pooley, 2006,162). Vielmehr wird man differenzieren müssen und angeben, welche der raumzeitlichen Teile *zeitliche* sind.

Desselben vorliegen kann, da kann auch keine I-Relation bestehen, da beides *gleichermaßen* durch die Kausalbeziehung unterfüttert sein muss. Die in der SRT herrschende Bezugssystem-Abhängigkeit der räumlichen Gestalten raumartig ausgedehnter, persistierender Objekte spricht demzufolge nicht nur gegen den Endurantismus, sondern gleichermaßen gegen den Perdurantismus.

Dies kann man als ein Argument gegen die zugrunde liegende eternalistische Hypothese werten: Auf ihrer Basis lässt sich in der SRT *nicht* beschreiben, auf welche Weise Objekte in der Raum-Zeit *zeitlich* ausgedehnt sind.

Literaturverzeichnis

Balashov, Yuri: "Persistence and Space-Time: Philosophical Lessons of the Pole and Barn". The Monist, 83, 2000. S. 321-340

Gibson, Ian/Pooley, Oliver: "Relativistic Persistence". Philosophical Perspectives, 20, Metaphysics, 2006. S. 157-198

Gilmore, Cody: "Where in the relativistic world are we?". Philosophical Perspectives, 20, Metaphysics, 2006. S. 199-236

Heller, Mark: "Things Change, Philosophy and Phenomenological Research". LII, 3, 1992. S. 695-704

Lewis, David: "Survival and Identity". Philosophical Papers, I, OUP, Oxford, 1983/1976. S. 55-77

Miller, Kristie: "Enduring Special Relativity". Southern Journal of Philosophy, 42(3), 2004. S. 349-370

Sider, Theodore: "Four-Dimensionalism. An Ontology of Persistence and Time". Clarendon Press, Oxford, 2001

Manipulationist and Regularity Theories of Constitution

Jens Harbecke
jens.harbecke@uni-wh.de
The Cohn Institute for History and Philosophy of Science and Ideas,
Tel Aviv, Israel

Abstract/Zusammenfassung

This paper reviews the manipulationist theory of the notion of "constitution" as it has been defended by proponents of the mechanistic approach to neurobiological explanations. It shows that this account suffers from at least two weaknesses. In particular, it is unclear how an isolated intervention of the candidate constituter should be possible with respect to coincident mechanisms. As a consequence, the notion of constitution is trivialized as any two coincident mechanisms turn out to be related by constitution. Moreover, it is pointed out that the definition has a modal ingredient that may render the corresponding interpretations of pertinent neurobiological explanations vague.
In a second step, a regularity account of constitution is sketched. This approach construes constitution on the basis of regular co-instantiations of types that conform to certain minimalization conditions. In a final comparison, the regularity account is described as superior to the manipulationist account.

Dieser Artikel analysiert die Manipulations-Theorie der Konstitutions-Relation, wie sie innerhalb des mechanistischen Ansatzes in der Philosophie der Neurobiologie entwickelt worden ist. Es wird aufgezeigt, dass die genannte Theorie mindestens zwei Schwachpunkte aufweist. Zum einen ist unklar, wie eine isolierte Intervention mit Hinsicht auf einen konstituierenden Mechanismus möglich sein soll, ohne dass zugleich der konstituierte Mechanismus manipuliert wird. Die Konsequenz dessen ist eine Trivialisierung des Konstitutionsbegriffes, da damit jegliche zwei ko-instantiierten Mechanismen einander konstituieren. Zum anderen gebraucht die genannte Definition modale Begriffe, wodurch die einzelnen Interpretationen der entsprechenden neurobiologischen Erklärungen als vage charakterisiert werden müssen.
In einem zweiten Schritt entwickelt der Artikel eine Regularitäts-Theorie des Konstitutions-Begriffs. Dieser Ansatz bestimmt den Begriff der Konstitution auf der Basis regulärer Ko-Instantiierungen von Typen, die bestimmten Minimalisierungs-Bedingungen genügen. Die Regularitäts-Analyse erweist sich als adäquat für die Rekonstruktion prototypischer Erklärungen der Neurobiologie und damit als dem Ansatz der Manipulations-Theorie überlegen.

1. Introduction

The concept of "constitution" plays a central role in recent works in the philosophy of the special sciences, notably in the philosophy of neurobiology. For instance, one of the main claims of the now popular "mechanistic approach" to

neurobiology (Machamer et al. 2000; Craver and Darden 2001; Craver 2007) is that informative neurobiological explanations are essentially "constitutive explanations". An example often used to illustrate this hypothesis is the now well established explanation of spatial learning in rats.[1] According to this explanation, the process of a development of spatial memory in these animals is (partially) constituted by an activity of NMDA receptors in the rat's hippocampus (Davis et al. 1992). If it is true that all informative neurobiological explanations mirror the structure of this explanation in the sense that they all specify a constitutive connection between relevant types, then obviously the explanatory content of neurobiological theory is transparent only if an informative definition of the two-place predicate "… constitutes…", or of corresponding terms such as "…plays a crucial role in…"[2], "…implements…", "…necessitates…" etc. is available. The philosophical literature currently contains two main proposals for an adequate analysis of the notion of constitution the first of which can be characterized as a "manipulationist" (Craver 2007, 152/53)[3] and the second of which can be described as a "regularity" (Harbecke *manuscript*) theory of constitution.

The aim of this paper is twofold: It first discusses the manipulationist theories of constitution and points out its main deficiencies (section 2). Afterwards it outlines the main ideas of a regularity theory of constitution and characterizes it as superior to the manipulationist account (section 3). The paper concludes with some thoughts on possible extensions of the research on the regularity account of constitution (section 4).

2. The manipulationist account

The most explicit statement of the manipulationist account of constitution is found in Craver (2007, 139-162). The author intends his account as a clarification of an important detail of the mechanistic approach. Consequently, he subscribes to the general contention that neurobiological explanations essentially

1 The example is discussed, for instance, in Bickle 2003, ch. 3-5; Churchland and Sejnowski 1996, ch. 5; Craver and Darden 2001, 115-119; Craver 2002, sec. 2; and Craver 2007, 165-170.

2 For instance, Davis et al. (1992) claim cautiously that "… hippocampal NMDA receptor activation … plays a crucial role in certain types of learning." (1992, 32). However, the authors would most likely agree that "…plays a crucial role in…" and "…constitutes…" should be interpreted synonymously in this context.

3 Craver does not explicitly call his theory of constitution "manipulationist"; however, he does describe it as being inspired by the manipulationist, or interventionist, theory of causation (Craver 2007, 152) as it has been defended, for instance, by Woodward (2003), and the former clearly shares the fundamental spirit of the latter. We take this to justify the tag "manipulationist" for Craver's theory of constitution (cf. also Harbecke *manuscript*, sec. 3).

specify certain mechanisms that are related by the constitution relation. In summarizing the main idea, Craver defines a mechanism as constitutively relevant to the behavior of a mechanism as a whole "…when one can wiggle the behavior of the whole by wiggling the behavior of the [constituent] and one can wiggle the behavior of the [constituent] by wiggling the behavior of the whole." (2007, 153)[4] The notion of "wiggling" a mechanism by "wiggling" another one has its roots in Woodward's manipulationist, or interventionist, theory of causation (2003, 26, 53, 91, 143). Woodward uses this colloquial expression to characterize an ideal experimental intervention designed to identify causal relevancies.

A slightly more formal formulation of the same idea offered by Craver is the following one. According to the author, any mechanism, say *x*'s φ-ing, is constitutionally relevant to, or simply constitutes, a second mechanism, say *y*'s ψ-ing, if, and only if:

> (i) [*x*] is part of [*y*]; (ii) in the conditions relevant to the request for explanation there is some change to [*x*]'s φ-ing that changes [*y*]'s ψ-ing; and (iii) in the conditions relevant to the request for explanation there is some change to [*y*]'s ψ-ing that changes [*x*]'s φ-ing. (2007, 153)

In this passage, it is the term "change" that alludes to Woodward's notion of an "intervention". In order to fully grasp the content of Craver's theory, it is therefore worthwhile to review the main points of Woodward's interventionist theory of causation. As is customary, Woodward distinguishes between singular and general causation, but takes the latter to be the primary *analysandum*. In this sense, (type-level) causation is a relation between value-differentiated types. Causal relations in the manipulationist sense are always relative to a causal model which is partly characterized by a variable set *V*. Woodward then defines direct causation and contributing causation in the following way:

> A necessary and sufficient condition for *X* to be a (type-level) *direct cause* of *Y* with respect to a variable set *V* is that there be a possible intervention on *X* that will change *Y* or the probability distribution of *Y* when one holds fixed at some value all other variables Z_i in *V*. A necessary and sufficient condition for *X* to be a (type-level) contributing cause of *Y* with respect to variable set *V* is that (i) there be a directed path from *X* to *Y* such that each link in this path is a direct causal relationship; (…) and that (ii) there be some intervention on *X* that will change *Y* when all other variables in *V* that are not on this path are fixed at some value. (Woordward 2003, 59)

The notion of a "possible intervention" is obviously at the heart of Woodward's definition. Note that this intervention is required to be *isolated* in the sense that all other variables contained in the models are to be held fixed. This condition is intended to prevent a manipulation of *Y* through a path that does not go through

4 Craver prefers speaking of "components" instead of "constituents". However, the intended reference of the terms can be presupposed as identical such that nothing substantial is lost by the replacement undertaken here for the sake of terminological unity.

X and which could falsely lead to the conclusion that there exists a causal relation between *X* and *Y*. It follows that, if there is no possible isolated intervention on a candidate cause variable then there is no (type-level) causation between the candidate cause variable and the relevant effect variable.

Craver's account of constitution bears obvious analogies to Woodward's theory of causation. Of course, Craver's terminology of "*x*'s φ-ing" and "*y*'s ψ-ing" makes it not immediately clear whether the author intends to define singular constitution between token events or general constitution between types. However, the fact the definition is intended for an adequate reconstruction of neurobiological explanations strongly suggests that what Craver has in mind is general constitution as paradigm neurobiological explanations such as the one hinted at in section 1 are primarily claims about the relationships of natural types. It may also not be immediately clear whether Craver intends the term "change" to refer to an isolated intervention performed on the candidate constituting type changing the constituted type in analogy to Woodward's notion. However, if the isolation were not presupposed, many scenarios seem possible in which an intervention on one mechanistic type leads to a change on the second type without the types being related by constitution. Given these characteristics of Craver's account, the question arises whether it is an adequate reconstruction of the constitution relation as it is referred to in paradigm neurobiological explanations.

As concerns Woodward, the strength of his theory clearly lies in the fact that it captures well a general methodology for isolating type-level causal relationships that is widely applied in the special sciences. However, as a metaphysical definition of causation it is clearly unsatisfactory. For instance, a general point made by Pearl on causal models and interventions is that, since an intervention must always be external to a relevant causal model and cannot be a variable in the model itself, there are no causal relations when the cause and effect are very large types. This is the case, for instance, when the relevant types are total subsequent stages of a universe such that there is nothing external any more. In contexts like these "…causality disappears because interventions disappear – the manipulator and the manipulated lose their distinction." (Pearl 2000, 350) Furthermore, as de Regt has pointed out, the manipulationist fails to provide a genuine definition of causation "…because causal relations are defined via the notion of an intervention, which is itself a causal notion" (2004). Finally, as Baumgartner (2009) has shown, the manipulationist account is unable to avoid certain definitional regresses due to the fact that the theory interdefines causation and interventions. For instance, according to the manipulationist theory, if a variable *X* is a cause of a variable *Y*, there must be a potential isolated intervention I_1 on *X* that induces a change on *Y* via *X*. This implies that, whatever else I_1 may be, it will have to be a cause of *X*, which implies that there must be a further potential isolated intervention I_2 on I_1 that induces a change on X via I_1. This makes the regress obvious.

If the above reconstruction of Craver's theory of constitution is correct, it can be expected that the account inherits the virtues of Woodward's theory. However, there will also be a worry that the theory suffers from analogous vices. Craver clearly intends his account of constitution in a metaphysical sense since the explanations analyzed with reference to the definition are intended as ontological claims. If so, however, it seems obvious that very large mechanisms cannot have constituters. In a way analogous to the manipulationis account of causation, if no intervention external to the relevant model is possible, there is no constitution. Yet, this possible criticism of Craver's account is a weak one if at all. In its normal applications, the account will only be concerned with middle-sized mechanisms such as the one occurring in the explanation of spatial memory formation in rats and so it does not have to bother with extravagantly large mechanistic types. The account also does not seem to suffer directly from a definitional regress. Constitution is defined over the notion of an intervention, which is a causal notion. Only if the kind of causation presupposed should be the one specified by Woodward, the total theory suffers indirectly from the regresses mentioned above.

However, the following considerations do pose an immediate problem for a manipulationist account of constitution. Whatever else the connection between the relevant mechanisms may be, their instances are clearly presupposed as always occurring in the same place and time. This fact is obvious from Craver's illustrations of the explanation of spatial memory formation in rats (cf. 2007, 166). The co-instantiation of the *relata* clearly distinguishes constitution from causation. However, if the constituting and the constituted mechanisms have their instantiations always in the same place and time, it becomes difficult to understand how there could be an intervention on the candidate constituting mechanism that was not an intervention on the relevant constituted mechanism as well. The presupposition within the interventionist framework of causation is, of course, that the intervention on the candidate cause is not simultaneously an intervention on the effect. The effect's change does not coincide with the intervention in question but follows it. In contrast, no such difference in time and place is possible if the relevant mechanisms coincide. Unfortunately, this fact trivializes the manipulationist account of constitution. Any two mechanisms whatsoever will be related by constitution as long as they regularly coincide.

A second disadvantage of the manipulationist definition of constitution stems from the fact that with the concept of a "possible intervention" it involves certain modal notions. This fact is not immediately obvious from Craver's formulation quoted above as the author only requires that "…there is some change to [*x*]'s φ-ing…" without explicitly using the word "possible". However, the number of mechanisms related by constitution would be very few if constitution was defined over actual interventions performed at some point in history. Hence, Craver's definition should be read as implicitly using the modal notion of a

"possible change". But since modal claims are notoriously vague, the actual explanations analyzed through a so-defined notion of constitution become vague as well. This consequence should be avoided if possible and it remains a deficit of the manipulationist definition.

The insights discussed in the previous paragraphs motivate the search for an alternative account that does not rely on the notion of an isolated intervention and that can be formulated with recourse to modal notions. The following section aims to provide such an account.

3. The regularity theory

This section presents what can be described as a "regularity account" of constitution. According to this theory, the notion of constitution can adequately be defined in an extensional first-order language and, therefore, in non-modal terms. Furthermore, the regularity theory can forgo the notion of an isolated intervention and it involves only the notions of a minimalized regularity and of the part-whole relation, both of which are independent of any actual or possible manipulator.

From the example of a neurobiological explanation hinted at in section 1 it is obvious that constitution in scientific contexts describes a regular co-occurrence of event types, where any event type *A* should just be understood as the set of all token events, or token *n*-tuples of events, in the extension of a relevant predicate, such as "…is a tying of a shoe", "…is an activity of NMDA receptors" etc.[5] Any type *A* has a negation ¬*A*, which should be understood as the complementary set of *A*. Harbecke (*manuscript*) has developed a detailed definition of constitution between intrinsic types. The following paragraphs develop a definition that mirrors the original one, but that in addition also allows for constitutional relations between relations.

Whatever else the notion of constitution involves, its application in paradigm neurobiological explanations makes clear that it makes at least a sufficiency claim of certain types for certain other types. A simple definition based on this insight can be formulated as follows.

> *Constitution as Sufficiency*: For any two types *A* and *B*, *A* constitutes *B* if, and only if, the following holds: if an *n*-tuple of individuals $<x_1,\ldots,x_n>$ instantiates *A*, then $<x_1,\ldots,x_n>$ instantiates *B* as well, or simply: $A \rightarrow B$.

5 The focus on types makes clear that the mentioned debate on the "constitution of kinds" is not identical to the debate on "material constitution" (cf. Baker 1997; Rea 1997; Wasserman 2004, 2009). Material constitution is understood as a first-order relation between objects, and not as a second-order relation between types. That the two relations have important connections is undisputed, but we will not address these here.

In an obvious way this formulation is useless for an adequate definition of constitution. For one thing, even if NMDA-receptor activity constitutes the formation of spatial memory in rats, the former can hardly be considered sufficient for the latter all by itself (cf. Davis et al. 1992, 32). Generally speaking, only complex bundles of properties are usually constitutionally sufficient for some relevant type.

Secondly, there are many regularities conforming to the above definition that are not interpretable as specifying constitutional relationships. Some of these regularities are due to the law of monotony: An antecedent of any conditionals can always be supplemented by further conjuncts *salva veritate*. If NMDA-receptor activity (call it '*F*') in conjunction with certain further neurophysiological types *G*, *H*, and *I* is sufficient for spatial memory formation, then *FGHI* conjoined with the type "…is made of cheese" is sufficient for spatial memory formation, too. However, it would simply be ridiculous to consider the type "…is made of cheese" as partly constituting the type "…is a spatial memory formation process in rats". Finally, it is not clear that the constitutionally related types are always instantiated by the same individuals and that constituting and constituted types are always of the same degree.

These three insights induce already a certain complexity on the regularity definition of constitution demanding certain notational simplifications. In order to express partial constitution of a type by another, we will use placeholders 'X_1', 'X_2' etc. to designate $0 \leq n < \infty$ elements of complex constituters that are not yet identified or simply neglected, and we use placeholders 'χ_1', 'χ_2' etc. to designate $0 \leq n < \infty$ ordered sets of individuals (with a finite number of elements). In order to block regularities from being constitutionally interpretable that result from a simple supplementation operation of the regularity's antecedent, the complex constitutors are demanded to be minimally sufficient for the constituted type.

> *Constitution as Minimal Sufficiency*: For any two types *A* and *B*, *A* constitutes *B* if, and only if, the following holds: i) if an *n*-tuple of individuals $<x_1,\ldots,x_n>$ instantiates *A* and co-occurs with a set χ of ordered sets of individuals each instantiating exactly one type in a set *X* of types, then *B* is instantiated as well by some *m*-tuple of individuals $<y_1,\ldots,y_m>$, and ii) for any set *W* of types such that $W \subset AX$ it is not the case that, if it is instantiated by some set χ' of ordered sets of individuals, then *B* is always instantiated as well by some *m*-tuple of individuals $<y_1,\ldots,y_m>$; or simply $AX \rightarrow_{min} B$.

This definition already seems to capture a substantial portion of the meaning of claims such as "the formation of spatial memory in rats is (partially) constituted by NMDA-receptor activity". The idea of this explanation seems to be that if NMDA-receptor activity occurs in, or is instantiated by, certain neurons in the rat's hippocampus, then in a normal rat in normal circumstances a learning process is instantiated in the rat, or by the rat's brain. Furthermore, a claim im-

plicitly intended is that, if some of the normal circumstances or the NMDA-receptor activity is not instantiated, then the learning process does not occur. In other words, the explanation mentions a minimally sufficient condition for the *explanandum*. This minimality constraint is captured by condition ii). Consequently, *Constitution as Minimal Sufficiency* seems to be a good candidate as a definition of constitution.

Unfortunately, the definition still suffers from at least four deficiencies. First of all, the notion of a "co-occurrence" appearing in condition i) is left unspecified despite its central role for the definition. Depending on how liberally the term is interpreted, the falling of an 8-ball can constitute the explosion of an atomic bomb given that, whenever there was an actual explosion of an atomic bomb in history, there also happened to occur the falling of an 8-ball somewhere in the world. Apart from these accidental cases, there are also many lawful regularities conforming to the definition that are not constitutional in nature. In particular, many causal regularities can be claimed to satisfy this definition (cf. Baumgartner 2008, 331/32). Moreover, many mechanistic types do not have only a single minimally sufficient constituter, but many. For instance, Urban and Barrionuevo (1996) argue that memory formation sometimes occurs without NMDA-receptor activity[6] but is sometimes constituted by different neural mechanisms.

A problem arising in the context of alternative constituters is well known from analogous regularity analyses of causation, where it is referred to as the problem of "spurious regularities" (cf. Baumgartner 2008, 338-340). To illustrate the general structure of this problem, suppose that a complex type *A* is a minimally sufficient constituter of both *B* and *C* and suppose that *B* is not a constituter of *C*. Suppose further that a complex type *D* non-identical to *A* is the only alternative minimally sufficient constituter of *B*. Then the conditional $\neg D \,\&\, B \rightarrow_{min} C$ conforms to the definition of *Constitution as Minimal Sufficiency*. However, by hypothesis, *B* is *not* a constituter of *C*. Hence, not all regularities conforming to the definition *Constitution as Minimal Sufficiency* are constitutionally interpretable. To avoid this problem, the definition needs to exclude a redundancy of sufficient constituters (in the example, the redundancy consists in the fact that, whenever $\neg D \,\&\, B$ occurs, *A* occurs also, but not vice versa) in order for the regularities conforming to the definition to be constitutively interpretable.

A definition answering to all these demands would specify that a type *A* constitutes a second type *B* if, and only if, *A* is a necessary part of a minimally sufficient condition of *B* where this minimally sufficient condition is a disjunct in a disjunctive minimally necessary condition of *B*, such that all the relevant indi-

6 Specifically, they argue that the phenomenon of longterm potentiation can occur without NMDA-receptor activity, where LTP is itself described as a constituter of memory formation.

viduals instantiating a minimally sufficient condition of *B* on some occasion stand to the individuals instantiating *B* on that occasion in some coordination relation. The coordinating relation is introduced to distinguish constitutional from causal regularities. One interpretation serving this purpose identifies the coordinating relation with the mereological, or part-whole, relation[7] securing that the individuals instantiating a minimally sufficient condition for a type *B* jointly occupy the same space and time as the individual(s) instantiating *B*. This distinguishes constitutional regularities from causal regularities, as the individuals instantiating causal regularities are commonly demanded to be spatio-temporally distinct (cf. Baumgartner 2008, 330). An explicit definition of this general idea that avoids the terms "necessary" and "sufficient" can be attained by using '*P*' to designate the mentioned coordination relation and by using '$X_1 \vee X_2 \vee \ldots \vee X_n$' to designate a disjunction of conjunctions of types. A possible formulation is the following one (note that conditions i) and ii) are almost identical to the first two conditions of the definition of *Constitution as Minimal Sufficiency*).

> *Constitution*: For any two types *A* and *B*, *A* constitutes *B* if, and only if, the following holds: i) if an *n*-tuple of individuals $<x_1,\ldots,x_n>$ instantiates *A* and each ordered set of individuals in a set χ instantiates exactly one type in a set *X* of types, then *B* is instantiated as well by some *m*-tuple of individuals $<y_1,\ldots,y_m>$, and ii) for any set *W* of types such that $W \subset AX$ it is not the case that, if it is instantiated by some set χ' of ordered sets of individuals, then *B* is always instantiated as well by some *m*-tuple of individuals $<y_1,\ldots,y_m>$; and iii) for every individual x_i in the ordered set $<x_1,\ldots,x_n>$ instantiating *A*, such that $i = 1,\ldots,n$, there is an individual y_j in the ordered set $<y_1,\ldots,y_m>$ instantiating *B* such that $j = 1,\ldots,m$ such that Px_iy_j, and for every individual *z* in each of the ordered sets of χ instantiating the types in *X*, there is an individual y_j in $<y_1,\ldots,y_m>$ such that Pzy_j; and vi) *AX* is part of a disjunction $AX \vee X_1 \vee \ldots \vee X_n$ of conjunctions of types such that $AX \vee X_1 \vee \ldots \vee X_n$ is minimally necessary for *B*.

This definition seems rather complex at first sight. However, its superficial complexity evaporates once it is noticed that condition iii) merely demands that the individuals referred to by conditions i) and ii) instantiating *AX* stand in a specific coordination relation *P* to the individuals instantiating *B* if the sufficiency expressed by $AX \rightarrow_{min} B$ should be interpretable as singling out a genuine constitution relation. Condition iv) merely minimalizes the number of alternative sufficient constituters of the type in question.

With these two further conditions *Constitution* can avoid the four deficiencies diagnosed for *Constitution as Minimal Sufficiency*. The definition no longer uses the notion of "co-occurrence" but involves only a coordination relation *P* that, if interpreted as the part-whole relation, has a clear interpretation. Simultaneously,

7 One possibility to cash this out would be to presuppose mereology as defined by the system *GEM* (Varzi 2003).

the relation *P* interpreted in this way excludes all causal regularities from being classifiable as constitutional relations since causes are not parts of their effects. They are not parts of their effects because parts occur at the same time as their wholes whilst causes essentially occur at different times than their effects. Moreover, the definition now allows constituted mechanistic types to have not only a single minimally sufficient constituter, but many. Finally, condition iv) helps to exclude the interpretation of spurious regularities as constitutive regularities. This result is achieved in the following way. In the example sketched above, the constitutive explanation of the type *C* would take the following form: $A \vee \neg D \,\&\, B \rightarrow_{min} C$. However, this conditional does not meet condition iv) of *Constitution*. Since, whenever $\neg D \,\&\, B$ occurs, so does *A*, it is not the case that the disjunction $A \vee \neg D \,\&\, B$ is minimally necessary for *C*. Since the occurrence of $\neg D \,\&\, B$ implies the occurrence of *A*, but not vice versa (since *D* can still be instantiated if A is), it is even possible to determine $\neg D \,\&\, B$ as the redundant disjunct. In this way, *Constitution* avoids the problem stemming from spurious constitutive regularities.

As it turns out, even this complex definition probably requires certain amendments to fully and adequately capture the content of the notion of "constitution" in contexts such as explanations in neurobiology. For instance, it may be objected that the definition still trivializes the relation of constitution as any empty type regularity conforms to the definition. At the same, there may remain worries that the definition now excludes too many genuine constitutive regularities by requiring that every individual (partly) instantiating a constituting type be part of *one* individual (partly) instantiating a constituted type. Sometimes the individuals of the constituters may only be a part of two fused individuals of the constituted type. However, for now we will leave these complications aside and contend that the definition of *Constitution* analyzes the notion as it is applied in paradigm explanations of neurobiology to a satisfactory extent.

If again we take as a reference point the explanation of the formation of spatial memory in rats, we can now reconstruct its content as follows. Recall that, according to this explanation, the formation of spatial memory in rats is (partially) constituted by an activity of NMDA receptors in the rat's hippocampus (cf. section 1). *Constitution* seems to capture adequately what is claimed here. First of all, the activity of NMDA receptors is clearly not itself taken as sufficient for an occurrence of memory formation. A range of other mechanisms and background conditions are implicitly presupposed as forming a condition, of which NMDA receptor activity is only a necessary part (conditions i) and ii)). Furthermore, the relevant instances talked about are considered as intimately related. The part-whole relation appears to express what is intended here (condition iii)). Finally, the explanation leaves open whether there are other mechanisms that can induce a learning process just as well. It is well-known today that the removal of the hippocampus in rats as well as in humans leads to a heavy im-

pairment of learning abilities. However, as mentioned above, phenomena linked to memory formation have been observed in other parts of the hippocampus that do not involve an NMDA receptor activity. If this hypothesis of an alternative constituter should be confirmed, a complete explanation of spatial memory would have to mention this fact by stating a disjunction of minimally sufficient conditions for a constituted mechanism. At the same time, any redundancies would have to be avoided in a complex explanation of this kind. In other words, such an explanation would have to meet condition iv). These points show that *Constitution* is a highly adequate definition of constitution as it is used in paradigm neurobiological explanations. Consequently, *Constitution* is a more powerful definition of constitution than the one provided by the manipulationist account.

4. Conclusion

This paper first reviewed the manipulationist theory of constitution as it has been defended by proponents of the mechanistic approach to neurobiological explanations. It showed that this account suffers from at least two weaknesses. Firstly, it was unclear how an isolated intervention of the candidate constituter should be possible with respect to coincident mechanisms. As a consequence, the notion of constitution was trivialized as any coincident mechanisms would turn out as being related by constitution. Secondly, it was pointed out that the definition had a modal ingredient that may render the corresponding interpretations of neurobiological explanations vague.

In a second step, a regularity account of constitution was sketched. This account construed constitution on the basis of regular co-instantiations of types that conformed to certain minimalization conditions. The regularity account was described as superior to the manipulationist account.

As it was pointed out towards the end of section 3, even the relatively complex formulation of *Constitution* may require certain extensions and adjustments. To make these precise will constitute a challenge for further research on the regularity account of constitution. Furthermore, little has been said on a corresponding methodology to determine constitutional explanations. Questions pertaining to this topic will be a matter of future research.

References

Baker, L.: "Why constitution is not identity". The Journal of Philosophy, 94(12), 1997. S. 599-621

Baumgartner, M.: "Regularity theories reassessed". Philosophia, 36(3), 2008. S. 327-354

Baumgartner, M.: "Interdefining Causation and Intervention". Dialectica, 63(2), 2009. S. 175-194

Bickle, J.: Philosophy and neuroscience: A ruthlessly reductive account. Kluwer, Doordrecht, 2003

Churchland, P./T. Sejnovski : The computational brain. MIT Press, Boston, 1996

Craver, C.: "Interlevel experiments and multilevel mechanisms in the neuroscience of memory". Philosophy of Science, 69(3), 2002. S. 83-97

Craver, C.: Explaining the brain. Oxford University Press, New York, 2007

Craver, C./L. Darden: "Discovering mechanisms in neurobiology". In: *P. Machamer, R. Grush, and P. McLaughlin (Eds.)*: Theory and method in the neurosciences. University of Pittsburgh Press, Pittsburgh, 2001. S. 112–137

Davis, S, S. Butcher, and R. Morris: "The NMDA receptor antagonist D-2-amino-5-phosphonopentanoate (D-AP5) impairs spatial learning and LTP in vivo at intracerebral concentrations comparable to those that block LTP in vitro". Journal of Neuroscience, 12(1), 1992. S. 21-34

De Regt, H.: Review of James Woodward, 'Making things happen'. Notre Dame Philosophical Reviews, 2004, http://ndpr.nd.edu/ review.cfm?id=1455

Harbecke, J.: "Levels of mechanism in neurobiological explanations". manuscript

Machamer, P./L. Darden/C. Craver: "Thinking about mechanisms". Philosophy of Science, 67(1), 2000. S. 1-25

Pearl, J.: Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, 2000

Rea, C.: "Supervenience and Co-location". American Philosophical Quarterly, 34(3), 1997. S. 367-375

Urban, N./G. Barrionuevo: "Induction of hebbian and non-hebbian mossy fiber long-term potentiation by distinct patterns of high-frequency stimulation". Journal of Neuroscience, 16(13), 1996. S. 4293-4299

Varzi, A.: "Mereology". In: *Zalta, E. N. (Ed.)*: Stanford Encyclopedia of Philosophy, 2003/Summer 2009 edition, http://plato.stanford.edu/archives/sum2009/entries/mereology/

Wasserman, R.: “The constitution question”. Nous, 38(4), 2004, S. 693-710
Wasserman, R.: “Material constitution”. In: *Zalta, E. N. (Ed.)*: Stanford Encyclopedia of Philosophy, Summer 2009 edition, http://plato.stanford.edu/entries/material-constitution/
Woodward, J.: Making things happen: A theory of causal explanation. Oxford University Press, New York, 2003

On Formalizing De Morgan's Argument[1]

Timm Lampert
lampertt@staff.hu-berlin.de
Humboldt Universität, Berlin

Abstract/Zusammenfassung

This paper compares several models of formalization. It articulates criteria of correct formalization and identifies their problems. All of the discussed criteria are so called "semantic" criteria, which refer to the *interpretation* of logical formulas. However, as will be shown, different versions of an implicitly applied or explicitly stated criterion of correctness depend on different understandings of "interpretation" in this context.
In particular, I will discuss the benefits and problems of the following criteria of correctness:

- *criterion of verbalization* (CRVERB): $\phi(A)$ is a correct formalization of a proposition *A* if and only if the free verbalization of *A*, $V(\phi(A))$, and *A* are equivalent (have the same meaning).
- *varying-$\mathfrak{I}$ criterion* (V$\mathfrak{I}$CR): $\phi(A)$ is a correct formalization of a proposition *A* if and only if the schematization of *A*, *sch(A)*, and $\phi(A)$ have the same truth value relative to all interpretations / realizations.
- *fixed-$\mathfrak{I}$ criterion* (F$\mathfrak{I}$CR): $\phi(A)$ of a logical language L with a fixed interpretation $\mathfrak{I}$, <L, $\mathfrak{I}$>, is a correct formalization of a proposition *A* if and only if $\Phi(A)$ is an effective translation of *A* or of a rephrasing of *A* in respect to <L, $\mathfrak{I}$>.
- *TC*: $\phi(A)$ is a correct formalization of proposition *A* if and only if $\phi(A)$ and *A* have the same truth values with respect to all interpretations in terms of conditions of truth or falsehood of *A* that are suitable according to the realization.
- *TC'*: $\phi(A)$ is a correct formalization of a proposition *A* if and only if $\phi(A)$ and *A* have the same truth values with respect to suitable realizations that do not restrict the space of possible interpretations in terms of conditions of truth and falsehood of *A*.

I will argue that the formal representation of informal reasoning highly depends on the implicitly assumed notion of a correct formalization. This, I will demonstrate by referring to the so called "De Morgan argument" ("All horses are animals. Therefore: All heads of horses are heads of animals.").

Der Artikel vergleicht verschiedene Modelle der Formalisierung. Er stellt unterschiedliche Kriterien der Formalisierung vor und identifiziert deren Probleme. Alle diskutierten Kriterien sind sogenannte „semantische" Kriterien, die sich auf *Interpretationen* logischer Formeln beziehen. Es wird gezeigt, dass die unterschiedlichen Kriterien auf einem unterschiedlichen Verständnis von „Interpretationen" logischer Formeln beruhen.
Insbesondere werden die Vorzüge und Probleme der folgenden Kriterien diskutiert:

1 I would like to thank Michael Baumgartner, with whom I wrote several papers on the theory of formalization, as well as Georg Brun for discussions on the topic. I am also grateful to Robert Wengert for comments on this paper.

- *Verbalisierungskriterium* (CRVERB): $\phi(A)$ ist eine korrekte Formalisierung einer Aussage *A* genau dann, wenn die freie Verbalisierung von *A*, $V(\phi(A))$, und *A* äquivalent sind (denselben Sinn haben).
- *variierendes-$\mathfrak{I}$ Kriterium* (V$\mathfrak{I}$CR): $\phi(A)$ ist eine korrekte Formalisierung einer Aussage *A* genau dann, wenn die Schematisierung von *A*, *sch(A)*, und $\phi(A)$ denselben Wahrheitswert relative zu allen Interpretationen haben.
- *festes-$\mathfrak{I}$ Kriterium* (F$\mathfrak{I}$CR): $\phi(A)$ einer logische Sprache L mit einer festen Interpretation $\mathfrak{I}$, $<L, \mathfrak{I}>$, ist eine korrekte Formalisierung einer Aussage *A* genau dann, wenn $\phi(A)$ relativ zu $<L, \mathfrak{I}>$ eine effektive Übersetzung von *A* oder einer Paraphrase von *A* ist.
- *TC*: $\phi(A)$ ist eine korrekte Formalisierung einer Aussage *A* genau dann, wenn $\phi(A)$ und *A* denselben Wahrheitswert in Bezug auf alle entsprechenden Interpretationen haben. Diese entsprechenden Interpretationen sind Bedingungen der Wahrheit und Falschheit von *A*, die gemäß der Legende sinnvoll sind.
- *TC'*: $\phi(A)$ ist eine korrekte Formalisierung einer Aussage *A* genau dann, wenn $\phi(A)$ und *A* denselben Wahrheitswert relativ zu einer entsprechenden Legende haben, die nicht den Raum möglicher Interpretationen, verstanden als Bedingungen der Wahrheit und Falschheit von *A*, beschränkt.

Es wird gezeigt, dass die formale Repräsentation informellen Argumentierens von implizit vorausgesetzten Kriterien logischen Formalisierens abhängt. Dies wird an dem sogenannten „De Morgan Argument“ illustriert („Alle Pferde sind Tiere. Also: Alle Pferdeköpfe sind Tierköpfe.“)

1. Introduction

Logical formalization is a standard tool for reconstructing informal arguments as well as for the logical analysis of ordinary language. The question of identifying criteria for correct formalization arises as soon as one considers alternative formalizations. The recent debate on criteria of adequate formalization demonstrates that this question is more difficult to answer than one might suppose from the familiar practice of formalization, cf. Brun(2003) as well as Blau(1977), Epstein(1990), Epstein(1994), Sainsbury(1993), Löffler(2006), Kleinknecht(2008), Baumgartner/Lampert(2008) and Lampert/ Baumgartner(2010). In particular, two questions arise: (i) how should established formalizations be reconstructed? and (ii) are the reasons underlying these formalizations persuasive? In this paper, these questions are discussed with respect to formalizations of the following well-known and allegedly trivial argument:

Premise (P): Horses are animals.
Conclusion (C): Heads of horses are heads of animals.

This argument is used in standard textbooks of logic to prove the insufficiency of monadic predicate logic. It is called “De Morgan's Argument” although De Morgan never used it in this form, cf. Merrill(1977) and Brun(2003), p. 189, footnote 1. In this paper, I refer to the discussion of its established formalization, which can be found in standard textbooks of logic such as Copi(1979), p. 131f.,

Lemmon(1998), p. 131f., Quine(1982) p. 168, 173, 251 and Suppes(1999), p. 93f. It is as follows:

ϕ(P): $\forall x\ (Hx \rightarrow Jx)$
ϕ(C): $\forall x(\exists y\ (Hy \wedge Ixy) \rightarrow \exists y(Jy \wedge Ixy))$

Realization 1:
Hx: x is a horse.
Jx: x is an animal.
Ixy: x is a head of y.

A formalization assigns a logical formula plus a realization to ordinary propositions. A realization interprets the categorematic parts of the formula, namely, its names, propositional variables and predicates.

According to the standard formalization, De Morgan's argument is valid. However, this formalization is hardly ever argued for. The detailed reasons for it are not obvious. In particular, the formalization of the conclusion is questioned. Presuming *realization 1*, Wengert(1974) suggests the following formalization:

$\phi(C)'$: $\forall x \forall y((Hy \wedge Ixy) \rightarrow (Jy \wedge Ixy))$

As will be seen, this alternative formalization raises the question of how extensively the formalization should take into account the relation of the concepts *horse* and *animal*. This is considered by Brun(2003). He discussed both of the mentioned alternatives and suggested a model of formalization according to which they are both correct. As soon as one acknowledges that questions of logical formalization imply conceptual analysis, it can also be questioned whether the formalization of the premise in terms of a formally invalid general implication is correct. At the end of this paper, an alternative formalization will be provided that formalizes the conclusion as well as the premise of De Morgan's argument in terms of a tautology.

All of the different formalizations are based upon different formalization models. These models imply different aims, differences within semantics and different formalization criteria. These differences will be spelled out in the following sections. I will only consider first-order logic formalizations without identity. Furthermore, I will confine the discussion to criteria for the correct formalization of single propositions. Criteria of correct formalization may be defined either depending on relations of implication or depending on interpretations of first-order formulae. With respect to the former, one distinguishes between the correctness and completeness of a formalization, cf. Baumgartner/Lampert(2008), p. 103-111. I will abstain from addressing completeness as well as from defining correctness with respect to relations of implication. Instead, I will only consider criteria of correctness with respect to interpretations of first-order formulae. The criteria of correctness of the different models of formalization discussed here differ with respect to the understanding of what is

meant by an interpretation of a logical formula. In addition to the criteria of correctness, the literature discusses criteria that refer to a certain structural similarity of the surface of the formalized proposition and its formalization. I will abstain from addressing these kinds of criteria as well. I will begin with Wengert's position, as it makes clear that the traditional formalization of De Morgan's argument requires justification.

2. Interpretations as free verbalizations

Who says "Heads of horses are heads of animals" does not merely mean *V(ϕ(C))* but *V(ϕ(C)')*:

V(ϕ(C)): "Every head of some horse is a head of some animal."
V(ϕ(C)'): "Every horse that has a head is an animal that has that head."

In contrast to *V(ϕ(C)')*, *V(ϕ(C))* is still true even if some (or all) horses were not animals but the heads of those beings were still the heads of animals. This difference also exists between the formalizations *ϕ(C)* and *ϕ(C)'*. In contrast to *ϕ(C)'*, there are models of *ϕ(C)* in which not all objects that satisfy *H* and are related to some object by *I* also satisfy *J*. *V(ϕ(C))* is a free verbalization of *ϕ(C)*, and *V(ϕ(C)')* is a free verbalization of *ϕ(C)'*. By use of the realizations, free verbalizations translate the logical formula into comprehensive ordinary propositions. From the fact that *(C)* is not equivalent to *V(ϕ(C))* but to *V(ϕ(C)')*, Wengert drew the conclusion that *(C)* is correctly formalized by ϕ(C)' and not by ϕ(C). He implicitly put forth the following criterion of correct formalization:

> *criterion of verbalization* (CRVERB): *ϕ(A)* is a correct formalization of a proposition *A* if and only if the free verbalization of *A*, *V(ϕ(A))*, and *A* are equivalent (have the same meaning).

According to this criterion, an "interpretation of a formula" is understood as the free verbalization of the formula by use of the realization. By "equivalence", it is not only meant that *A* and *V(ϕ(A))* have the same truth value. This condition is also satisfied in the case of (C) and *Vϕ(C))*. Instead, "equivalence" is meant in terms of "identity of meaning (content)". This kind of equivalence clearly depends on both the concepts contained in *A* as well as on the realization. Wengert alludes to this by comparing (C) with proposition (R), which is similar with respect to its grammatical structure:

(R): Children of mothers are children of fathers.

In this case, the following formalization is correct according to CRVERB:

ϕ(R): $\forall x(\exists y\ (Hy \wedge Ixy) \rightarrow \exists y(Jy \wedge Ixy))$

Realization 2:
Hx: *x* is a mother.
Jx: *x* is a father.
Ixy: *x* is a child of *y*.

The free verbalization is as follows:

V(ϕ(R)): "Every child of some mother is a child of some father."
V(ϕ(R)) has the same meaning as (R). In contrast, the replacement of *ϕ(R)* with *ϕ(C)'* (= *ϕ(R)'*) results in an incorrect formalization, because in this case, *V(ϕ(R)')* would result in "Every mother that has a child is a father that has that child". Or, as Wengert(1974), p. 166 stated briefly, "Everybody's mother is his father". Obviously this is not what is meant by (R).

According to CRVERB, different logical formalizations of grammatically similar propositions might be correct. According to this point of view, one of the benefits of formalization is that it can express differences in logical form that are not expressed by the syntax of ordinary language.

The following example is compatible with both formulae, *ϕ(C)* and *ϕ(C)'*, in addition to the respective realization (cf. p. 20):

(S): Bets on winning numbers are bets on prime numbers.

(S) is ambiguous. CRVERB makes the two possible interpretations explicit: Either it means "Every bet on some winning number is also a bet on some prime number" or it means "Every bet on a winning number is a bet on a prime number". The benefit of logical formalization according to CRVERB is found in expressing this difference in meaning through a syntactic difference in logical formulae.

In fact, CRVERB is frequently used as a tool for judging formalizations. However, CRVERB faces the following problems:

problem 1: No mechanical procedure exists to generate free verbalizations. One therefore runs the risk of generating a verbalization in light of the proposition to be formalized. In consequence, CRVERB cannot serve as an independent criterion. Even referring to an explicit paraphrase of the logical formula would not achieve a criterion. This is because it is generally not possible to judge the equivalence of a proposition to be formalized and an explicit paraphrase of a logical formula. In fact, the question "which of the alternative paraphrases is equivalent to *A*?" is as problematic as the question "which of the paraphrased alternative formalizations is correct?"

problem 2: Judgments of equivalence or identity of meaning are presumed to be primitive. However, as will be seen with respect to alternative formalizations of De Morgan's argument, such judgments require further explication or justification.

As Brun (2003), p. 200 notes, one runs the risk of shifting the problem of identifying a correct formalization to the problem of identifying a correct free verbalization. As long as no algorithm is available to precisely define how to generate verbalizations, one cannot speak of a *criterion* in a strict sense. Rather, one might speak of a device that does not yield a definite decision in any case.

However, CRVERB expresses a certain intuition: correct formalizations should not only correspond to the truth value of the proposition *A* to be formalized, but also to the *meaning* of *A*. According to this understanding, the logical formula represents the *form of the meaning* of *A*. The problem of CRVERB is how to explain the meaning of a proposition in this context. I will come back to this problem in sections 4 and 5.

3. Interpretations as realizations

The application of CRVERB to (C) of De Morgan's argument raises the question of how the traditional formalization of (C) by means of $\phi(C)$ can be justified. To argue that this formalization is based on CRVERB but an understanding of (C) in terms of $V(\phi(C))$ is not persuasive. Wengert drew the conclusion that the traditional formalization cannot be justified and must be given up in favor of $\phi(C)'$. Brun likewise did not provide an argument in favor of $\phi(C)$ and against $\phi(C)'$. In contrast to Wengert and Brun, this section shows how to reconstruct the traditional practice. I will distinguish between two different strategies of formalization that both justify the formalization of (C) by $\phi(C)$. The former is typical for philosophical traditions within logic, while the latter applies common strategies of formalization within mathematical logic. Although this strategy might not be as evident as others in the context of formalizing De Morgan's argument, I refer to it in order to illustrate different aims and criteria of logical formalization.

3.1 Varying interpretations

Within the philosophical tradition of logic textbooks, the aim of formalization is mostly to evaluate the formal validity of informal arguments. Examples of formalizations within this tradition go hand in hand with the distinction between form and content. Typically one refers to plain informally valid arguments, such as certain types of Aristotelian syllogisms, to motivate predicate logic as a necessary enhancement of propositional logic. Then, one notes that the validity of those arguments does not depend on the specific content of the concepts. This is done by replacing the predicates of the argument with other predicates with a different meaning. If one abstains from the specific predicates, one yields a schematization of the informal arguments. According to the standard view, the logical form of the proposition corresponds to this schematization. The aim of

formalization in this tradition is to abstain from the specific meaning of the predicates. This is incompatible with CRVERB, as this criterion refers to the specific meaning of the predicates within the realization in order to judge the identity of meaning of *A* and *V(ϕ(A))*. According to the traditional view, a given realization is just one of many other possible interpretations of *ϕ(A)* that must be considered. In order to judge the correctness of a formalization, one must refer to varying interpretations / realizations. The realizations / interpretations regulate how to substitute propositional variables, predicates and names within the schematization of the proposition to be formalized. One thereby compares the schematization of *A*, *sch(A)*, and *ϕ(A)* in accord with varying realizations / interpretations. In contrast to CRVERB, one refers to purely extensional judgments of equivalence: *sch(A)* and *ϕ(A)* must have the same truth value according to their respective interpretations. The interpretations of names, $\mathfrak{I}(t)$, of propositions, $\mathfrak{I}(A)$, and of predicates, $\mathfrak{I}(\varphi)$, are identified with their respective extensions: $\mathfrak{I}(t)$ is an object, $\mathfrak{I}(A)$ is a truth value, and $\mathfrak{I}(\varphi)$ is a class of objects.

Let us illustrate this approach by considering the formalization of (C). Like (R) and (S), (C) is an instance of *sch(C)*:

sch(C): All Xs of Ys are Xs of Zs.

Strictly speaking, schematizations may presume changes in structure and wording of the respective propositions, cf. p. 11 below. Considering the correctness of *ϕ(C)'* for (C), it does not suffice to evaluate whether *sch(C)* and *ϕ(C)'* have the same truth value according to *realization 1*. In fact, one may admit that (C) is to be understood in terms of *V(ϕ(C)')*. However, judging the correctness of *ϕ(C)'* also requires considering whether *sch(C)* and *ϕ(C)'* have the same truth value according to *realization 2*. This is not the case; (R) is true while *realization 2* is not a model of *ϕ(C)'*. For this reason, *ϕ(C)'*is not a correct formalization of (C). In contrast, the truth value of *sch(C)* corresponds to the truth value of *ϕ(C)* relative to both realizations. Whereas *ϕ(C)* represents the form of (C), *ϕ(C)'* does not represent the form of (C); *ϕ(C)'* does not correspond to *sch(C)*. One erroneously considers semantic aspects if one chooses *ϕ(C)'*. However, this model of formalization requires that logical formalization must be independent of the specific meaning of the categorematic parts of propositions.

The traditional formalization of De Morgan's argument is based on the following criterion:

> *varying-$\mathfrak{I}$ criterion* (V$\mathfrak{I}$CR): *ϕ(A)* is a correct formalization of a proposition *A* if and only if the schematization of *A*, *sch(A)*, and *ϕ(A)* have the same truth value relative to all interpretations / realizations.

In contrast to CRVERB, the intention of this criterion is to evaluate the correctness of formalizations merely with respect to formal equivalence. Any implication based on conceptual relations must not be considered. De Morgan's argument can still be proven as (formally) valid according to this approach. Howev-

er, the class of arguments that are proven as valid according to this strategy is significantly smaller than it is according to formalization strategies not based on schematization. The following example illustrates this:

argument B: Heads of horses are heads of animals. The horse Fury has a head. Therefore: Fury is an animal.

According to a pre-theoretic, informal understanding of validity, *argument B* is valid. This means that the truth of the premises is incompatible with the falsehood of the conclusion. How to explicate this validity is another question. The traditional strategy of formalization in compliance with VℑCR does not make it possible to explicate *argument B* as a valid argument in terms of first-order logic. The reason for this is that its schematization allows for invalid instances such as the following:

argument C: Children of mothers are children of fathers. The mother Jane has a child. Therefore: Jane is a father.

In contrast, CRVERB allows one to prove the validity of *argument B* by means of first-order logic without thus implying that *argument C* is valid. This shows that VℑCR goes hand in hand with a loss of the power to explicate the validity of arguments by means of first-order logic.

Other examples that are praised as achievements of the logical analysis of ordinary language show that VℑCR trivializes the problem of formalizing ordinary propositions. Consider, for example, Davidson's analysis of action sentences:

argument D: Ann strolls slowly. Therefore: Ann strolls.

Davidson's strategy makes it possible to formalize this argument as valid according to first-order logic:

ϕ(D): $\exists x\,(Fx \wedge Gx \wedge Ixa) \vdash \exists x\,(Fx \wedge Ixa)$

Realization 3:
Fx: *x* is a stroll.
Gx: *x* is slowly.
Ixy: *x* is conducted by *y*.
a: Ann.

However, this formalization is not correct according to VℑCR. This can be seen by replacing “slowly” with “allegedly”.

There are two strategies for avoiding this objectionable consequence of trivializing formalization: (i) the limitation of admissible interpretations and (ii) standardization. Strategy (i), for example, is applied if one rules out the substitution of “slowly” by “allegedly” in *argument D*. The same strategy is applied in the following case:

argument E: John loves a human. Therefore: Some human exists.
argument F: John seeks a unicorn. Therefore: Some unicorn exists.

Argument E is valid, whereas *argument F* is not. The validity of *argument E* can be proven by the following formalization:

ϕ(E): $\exists x\,(Fx \wedge Gax) \vdash \exists x\,Fx$

Realization 4:
Fx: *x* is a human.
Gxy: *x* loves *y*.
a: John.

According to CRVERB, this is a correct formalization of *argument E*. However, *ϕ(E)* with a respective realization is not a correct formalization of *argument F*, because "John seeks a unicorn" and "Some unicorn exists that John seeks" are not identical in meaning. In contrast to the latter, the former does not imply the existence of some unicorn.

According to V∃CR, however, one cannot make this argument. The replacement of "love" with "seek" and "human" with "unicorn" results in an invalid argument. Thus, *ϕ(E)* cannot be a correct formalization according to V∃CR unless one rules out this alternative interpretation. Indeed, this interpretation is commonly excluded by arguing that "to seek" is not a predicate obeying the principle of extensionality. This is demonstrated by the fact that "*x* seeks *y*" might be true even in case that no *y* exists to be sought. Yet, this raises the problem of how to define a criterion for distinguishing between admissible and inadmissible interpretations. This question also becomes relevant with respect to the possibility of formalizing paradoxes. Is "*x* is a member of *x*", for example, an admissible predicate as one assumes by formalizing Russell's Paradox within predicate logic? The question is whether the two commonly presumed semantic principles, namely, the principle of extensionality and the principle of bivalence, suffice to rule out inadmissible interpretations. I will return to this question in section 6.

In addition to strategy (i), strategy (ii), namely standardization, is indispensable for applying V∃CR. Standardization indicates the rephrasing of ordinary propositions to the effect that their surface grammar becomes more similar to logical formulae. For example, one may rephrase *argument D* as follows: "Some event *x* exists such that *x* is a stroll and *x* is slowly and *x* is conducted by Ann. Therefore: Some event *x* exists such that *x* is a stroll and *x* is conducted by Ann." However, the question of a correct formalization is clearly shifted to the question of correct standardization. It is also apparent that standardization is necessary if one refers to schematization. Consider, for example, (C), (R) and (S): rephrases are necessary to schematize all of these by "All Xs of Ys are Xs of Zs".

If one dispensed with strategy (ii), only arguments that are explicit paraphrases of formally valid inferences could be proven as logically valid. As a

consequence, formalization would become an unprofitable endeavor. Nearly all pre-theoretically valid arguments had to be classified as "semantically valid". This is not in fact what is done. In sum, if one applies V$\mathfrak{I}$CR, strategies (i) and / or (ii) are almost always used. However, if this is conceded, one can hardly sustain the distinction of formally and semantically valid arguments. Instead, one should regard reducing the validity of arguments, if possible, to formally valid formalizations as a first task of logical formalization.

The problems of V$\mathfrak{I}$CR result from the reference to schematizations, which are indispensable if one intends to pass judgment upon the correctness of formalizations due to varying interpretations. Like CRVERB, all of the following models of formalization do not depend on schematizations and do not vary realizations.

3.2 Fixed interpretations

As in the philosophical tradition of logic, mathematical logic refers to realizations in terms of interpretations and presumes a purely extensional conception of interpretations. However, in contrast to the philosophical tradition, it is referred to a logical language that assigns a fixed interpretation not only to so-called "logical constants" but to *all* constituents of the formal language. The question of formalization thus becomes a question of translation (or encoding).

The objective of this model of formalization is to prove the truth of propositions by their derivation from axioms (or the falsity of propositions by deriving their negation from axioms). This objective would be satisfied if any true proposition is formalized by $P \vee \neg P$ (and any false one by $P \wedge \neg P$). However, the problem is that this would presume knowledge of the truth values. This is exactly what one wants to find by deriving the proposition or its negation within an axiomatic system. To do so, one must translate the proposition into a proposition of the presumed logical language. A logical language L is understood as a pair consisting of the recursively defined formulae, *L*, and the fixed interpretation, $\mathfrak{I}$, of all expressions of the alphabet of *L*. Formalization consists in translating ordinary propositions into propositions of a presumed logical language with a limited vocabulary. Thus, the objective is to prove the truth or falsehood of as many propositions that can be expressed by L as possible. These proofs are carried out within an axiomatic theory T that comprises only a limited number of non-logical axioms. Gödel has shown that this ideal cannot be realized to its full extent for basic arithmetic if basic arithmetic is formalized according to this model of logical formalization. In order to address this conception of formalization, predicates of ordinary language must first be expressed by predicates of L. This is done according to the following criterion:

express-criterion (ECR): A predicate *Px* of ordinary language is expressed by $\varphi(x)$ of L if and only if for all *x*
if an object satisfies *Px*, then $\varphi(x)$ is true,
if an object does not satisfy *Px*, then $\neg\varphi(x)$ is true.

The application of ECR does not presume that the truth value of *Px* is known for every object. Rather, it is presumed that one can judge whether *Px* and $\varphi(x)$ have the same extension according to $\mathfrak{I}$.

In the case of formalizing De Morgan's argument along these lines, one must presume a logical language with the predicates *Hx*, *Jx* and *Ixy* and their fixed interpretations in terms of *realization 1*. Thus, $\mathfrak{I}(H)$ is the extension of "_ is a horse", $\mathfrak{I}(J)$ is the extension of "_ is an animal", and $\mathfrak{I}(I)$ is the extension of "_ is a head of _". According to ECR and L, the predicates "*x* is a head of a horse" and "*x* is a head of an animal", which occur in (C), are to be translated as follows:

x is a head of a horse $=_{\text{Def}} \exists y\,(Hy \wedge Ixy)$
x is a head of an animal $=_{\text{Def}} \exists y\,(Jy \wedge Ixy)$

In contrast, $\forall y(Hy \wedge Ixy)$ or $\forall y(Jy \wedge Ixy)$ would not satisfy ECR.

To translate (C) into an expression in L, (C) must be rephrased as a proposition that can be effectively translated to L. (C) is of the form "Ys are Zs". Propositions of this form are translated to general implications of the form "For all *x*, if *x* is *Y*, then *x* is *Z*". This results in the following rephrase of (C), (C*) "For all *x*, if *x* is a head of a horse, then *x* is a head of an animal". To translate (C*) to L, the definitions of "*x* is a head of a horse" and "*x* is a head of an animal" must be applied. This results in $\phi(C)$: $\forall x\,(\exists y\,(Hy \wedge Ixy) \rightarrow \exists y\,(Jy \wedge Ixy))$. This is an effective translation of (C**) "For all *x*, if some *y* exists such that *y* is a horse and *x* is a head of *y*, then some *y* exists such that *y* is an animal and *x* is a head of *y*". (C**) is obtained from (C) by expressing (C) by means of the vocabulary of L. The formalization of (C) by $\phi(C)$ is based on the following criterion:

fixed-$\mathfrak{I}$ *criterion* (F$\mathfrak{I}$CR): $\phi(A)$ of a logical language L with a fixed interpretation $\mathfrak{I}$, <L, $\mathfrak{I}$>, is a correct formalization of a proposition *A* if and only if $\Phi(A)$ is an effective translation of *A* or of a rephrasing of *A* in respect to <L, $\mathfrak{I}$>.

$\phi(C)$ (or $\neg\phi(C)$) is to be proven by derivation from axioms. For this sake, it suffices to introduce $\forall x(Hx \rightarrow Jx)$ as a non-logical axiom. This axiom expresses the relation of $\mathfrak{I}(H)$ and $\mathfrak{I}(J)$. The class of horses is a subclass of the class of animals. As a consequence, (C) is true as $\phi(C)$ is derivable from the axiom $\forall x\,(Hx \rightarrow Jx)$. Thus, the following criterion is satisfied:

capture-criterion (CCR): A proposition *A* is captured by an axiomatic theory T if and only if
if *A* is true, then $T \vdash \phi(A)$,
if *A* is false, then $T \vdash \neg\phi(A)$.

Thus, the aim of this model is satisfied in the case of De Morgan's argument as $\phi(C)$ follows from T, namely, $\forall x\,(Hx \rightarrow Jx)$.

This model of formalization does not intend to represent the meaning of a proposition or its truth conditions. This can be seen by the fact that ECR is still satisfied if arbitrary formulae that are true according to the fixed interpretation are added by conjunction. "*x* is the head of an animal", for example, could also be expressed by $\exists y(Jy \wedge Ixy) \wedge \exists y \exists z(Hz \wedge Iyz)$. The same holds in the case of CCR and theorems of T that are added by conjunction. ECR and CCR are purely based on extensional considerations, cf. Smith(2007), p. 33-36.

In contrast to philosophically motivated models of formalization, the use of formalization within mathematical logic is not motivated by logical analysis of ordinary language or by proofs of the formal validity of informal arguments. Proofs of validity are only considered within the framework of an axiomatic theory T. If some informally valid argument cannot be proven as formally valid, this is not a deficiency of the formalization but rather of T. For example, the conclusion of *argument B*, cf. p. 9, follows trivially from the second premise and the axiom that all horses are animals. This can be shown by the fact that the formalization of the second premise implies *Ha* (= "Fury is a horse"). From this and the formalization of the axiom, $\forall x(Hx \rightarrow Jx)$, the formula *Ja* (= "Fury is an animal") follows. Within the framework of T, the problem does not arise that the conclusion is not derivable. As it is not referred to schematization, the problem of distinguishing between formally and semantically valid arguments does not arise either.

As a matter of fact, philosophical requirements of a theory of formalization are not fulfilled by this mathematical model of formalization. One reason for this is that the problem of formalization is trivialized due to the presumption of an effective translation of the propositions in questions or of their rephrasing. Controversial formalizations of propositions that cannot be effectively translated into a logical language cannot be resolved by referring to F$\mathfrak{I}$CR. For example, one cannot argue with respect to F$\mathfrak{I}$CR why "Smith died because he ate tomato sorbet" is not correctly formalized by $P \wedge Q$ with "$\mathfrak{I}(P)$ = Smith died" and "$\mathfrak{I}(Q)$ = Smith ate tomato sorbet". However, this model of formalization does not claim to fulfill such aims of other models of formalization. Only propositions that can be expressed within the vocabulary of L and that can be translated effectively to L are considered. By making use of logical formalization within mathematics, the only propositions that are considered are more or less standardized and are capable of being effectively translated to a logical language with a suitable vocabulary. A problem of formalization in which one has to identify the correct logical form of the propositions in the first place is not recognized within the framework of this model.

However, as I will explain in section 6, this also marks the problem of this model of formalization, even in applying it to mathematical propositions. The

grammatical form of declarative sentences is seen as sufficient reason for their capability of being true or false and thus of their capability of being logically formalized. Whatever can be effectively translated to a logical language seems to be logically correct. Logical formalization is not a means of identifying fallacies that stem from taking the apparent form of ordinary propositions as their real, logical form. As we will see, it cannot be excluded on this basis that paradoxes relying on meaningless, inadmissible interpretations are expressed by apparently meaningful logical formulae. This problem is shared by F$\mathfrak{I}$CR and V$\mathfrak{I}$CR. Neither one provides a sufficient criterion for distinguishing admissible and inadmissible interpretations. As a consequence, they cannot distinguish between proofs by reduction, which prove the incompatibility or falsehood of axioms under the presumption of a correct formalization, and paradoxes, which rely on incorrect logical formalizations of meaningless propositions. We will come back to this in section 6.

4. Interpretations as restricted truth conditions

The following two models of formalization to be discussed in this and the following section adhere to the philosophical tradition of formalizing ordinary propositions independent of a logical language with a fixed interpretation and independent of axioms of a theory T. As opposed to the philosophical tradition of logic textbooks, the aim of these two models is not restricted to the proof of the validity of informal arguments. Instead, they are concerned with the logical analysis of ordinary language. They share the intuition underlying CRVERB: logical formalization serves to logically analyze the meaning of ordinary propositions. However, this is made precise within a framework of semantics that understands interpretations of a logical formula as an expression of *truth conditions* of meaningful propositions. In this respect, they differ from the semantics underlying the models described in the previous section. Both of the following models presume that formal and semantic validity cannot be reasonably distinguished. As a consequence, they account for the internal relation between the concepts *horse* and *animal* by formalizing De Morgan's argument.

In contrast to Wengert and the tradition of logical textbooks, Blau and Brun articulated criteria of formalization. Their criterion of correctness is as follows, cf. Brun(2004), p. 208:

> *TC*: $\phi(A)$ is a correct formalization of proposition *A* if and only if $\phi(A)$ and *A* have the same truth values with respect to all interpretations in terms of conditions of truth or falsehood of *A* that are suitable according to the realization.

TC differs from V$\mathfrak{I}$CR in essentially two respects: (i) it refers to interpretations in terms of conditions of truth and falsehood, and (ii) it understands the truth conditions / interpretations as a function of the realization, cf. Brun(2003), p.

209-211. While VℑCR considers the truth value of *different* propositions, *TC* refers to truth values of the *same* proposition with respect to *different* conditions. This determines a conception of interpretations that deviates significantly from that presumed in VℑCR or FℑCR. Interpretations articulate truth conditions of formulae and propositions. They represent *possible* extensions of categorematic parts, namely, possible extensions of predicates, possible truth values of atomic propositions and possible references of names. In contrast, VℑCR and FℑCR presume that any interpretation refers to the actual extension of a predicate, atomic proposition or name. For example, if one interprets *P* by "Paris is the capital of France", then $\mathfrak{I}(P) = T$ according to traditional semantics, as Paris is indeed the capital of France. According to *TC*, however, $\mathfrak{I}(P)=T$ and $\mathfrak{I}(P)=F$ are two possible truth values of the atomic proposition "Paris is the capital of France". Thus, according to this semantics, the interpretations vary in terms of possible extensions of predicates, atomic propositions and names, while the predicates, atomic propositions and names are assigned to fixed expressions by the realization. According to this understanding, the question in each case is how the truth value of the proposition to be formalized and the truth value of the formula depend on the respective possible extensions of the categorematic constituents.

According to Blau's and Brun's model of formalization, it may happen that certain extensions of the categorematic parts are unsuitable interpretations, as they constitute unsuitable conditions of truth or falsehood due to their meaning. For example, the realizations of the formalizations $\phi(C)$ and $\phi(C)'$ both refer to the concepts *horse* and *animal*. Due to the meaning of these two concepts, interpretations in which the class of horses is not a subclass of the class of animals are unsuitable. This takes into account the conceptual, internal relation of horses and animals according to which it is impossible that some horse is not an animal. This is no possible condition that allows one to judge the truth or falsehood of (C). Interpretations in terms of varying possible extensions of fixed concepts (propositions, names) only determine consistent conditions of truth or falsehood if "unsuitable interpretations" are not taken into account. That is why *TC* refers to suitable interpretations as a function of the realizations. From this, it follows that, in the case of formalizing (C), the logical difference between $\phi(C)$ and $\phi(C)'$ is of no consequence: the interpretations that are models of $\phi(C)$ but not models of $\phi(C)'$ are unsuitable. For this reason, Brun qualifies both $\phi(C)'$ and $\phi(C)$ as correct formalizations of (C). Both $\phi(C)$ and $\phi(C)'$ have the same truth value that (C) has with respect to all suitable interpretations.

Compared to CRVERB, this conception has the advantage that it explicates the equivalence or identity of meaning of the formalization and of the proposition in question in terms of identity of truth conditions. Regarding VℑCR, *TC* does not refer to the problematic schematization of ordinary propositions. However, the formalization of (C) reveals a fundamental problem of this conception.

Non-equivalent formulae, such as $\phi(C)$ and $\phi(C)'$, are equivalent with respect to the restricted space of possible interpretations. This shows that this conception is incompatible with the traditional interpretation of first-order formulae. Finally, not only the interpretations but also the truth conditions and logical relations of first-order formulae become a function of the realizations and, thus, of the proposition to be formalized. In fact, one does not explicate the truth conditions and logical relations of ordinary propositions by means of formalizations *within first-order logic*. This problem is labeled "the problem of suitable interpretation"', cf. Baumgartner/Lampert(2008).

Another problematic consequence of the impossibility of identifying only one correct formula out of multiple non-equivalent formulae is the impossibility of deciding upon the validity of arguments in certain cases. For example, this problem arises in the case of formalizing *argument B*, cf. p. 9. The formalization of (C) by $\phi(C)$ results in an invalid formalization (according to standard first-order logic), while the formalization of (C) with $\phi(C)'$ results in a valid formalization. A "problem of validity" arises from this and further assumptions of this model of formalization, namely, the problem of identifying the validity of certain valid arguments, cf. Lampert/Baumgartner(2010). This problem complements the so-called "problem of invalidity" identified by Massey(1975). This problem results from the fact that formally invalid formalizations may be correct for valid arguments, while no criterion is available to constrain the class of potentially correct formalizations to be finite. Thus, it is impossible to conclude the invalidity of the formalized argument from a correct and invalid formalization. In the following section, I propose a modification of *TC* that solves all of the aforementioned problems.

5. Interpretations as unrestricted truth conditions

The "problem of suitable interpretations" arises from the fact that the space of possible interpretations is restricted by logical dependencies that are due to the realization's concepts. This problem does not arise if one claims that the concepts of a realization must be logically independent. This means that the space of possible interpretations is unrestricted. I call realizations satisfying this condition "suitable realizations". The semantics based on interpretations in terms of conditions of truth or falsehood must rely on their logical independence. Traditional semantics are based on the principles of bivalence and extensionality. The principle of bivalence states that any proposition is either true or false. The principle of extensionality states that the truth or falsehood of any proposition depends on nothing but the extension of the categorematic parts. Within semantics referring to truth conditions, these two principles must be complemented by the principle of logical independence. This principle claims that all interpretations

that may be generated by combinatorial means are also possible. Thus, if *n* propositional variables occur in the realization, 2^n interpretations are admissible. If *u* predicates with arity $k_1 \ldots k_u$ occur in the realization, $2^{i^{k1}} + \ldots + {}^{i^{ku}}$ interpretations must be possible with respect to a domain with *i* objects. Any name can be interpreted by any of these *i* objects. According to this understanding, the principle of independence implies the principle of bipolarity, which states that any atomic proposition *may* be either true or false. The principle of independence is indispensable for any theory of formalization based upon both: (i) semantics of truth conditions and (ii) standard first-order logic (without identity). Within this conception, the suitable realizations assign meaningful expressions to the categorematic parts of the formulae. These assignments are fixed, while their interpretations in terms of their possible extensions vary without any restrictions. Only truth conditional semantics relying on the principle of independence make a logical analysis of ordinary language possible, which reduces and explicates informal logical dependencies within ordinary language to formal logical dependencies of first-order formulae. To satisfy this claim, no internal, logical dependencies must exist between the categorematic parts of the realizations.

The criterion of correctness established by this model results from modifying *TC*:

> *TC'*: $\phi(A)$ is a correct formalization of a proposition *A* if and only if $\phi(A)$ and *A* have the same truth values with respect to suitable realizations that do not restrict the space of possible interpretations in terms of conditions of truth and falsehood of *A*.

Let us illustrate the application of this criterion by formalizing De Morgan's argument. Like Brun, it is assumed that the concept *horse* contains the concept *animal*. In contrast to all other models of formalization, *TC'* thus rules out realizations that contain both of these two concepts. Rather, *TC'* reduces this conceptual implication to a formal one. For the following, it is assumed that horses are defined as animals with a certain differentia specifica, say "tiptoeing equid". We also presume that it is possible to be a tiptoeing equid but not an animal. This may, in fact, be false, but it shall not be excluded by the meaning of the concepts. According to these assumptions, the following formalization of De Morgan's argument is correct according to *TC'*:

(P): Horses are animals.
$\phi(P)_{TC'}$: $\forall x((Gx \wedge Jx) \rightarrow Jx)$.

(C): Heads of horses are heads of animals.
$\phi(C)_{TC'}$: $\forall x((Gx \wedge Jx \wedge Ix) \rightarrow (Jx \wedge Ix))$

Suitable realization:
Gx: *x* is a tiptoeing equid.
Jx: *x* is an animal.
Ix: *x* has a head.

It would also be possible to replace *I*x with the dyadic predicate *Iyx*, representing "*y* is a head of *x*". Regardless of whether *y* would be bound by an existential quantifier or a universal quantifier, the resulting formula would be a tautology. However, in contrast to all of the other proposals for formalizing De Morgan's argument, it is not necessary to introduce a dyadic predicate to derive (C). According to *TC'*, both the premise and the conclusion of De Morgan's argument are tautologies if one presumes that horses are defined as animals. This seems unexpected, as De Morgan's argument suggests that (C) follows from (P) and not from any premise. However, the presumed conceptual relation between the concepts *horse* and *animal* is already contained in the conclusion. This is taken into account by $\phi(C)_{TC'}$ by virtue of representing the concept *horse* by $Gx \wedge Jx$ and the concept *animal* by Jx. According to this understanding, it is still adequate to say that heads of horses are heads of animals *because* horses are animals. However, this justification is not expressed by a formally valid inference from a premise that is assumed to be true. Instead, this internal relation is expressed within the formalization of (C). Thus, (P) is not a falsifiable proposition on which the truth of the conclusion depends. Instead, (P) articulates a conceptual relation that must be taken into account if the truth conditions of (C) are to be correctly represented.

In contrast, the similar propositions (R), cf. p. 5, and (S), cf. p. 6, are not to be formalized by tautologies. A *TC'*-correct formalization of (R) is the following:

(R): Children of mothers are children of fathers.
$\phi(R)_{TC'}$: $\forall x(\exists y\,(Hy \wedge Ixy) \rightarrow \exists y(\neg Hy \wedge Ixy))$

Suitable realization:
Hx: *x* is a woman.
Ixy: *x* is a child of *y*.

This formalization presumes that no object exists that (i) is neither a woman nor a man and (ii) is a woman and a man. This is plausible on the basis of human beings as domains. Thus, men are definable by "*x* is not a woman".

(S): Bets on winning numbers are bets on prime numbers.
$\phi(S)_{TC'}$: $\forall x \forall y((Hy \wedge Ixy) \rightarrow (Jy \wedge Ixy))$

Suitable realization:
Hx: *x* is a winning number.
Jx: *x* is a prime number.
Ixy: *x* is a bet on *y*.

This formalization expresses that (S) means that any bet on a winning number is also a bet on a prime number. In this sense, the concepts *winning numbers* and *prime numbers* are logically independent, but as a matter of fact all winning numbers are prime numbers. On the other hand, if (S) means that whoever bets on a winning number also bets on a prime number, the following formalization with the respective realization would be *TC'*-correct:

$\phi(S)_{TC'}$: $\forall x(\exists y\,(Hy \wedge Ixy) \rightarrow \exists y(Jy \wedge Ixy))$

Thus, one should illustrate the limits of monadic first-order logic by referring to an argument based on (S) rather than referring to De Morgan's argument.

The key point is that the respective formalization is understood as an explication of the truth conditions of the proposition to be formalized. It is by no means presumed that the formalization of an ordinary proposition must be unambiguous. Rather, logical formalization presents a means for unambiguously expressing the respective meaning. It is also not assumed that the mentioned *TC'*-correct formalization of De Morgan's argument is the only possible correct explication of (P) and (C). Rather, the formalization depends on the assumed relation of the concepts *horse* and *animal*. If one does not assume that these two concepts are logically dependent, *TC'* claims a different formalization. If it is assumed that it is not meaningless that horses exist that are not animals, the common formalizations might well be *TC'*-correct. With this assumption, Wengert's formalization, $\phi(C)'$, is *TC'*-correct if it means that any horse with a head is an animal with that respective head. The traditional formalization, $\phi(C)$, is correct if it simply means that whenever something is the head of a horse it is also the head of an animal. Of course, the stronger claim follows given the assumption that all horses are animals. However, this does not mean that this stronger claim is inferred.

The purpose of logical formalization according to this model is to explicate the truth conditions of ordinary propositions. This model explains precisely what is meant by a formal representation of the meaning of propositions; namely, the representation of their truth conditions by a logical formula according to suitable realizations. It is important to note that the space of possible interpretations is defined independent of and prior to the categorematic parts of suitable interpretations. The form of propositions representable by first-order logic is thus not defined by ordinary language; one needs to refer neither to the grammatical surface nor to some assumed deep structure of ordinary propositions. Instead, the form of propositions is provided by first-order logic, namely, by its ability to represent truth conditions. Logical formalizations determine whether some ordinary proposition has the form of a proposition with precisely defined truth conditions according to logic.

The vagueness of ordinary language is no objection against the conception of logical formalization in terms of the logical analysis of ordinary language. Rather, this vagueness motivates the task of logical formalization. It is also not presumed that ordinary propositions must have definite truth conditions with respect to context and speaker. The purpose of logical formalization is to express possible meanings of ordinary propositions. This is compatible with the fact that no purported formalization expresses what is meant by some proposition. In this case, what is not meant by the proposition is at least clarified. This leaves the question open as to whether it is at all possible to express the meaning of a

proposition within first-order logic. In any case, logical formalization fulfils the purpose of explicating the meaning of propositions by providing the means to precisely explain their truth conditions.

TC' overcomes the problems of the other models of formalization. In contrast to *TC*, the problems of validity and invalidity are solved in addition to the problem of suitable interpretations. This is because by claiming identical truth conditions with respect to suitable realizations, valid arguments are formalized correctly if and only if their formalization is formally valid. The problems of VℑCR result from the reference to schematization. The semantics according to which interpretations represent conditions of the truth and falsehood of propositions make this reference superfluous. In contrast to VℑCR, *TC'* reduces informal logical dependencies of ordinary concepts and propositions to formal relations within logic. Furthermore, in contrast to VℑCR and FℑCR, TC' is not in need of a criterion for admissible interpretations. Any interpretation that is possible according to combinatorial means is also admissible, and it represents a condition that allows one to judge the truth value of a proposition. If this claim is not satisfied, then the claim of suitable realizations is also not satisfied. As a consequence, the formalization is not *TC'*-correct. If some proposition does not have truth conditions that can be represented within logic according to *TC'*, it is meaningless according to logic. The crucial advantage of the truth conditional semantics, as compared to the semantics that VℑCR and FℑCR rely on, is that they do not refer to *actual* extensions, but more basically to *possible* extensions. Thus, it is possible to use logical formalization as a means of determining the meaning of propositions rather than simply assuming it. Furthermore, standardizations are not assumed but instead result from paraphrases of *TC'*-correct formalizations. In contrast to CRVERB, judgments of equivalence are not assumed to be primitive. Rather, they are based upon explications of truth conditions, which, in turn, refer to the construction of possible interpretations. It is even possible to refer to an effective, mechanical procedure to generate verbalizations in terms of explications of truth conditions. At least this is possible in so far as it is possible to not only enumerate single models (conditions of truth) and counter-models (conditions of falsehood), but also to identify the class of models and counter-models by certain distributive normal forms, cf. Lampert(2006).

The presumed semantics of *TC'* as well as the objective of explaining the truth conditions of ordinary propositions are rooted in the philosophical tradition of logical analysis of ordinary language. The most decisive articulation of this conception can be found in Wittgenstein's *Tractatus*. The main reason that this model of formalization is rarely articulated and is not applied to all consequences lies in the strong claim of suitable realizations. The principle of logical independence seems to be unsatisfiable with respect to logical dependencies of ordinary language expressions. Thus, like VℑCR, *TC'* runs the risk of unduly restricting the realm of informal arguments that can be formalized within first-

order language. For example, should one decline to formalize “All men are beings. All beings are mortal. Therefore, all men are mortal” by means of a syllogismus barbara merely because the respective realization is not suitable? Or should one question the formalization of an inference by means of a modus ponens simply because, by adding an arbitrary premise, the principle of logical independence is no longer satisfied? Finally, does *TC'* not rely on the unrealistic presumption that propositions of ordinary language can be analyzed as a truth function of logically independent, bipolar atomic propositions?

Wittgenstein dealt with questions like these in his *Notebooks from 1914-16*. How can logic be applied to ordinary language, he asked, if one cannot carry out the complete analysis of ordinary propositions, cf., for example, Wittgenstein(1995), diary entries from 3.9.-7.9.1914, 11.10.1914, 20.6.-22.6.1915? On the one hand, Wittgenstein assumed in the *Tractatus* that the real, logical form of ordinary propositions can only be revealed by analyzing them in a truth function of logically independent, bipolar atomic propositions, cf. Wittgenstein(1995), remark 5. On the other hand, he assumed that logic can be applied to unanalyzed propositions in so far as their categorematic parts can be treated as if they were primitive and logically independent, cf., for example, Wittgenstein(1995), entries form 11.10.1914[2], 21.6.1915[10]. The mentioned disagreeable consequences can be avoided if the demand of reducing informal logical dependencies to formal logical relations is met in a pragmatic and context-dependent way. From the possibility of identifying further informal logical dependencies by formal ones, it does not follow that one must do so. One may well treat expressions of the realization as if they were primitive and logically independent, even if they might be capable of some further analysis. The application of *TC'* achieves neither more nor less than making explicit the internal relations that follow from this assumption. Any further detailed analysis can only reveal more internal relations and thus invoke a more thorough understanding of the meaning of the proposition to be formalized. However, no further detailed analysis can revise identified internal relations that are already identified by a superficial analysis. For example, one can accept the traditional formalizations of De Morgan's argument, $\phi(C)$ or $\phi(C)'$, as an expression of the internal relation between the premise (P) and the conclusion (C) that results even if one does not consider the relations between the concepts *horse* and *animal* as internal. A more thorough analysis considers this relation within the formalization of the conclusion (C). This results in $\phi(C)_{TC'}$, which does not derive the conclusion of the argument from some external relation of the class of horses and the class of animals, but rather from the internal relation of the concepts *horse* and *animal*. This does not reject the justification of (C) by (P) but makes explicit that this conceptual relation is already implied by (C) itself.

According to this model, the benefit of logical formalization consists in making explicit the truth conditions of ordinary propositions. How far one abstains

from implied logical dependencies depends on the context and the aim of the respective formalization. Finally, by means of Russell's Paradox, we will illustrate the importance of relativizing inferences from formalizations to presumed criteria of formalization and to the grade of analysis.

6. Inadmissible interpretations

Within mathematical logic, Russell's Paradox is mostly understood as a refutation of the unrestricted comprehension axiom schema. Expressed within ordinary language, this axiom of "naïve" set theory is as follows:

> UCAS: There exists a set *y* whose members are precisely those objects that satisfy the propositional function *φ(x)*.

According to FℑCR and a logical language with the dyadic predicate $x \in y$ and its fixed interpretation as "*x* is a member of *y*", UCAS is to be formalized as follows:

> ϕ(UCAS): $\exists y \forall x(x \in y \leftrightarrow \varphi(x))$

The replacement of *φ(x)* with "*x* is not a member of itself" and $\neg x \in x$, respectively, results in Russell's Paradox:

> UCAS*: There exists a set *y* whose members are precisely those objects that satisfy the propositional function *x is not a member of itself*.
> ϕ(UCAS*): $\exists y \forall x(x \in y \leftrightarrow \neg x \in x$

From ϕ(UCAS*), a contradiction follows. In the framework of modern set theory, this is taken as a sufficient reason to conclude that UCAS is false. Russell's Paradox apparently demonstrates that there are concepts that do not define sets. According to FℑCR, this reasoning is conclusive. According to modern mathematical logic, Russell's Paradox refutes naïve set theory and calls for a different axiomatic system such as ZF that does not allow for Russell's Paradox due to the axiom of separation. Likewise, according to VℑCR, ϕ(UCAS*) is correct, and thus, UCAS* is contradictory.

Russell himself argued this way in *Principles of Mathematics*, Russell(1992), p. 102f. However, ϕ(UCAS*) is incompatible with the view that Russell and Whitehead advanced in *Principia Mathematica*, cf. Whitehead(1910), p. 75. Here, they analyzed $\in$ as an incomplete symbol. This analysis is incompatible with the understanding of $\in$ as a primitive symbol in terms of a dyadic predicate. According to the point of view put forth in the *Principia*, not all interpretations of $x \in y$ are admissible.

Within the framework of a theory of formalization, such a critique is opened up by *TC'*. $\in$ does not satisfy this criterion as its interpretation is not unrestricted. At least some interpretations do not lead to meaningful, bipolar propositions.

Thus, correctly understood, Russell's Paradox is not a reduction to absurdity of the unrestricted comprehension axiom schema. Instead, the fault of the paradox lies in the formalization of "*x* is a member of *y*" in terms of a dyadic, logical predicate that articulates a primitive relation between objects. $x \in y$ is not a proper propositional function identifying sets. According to this point of view, the solution of the paradox consists in an alternative formalization of propositions about sets, for example propositions such as UCAS. However, it is not necessary to abandon UCAS. The paradox, in terms of a reduction to absurdity of an assumption that seemed to be true, is due to a mistaken formalization.

Only by *TC'* can one distinguish between two kinds of "absurdity": (i) the absurdity relying on a proof by reduction and (ii) the absurdity relying on an inadequate logical formalization. Without *TC'*, there is no sufficient criterion for distinguishing between meaningful but inconsistent propositions and meaningless propositions that cannot be formalized within logic. In the first case, we have a contradiction within the axiomatic system, while in the second case, we have a contradiction to logic. Put more precisely, we have a contradiction to the principles of a logic that provides the means for representing propositions with well-defined conditions of truth and falsehood. In this case, the propositions to be formalized cannot be true because well-defined interpretations in terms of meaningful conditions of truth and falsehood have not yet been specified. In the former case, in contrast, the propositions cannot be true because all interpretations are counter-models.

What seems to be a proof by reduction according to a superficial analysis may shows up as a paradox according to a more thorough logical analysis. In this case, it is not the falsity of some axiom that is proven, but rather the impossibility to represent it by a proposition with well-defined truth conditions according to first-order logic. In the case of Russell's Paradox, it is proven that the so-called "relation of membership" cannot be represented by an atomic propositional function expressing an external, primitive relation between objects. Such a relation presumes that the objects satisfying the relation are identifiable independent of the relation. This criterion is not satisfied in the case of membership if a set is considered to be (or not to be) its own member. Such an understanding contradicts the principle of logical independence. In consequence, one cannot represent "*x* is (not) a member of *y*" by means of a propositional function $\varphi(x)$. Thus, it is the substitution of $\varphi(x)$ with $\neg x \in x$ that must be rejected in the first place. Such a substitution mistakenly confuses the grammatical form with the logical form of propositions. According to this kind of critique, the meaning of UCAS cannot be represented by some logical formula and even less so by the non-tautologous formula ϕ(UCAS). This is due to the fact that it does not make sense to assume well-defined propositional functions in terms of first-order logic that do not identify sets. Within this conception, it is not even possible to interpret a proper propositional function of logic such that it does not identify a set.

Whatever does not identify a set is not a proper propositional function. To be a member of a set means to satisfy some propositional function identifying that set. Thus, what one intends to say by (UCAS) is quite right. Yet, (UCAS) cannot be articulated as a meaningful proposition within the logical symbolism because doing so would presume mistakenly that the relation between sets and propositional functions is external. The problems of "naïve" set theory do not arise before one makes use of a logical formalization naively expressing membership as a primitive relation between objects. Only a superficial analysis, "bedeviled" by surface grammar of ordinary language, makes such a deficient logical formalization of set theory possible.

Within the framework of a theory of formalization, it is irrelevant to consider whether such a critique of the logical formalization of set theory is adequate. What is more important is that it is possible. It should not be excluded by presuming some model of formalization without discussing and arguing against its alternatives. In contrast to V$\mathfrak{I}$CR and F$\mathfrak{I}$CR, *TC'* articulates assumptions of logical formalization concerning the alleged meaning of ordinary propositions. Whether these assumptions are valid is unimportant for a theory of formalization. Yet, it is important to identify them as assumptions that can be questioned. Only *TC'* makes it possible to use logical formalization as a means of logically analyzing the meaning of propositions. In contrast, V$\mathfrak{I}$CR and F$\mathfrak{I}$CR presume that grammatically well-formed propositions have a truth value without considering their truth conditions. However, in order to assume that some proposition *is* true or false, it must be *capable* of being true or false. V$\mathfrak{I}$CR and F$\mathfrak{I}$CR do not consider this priority of the meaning of propositions over their truth value. Thus, they cannot rule out that nonsense is represented by logic.

References

Baumgartner, M./ Lampert, T.: "Adequate formalization". Synthese, 164, 2008. S. 93–115

Blau, U.: Die dreiwertige Logik der Sprache. de Gruyter, Berlin, 1977

Brun, G.: Die richtige Formel. Philosophische Probleme der logischen Formalisierung. Ontos, Frankfurt/M., 2004

Copi, I.: Symbolic Logic. Macmillan, New York, 1979

Epstein, R. L.: The Semantic Foundations of Logic: Propositional Logic. Kluwer, Dordrecht, 1990

Epstein, R. L.: The Semantic Foundations of Logic: Predicate Logic. Oxford University Press, Oxford, 1994

Kleinknecht, R.: "Probleme des Formalisierens: Zum Verhältnis zwischen natürlichen und formalen Sprachen". In: *Kreuzbauer, G./Gratzl, N./Hiebl, E. (Hrsg.):* Rhetorische Wissenschaft: Rede und Argumentation in Theorie und Praxis. LIT-Verlag, Wien, 2008. S. 163–178

Lampert, T.: "Explaining formulae of first order logic". Ruch Filozoficzny, 63, 2006. S. 459–480

Lampert, T. and Baumgartner, M.: "The problems of (in)validity proofs". Grazer Philosophische Studien, 80, 2010. S. 79-109

Lemmon, E. J.: Beginning Logic. Hackett, Indianapolis, 1998

Löffler, W.: "Spielt die rhetorische Qualität von Argumenten eine Rolle bei deren logischer Analyse? Überlegungen zum Verhältnis von Argumentationstheorie und formaler Logik". In: *Kreuzbauer,* G./*Dorn, G. (Hrsg):* Argumentation in Theorie und Praxis. Salzburger Beiträge zu Rhetorik und Argumentationstheorie, Band 1. LIT, Wien, 2006. S. 115–130

Massey, G. J.: "Are there any good arguments that bad arguments are bad?", Philosophy in Context, 4, 1975. S. 61–77

Merrill, D. D.: "On de morgan's argument". Notre Dame Journal of Formal Logic, 18, 1977. S. 133–139

Quine, W. v. O.: Methods of Logic. Harvard University Press, Cambridge, 4. edition, 1982

Russell, B.: The Principles of Mathematics. Routledge, London, 2. edition, 1992

Sainsbury, R. M.: Logical Forms. Blackwell, Oxford, 1993

Smith, P.: An Introduction to Gödel's Theorems. Cambridge University Press, Cambridge, 2007

Suppes, P.: Introduction to logic. Mineola, Dover, 1999

Wengert, R. G.: "Schematizing de Morgan's argument". Notre Dame Journal of Formal Logic, 1, 1974. S. 165–166

Whitehead, A. N. and Russell, B.: Principia Mathematica. Cambridge University Press, Cambridge, 1910

Wittgenstein, L.: Tractatus logico-philosophicus, Werkausgabe Band 1. Suhrkamp, Frankfurt/M., 1995

Was ist eine Theoriensynthese?

Tilmann Massey
massey@web.de
LMU München

Abstract/Zusammenfassung

In this overview on an on-going investigation the concept of *theoretical synthesis* is examined. Such an integration of different fields of research seems to have happened particularly in the history of biology. A classic example would be the „evolutionary synthesis“ (or „modern synthesis“) of the 1930s and 1940s, during which – roughly spoken – darwinism and genetics merged. Recently there is much talk about a synthesis of evolutionary and developmental biology („evo-devo“). A general *(formal) meta-theoretical explication* of the concept of theoretical synthesis, however, is still missing.

First it has to be checked whether results from history of science really suggest the existence of syntheses of *theories* (and not of other entities, like disciplines, research programs, etc.). Using the „modern synthesis“ as a case study, a look into its historiography is revealing a plurality of interpretations. All in all, how-ever, the historical studies tend towards the view that *(fragments of) theories* were indeed the merging entities – provided that a *pragmatically enriched* concept of theory is applied.

Once the historical existence of theoretical syntheses is accepted, for further investigations we are in need of a precise and detailed concept of empirical theories, viz. an elaborated metatheory. The structuralist approach (Sneed, Stegmüller, and others) seems to be particularly adequate to the study at hand for several reasons. So, in the second paragraph a precis of this approach will be given, including many simplifications and adjustments to the given problem. Then, a suggestion of how to tackle an explication of the concept of theoretical synthesis is provided. The central point is the *historical character* of a theoretical synthesis. This implies that an adequate treatment of the phenomenon of synthesis is possible only within an accordingly designed framework.

Gegenstand des vorliegenden Überblicks ist der Begriff der *Theoriensynthese*. Besonders in der Geschichte der Biologie kam es anscheinend öfter zur Integration verschiedener Forschungsbereiche. Das klassische Beispiel ist die „Evolutionäre Synthese“ (auch „Moderne Synthese“) der 1930er und 1940er Jahre, in der – grob gesagt – Darwinismus und Genetik fusionierten. Aktuell ist häufig die Rede von einer Synthese von Evolutionsbiologie und Entwicklungsgenetik („Evo-Devo“). Eine allgemeine *formale metatheoretische Explikation* des Theoriensynthesebegriffs liegt jedoch bislang nicht vor.

Zunächst muss geklärt werden ob wissenschaftsgeschichtliche Untersuchungen wirklich die Existenz von Fusionen von *Theorien* nahelegen (und nicht etwa von anderen Entitäten, wie Disziplinen, Forschungsprogrammen, o.ä.). Dazu wird ein historischer Fall – die „*Moderne Synthese*“ – betrachtet. Die dazu bereits vorliegenden historischen Studien erweisen sich in ihrem Ergebnis sehr heterogen, in der Gesamtheit verweisen die historiographischen Untersuchungen jedoch durchaus auf *Theorien(-fragmente)* als zu vereinigende Entitäten – sie verpflichten dabei allerdings auch auf einen *pragmatisch angereicherten* Theorienbegriff.

Akzeptiert man die historische Existenz von Theoriensynthesen, so benötigt man zur weiteren Untersuchung dieses Phänomens ein möglichst präzises und detailliertes Theorienkonzept, d.h. eine gut ausgearbeitete Metatheorie empirischer Theorien. Der strukturalistische Ansatz (nach Sneed, Stegmüller, und anderen) scheint aus mehreren Gründen für die vorliegende Untersuchung besonders geeignet zu sein. Im zweiten Teil des Aufsatzes soll also zunächst ein kurzer Abriss dieser Metatheorie gegeben werden, wobei Vereinfachungen und Anpassungen an die hier behandelte Fragestellung vorgenommen werden. Anschließend wird ein Vorschlag für die prinzipielle Herangehensweise an eine Explikation des Theoriensynthesebegriffs gegeben. Zentral dabei ist der *historische Charakter* einer Theoriensynthese. Dies impliziert, dass eine adäquate Behandlung des Phänomens der Theoriensynthese nur im Rahmen einer entsprechend ausgestatteten Metatheorie möglich ist.

1 Die „Moderne Synthese“ der Evolutionstheorie und ihre Historiographie

Die erste Hälfte des 20. Jhds. war für die biologischen Wissenschaften eine ereignisreiche Zeit. Kam es zunächst ab etwa 1900 zur Entwicklung der klassischen experimentellen Genetik (durch W. Bateson, T. H. Morgan, u.a.), so stellte sich bald darauf die Frage, was die Ergebnisse dieser genetischen Untersuchungen für evolutionäre Fragestellungen bedeutet. In der Folge von Darwin wurden zwar die Veränderlichkeit von Arten („Evolution als solche“) und die gemeinsame Abstammung aller Lebewesen („Deszendenzhypothese“) von den allermeisten Biologen akzeptiert, bezüglich der zugrundeliegenden Mechanismen herrschte jedoch Uneinigkeit. Gerade von den Genetikern der ersten Stunde wurde Darwins Vorschlag eines graduell-selektionistischen Modells, das Grundlage für die Arbeit vieler Systematiker und Feldbiologen war, angezweifelt. Ab etwa 1925 jedoch kam es verstärkt zu einer Vereinheitlichung, so dass zur Mitte des Jahrhunderts von einer neuen „synthetischen“, „modernen“ Evolutionstheorie gesprochen wurde. Was genau war passiert?

Die wissenschaftsgeschichtliche Aufarbeitung jener Ereignisse begann in den 1970er Jahren und dauert bis heute an. Während anfänglich der Fokus auf der Entstehung der Populationsgenetik lag (z. B. Provine 1971), kamen bald auch andere Aspekte zur Sprache, wie z.B. der Beitrag der Feldbiologen und Systematiker (vgl. Mayr 1980). Es folgte eine Zeit der kontroversen Diskussion, in der speziell im angelsächsischen Raum eine Vielfalt an Ideen und Ansätzen zur Interpretation der „Evolutionären Synthese“ vorgebracht wurden. Keiner der Ansätze konnte sich jedoch vollständig durchsetzen, so dass bald resigniert die „enigmatische und schwer fassbare Qualität“ (vgl. Smocovitis 1996, S. 43f.) der Synthese konstatiert wurde. Neuer Schwung kam in die Debatte erst wieder mit der Behandlung des Themas durch europäische Wissenschaftler. So legt die wissenschaftstheoretisch informierte Untersuchung von Weber (1998) die Grundlage für eine begriffsanalytische Herangehensweise; andere Autoren (wie etwa Reif et al. 2000 und Junker 2004) betonen die Internationalität der Synthese und

sehen die Gründe zumindest eines Teiles der Komplikationen in der einseitigen Ausrichtung der Historiographie auf englischsprachige Werke.

Was sind Beispiele für kontroverse Positionen? Uneinigkeit besteht etwa in der Beurteilung der zeitlichen Dauer des Syntheseprozesses. Hält Ernst Mayr (1980)[2] die Synthese für einen schnellen Paradigmenwechsel, so betonen im Gegensatz dazu viele Autoren deren langwierigen und verwickelten Charakter (u.a. Dobzhansky 1955,Provine 1980,Junker 2004). In Frage steht auch die Stabilität des Syntheseprodukts. Nach Mayr (1993) oder Reif et al. (2000) gab es kaum nachträgliche (prinzipielle) Veränderungen, Gould (1980) hingegen sieht dramatische Veränderungen. Sehr kontrovers wurde die Rolle der Populationsgenetik diskutiert. Betrachten etwa Wright (1960) und Provine (1971) die Populationsgenetik als weitgehend identisch mit der Evolutionären Synthese, so kam es im Folgenden zu einer kontinuierlichen Abschwächung in der Beurteilung der Rolle der Populationsgenetik – zunächst als „Kern" der Synthese (u.a. Beatty 1986), später als gleichberechtigter Teil neben anderen (Weber 1998). Ernst Mayr wiederum wies der Populationsgenetik von Anfang an (Mayr 1959) nur eine geringe Bedeutung bezüglich der synthetischen Evolutionstheorie zu und behielt diese Position mehr oder weniger Zeit seines Lebens bei.

Am Wichtigsten für uns ist nun die Frage nach dem „Wesen" der Ereignisse im oben abgegrenzten historischen Rahmen, also von welcher *Art* der Prozess war. Folgende Positionen (und Unterpositionen) lassen sich ausmachen:

- Es handelte sich um eine Synthese, also eine Fusion von zwei (oder mehreren) Entitäten; diese Entitäten sind:
  - Theorien/Theorienfragmente: Betonung der begrifflichen Basis
    (die meisten Autoren, besonders Weber 1998)
  - Forschungsprogramme, -traditionen: Betonung der pragmatischen Aspekte
    (u.a. Mayr 1980)
  - Disziplinen, „fields": Betonung der Methodik
    (u.a. Darden 1991)
  - ein mathematisch-deduktives Gerüst einerseits und eine Menge an Beobachtungsdaten andererseits
    (Beatty 1986, dies ist eher eine Interpretation des Syntheseprodukts, weniger des Prozesses)
- es handelt sich um *keine* Synthese, sondern
  - eine „constriction" (Provine 1992): Ausschluss von Variablen
  - eine „restoration" (Ghiselin 2001): Vollendung des darwinschen Projekts
  - es gibt einen Zusammenhang mit dem „Unity of Science Movement" (Smocovitis 1996): Verweis auf den „Zeitgeist"

2 Die folgenden Angaben zu den Autoren erheben keinen Anspruch auf Vollständigkeit. Es wurden exemplarisch Autoren gewählt, die die jeweilige Position besonders pointiert vertreten haben.

Auch wenn vielleicht durch obige pointierte Auflistung der Eindruck einer unübersichtlichen Meinungsvielfalt entsteht, so gibt es doch – betrachtet man den Lauf der nun bald vierzigjährigen Geschichtsschreibung mit etwas Distanz – durchaus bestimmte einheitliche Tendenzen in der Beurteilung der damaligen Ereignisse. Positive Resultate der Historiographie, die wir im Folgenden als Basis für weitere Überlegungen als gegeben akzeptieren, waren u.a.:

- Der frühe Hinweis auf die Prozess-Produkt-Ambiguität des Begriffs „Synthese". Es muss ggf. unterschieden werden, ob vom diachronischen Syntheseprozess oder von dem fertigen Syntheseprodukt (meist als Theorie interpretiert) die Rede sein soll.[3]
- Die Anerkennung des internationalen Charakters der Synthese und die damit verbundene Ausweitung der bearbeiteten Quellen
- Die Isolierung zweier wesentlicher Hauptkomponenten der Modernen Synthese:
    1. Es kam zu bedeutenden *begrifflichen Änderungen* resultierend in einer *begrifflichen Vereinheitlichung*[4]
    2. *Pragmatische Aspekte* spielten eine nicht zu unterschätzende Rolle[5]

Die begrifflichen Änderungen betrafen u. a. den Mutationsbegriff, den Variationsbegriff, sowie die Einführung der Unterscheidung Genotyp-Phänotyp und eines speziellen Populations- und Artbegriffs. Zusätzlich kam es auch zu bestimmten Änderungen bezüglich der Anwendungsbereiche, z. B. der Übertrag von anfänglich reinen Labormethoden auf natürliche Populationen. Begriffliche Zusammenhänge und deren Anwendungsbereiche werden in der Wissenschaftstheorie üblicherweise im Rahmen der Untersuchung wissenschaftlicher *Theorien* behandelt. Akzeptiert man also einmal die Existenz von begrifflichen Änderungen und die Vereinheitlichung vormals disparater Begriffssysteme zu jener Zeit, liegt es also auf der Hand das reichhaltige diesbezügliche Instrumentarium der Wissenschaftstheorie zu nutzen und die Geschichte der Modernen Synthese mit metatheoretischen Methoden zu analysieren. Die Fragestellung lautet – unter Berücksichtigung weiter unten ausgeführter Bedingungen – dann: Wie funktioniert eine Zusammenführung von Begriffssystemen? Oder anders ausgedrückt: Was ist eine Theoriensynthese?

---

3 Im Weiteren soll folgende Terminologie verwendet werden: „Synthetischer Darwinismus" bezeichnet nach einem Vorschlag von Junker (2004) das Produkt, „(Evolutionäre) Synthese" den Prozess, „Moderne Synthese" bleibt allgemeiner Überbegriff.

4 Für eine sehr detaillierte Untersuchung dieses Aspekts siehe besonders den ersten Teil von Weber (1998).

5 Als zusätzliche Anmerkung sei hier erwähnt, das pragmatische Aspekte auch bei der Rezeptionsgeschichte sowie bei der Geschichtsschreibung der Modernen Synthese selbst eine große Rolle spielten (vgl. dazu z. B. die Darstellung Provines (1992, S. 169ff.) der Eitelkeiten der beteiligten Wissenschaftler, die z. T. große Anstrengungen unternahmen, um als „(Mit-)Architekt" des synthetischen Darwinismus wahrgenommen zu werden).

Als wichtigster Aspekt der zweiten Komponente („Pragmatik“) könnte die Überwindung bestimmter Antipathien zwischen verschiedenen damaligen „scientific communities“ gelten. Ernst Mayr sieht die „gegenseitige Erziehung“ von experimentell arbeitenden („experimentalists“) und feldbiologisch („naturalists“) arbeitenden Wissenschaftlern als wesentlich an (vgl. Mayr 1980, *passim*). Neben der oben erwähnten begrifflichen Vereinheitlichung kam es also auch zu einer personellen Vereinheitlichung, dergestalt, dass nach Vollzug der Synthese eine neue, einheitliche wissenschaftliche Gemeinschaft entstand. Deren Mitglieder hatten das Gefühl an einem „gemeinsamen Projekt“ beteiligt zu sein, obwohl sie sich durchaus für unterschiedliche Bereiche der Evolutionsbiologie interessierten und obwohl es in Bezug auf einen weiteren pragmatischen Aspekt, nämlich die unterschiedlichen Methodiken[6], *nicht* zu einer Vereinheitlichung kam. Diese *Überwindung eines Kommunikationsproblems* scheint, wenn man der Geschichtsschreibung folgt, ein zu wichtiger Bestandteil der Evolutionären Synthese gewesen zu sein, als dass er in einer umfassenden Analyse fehlen könnte.

Was für Anforderungen muss man also an eine Metatheorie stellen, in der man die Moderne Synthese adäquat bearbeiten kann? Nun, zum einen sollte diese Metatheorie ein differenziertes Instrumentarium zur Darstellung unterschiedlichen begrifflichen Beziehungen bereithalten. Des weiteren sollten pragmatische Aspekte darstellbar sein, wir benötigen eine *pragmatisch angereicherten Metatheorie*. Ein letztes Desiderat wird schließlich bereits durch den Prozesscharakter der Evolutionären Synthese nahegelegt: es soll die Behandlung diachronischer Probleme möglich sein. Insgesamt wird also ein *pragmatisch angereicherter, diachronischer Theoriebegriff* benötigt. Im folgenden Teil soll kurz angerissen werden, wie, aufbauend auf einer solchen geeigneten Metatheorie, das Phänomen der Theoriensynthese angegangen werden könnte.

2 Was ist eine Theoriensynthese?

Ausgangspunkt für eine Beantwortung dieser Frage ist die Wahl einer obigen Kriterien genügenden Metatheorie. Der strukturalistische Ansatz (nach Sneed 1971, Stegmüller 1979, und anderen) scheint v.a. aus folgenden Gründen für die vorliegende Untersuchung besonders geeignet zu sein:

1. Der semantische Zugang des Ansatzes erleichtert die Behandlung von Theorien, die nicht in axiomatischer (bzw. sonst stark mathematisierter) Form vorliegen.
2. Relationen zwischen Theorien(-bestandteilen) lassen sich leicht einführen und klar darstellen
3. Es gibt bereits Ansätze einer diachronisch-pragmatischen Ausarbeitung

6 So wenden z. B. Genetiker, systematische Zoologen, systematische Botaniker, Paläontologen, usw. sehr verschiedene Methoden an; trotzdem waren die *theoretischen* Streitigkeiten nach der Etablierung des synthetischen Darwinismus (weitgehend) verschwunden.

Wir kürzen die Einführung in den strukturalistischen Begriffsapparat stark ab und betrachten nur die für die vorliegende Fragestellung relevanten Aspekte[7]. Der Begriff der „Theorie“ wird im Strukturalismus als mehrdeutig erachtet und es werden folglich verschiedene Typen von Theorien auch nomenklatorisch unterschieden. Die kleinsten epistemologischen Einheiten bilden die sog. TheorieElemente (T), bestehend aus mehreren „slots“, die beispielsweise die Gesetze, den Begriffsrahmen, eine nicht-theoretische Ebene, Verbindungen zu anderen Theorie-Elementen, den intendierten Anwendungsbereich der Theorie, u.s.w. repräsentieren.[8] Theorie-Elemente können zu anderen Theorie-Elementen in bestimmten Relationen stehen. Unterscheidet sich ein Theorie-Element z. B. nur durch stärkere Gesetze und einen engeren Anwendungsbereich von einem Zweiten, ist aber in sonstiger Hinsicht (und insbesondere bzgl. des begrifflichen Gerüsts) gleich, so steht es zu diesem in der sog. „Spezialisierungsrelation“[9]. Mengen an Theorie-Elementen, die untereinander in der Spezialisierungsrelation stehen, bilden ein „Theorie-Netz“ (N) – ein weiterer Theorientyp, den der strukturalistische Ansatz identifiziert. Für uns am Wichtigsten ist aber ein weiterer Typ, der in Balzer et al. (1987, Kap.5) als Theorie-Evolution eingeführt wird, hier aber „Theorie-Entwicklung“ (D) heißen soll.[10] Es handelt sich dabei um eine in bestimmter Weise definierte diachronische Folge von Theorie-Netzen, deren Elementanzahl nicht fest ist (es können also mit der Zeit Theorie-Elemente hinzukommen oder wegfallen; radikale Brüche sind jedoch nicht erlaubt).[11] Eine Theorie wird hier also als genidentische Entität angesprochen, ähnlich einer Person oder einer natürlichen Sprache. Gleichzeitig können pragmatische primitive Begriffe eingeführt werden. Der Einfachheit halber beschränken wir uns hier auf den Begriff der „Generation“ (G), womit die einem einzelnen Theorie-Netz N (aus D) zugeordnete Menge an Wissenschaftlern gemeint ist.[12] Wir halten also fest: ein

7 Eine kurze Einführung in die strukturalistische Theorienkonzeption inklusive Angaben zu weiterführender Literatur findet sich in Balzer & Moulines (1996), eine systematische und ausführliche Darstellung ist Balzer et al. (1987).

8 Genau genommen handelt es sich bei den „slots“ um Modellmengen, die jeweils charakteristische Bedingungen erfüllen (also z.B. die Bedingung die Gesetze zu erfüllen, etc.). Diese übliche Einführung von Theorien als Modellklassen wird im Folgenden übergangen, steht aber natürlich im Hintergrund.

9 Auf diese Weise können je nach Interesse und Fragestellung leicht weitere Relationen eingeführt und diskutiert werden, z. B. Reduktions-, Äquivalenz-, Supervenienzrelationen, etc.

10 Die Umbenennung erfolgt in erster Linie um Verwechslungen in Bezug auf die biologische Evolutionsthematik auszuschließen, aber auch weil - zumindest im Deutschen - „Entwicklung“ das Wort mit den passenderen Konnotationen ist.

11 Dazu ist die Einführung von diachronischen primitiven Begriffen wie „historische Periode“ und eine „historische Präzedenzrelation“ notwendig. Für Details s. Balzer et al. (1987, Kap.5) und Moulines (1991).

12 Die pragmatische Anreicherung wird hier nur sehr verkürzt dargestellt; doch auch die detailreicheren Varianten in Balzer et al. (1987, Kap.5) und Moulines (1991) lassen noch Erweiterungspotential erkennen. Die weiterführende Ausarbeitung eines umfassenden

Theorie-Element besteht für uns nun aus einem begrifflichen Rahmen, Gesetzen, einer Menge an intendierten Anwendungen und einer Menge an Wissenschaftlern. Eine Theorie-Entwicklung (man kann hier auch an eine Forschungstradition, o.ä. denken) besteht aus einer Folge von Spezialisierungsnetzen, deren Bestandteile obige Theorie-Elemente sind.

Doch nun zurück zur Frage nach der Theoriensynthese. Eine Intuition wie eine solche Fusion in etwa zu fassen ist vermittelt Larry Laudan (1977, S. 103):

> „There are times when two or more research traditions, far from mutually undermining one another, can be amalgamated, producing a synthesis, which is progressive with respect to both the former research traditions"

Die grundlegende Idee ist nun die Klärung des Syntheseprozessbegriffs anzugehen, indem man Folgen von Mengen von Theorie-Entwicklungen betrachtet. Wir wollen uns außerdem auf den Fall konzentrieren, in dem auf zwei Theorie-Entwicklungen eine folgt und zwar so, dass sich das erste Element N'_1 der nachfolgenden D mit den letzten Elementen N^1_n , N^2_n der vorausgehenden Ds überlappt. Dies trifft jedoch auf sehr viele Ds zu: wie kann man also jetzt eine solche Eingrenzung vornehmen, dass nur noch die intendierten Fälle übrigbleiben? Dazu muss man die Relation bestimmen, in denen die sich überlappenden Theorien-Netze stehen müssen. Ein Vorschlag, der auf den ersten Teil der Untersuchung Bezug nimmt, wäre nicht-leere Schnittmengen der intendierten Anwendungen und der Wissenschaftler von N'_1 zu jedem vorsynthetischen Theorie-Netz zu fordern. Dies entspricht einer Überschneidung der Anwendungsgebiete und der Existenz einiger (evtl. weniger) „bridge-builders" (vgl. Mayr 1980). Außerdem sollten zumindest einige der Grundbegriffe aus N'_1 einen Bezug zu denen jedes vorsynthetischen Theorie-Netzes haben. Dieser Bezug kann in einer Identität oder auch in einer Ableitung bestehen. Da sich insgesamt ein neues Begriffsgerüst ergibt, liegt es Nahe, dass sich auch neuartige gesetzesartige Zusammenhänge ergeben. Doch soll dies für diese Syntheseteilrelation auch stets gefordert werden?

Wichtiger als diese Details zu klären, war es die Grundintention klarzumachen. Hat man nämlich einmal eine geeignete *Syntheseteilrelation* genau defniert, so ist der *Syntheseprozess* – wie oben angedeutet – leicht explizierbar. Eine *synthetische Theorie(-Entwicklung)* ließe sich dann einfach als Produkt eines Syntheseprozesses bestimmen.[13] Wesentliches Ergebnis scheint also zu sein, dass die Rede von einer „synthetischen Theorie" implizit einen diachronischen

---

pragmatischen Ansatzes auf Basis des Strukturalismus einschließlich der Aufstellung von operationellen Kriterien für „Wissenschaftler", etc., ist Thema einer zukünftigen Arbeit.

13 Das Wort „Produkt" suggeriert eine starre, nicht mehr veränderliche Qualität; dies ist aber in diesem Fall irreführend und hat meiner Einschätzung nach zu der oben erwähnten Kontroverse über die Stabilität des synthetischen Darwinismus beigetragen. Das „Produkt" ist in der hier vorgeschlagenen Explikation eine Theorie-Entwicklung, also eine diachronische Entität!

Theorienbegriff voraussetzt. Eine Theoriensynthese ist ein *geschichtliches Phänomen* und *nur* in einem entsprechend ausgestatteten metatheoretischen Begriffsrahmen behandelbar. Es gibt keine *intrinsische* Eigenschaft die eine beliebige Theorie zu einer synthetischen Theorie macht, dies kann ausschließlich die spezifische Geschichte der Theorie leisten. Die Verkennung dieses Umstandes hat maßgeblich zur Konfusion bezüglich der Historie der „Modernen Synthese“ geführt.

Literaturverzeichnis

Balzer, W./Moulines, C. U. (Hrsg.): Structuralist Theory of Science: Focal Issues. New Results. de Gruyter, Berlin, 1996

Balzer, W./Moulines, C. U./Sneed, J. D: An Architectonic for Science. D. Reidel, Dordrecht, 1987

Beatty, J.: "The Synthesis and the Synthetic Theory". In: *Bechtel, W. (Hrsg.):* Integrating Scientific Disciplines. Nijhoff, Dordrecht, 1986. S. 125–135

Bechtel, W. (Hrsg.): Integrating Scientific Disciplines. Nijhoff, Dordrecht, 1986

Darden, L.: Theory Change in Science: Strategies from Mendelian Genetics. Oxford Univ. Press, Oxford, 1991

Dobzhansky, T.: "A review of some fundamental concepts and problems of population genetics". Cold Spring Harbour Symposium on Quantitative Biology, 20, 1955. S. 1–15

Ghiselin, M. T: "Evolutionary Synthesis from a cosmopolitan point of view: a commentary on the ideas of Reif, Junker and Hossfeld". Theory in Bioscience, 120, 2001. S. 166–172

Gould, S. J.: "Is a new and general theory of evolution emerging?". Paleobiology, 6, 1980. S. 119–130

Junker, T.: Die zweite Darwinsche Revolution: Geschichte des synthetischen Darwinismus in Deutschland 1924 bis 1950. Basilisken-Presse, Marburg/Lahn, 2004

Karl-Marx-Universität (Hrsg.): Untersuchungen zur Logik und zur Methodologie 8. Leipzig, 1991

Laudan, L.: Progress and Its Problems: Toward a Theory of Scientific Growth. University of California Press, Berkeley, 1977

Mayr, E.: "Where are we?". Cold Spring Harbour Symposium on Quantitative Biology, 24, 1959. S. 1–14

Mayr, E.: "Prologue: Some Thoughts on the History of the Evolutionary Synthesis". In: *Mayr, E./Provine, W. B. (Hrsg.):* The Evolutionary Synthesis: Perspectives on the Unification of Biology. Harvard Univ. Press, Cambridge (MA), 1980. S. 1–48;

Mayr, E.: "What was the evolutionary synthesis?", Trends in Ecology and Evolution 8, 1993. S. 31–34;

Mayr, E./Provine, W. B. (Hrsg.): The Evolutionary Synthesis: Perspectives on the Unification of Biology. Harvard Univ. Press, Cambridge (MA), 1980

Moulines, C. U.: "Pragmatisch-diachronische Aspekte der Wissenschaftstheorie". In: Karl-Marx-Universität, 1991. S. 1-21

Provine, W. B.: The Origins of Theoretical Population Genetics. University of Chicago Press, Chicago, 1971

Provine, W. B.: "Epilogue". In: *Mayr, E./Provine, W. B. (Hrsg.):* The Evolutionary Synthesis: Perspectives on the Unification of Biology. Harvard Univ. Press, Cambridge (MA), 1980. S. 399-412

Provine, W. B.: "Progress in Evolution and Meaning in Life". In: *Waters, C. K./van Helden, A. (Hrsg.):* Julian Huxley: Biologist and Statesman of Science. Rice Univ. Press, Houston, 1992. S. 165–180

Reif, W.-E./Junker, T./Hoßfeld, U.: "The synthetic theory of evolution: general problems and the German contribution to the synthesis". Theory in Bioscience, 119, 2000. S. 41–91

Smocovitis, V. B.: Unifying Biology: The Evolutionary Synthesis and Evolutionary Biology. Princeton Univ. Press, Princeton, 1996

Sneed, J. D.: The Logical Structure of Mathematical Physics. D. Reidel, Dordrecht, 1971

Stegmüller, W.: The Structuralist View of Theories: A Possible Analogue of the Bourbaki Programme in Physical Science. Springer, Berlin, 1979

Waters, C. K./van Helden, A. (Hrsg.): Julian Huxley: Biologist and Statesman of Science. Rice Univ. Press, Houston, 1992

Weber, M.: Die Architektur der Synthese: Entstehung und Philosophie der modernen Evolutionstheorie. de Gruyter, Berlin, 1998

Wright, S.: "Genetics and twentieth century Darwinism: A review and discussion". American Journal for Human Genetics, 12, 1960. S. 365–372

Pluralistische Interpretationen von Wahrscheinlichkeit

Wolfgang Pietsch
pietsch@cvl-a.tum.de
Lehrstuhl für Philosophie und Wissenschaftstheorie, TU München

Abstract/Zusammenfassung

While there has been extensive work on the 'pure' accounts of probability, i.e. the classical, the logical, the subjective, the frequency and the propensity interpretations, pluralist views have attracted much less attention. We count as pluralist all those views in which the two traditional applications of probability theory, i.e. statistics and induction, are not treated by one and the same concept. According to this definition many important accounts of probability in the 20th century are pluralist, including for example those of Carnap, Popper, and Lewis. The study of pluralist accounts helps to answer the following questions: Which of the 'pure' views are compatible with each other and which are not? Do we need different probability concepts for different applications, e.g. for the natural and for the social sciences or for statistics and for induction? In this essay, a classification of pluralist views will be suggested by distinguishing two main types: (i) *Empirical dualism* starts from the empirical use of probability in the description of repeatable events. Empirical dualists mostly deny that there exists an inductive concept that satisfies the probability axioms. (ii) By contrast, *rational dualism* starts from the epistemic function of probability, while nevertheless acknowledging the need for a probability concept dealing with repeatable events. In the spirit of logical empiricism, rational dualists split probability into a logical and an empirical concept, both of which satisfy the probability axioms. On the basis of this classification, pluralism about probability is then subjected to some general criticism. If, as in rational dualism, the two concepts are very similar, then the pluralist position violates a methodological imperative to unify. By contrast, empirical dualism, in which the two concepts are held to be quite distinct, ignores that there exists a genuine continuum of applications of probability theory.

Während die ‚reinen' Wahrscheinlichkeitsinterpretationen, d.h. die klassische, die logische, die subjektive, die frequentistische und die Propensitätsinterpretation, in einer umfangreichen Literatur abgehandelt werden, haben pluralistische Positionen bisher weit weniger Aufmerksamkeit erfahren. Zu letzteren zählen wir all jene Positionen, welche für die beiden traditionellen Anwendungsfelder der Wahrscheinlichkeitstheorie unterschiedliche Konzepte vorsehen, d.h. für die mathematische Statistik und für die Bestätigungstheorie. Nach dieser Definition erweisen sich eine Reihe einflussreicher Wahrscheinlichkeitsauffassungen im zwanzigsten Jahrhundert als pluralistische Positionen, zum Beispiel diejenigen von Carnap, Popper und Lewis. Durch die Auseinandersetzung mit dem Pluralismus lassen sich unter anderem folgende Fragen beantworten: Welche der ‚reinen' Interpretationen sind miteinander vereinbar? Braucht man unterschiedliche Konzepte für unterschiedliche Anwendungen, zum Beispiel für die Natur- und die Sozialwissenschaften? In dem vorliegenden Aufsatz wird eine Klassifikation pluralistischer Positionen vorgenommen, welche zwei Haupttypen unterscheidet: (i) Der *empirische Dualismus* beginnt beim empirischen Gebrauch von Wahrscheinlichkeit, also bei der Beschreibung von wiederholbaren Ereignissen. Empirische Dualisten verneinen im All-

gemeinen, dass es ein induktives Konzept gibt, welches ebenfalls den Wahrscheinlichkeitsaxiomen genügt. (ii) Der *rationale Dualismus* beginnt hingegen bei der epistemischen Funktion von Wahrscheinlichkeit, wobei generell zugestanden wird, dass Wahrscheinlichkeit auch für relative Häufigkeiten eine Rolle spielt. Im rationalen Dualismus erfüllen beide Konzepte die Wahrscheinlichkeitsaxiome, jedoch wird scharf zwischen der empirischen und der logischen Rolle unterschieden. Auf der Grundlage dieser Klassifikation pluralistischer Interpretationen werden diese dann einer Kritik unterworfen. Wenn – wie im rationalen Dualismus – die beiden Konzepte sehr ähnlich sind, missachten pluralistische Positionen generell den methodischen Imperativ zur Vereinheitlichung. Im empirischen Dualismus, in welchem die beiden Konzepte relativ große Unterschiede aufweisen, bleibt dagegen unberücksichtigt, dass es ein regelrechtes Kontinuum von Wahrscheinlichkeitsanwendungen gibt.

1. Einleitung

Glücksspiel, Bevölkerungsstatistiken und die Verlässlichkeit von Gerichtsurteilen, das waren lange Zeit die wichtigsten Anwendungsbereiche der Wahrscheinlichkeitstheorie. Hingegen spielte in den grundlegenden Erfahrungswissenschaften wie der Physik Wahrscheinlichkeit nur ganz am Rande eine Rolle, beispielsweise bei der Abschätzung von Fehlern. Weitgehend unangefochten behauptete sich dort der metaphysische Satz vom zureichenden Grunde, nach welchem jedes noch so unbedeutende Geschehnis eine vollständige Ursache hat. Neben diesem Satz hatte der blinde Zufall keinen Platz und damit konnten auch die ersten Prinzipien und die fundamentalen Gesetze nicht wahrscheinlichkeitstheoretischer Natur sein.

Gegen Ende des 19. Jahrhundert rückt dann der Wahrscheinlichkeitsbegriff auch in den grundlegenden Erfahrungswissenschaften ins Zentrum der Aufmerksamkeit. Im gleichen Zug verliert der Satz vom zureichenden Grund seine beherrschende Stellung. Es ist vielleicht nicht übertrieben zu sagen, dass die zentrale Rolle von Wahrscheinlichkeit die wichtigste Neuerung in der wissenschaftlichen Methode des zwanzigsten Jahrhunderts darstellt. Das gilt insbesondere für die Physik und für die Biologie. In ersterer erlangt der Wahrscheinlichkeitsbegriff zunächst Bedeutung im Rahmen von Ludwig Boltzmanns Überlegungen zum Entropiekonzept. Viel radikaler ist dann die quantenmechanische Einführung eines objektiven Zufalls, der durch keine noch so genaue Beobachtung der jeweiligen Umstände eliminierbar ist. In der Biologie lösen sich zur gleichen Zeit die einstmals scharfen Grenzen zwischen den Arten auf, an deren Stelle eine von zufälliger Mutation geleitete Entwicklung der Lebewesen tritt.

Mit diesem Bedeutungsgewinn geht eine neuerliche Auseinandersetzung mit den philosophischen und konzeptionellen Grundlagen von Wahrscheinlichkeit einher. Am auffälligsten ist vielleicht die Aufsplitterung des Konzepts. Die einst dominierende klassische Interpretation – die diesen Namen erst erhält, als sie kaum jemand mehr ernst nimmt – wird allgemein verworfen. Sie wird jedoch nicht durch ein einziges allgemein akzeptiertes Konzept ersetzt. Stattdessen fin-

det sich eine Fülle neuer Auffassungen. Im Allgemeinen werden zwei große Strömungen unterschieden. Auf der einen Seite ordnen epistemische Auffassungen den Wahrscheinlichkeitsbegriff der Theorie des Wissens und der Erkenntnis zu (gr. episteme: Wissen, Wissenschaft). Auf der anderen Seite finden sich ontische Auffassungen (gr. ontos: das Seiende), welche Wahrscheinlichkeiten zum Inventar der Welt rechnen vergleichbar mit Atomen, Quarks oder Feldern.

Ein aufmerksamer Beobachter stellt bald fest, dass sich die Trennlinie zwischen den unterschiedlichen Auffassungen grob entlang von Fächergrenzen ziehen lässt. Für die objektive Bedeutung von Wahrscheinlichkeit interessieren sich hauptsächlich Naturwissenschaftler, insbesondere Physiker oder an physikalischen Fragen interessierte Philosophen. Vereinfacht gesagt liegt die Ursache darin, dass Naturwissenschaftler subjektiven Einflüssen in der Wissenschaft grundsätzlich argwöhnisch gegenüberstehen. Andererseits müssen sie natürlich der zunehmenden Bedeutung wahrscheinlichkeitstheoretischer Aussagen in den Kernbereichen physikalischer Theorien gerecht werden. Für Sozialwissenschaftler, insbesondere für Ökonomen, steht der Mensch mit seinem subjektiven Denken und Handeln ohnehin seit jeher im Mittelpunkt ihres Denkens. Subjektive Einflüsse sind aus ökonomischen Theorien nicht vollständig wegzudenken und so finden sich unter Sozialwissenschaftlern nur wenige Einwände gegen einen subjektiven oder epistemischen Charakter von Wahrscheinlichkeit.

Diese Arbeitsteilung der verschiedenen Wahrscheinlichkeitskonzepte ist natürlich nicht unbemerkt geblieben. Eine natürliche Reaktion scheint in einem Pluralismus bezüglich Wahrscheinlichkeit zu bestehen. Verschiedene Wahrscheinlichkeitskonzepte erfüllen demnach unterschiedliche Funktionen und decken unterschiedliche Anwendungsbereiche ab. Obwohl zumindest in der Wissenschaftstheorie ein derartiger Pluralismus weit verbreitet ist, haben pluralistische Positionen bisher nur wenig systematische Aufmerksamkeit in der Literatur erfahren. Nach meinem Kenntnisstand findet sich noch nicht einmal eine Klassifikation der verschiedenen pluralistischen Auffassungen, obwohl eine solche die notwendige Voraussetzung für eine adäquate Kritik des Pluralismus darstellt. Diesen zwei Problemstellungen, Klassifikation und Kritik, gilt das Augenmerk des vorliegenden Aufsatzes. In Anbetracht der Komplexität des Themas können sie allenfalls überblicksartig behandelt werden.

Als erstes aber führen wir in Abschnitt zwei in die Debatte um die Interpretation von Wahrscheinlichkeit ein. Jeweils anhand eines paradigmatischen Vertreters werden die verschiedenen Auffassungen umrissen: die klassische, die logische und die subjektive Interpretation, die allesamt zu den epistemischen Auffassungen zählen, sowie die frequentistische Intepretation und die Propensitätsauffassung, die sich grob zu den ontischen Interpretationen rechnen lassen.

Im dritten Abschnitt wird zuerst einmal definiert, was unter einer pluralistischen Auffassung zu verstehen ist. Danach erweist sich zum Beispiel die Propensitätstheorie als pluralistisch. Wir schlagen dann eine grobe Klassifikation

der pluralistischen Auffassungen vor, welche im weitesten Sinne der Trennung in epistemische und ontische Interpretationen analog ist. Der *rationale Dualismus* nimmt insgesamt eine eher epistemische Perspektive ein, während der *empirische Dualismus* den Fokus auf die empirische Rolle von Wahrscheinlichkeit legt.

Nachdem wir eine Klassifikation vorgenommen haben und damit die grundsätzlichen Annahmen pluralistischer Positionen geklärt haben, können wir im Abschnitt vier diese pluralistischen Auffassungen einer generellen Kritik unterziehen. Zum einen werden wir uns genauer ansehen, inwieweit die verschiedenen Wahrscheinlichkeitsinterpretationen bezüglich ihrer metaphysischen Annahmen überhaupt miteinander vereinbar sind. Ein weiterer Ansatzpunkt ergibt sich in der Frage, ob nicht eine allgemeine methodische Maxime, wonach bei konzeptioneller Ähnlichkeit auch vereinheitlicht werden muss, einem wahrscheinlichkeitstheoretischen Pluralismus entgegensteht. Schließlich scheint das Kontinuum wahrscheinlichkeitstheoretischer Anwendungen, welches aus der kontinuierlichen Natur des Ähnlichkeitsbegriffs resultiert, nur schwierig mit einem Pluralismus vereinbar.

2. Ein kleiner Überblick über die wichtigsten Wahrscheinlichkeitsauffassungen

a. Die klassische Interpretation

Unter der *klassischen Interpretation* versteht man heute die bis zum Ende des 19. Jahrhunderts dominierende Auffassung von Wahrscheinlichkeit. Die vielleicht einflussreichste und umfassendste Darstellung findet sich im „Philosophischen Versuch über die Wahrscheinlichkeit" von Pierre Simon de Laplace (1814/1996). Bei Laplace fügt sich das Wahrscheinlichkeitskonzept nahtlos ein in sein extrem mechanistisches Weltbild, in dem das Prinzip vom zureichenden Grunde ohne Einschränkung gilt und damit alles vorherbestimmt ist. Einem hypothetischen Geist, der den Zustand der Welt in einem bestimmten Moment und dazu alle Gesetz der Natur erfassen könnte, würde nichts unbekannt bleiben und Zukunft wie Vergangenheit offen vor Augen liegen. Dieser Laplacesche Determinismus ist eine zentrale Voraussetzung seiner epistemischen Wahrscheinlichkeitsauffassung. Wenn nämlich in der Natur alles vorherbestimmt ist, dann kann es keinen objektiven Zufall geben. Wahrscheinlichkeitstheorie kann also keine Theorie eines objektiven Zufalls sein, sondern nimmt ihren natürlichen Platz in einer Theorie des Wissens und der Erkenntnis ein. Nur wenn uns die genauen Ursachen aufgrund unserer menschlichen Beschränktheit verschlossen bleiben, dann müssen wir auf wahrscheinlichkeitstheoretische Konzepte zurückgreifen.

Laplace definiert Wahrscheinlichkeit als „Verhältnis der Zahl der günstigen Fälle zu der aller möglichen Fälle" (1814/1996, 7). Um so einer Definition empirischen Gehalt zu geben, das heißt um sie anwendbar zu machen, muss natürlich unbedingt geklärt werden, was genau gleich mögliche Fälle sind und wie man erkennt, dass solche vorliegen. Da die Wahrscheinlichkeitstheorie für Laplace zur Theorie des Wissens gehört, verwundert es kaum, dass epistemische Überlegungen diese Aufgabe übernehmen. Mögliche Fälle sind danach solche, „über deren Existenz wir in gleicher Weise unschlüssig sind" (4). Diese Regel ist unter der etwas spöttischen Bezeichnung „Prinzip vom *un*zureichenden Grunde" bekannt geworden. Keynes, auf den wir gleich noch zu sprechen kommen, hat dagegen die Bezeichnung *Indifferenzprinzip* vorgeschlagen, welche sich später allgemein durchgesetzt hat.

Nehmen wir den Wurf eines Würfels und fragen nach der Wahrscheinlichkeit, dass eine gerade Zahl geworfen wird. Solange uns keine Informationen vorliegen, dass der Würfel bestimmte Augenzahlen bevorzugt, müssen wir nach dem Indifferenzprinzip allen Augenzahlen die gleiche Wahrscheinlichkeit zuordnen. Damit ergibt sich die Wahrscheinlichkeit für eine gerade Augenzahl als Verhältnis der günstigen zu den möglichen Fällen zu genau 1/2.

Die klassische Auffassung ist nach langer Dominanz in der zweiten Hälfte des neunzehnten Jahrhunderts zunehmend in die Kritik geraten. Erst ihr Bedeutungsverlust ermöglichte im zwanzigsten Jahrhundert die Entwicklung und Aufstellung neuer Wahrscheinlichkeitskonzepte. Manche verwiesen zum Beispiel darauf, dass uns in vielen Fällen gar keine gleich möglichen Fälle zur Verfügung stehen. So behauptet von Mises, dass der Fall eines falschen Würfels der Laplaceschen Definition gar nicht zugänglich ist (1919/1964, 58). Solch oberflächliche Kritik zeugt allenfalls von einer grundsätzlichen Weigerung, sich im Detail mit den Laplaceschen Vorstellungen auseinanderzusetzen, da Laplace beispielsweise den Fall einer ‚falschen' Münze ausführlich behandelt (1814/1996, 41-44).

Unter Beschuss geriet vor allem auch das Indifferenzprinzip im Zusammenhang mit den Bertrandschen Paradoxien (z.B. von Mises 1928/1981, 77-79; Keynes 1921, Kap. 4). Das sind Beispiele für Fälle, in denen das Indifferenzprinzip widersprüchliche Ergebnisse liefert beziehungsweise in denen das Indifferenzprinzip auf unterschiedliche Weise angewendet werden kann. Diese Kritik übersieht oftmals, dass Wahrscheinlichkeit bei Laplace zur Theorie des Wissens gehört und es damit gar keine objektiven Werte von Wahrscheinlichkeit gibt, welche durch das Indifferenzprinzip eindeutig bestimmt werden müssen. Anders gesagt setzen Kritiker am Indifferenzprinzip generell bereits ein ontisches Konzept von Wahrscheinlichkeit voraus und führen durch diese Annahme die Laplaceschen Vorstellungen auf einen nur vermeintlichen Widerspruch.

b. Die logische Interpretation

Von den modernen Wahrscheinlichkeitskonzepten im 20. Jahrhundert steht die *logische Interpretation* der klassischen Auffassung am nächsten. Auch in dieser Interpretation geht es hauptsächlich um die Frage, wie der Mensch Wissen zusammenfasst und bewertet. Logische Wahrscheinlichkeit betrifft also nie Ereignisse in der Welt selber, sondern immer nur *Aussagen* über Ereignisse.

Nach der logischen Auffassung liefert die Wahrscheinlichkeitstheorie ein Werkzeug zur Umformulierung und Verarbeitung von Wissen und übernimmt damit eine ähnliche Funktion wie die deduktive Logik – daher auch die Namensgebung. John Maynard Keynes, einer der bekanntesten Vertreter der logischen Auffassung, hat das folgendermaßen formuliert: „The Theory of Probability is concerned with that part [of our knowledge] which we obtain by argument, and it treats of the different degrees in which the results so obtained are conclusive or inconclusive." (1921, 3) Logische Wahrscheinlichkeiten quantifizieren also die Stärke von Argumenten. Aus dieser Sicht stellt die Wahrscheinlichkeitstheorie eine Erweiterung der deduktiven Logik dar, welche bekanntlich nur solche Argumente behandelt, die absolut sichere Ergebnisse liefern oder Widersprüche aufzeigen. Anders gesagt behandelt die deduktive Logik nur solche Herleitungen, welche die Wahrscheinlichkeit eins oder null der zu untersuchenden Aussage implizieren.

John Maynard Keynes war einer der einflussreichsten Ökonomen des zwanzigsten Jahrhunderts. Seine Vorstellungen zur Einflussnahme des Staates in wirtschaftliche Prozesse, die im ausgehenden zwanzigsten Jahrhundert bereits für längst überholt galten, sind durch die Weltwirtschaftskrise zuletzt wieder in den Vordergrund gerückt. Es ist wichtig, die Wechselwirkungen zwischen Keynes' Wirtschaftstheorie und seinem Wahrscheinlichkeitskonzept nicht aus den Augen zu verlieren. Die Komplexität menschlicher Interaktionen und der Einfluss des Menschen auf das wirtschaftliche Gesamtsystem machen von vornherein alle Erwartungen zunichte, die Ökonomie als eine vollständig objektive Naturwissenschaft zu formulieren, wie es beispielsweise der Anspruch der Physik ist. Für einen Ökonomen ist es damit nicht weiter problematisch, Wahrscheinlichkeitsaussagen einen teilweise subjektiven Charakter zuzugestehen.

Logische Wahrscheinlichkeiten sind in dem Sinne subjektiv, dass sie nicht objektive Naturvorgänge an sich sondern lediglich Aussagen über die Natur betreffen. Sie geben Glaubensgrade bezüglich dieser Aussagen an. Wie Wissen allgemein sind sie damit an einen Informationsträger und im weitesten Sinne an ein Subjekt gebunden. Andererseits haben logische Wahrscheinlichkeiten einen objektiven Charakter, indem sie keine beliebigen Glaubensgrade irgendeines beliebigen Individuums bezeichnen. Sie gehören vielmehr zu einer allgemeinen Theorie von Rationalität, welche ganz unabhängig vom einzelnen Subjekt gilt.

Sie geben die Bewertung von Wahrscheinlichkeiten in einer festen Wissenssituation bindend vor.

Logische Wahrscheinlichkeiten geben also den *rationalen Glaubensgrad* an, den man aufgrund eines bestimmten relevanten Wissens einer Aussage zuordnen muss. Logische Wahrscheinlichkeiten sind auf diese Weise immer relational, das heißt bedingte Wahrscheinlichkeiten. Wie in der klassischen Theorie nimmt für die numerische Zuordnung von Wahrscheinlichkeiten das Indifferenzprinzip eine zentrale Rolle ein. Lässt das gegebene Wissen keine Bevorzugung der einen oder der anderen Aussage zu, dann müssen gleiche Wahrscheinlichkeiten zugewiesen werden.

Betrachten wir als kleines Anwendungsbeispiel wieder die Frage, welche Wahrscheinlichkeit bei einem Würfelwurf eine gerade Augenzahl hat. Zuerst muss man eine Einschätzung vornehmen, dass alles uns gegebene Wissen indifferent gegenüber den sechs verschiedenen Augenzahlen ist. Man muss dann nach dem Indifferenzprinzip allen Augenzahlen dieselbe Wahrscheinlichkeit zuordnen und kommt so wieder auf das Ergebnis 1/2. Am Anfang steht also eine Relevanzeinschätzung, dass solche Unterschiede wie die Einkerbungen der Augenzahlen letztlich für die Wahrscheinlichkeiten nicht von Bedeutung sind.

c. Die subjektive Interpretation

Die *subjektive Auffassung* steht wie die klassische und die logische Theorie in der Tradition epistemischer Auffassungen. Sie ist allerdings in vieler Hinsicht radikaler. Vor allem besteht für den Einzelnen eine viel größere Freiheit, beliebige Wahrscheinlichkeitswerte zuzuweisen. Der vielleicht wichtigste Vertreter der subjektiven Auffassung, der italienische Mathematiker Bruno de Finetti, war als Persönlichkeit in mancher Hinsicht nicht weniger radikal. So liebäugelte er beispielsweise als junger Mann mit dem Faschismus: „That fascism represents the relativistic attitude in politics [in analogy with his approach to probability] as against the staticity of empty doctrinaire ideologies has been stated by Mussolini himself" (1931/1989, 223).

Nach der subjektiven Auffassung sind Wahrscheinlichkeiten nicht mehr rationale Glaubensgrade, sondern Glaubensgrade simpliciter. Das subjektive Element tritt im Vergleich zur logischen Auffassung viel stärker in den Vordergrund, die Rationalitätseinschränkungen sind viel schwächer. Selbst bei gegebenem Wissen sind die Wahrscheinlichkeiten nunmehr nicht mehr eindeutig zuzuordnen. Vielmehr liegt überhaupt kein Widerspruch darin, wenn unterschiedliche Individuen in derselben Wissenssituation unterschiedliche Wahrscheinlichkeitswerte zuweisen. Lediglich eine gewisse innere Konsistenz müssen diese Wahrscheinlichkeiten zeigen, indem sie den Axiomen der Wahrscheinlichkeitstheorie Genüge tun. Zum Beispiel sollten alle Wahrscheinlichkeitswerte zwi-

schen Null und Eins liegen, die Summe sich gegenseitig ausschließender Ereignisse sollte Eins nicht übersteigen und so weiter.

Eine besonders anschauliche Metapher, die die subjektive Auffassung treffend umreißt, ist, dass es sich bei allen Wahrscheinlichkeitsaussagen um subjektive Wetten handelt. Setzen zwei Leute bei einem Rennen auf unterschiedliche Pferde, so kann vor Abschluss des Rennens auch nicht die Rede davon sein, dass einer richtig setzt und der andere falsch. Inkonsistent kann so eine Wette nur sein, wenn ein Spieler so wettet, dass er in jedem Fall verliert. Solch eine Wette nennt man ein *Dutch Book*. Die minimalen Rationalitätsanforderungen der subjektiven Theorie ergeben sich aus der Forderung, dass jeder Spieler solche Dutch Books vermeiden muss. Aus dieser Bedingung lässt sich vor allem eine primitive Wahrscheinlichkeitsaxiomatik herleiten.

Das Indifferenzprinzip der logischen Theorie verliert damit seine Bedeutung. Ein Vertreter der subjektiven Auffassung kann den unterschiedlichen Augenzahlen eines Würfels prinzipiell beliebige Wahrscheinlichkeiten zuordnen, ohne irgendwelche Rationalitätsbedingungen zu verletzen. Allerdings muss er sich Gedanken darüber machen, warum sich Wahrscheinlichkeiten letztlich nach stabilen relativen Häufigkeiten richten sollten, wenn diese bekannt sind.

d. Die frequentistische Interpretation

In der *frequentistischen Auffassung* rückt der Zusammenhang zwischen Wahrscheinlichkeit und relativer Häufigkeit in den Vordergrund. Da relative Häufigkeiten dasjenige sind, was an Wahrscheinlichkeiten der unmittelbaren Beobachtung zugänglich ist, ist es kaum verwunderlich, dass eine solche Auffassung positivistischen und empiristischen Anschauungen nahe steht. Und so zählen mit Richard von Mises und Hans Reichenbach zwei zentrale Figuren des logischen Empirismus zu den wichtigsten Vertretern der frequentistischen Auffassung. Beide gehörten der Gesellschaft für empirische Philosophie an, dem Berliner Pendant zum Wiener Kreis.

Die frequentistische Auffassung gehört mit der gleich zu besprechenden Propensitätsauffassung zu den ontischen Interpretationen von Wahrscheinlichkeit. Bei diesen geht es um die Frage, was uns Wahrscheinlichkeiten über die wirklichen Verhältnisse in der Welt mitteilen. Wahrscheinlichkeiten werden gleichsam zu Entitäten, die unabhängig von Subjekten die Welt bevölkern. Der Bezug der Wahrscheinlichkeitstheorie zur Theorie des Wissens gerät weitgehend aus dem Fokus. Solche ontischen Auffassungen werden gemeinhin von Naturwissenschaftlern vertreten, insbesondere aus den Grundlagenwissenschaften wie der Physik. Diese sind seit jeher auf der Suche nach objektivem Wissen über die Welt und stehen subjektiven Elementen in der Wissenschaft extrem skeptisch gegenüber. Wieder ist also die Wahrscheinlichkeitsauffassung unmittelbar in einer entsprechenden Weltsicht verankert.

Wichtige Vertreter ontischer Wahrscheinlichkeitsinterpretationen gehören zu den Vordenkern indeterministischer Strömungen in den Naturwissenschaften. Der hehre Anspruch des Satzes vom zureichenden Grunde wird aufgegeben. Damit dürfen auch statistische Zusammenhänge als fundamentale Naturgesetze akzeptiert werden. So schreibt von Mises: „Es wird sich [...] klar herausstellen, dass [die Wahrscheinlichkeitstheorie] die sichere Grundlage für eine befriedigende Beschreibung einer umfassenden Klasse von Naturvorgängen bietet oder [...] dass man, in zahlreichen Gebieten menschlicher Betätigung und Gedankenarbeit, von statistischen Aufnahmen ausgehend, über einen wissenschaftlich geläuterten Wahrscheinlichkeitsbegriff zur Erkenntnis der Wahrheit gelangen kann." (1928, 178-179)

Das zentrale Konzept in von Mises' Wahrscheinlichkeitstheorie ist das so genannte *Kollektiv*. Bei einem Kollektiv handelt es sich um „eine unendliche Folge gedachter Dinge, die wir kurz als ‚Elemente' e_1, e_2, e_3 ... bezeichnen. Jedem Element sei als ‚Merkmal' ein bestimmtes Wertsystem der k reellen Veränderlichen k_1, k_2, k_3 ... oder ein Punkt des k-dimensionalen ‚Merkmalraumes' zugeordnet, wobei nicht alle Elemente und auch nicht alle bis auf endlich viele dasselbe Merkmal aufweisen sollen." (1919/1964, 60)

Zusätzlich müssen zwei Bedingungen erfüllt sein, die von Mises als *Grenzwertaxiom* und als *Regellosigkeitsaxiom* bezeichnet. Ersteres fordert, dass die relativen Häufigkeiten der einzelnen Merkmale einen festen Grenzwert besitzen, wenn die Anzahl der Elemente im Kollektiv gegen unendlich geht. Diese Grenzwerte bezeichnen dann einfach die Wahrscheinlichkeiten. Das Regellosigkeitsaxiom fordert, dass es keine systematische Vorschrift geben darf, nach welcher die Merkmale im Kollektiv verteilt sind. Dieses Postulat, was von Mises anschaulich auch als „Axiom von der Unmöglichkeit eines Spielsystems" bezeichnet, unterstreicht noch einmal seine indeterministische Grundhaltung. Ein Determinist würde hingegen davon ausgehen, dass grundsätzlich eine solche Vorschrift immer existiert, sie uns nur in vielen Fällen unbekannt ist.

In unserem Würfelbeispiel würden die Elemente im Kollektiv die einzelnen Würfe sein. Wir hätten es mit einem ein-dimensionalen, diskreten Merkmal zu tun, welches die Augenzahlen von eins bis sechs bezeichnet. Um auf die Wahrscheinlichkeitswerte zu kommen, müssten wir von den idealisierten Annahmen ausgehen, dass wir es mit einer unendlichen Anzahl von Würfen zu tun haben und dass es keine Regel gibt, die uns erlaubt die Würfe in den einzelnen Fällen vorherzusagen. Es mag kaum überraschen, dass diese beiden Idealisierungen den Hauptansatzpunkt für Kritik darstellen und sicherlich letztlich auch zum weitgehenden Verschwinden der von Misesschen Position beigetragen haben. Andererseits hat von Mises mit Recht darauf verwiesen, dass Idealisierungen in den Naturwissenschaften allgegenwärtig sind. Zudem hat von Mises mit Kolmogorow einen bedeutenden Fürsprecher seiner Auffassung gefunden (Kolmogorow 1956, 3).

e. Die Propensitätsauffassung

Die *Propensitätsauffassung* stellt eine Weiterentwicklung der frequentistischen Position dar und gehört damit in die Reihe ontischer Interpretationen des Wahrscheinlichkeitsbegriffs. Karl Popper, der vielleicht wichtigste Vertreter dieser Auffassung, nennt eine zweifache Motivation für seine Weiterentwicklung des Frequentismus (Popper 1959, 27). Zum einen ist da die Tatsache, dass die frequentistische Auffassung nur in Bezug auf Kollektive Sinn macht. Es ist demnach nicht möglich, einem einzelnen Ereignis unmittelbar eine Wahrscheinlichkeit zuzuweisen. Damit erweist sich diese Auffassung als bedeutend schwächer zum Beispiel im Vergleich mit der logischen Theorie, welche natürlicherweise einer Aussage wie „*Dieser* Würfel zeigt im *nächsten* Wurf Augenzahl sechs." eine Wahrscheinlichkeit zuordnen kann. Anders gesagt gibt es nach der frequentistischen Auffassung keine so genannten *Einzelfall-Wahrscheinlichkeiten*. Popper sah das als Schwäche der frequentistischen Theorie an, zumal solche Einzelfall-Wahrscheinlichkeiten in der Quantentheorie allgegenwärtig sind, unserer grundlegenden Theorie von der Natur: zum Beispiel wenn es um den Zerfallszeitpunkt eines einzelnen Atoms geht. Vehement traten die Mitglieder der so genannten Kopenhagener Schule, darunter Niels Bohr und Werner Heisenberg, dem Ansinnen Albert Einsteins entgegen, die Wahrscheinlichkeiten der Quantentheorie im Bezug auf ein Kollektiv oder Ensemble zu interpretieren. Die Rolle von Wahrscheinlichkeit in der orthodoxen Quantentheorie war neben den Einzelfall-Wahrscheinlichkeiten die zweite wichtige Motivation für Poppers Propensitätsauffassung.

Das sperrige Wort „Propensität" bezeichnet eine natürliche Neigung oder Tendenz, dass ein bestimmtes Ereignis eintritt. Popper verweist auf den aufschlussreichen Vergleich mit Kräften. Auch Kräfte bezeichnen nicht grundsätzlich real existierende Größen, sondern nur die Disposition, dass sich Körper in einem Kraftfeld auf eine bestimmte Art und Weise beschleunigen würden. Popper zufolge geben Propensitäten also die Tendenz an, dass sich bestimmte Möglichkeiten realisieren. Sie sind durch die relevanten Rahmenbedingungen eines Experiments vollständig bestimmt. Auf lange Sicht sind die Propensitäten für die relativen Häufigkeiten verantwortlich, die bei ausreichender Wiederholung des Experiments beobachtet werden.

In der Propensitätsauffassung richtet sich das Augenmerk nun also weg von den relativen Häufigkeiten der Ereignisse hin zu den speziellen experimentellen Rahmenbedingungen, unter denen diese relativen Häufigkeiten erzeugt werden. Im Beispiel des Würfels würden die relevanten Rahmenbedingungen folgende Informationen umfassen: die Beschaffenheit des Würfels, des Bechers und des Tisches, auf dem der Würfel ausrollt, sowie die Art und Weise wie der Würfel geworfen wird. Diese Rahmenbedingungen erzeugen dann gewissermaßen die Wahrscheinlichkeiten.

Im Zusammenhang mit ontischen Interpretationen muss zum Schluss noch einmal die Tatsache betont werden, dass ontische Wahrscheinlichkeitskonzepte keine grundsätzliche Bedeutung für eine Theorie des Wissens oder der Erkenntnis haben. Insbesondere sind ontische Konzepte nicht in der Lage Bestätigungsmaße für Hypothesen zu liefern oder die Verlässlichkeit von induktiven Schlüssen zu bewerten. So ist es nur konsistent, dass Popper einer der großen Kritiker der induktiven Methode und ein Verfechter eines falsifikationistischen oder streng hypothetisch-deduktiven Ansatzes ist. Diese Wechselwirkung zwischen Wahrscheinlichkeitskonzept und wissenschaftlicher Methode ist natürlich kein Einzelfall und bei allen Positionen in stärkerem oder schwächerem Maße zu beobachten.

3. Pluralistische Positionen bezüglich Wahrscheinlichkeit

a. Was ist Pluralismus?

Den unterschiedlichen Interpretationen liegen offenbar verschiedene Vorstellungen zugrunde, was Wahrscheinlichkeiten behandeln. Besonders markant ist dabei die Aufteilung in epistemische und ontische Wahrscheinlichkeiten. Erstere stellen vor allem Bestätigungsgrade von Hypothesen oder Glaubensgrade dar, während letztere vor allem relative Häufigkeiten von Ereignissen oder objektiven Zufall behandeln. Diese Zweiteilung der Anwendungsfelder legt nahe, Pluralismus bezüglich Wahrscheinlichkeit folgendermaßen zu definieren: *Pluralismus in Bezug auf Wahrscheinlichkeit bezeichnet die Position, dass die zwei traditionellen Anwendungsfelder der Wahrscheinlichkeitstheorie, also induktives Schließen und relative Häufigkeiten, nicht von ein und demselben Konzept abgedeckt werden.* Es sei angemerkt, dass diese Definition nicht annimmt, dass beide Konzepte notwendigerweise die Wahrscheinlichkeitsaxiomatik erfüllen müssen. Unmittelbar einsichtig ist jedenfalls, warum es sich bei pluralistischen Positionen fast immer um einen Dualismus handelt.

Aufgrund dieser Definition entpuppen sich nun einige der im letzten Abschnitt vorgestellten Wahrscheinlichkeitsinterpretationen als pluralistische Positionen. Das betrifft vor allem Vertreter der ontischen Auffassungen wie Popper und bis zu einem gewissen Grad auch von Mises, wenn sie eine Rolle von Wahrscheinlichkeit für Induktion und Bestätigung grundsätzlich bestreiten. Popper zum Beispiel reserviert den Wahrscheinlichkeitsbegriff allein für seine Propensitäten. An die Stelle von induktiven Wahrscheinlichkeiten tritt bei ihm das Konzept der *Bewährung* von Hypothesen, welches explizit nicht den Wahrscheinlichkeitsaxiomen genügt (Popper 1934/2005, Kap. X und Anhang *IX). Hingegen sind Positionen wie die von Laplace, Keynes oder de Finetti keine pluralistischen Positionen, da sie das ganze Anwendungsspektrum abdecken, re-

lative Häufigkeiten eingeschlossen. Allgemein gilt, dass pluralistische Positionen eine indeterministische Grundhaltung voraussetzen und damit die Existenz objektiven Zufalls zugeben, da für einen strengen Deterministen Wahrscheinlichkeiten grundsätzlich epistemischer Natur sein müssen.

Insgesamt ergeben sich verschiedene Abstufungen und Grade von Pluralismus, abhängig davon wie stark die Konzepte sich voneinander unterscheiden. Am einen Ende des Spektrums findet man Positionen, bei denen in sehr unterschiedlichen Anwendungsbereichen weitgehend mit demselben Konzept gearbeitet wird. Hier sollte man generell nicht von Pluralismus sprechen. Am anderen Ende finden sich Positionen wie diejenige von Popper mit seinen sehr unterschiedlichen Begriffen Propensität und Bewährung, wobei letzteres nicht einmal die Wahrscheinlichkeitsaxiomatik erfüllt. Wir werden in den nächsten beiden Abschnitten zwei pluralistische Positionen kennenlernen, von denen der empirische Dualismus als Pluralismus weit ausgeprägter ist als der rationale Dualismus.

b. Empirischer Dualismus

Unter der Bezeichnung *empirischer Dualismus* lassen sich eine Reihe ontischer Positionen zusammenfassen. Die Namensgebung soll nahelegen, dass Vertreter dieser Auffassung bei den empirischen Aspekten von Wahrscheinlichkeit beginnen, also bei relativen Häufigkeiten oder bei Propensitäten. Eine dualistische Position ergibt sich in dem Augenblick, wenn die betreffenden Vertreter bestreiten, dass Wahrscheinlichkeit für eine Theorie des induktiven Schließens und der Bestätigung von Hypothesen verwendet werden kann. Damit bleibt nur ein Konzept, welches den Wahrscheinlichkeitsaxiomen genügt. Ein wie auch immer geartetes quantitatives Maß von Bestätigung erfüllt die Axiome explizit nicht. Karl Popper mit seinem Bewährungsbegriff ist ein paradigmatischer Vertreter.

Richard von Mises ist ein weiterer Kandidat für diesen empirischen Dualismus. So betont er wieder und wieder in „Probability, Statistics and Truth“, dass Wahrscheinlichkeit nichts mit Fragen zu tun hätte, ob Deutschland einmal mit Liberia einen Krieg führen wird (1928/1981, 9) oder ob eine Hochzeit den gewünschten Erfolg im Leben bringen wird (94). Er kritisiert heftig die logische Auffassung und legt nahe, dass Fragen der Bestätigung nicht zum Kerngeschäft der Wahrscheinlichkeitstheorie gehören (94-97). Andererseits bestreitet von Mises nicht, dass Wahrscheinlichkeitstheorie überhaupt für logische Überlegungen relevant sein kann: „I certainly do not wish to contest the usefulness of logical investigations, but I do not see why one cannot admit to begin with that any numerical statements about [...] degree of confirmation, etc., are actually statements about relative frequencies.“ (97)

Die Tatsache, dass von Mises logische Überlegungen auf relative Häufigkeiten zurückführen will, lassen ihn eher als wahrscheinlichkeitstheoretischen Mo-

nisten erscheinen. Allerdings beschränkt er damit empfindlich den Anwendungsbereich einer Bestätigungstheorie auf Basis von Wahrscheinlichkeit – nämlich auf solche Fragen, in welchen näherungsweise Kollektive vorliegen. Generell würde man aber von einer Theorie der Bestätigung erwarten, dass sie auch Fragen behandeln kann, wo dieses nicht der Fall ist – eben zum Beispiel ob Deutschland und Liberia Krieg führen werden. Letztlich müsste von Mises in einer vollständigen Bestätigungstheorie also ein neues Konzept einführen, welches dann explizit keine Wahrscheinlichkeit wäre. Das spricht dafür, ihn als empirischen Dualisten einzuordnen.

Der empirische Dualismus verlangt nach einem Argument, warum die Wahrscheinlichkeitstheorie induktives Schließen und Bestätigung nicht abdecken kann (oder zumindest nur Teilbereiche davon). Poppers Begründung in dieser Hinsicht ist tief in seiner falsifikationistischen Methodik verankert: „[Hypothesen mit hohem Grad der Prüfbarkeit] sind zugleich die in hohem Grade bewährbaren Hypothesen [...] Nun wissen wir aber, dass Prüfbarkeit dasselbe ist wie hohe (absolute) logische *Un*wahrscheinlichkeit." (1934/2005, 258; meine Kursivsetzung) Ergo, der Grad der Bewährung ist logischer Wahrscheinlichkeit entgegengesetzt und kann deshalb die Wahrscheinlichkeitsaxiomatik nicht erfüllen. Weil Wissenschaft nach Popper hypothetisch-deduktiv ist, dürfen Hypothesen im Allgemeinen keine Wahrscheinlichkeiten zugeordnet werden.

Ein weiteres Kennzeichen des empirischen Dualismus ist, dass eine systematische Theorie objektiven Zufalls vorausgesetzt werden muss. Nur dadurch gelingt es, eine rein empirische Rolle von Wahrscheinlichkeit zu begründen und von der Funktion von Wahrscheinlichkeit in der Bestätigungstheorie abzugrenzen. Bei von Mises ist der objektive Zufall durch das Regellosigkeitsaxiom bestimmt. Bei Popper und vielen anderen tragen unsere fundamentalen physikalischen Theorien, insbesondere die indeterministische Quantenmechanik, einen entscheidenden Teil zur Grundlegung des objektiven Zufalls bei.

Zusammenfassend zeichnet den empirischen Dualismus also aus, dass der Fokus auf die empirische Rolle von Wahrscheinlichkeit gesetzt wird. Eine Funktion von Wahrscheinlichkeit für die Bestätigung von Hypothesen wird entweder vollständig bestritten oder zumindest stark eingeschränkt. Damit wird ein neues Konzept notwendig, um den Bereich Bestätigungstheorie abzudecken, welches dann per definitionem nicht die Wahrscheinlichkeitsaxiomatik erfüllen kann. Die Plausibilität des empirischen Dualismus hängt vor allem davon ab, ob der weitgehende Ausschluss von Induktion und Bestätigung gerechtfertigt werden kann und inwieweit eine Theorie des objektiven Zufalls gelingt.

c. Rationaler Dualismus

Als *rationaler Dualismus* lassen sich eine Reihe von Positionen zusammenfassen, die traditionell eher den epistemischen Interpretationen zugerechnet wer-

den. Die Namensgebung soll darauf hindeuten, dass der Ausgangspunkt der rationale Aspekt von Wahrscheinlichkeit ist. Konsequenterweise findet man als Vertreter dieser Position im Vergleich zum empirischen Dualismus weit weniger Naturwissenschaftler. In unterschiedlicher Ausprägung könnte man zum Beispiel Rudolf Carnap, David Lewis und Frank Ramsey dazurechnen.

Vielleicht der Urvater des rationalen Dualismus ist Rudolf Carnap mit seiner Unterteilung in Wahrscheinlichkeit$_1$ und Wahrscheinlichkeit$_2$: „It seems to me that the number of explicanda in all the various theories of probability is neither just one nor about a dozen, but in all essential respects—leaving aside slight variations—very few, and chiefly two. [...] The two concepts are: (i) probability$_1$ = degree of confirmation; probability$_2$ = relative frequency in the long run." (1945, 517) Wahrscheinlichkeit$_1$ findet Anwendung im Bereich der induktiven Logik und damit in der Methodologie von Wissenschaft. Wahrscheinlichkeit$_2$ hingegen wird hauptsächlich in der mathematischen Statistik und ihren Anwendungen verwendet. Mit dieser strengen Aufteilung in eine empirische und logische Rolle von Wahrscheinlichkeit findet sich Carnap ganz in der Tradition des logischen Empirisimus, dessen Entwicklung er ja entscheidend mitgeprägt hat. Carnaps wichtigstes Tätigkeitsfeld war dabei die logische Wahrscheinlichkeit, die er allgemein für unterrepräsentiert hielt. Carnaps Fokus liegt also tatsächlich auf dem rationalen Aspekt von Wahrscheinlichkeit.

Anders als im empirischen gibt es im rationalen Dualismus grundsätzlich zwei Konzepte, die der Wahrscheinlichkeitsaxiomatik genügen. Das liegt vor allem daran, dass rationale Dualisten bei der epistemischen Rolle von Wahrscheinlichkeit beginnen und dass es andererseits nur schwerlich zu bestreiten ist, dass die empirischen relativen Häufigkeiten die Axiomatik erfüllen.

Ein Vertreter des rationalen Dualismus muss sich folglich über den konzeptionellen Zusammenhang zwischen den beiden Wahrscheinlichkeitsarten Gedanken machen. Diese konzeptionelle Brücke schlagen die so genannten *Probability Coordination Principles* (der Begriff stammt aus Strevens 1999). Das bekannteste darunter ist das *Principal Principle* von David Lewis (1986). David Lewis unterscheidet sich von Carnap darin, dass er dessen logische durch eine eher subjektive Wahrscheinlichkeit ('credence' oder 'degree of belief') und Carnaps relative Häufigkeiten durch Propensitäten ersetzt wissen will. Das Principal Principle zeigt dann den Zusammenhang zwischen diesen beiden Konzepten auf: „Let C be any reasonable initial credence function. Let t be any time. Let x be any real number in the unit interval. Let X be the proposition that the chance, at time t, of A's holding equals x. Let E be any proposition compatible with X that is admissible at time t. Then C(A|XE) = x." (Lewis 1986, 87) Als erlaubte ('admissible') Sätze fasst Lewis all das zusammen, was die Glaubensgrade nicht beeinflusst – beispielsweise historische Informationen. Letztlich weist uns das Principal Principle also an, unsere subjektiven Wahrscheinlichkeiten nach den objektiven zu richten, falls solche gegeben sind.

Bei Carnap finden sich bereits ähnliche Überlegungen über den Zusammenhang zwischen Wahrscheinlichkeit$_1$ und Wahrscheinlichkeit$_2$. In seiner intellektuellen Autobiographie schildert er seine Herangehensweise an Wahrscheinlichkeit$_1$ (Carnap 1963, 70-76). Er betont, dass diese neben ihrer Rolle in der Bestätigungstheorie insbesondere auch der Abschätzung von Wahrscheinlichkeit$_2$ dient. Im Zuge einer Ausarbeitung dieser Rolle von Wahrscheinlichkeit$_1$ widmet er sich einer detaillierten Kritik der frequentistischen Methoden in der angewandten Statistik.

Probability Coordination Principles setzen offenbar voraus, dass es etwas zu koordinieren gibt. Ein rationaler Dualist muss daher begründen, warum die empirische und die logische Rolle von Wahrscheinlichkeit unterschiedlicher Konzepte bedürfen. Das ist umso schwieriger als der rationale Dualist anders als der empirische zugesteht, dass beide Konzepte eng miteinander verwandt sind. Meistens verweist auch der rationale Dualist auf eine Theorie objektiven Zufalls, um die Probability Coordination Principles mit Leben zu füllen.

Zusammenfassend lässt sich sagen, dass der rationale Dualismus von der epistemischen Rolle von Wahrscheinlichkeit ausgeht, jedoch zusätzlich anerkennt, dass Wahrscheinlichkeit auch eine objektive und empirische Rolle spielt. Es gibt damit zwei unterschiedliche Konzepte, die beide die Wahrscheinlichkeitsaxiome erfüllen. Wegen dieses engen Zusammenhangs werden Probability Coordination Principles nötig. Um trotzdem eine grundsätzliche Trennung der beiden Rollen von Wahrscheinlichkeit zu rechtfertigen, benötigt der rationale Dualist im Allgemeinen eine Theorie objektiven Zufalls.

4. Kritik pluralistischer Positionen

a. Wider das Baukastendenken

Viele pluralistische Positionen scheinen nach dem Baukastenprinzip zusammengefügt. Dabei wird häufig übersehen, dass nicht alle verschiedenen Interpretationen leicht zusammenpassen. Beispielsweise ist eine Kombination aus de Finettis subjektiver Intepretation mit ontischen Wahrscheinlichkeitskonzepten wie der Propensitätstheorie nur schwer vorstellbar, weil erstere mit ihrem starken Antirealismus einer objektiven Rolle von Wahrscheinlichkeit grundsätzlich entgegensteht. Auch wenn epistemische Theorien wie bei Laplace aus einer stark deterministischen Grundüberzeugung entwickelt werden, ist eine Synthese mit ontischen Wahrscheinlichkeitskonzepten nur schwer vorstellbar. An anderer Stelle wird im Detail zu zeigen sein, wie eng metaphysische Grundannahmen bezüglich Determinismus und Realismus mit den entsprechenden Wahrscheinlichkeitsinterpretationen verknüpft sind. Weil diese Tatsache oft unberücksichtigt bleibt, wird die Vereinbarkeit der Wahrscheinlichkeitsinterpretationen all-

gemein überschätzt – wie beispielsweise in der folgenden Aussage von Alan Hájek, die für solch einen unvorsichtigen Pluralismus charakteristisch ist: „My bet, for what it is worth, is that we will retain at least three distinct notions of probability: one quasi-logical, one objective, and one subjective." (2007) Bemerkenswert ist in diesem Zusammenhang auch, dass über die gesamte historische Entwicklung gesehen pluralistische Positionen eher die Minderheit bilden. Die meisten Denker bevorzugten offenbar Interpretationen aus einem Guss.

b. Wissenschaft zielt auf Vereinheitlichung

Im rationalen Dualismus geht man von zwei sehr ähnlichen Konzepten aus, die insbesondere beide die Wahrscheinlichkeitsaxiomatik erfüllen. Es scheint in solch einer Situation methodisch gegeben auf eine Vereinheitlichung der Konzepte abzuzielen. Erst im zweiten Schritt wäre dann zu klären, wie aus einem einheitlichen Konzept unterschiedliche Aspekte folgen können. Das entspricht einer allgemeinen methodischen Maxime zur Vereinheitlichung wissenschaftlicher Begriffe. Wenn Gesetzmäßigkeiten verschiedene Anwendungsbereiche abdecken, aber kleine konzeptionelle Unterschiede aufweisen, sollte generell eine Synthese angestrebt werden. Wenn sich beispielsweise herausstellt, dass die Gesetze der Himmelsmechanik und der irdischen Mechanik eine sehr ähnliche Struktur haben und trotzdem im Detail wichtige Unterschiede aufweisen, so würde wohl kein Naturforscher ruhen ehe gezeigt ist, wie sie im unterschiedlichen Kontext aus einem einheitlichen Konzept folgen.

Im Umkehrschluss gilt, dass eine Erklärungslücke entsteht, wenn keine Vereinheitlichung der Konzepte angestrebt wird und wenn nicht gezeigt wird, wie die ähnlichen Konzepte aus einer gemeinsamen Wurzel folgen. Dann bleibt nämlich offen, warum die beiden Konzepte so starke strukturelle Ähnlichkeiten aufweisen. Im Fall des rationalen Dualismus bleibt zum Beispiel die Frage im Raum stehen, warum objektiver Zufall und induktives Schließen derselben Axiomatik folgen. Die Probability Coordination Principles liefern dabei keine Erklärung für die strukturellen Ähnlichkeiten, sondern lediglich eine Anleitung, wie die unterschiedlichen Wahrscheinlichkeitskonzepte aufeinander abzustimmen sind. Begnügt man sich also mit der pluralistischen Position, kann es qua Postulat keine Erklärung für die strukturellen Ähnlichkeiten der verschiedenen Wahrscheinlichkeitskonzepte geben.

c. Das Kontinuum wahrscheinlichkeitstheoretischer Anwendungen

Der rationale Dualismus scheitert also an der methodischen Maxime, dass grundsätzlich eine Vereinheitlichung von Konzepten angestrebt werden sollte, wenn sie eine große Ähnlichkeit aufweisen. Als Ausweg aus diesem Dilemma bietet sich an, einen derartigen engen Zusammenhang zwischen Bestätigungs-

maß und statistischer Wahrscheinlichkeit zu bestreiten. Das ist die Position des empirischen Dualismus. Wir erinnern uns in diesem Zusammenhang an Poppers Behauptung, dass Bewährungsgrade nicht einmal die Wahrscheinlichkeitsaxiomatik erfüllen.

Obwohl damit die Forderung nach Vereinheitlichung umgangen wird, ergibt sich nunmehr ein anderes Problem. Betrachtet man alle Anwendungen der Wahrscheinlichkeitstheorie, so findet sich nämlich keine deutliche Grenze zwischen rein objektiven und rein epistemischen Anwendungen. Vielmehr zeigt sich ein regelrechtes Kontinuum. Am einen Ende des Spektrums stehen scheinbar vollständig ontische Wahrscheinlichkeiten wie beispielsweise im Fall der Quantenmechanik. Im Weiteren sind jedoch Anwendungen auszumachen, in denen sich eine kontinuierliche Zunahme epistemischer Anteile beobachten lässt. Zufall im Glücksspiel scheint noch weitgehend objektiv zu sein und doch ist gut vorstellbar, dass sich bei immer besserer Kenntnis der Anfangsbedingungen der Ausgang der einzelnen Würfe immer genauer bestimmen lässt. Wenn man dann Bevölkerungsstatistiken untersucht, scheint der epistemische Anteil eine noch größere Bedeutung zu haben. Es ist klar, dass zum Beispiel bei einer Sterbestatistik der Tod des Individuums jeweils durch Einzelursachen genau bestimmt ist. Trotzdem bleibt ein statistischer Ansatz auf makroskopischer Ebene sinnvoll. Wenn man schließlich über Kriege mit Liberia redet, werden die Wahrscheinlichkeiten sehr ungewiss. Trotzdem ist nicht einzusehen, warum man nicht zumindest qualitative Intuitionen über Wahrscheinlichkeit zulassen sollte, schließlich lässt sich eine derartige Frage mit anderen historischen Situationen vergleichen. Der entscheidende Punkt ist folgender, dass die vergleichbaren Ereignisse immer unähnlicher der zu untersuchenden Aussage oder dem zu untersuchenden Ereignis werden. Diesem kontinuierlichen Spektrum von Ähnlichkeit entspricht ein kontinuierliches Spektrum von Wahrscheinlichkeitsanwendungen, dem der empirische Dualismus nicht gerecht werden kann.

Im empirischen Dualismus wird notwendigerweise eine scharfe Grenze gezogen, ab welchem Grad von Ähnlichkeit man nicht mehr von Wahrscheinlichkeit reden darf. So eine scharfe Grenze ist völlig unplausibel. Viel plausibler wäre eine ungefähre Grenze, ab welcher keine quantitativen Wahrscheinlichkeitswerte mehr verwendet werden sollten. Wirklich abgrenzen lässt sich nur der eine Fall, in welchem die relevanten Rahmenbedingungen der einzelnen Instanzen nicht nur sehr ähnlich, sondern in der Tat *identisch* sind. Dadurch könnte ein echter objektiver Zufall definiert werden. Aber solch eine Identität der relevanten Rahmenbedingungen ist empirisch prinzipiell nicht überprüfbar. Außerdem würde so ein objektiver Zufall schon die Wahrscheinlichkeiten beim Würfelspiel nicht mehr betreffen, da hier von einer Identität der relevanten Rahmenbedingungen grundsätzlich nicht mehr die Rede sein kann. Solch ein empirischer Dualismus würde also tatsächlich nur die Wahrscheinlichkeiten der Quantenmechanik aussondern. Das entspricht weder Poppers noch von Mises' Position.

Ein verwandter Kritikpunkt ergibt sich aus der Bemerkung, dass der empirische Dualismus streng zwischen Ereignis- und Hypothesenwahrscheinlichkeiten unterscheiden muss, wobei er dem zweiten Konzept grundsätzlich skeptisch gegenübersteht. Diese Unterteilung ist jedoch zutiefst zweifelhaft in Anbetracht der Tatsache, dass jede Ereigniswahrscheinlichkeit letztlich als Hypothesenwahrscheinlichkeit formuliert werden kann.

Allgemein gilt, dass der empirische Dualismus den Anwendungsbereich der Wahrscheinlichkeitstheorie sehr stark einschränkt. Von Mises beispielsweise lässt nur Glücksspiel, soziale Massenphänomene und die Anwendungen in der Physik zu (1928/1981, 8-10). Popper schließt jedwede Anwendung mit Bezug auf induktives Schließen und wissenschaftliche Methode aus. Im Zusammenhang mit *Humphrey's Paradox* wird genau diese Kritik auf die Spitze getrieben. Einerseits lässt die Propensitätstheorie keine Wahrscheinlichkeit für Hypothesen zu, andererseits sind diese ganz natürlich im Wahrscheinlichkeitsformalismus als inverse Wahrscheinlichkeiten vorgesehen. Es scheint, dass die Wahrscheinlichkeitstheorie damit nicht einmal mehr für die Fehlerrechnung in Betracht käme, die sich ja für die Verlässlichkeit von Hypothesen interessiert. Dieser gewaltsame Ausschluss einiger der fruchtbarsten Anwendungen der Wahrscheinlichkeitstheorie stellt den empirischen Dualismus vor große, vermutlich unüberwindbare Herausforderungen.

5. Zusammenfassung und Ausblick

Wir haben uns in diesem Essay an einer Klassifikation pluralistischer Auffassungen bezüglich Wahrscheinlichkeit versucht und insbesondere die Unterteilung in empirischen und rationalen Dualismus vorgeschlagen. Dies war die Voraussetzung dafür, dass wir im Folgenden eine grundlegende Kritik pluralistischer Positionen anbringen konnten. Die Abhandlung dieser diffizilen Fragestellungen konnte dabei nur überblicksartig geschehen. Zwischen den Zeilen liegt sicher noch so manches teuflische Detail verborgen. Allein die Frage nach der Existenz eines objektiven Zufalls, nicht zuletzt im Zusammenhang mit der Quantenmechanik, könnte ganze Bücher füllen. Dennoch bleibt die Hoffnung, dass unsere Kritik den pluralistischen Positionen einiges von ihrer intuitiven Plausibilität genommen hat. Der moderne Pluralismus bezüglich Wahrscheinlichkeit scheint nicht zuletzt der alten philosophischen Liebe zur Klassifikation zu entspringen. Dabei bleibt die naturwissenschaftliche Maxime zur Vereinheitlichung oft und in ungerechtfertigter Weise unberücksichtigt.

Literaturverzeichnis

Carnap, Rudolf : "The Two Concepts of Probability". Philosophy and Phenomenological Research, 5 (4), 1945. S. 513-532

Carnap, Rudolf: "Intellectual Autobiography". In: *Schilpp, Paul A. (Hrsg.):* The Philosophy of Rudolf Carnap. Open Court, La Salle, IL, 1. Auflage, 1963. S. 1-83.

de Finetti, Bruno: "Probabilism". Erkenntnis, 31, 1989, S. 169-223

Gallavotti, Maria C.: Philosophical Introduction to Probablity. CSLI, Stanford (CA), 1. Auflage, 2005

Gillies, Donald: Philosophical Theories of Probability. Routledge, London, 1. Auflage, 2000

Hájek, Alan: "Interpretations of Probability". In: *Zalta, Edward N. (Hrsg.):* The Stanford Encyclopedia of Philosophy, Spring 2010 Edition, http://plato.stanford.edu/archives/spr2010/entries/probability-interpret/

Keynes, John Maynard: A Treatise on Probability. Macmillan, London, 1. Auflage, 1921

Kolmogorow, Andrei N.: Foundations of the Theory of Probability. Chelsea Publishing, New York, 2. Auflage, 1956

de Laplace, Pierre S.: Philosophischer Versuch über die Wahrscheinlichkeit. Harri Deutsch, Frankfurt/M., 2. Auflage, 1996

Lewis, David K.: „A Subjectivist's Guide to Objective Chance". In: *Lewis, David K.*: Philosophical Papers, Vol. II., Oxford University Press, Oxford, 1. Auflage, 1986. S. 83-113

von Mises, Richard: "Grundlagen der Wahrscheinlichkeitsrechnung". In: *von Mises, Richard:* Selected Papers of Richard von Mises. Volume Two: Probability and Statistics, General. AMS, Providence, RI, 1. Auflage, 1964. S. 57-105

von Mises, Richard: Probability, Statistics and Truth. Dover, Mineola (NY), 2. Auflage, 1981

Popper, Karl R.: Logik der Forschung. Mohr Siebeck, Tübingen, 11. Auflage, 2005

Popper, Karl R.: "The Propensity Interpretation of Probability". The British Journal for the Philosophy of Science, 10 (37), 1959. S. 25-42

Ramsey, Frank P.: “Truth and Probability“. In: *Richard B. Braithwaite (Hrsg.):* Foundations of Mathematics and other Logical Essays. Kegan, Paul, Trench, Trubner & Co, London, 1. Auflage, 1931. S. 156-198

Strevens, Michael: “Objective Probabilities as a Guide to the World“. Philosophical Studies, 95, 1999. S. 243-75

The Non-Universality of Special Science Laws – How Quasi-Newtonian Laws avoid Lange's Dilemma

Alexander Reutlinger
areutlin@smail.uni-koeln.de
Universität Köln, Philosophisches Seminar

Abstract/Zusammenfassung

Laws in the special sciences are usually regarded to be non-universal. Due to their non-universality, theories of laws in the special sciences face a challenge: According to Lange's dilemma, they are either false or trivial. I argue that this challenge can be meet, if one distinguishes four dimensions of (non-)universality. In this paper, I focus on the dimension of non-universality with respect to external, disturbing factors. I argue that Lange's dilemma can be avoided with respect to this dimension of non-universality if one reconstructs laws in the special sciences as quasi-Newtonian laws.

Gesetze in den speziellen Wissenschaften werden gewöhnlich als nicht-universell bezeichnet. Aufgrund dieser Nicht-Universalität müssen sich Theorien über Gesetze in den speziellen Wissenschaften der folgenden Herausforderung stellen: Langes Dilemma zufolge sind solche Gesetze entweder falsch oder trivialerweise wahr. Ich argumentiere für die These, dass man Langes Dilemma umgehen kann, indem man vier Dimensionen der (Nicht-)Universalität von Gesetzen unterscheidet. In diesem Text werde ich mich auf eine Dimension der Nicht-Universalität beschränken – auf die Nicht-Universalität in Bezug auf externe Störfaktoren. Ich argumentiere für die These, dass Langes Dilemma in Bezug auf Gesetze, die nicht-universell in diesem Sinne sind, vermieden werden kann, wenn man Gesetze in den speziellen Wissenschaften als quasi-newtonsche Gesetze versteht.

1. Introduction: Why we need a Theory of Non-Universal Laws

Most philosophers are convinced that (fundamental) physics states universal laws, while the special sciences (e.g., biology, psychology, sociology, economics, medical science etc.) state non-universal or *ceteris paribus* laws (henceforth, cp-laws).[1] Paradigmatically, Barry Loewer describes the important differences between fundamental physical laws (he uses Newton's laws of motion as an example) and special science laws as follows:

1 Two clarifications: (1) I will use "non-universal laws" and "*ceteris paribus* laws" interchangeably. (2) My focus is on *law statements* rather than on laws themselves – thus, my aim is not to argue for any metaphysical claim.

> "The main relevant differences between fundamental dynamical laws and special science laws are these: The candidates for fundamental dynamical laws are (i) *global,* (ii) *temporally symmetric,* (iii) *exceptionless,* and (iv) *fundamental (not further implemented)* (v) *make no reference to causation.* In contrast, typical special science laws are (i*) *local,* (iii*) *temporally* asymmetric, (iii*) *multiply realized and implemented,* (iv*) *ceteris paribus,* and (v*) *often specify causal relations and mechanisms*." (Loewer 2008: 154, original emphasis)

In this paper, I will agree with Loewer that the dynamical laws in fundamental physics and in the special sciences differ in the way he describes.[2] Nonetheless, I will address explicitly only *some* of the features of special science laws, such as *being local*, *having exceptions* or *being ceteris paribus*. (I will leave aside features such as being temporally asymmetric, the possible multiple realization of special science laws, and the question whether and how fundamental physical laws relate differently to causation than special science laws.) Despite these differences, most philosophers believe that, in physics as well as in the special sciences, laws are important because they are statements used to explain and to predict phenomena, they provide knowledge how to successfully manipulate the systems they describe, and they support counterfactuals etc. Statements that play this role in the sciences I call *lawlike*. Note that, in the debate on laws of nature, lawlikeness is associated with universality (Braithwaite 1959, 301). I use "lawlike" differently: a general statement is lawlike if it is explanatory, of predictive use, successfully guides manipulation, and supports counterfactuals.

Let me provide two examples of special science laws. An example from economics is the *law of supply and demand*, which – in the words of John Roberts, a critic of cp-laws – states:

> "If the supply of a commodity increases (decreases) while the demand for it stays the same, then the price decreases (increases); if the demand for a commodity increases (decreases) while the supply remains the same, then the price increases (decreases)."[3]

Another example is the area law in island biogeography:

> "the equilibrium number S of a species of a given taxonomic group on an island (as far as creatures are concerned) increases [polynomially][4] with the islands area [A]: $S=cA^z$. The (positive-valued) constants c and z are specific to the taxonomic group and island group." (Lange 2000, 235f; see Lange 2002, 416f.)

Although I lack space to discuss them here, other vividly debated examples of special science laws are: in neuroscience, the Hodgkin-Huxley Model of the ac-

2 Many of the problems I will discuss in the paper would be even trickier if one disagreed with Loewer (and others) at this point. Some philosophers believe that even fundamental physics deals (at least in part) with non-universal laws. If this were the case, the issue of non-universal laws might turn out to be even more pressing.

3 Cf. Roberts (2004, 159), Kincaid (2004, 177).

4 Lange mistakenly writes "exponentially".

tion potential (Weber 2008: 997-1001) and the generalizations describing the mechanism of Long-Term Potentiation (Craver 2007: 65-72, 168); in psychology, generalizations describing learning and memory (Gadenne 2004: 107f); in economics, generalizations in models of economic growth (Kincaid 2009: 456f); and, in biology, Mendel's law of segregation and the Hardy-Weinberg law (Rosenberg & McShea 2008: 36). Generalizations like these are believed to be *lawlike*, although they are not universal generalizations.[5] But, traditionally[6], the most important feature of a law to understand its lawlikeness is *universality*. Furthermore, picturing lawlikeness mainly in terms of universality has lead many theories of causation and explanation to rely on universal laws. The major challenge for any theory of non-universal laws in the special sciences is to account for their apparent lawlike function (in the sense introduced above).

In this paper, I argue as follows: In section 2, I will introduce Lange's dilemma stating that special science laws are either false or trivially true. In section 3, I set up a theory of non-universal laws by distinguishing different meanings (or dimensions) of "non-universal". I assume that special science laws are universal regarding the first and second dimension, and they are non-universal regarding the third and fourth dimension of universality. In the paper, I focus on the third dimension of non-universality, i.e. the claim that special science laws are sensitive to external, disturbing factors. In section 4, I argue that – with respect to the third dimension of non-universality – special science laws should be reconstructed as quasi-Newtonian laws in order to avoid Lange's dilemma. In section 5, I address the question whether all possible disturbing factors are to be considered equally relevant. Building on ideas by Marc Lange (2000), I will argue that, from a pragmatic point of view, the relevance of disturbing factors can be ranked (and this seems to be an adequate description of scientific practice in the special sciences). In section 6, I conclude that the results of the preceding sections amount to the following necessary condition for a theory of special science laws: if L is a special science law, then L is backed up a methodology to describe disturbing factors, i.e. L is a quasi-Newtonian law.

5 It is certainly a matter of convention whether one would still want to use the term "law" for non-universal general statements. One can either use a new term for non-universal explanatory, general statements (e.g., Woodward and Hitchcock 2003 introduce the term "explanatory generalization"). Or, as I maintain in this paper, one can insist that if a statement plays a lawlike role then it shares sufficiently many properties with universal laws in order to be called a law. Note that Hitchcock and Woodward admit that their account may be read as a *reconceptualization* of lawhood (cf. Woodward and Hitchcock 2003, 3).

6 Cf. Lewis (1973, 73-76) and Armstrong (1983, 88-93).

2. A Challenge: Falsity or Triviality

A philosophical reconstruction of special science laws faces a severe problem, which can be articulated in the form of *Lange's Dilemma*[7]. Here is the first horn (*Falsity*): Strictly and literally speaking, special science laws are false because it is not the case that all Fs are Gs (if that is what the law says). For instance, the relationship between supply and price is not always as the law of supply says (or, as it seems to say *prima facie*), because an interfering factor might occur. In other words, special science laws that instantiate perfect regularities are – mildly put – "scarce" (Cartwright 1983, 45). Yet, if one supposes that the law of supply is to be formalized as a universally quantified conditional sentence. Then one counter-instance (due to a disturbing factor) to the universally quantified sentence means that it is false.

The second horn of Lange's Dilemma (*Triviality*) can be stated as follows: If laws in the special sciences are cp-laws, then they are trivially true. If we instead suppose that an implicit cp-clause is attached to the law then it seems to mean 'All Fs are Gs, *if nothing interferes*'. But then the cp-law in question is in danger to lack empirical content. It lacks empirical content because it seems to say nothing more than 'All Fs are Gs or it is not the case that all Fs are Gs'. If this is the correct theory of cp-laws in the special sciences, then cp-laws are analytically true sentences and, therefore, trivially true. Obviously, this is a bad result because laws of special science should be reconstructed as *empirical* statements – not as sentences being true in virtue of meaning of their components.

Note that the second horn is a more pressing problem than the first, because I have already given up the assumption that special science laws are universal (as presupposed in the first horn). In the recent debate, some philosophers take Lange's Dilemma as a reason to be entirely pessimistic about whether there really is a convincing explication of laws in the special sciences:

> „[...] there is no persuasive analysis of the truth conditions of such laws; nor is there any persuasive account of how they are saved from vacuity; and, most distressing of all, there is no persuasive account of how they meld with standard scientific methodology, how, for example, they can be confirmed or disconfirmed. *In sum, a royal mess*." (Earman and Roberts 1999: 470f, my emphasis)

So, to deal with this dilemma is clearly a central challenge.

3. Four Dimensions of Non-Universal Laws

It is the received view that, in the special sciences, laws appear to be non-universal – or, they are said to 'have exceptions'. But what does it mean to be uni-

7 Cf. Lange (1993, 235).

versal, and, respectively, non-universal? Surprisingly, in the recent debate on cp-laws this question is not answered in a systematic way.[8] The lack of a systematic approach is a major problem, because universality is an ambiguous concept. We may distinguish four meanings or *dimensions* of universality[9] with respect to a law statement:

- *First Dimension – Universality of space and time*: Laws are $universal_1$ iff they hold for all space-time regions.
- *Second Dimension – Universality of Domain of Application*: Laws are $universal_2$ iff they hold for all (kinds of) objects.
- *Third Dimension – Universality for External Circumstances*: Laws are $universal_3$ iff they hold under all external circumstances (i.e. circumstances that are not referred to by the law statement itself).[10]
- *Fourth Dimension – Universality with respect to the Values of Variables*: Laws are $universal_4$ iff they hold for all possible values of the variables[11] in the law statement. Universality in this sense acknowledges that laws usually are quantitative statements (and, thus, the predicates contained in these statements are to be conceived as variables ranging over a set of possible values).

In this paper, I will focus solely on the third dimension. For the sake of the argument, I will assume that laws in the special sciences are (a) universal in the first and second dimension and (b) non-universal with respect to the third and the fourth dimension of universality. This diagnosis with respect to the fourth dimensions amounts to a challenge: Any theory of cp-laws is obliged to explain for each dimension how a cp-law can be non-$universal_3$ and *still* play a lawlike role. Since I focus on the third dimension here, an important step towards meet-

8 Exceptions are Mitchell (2000) and Schurz (2002).

9 Cf. Hüttemann (2007, 139-41) and Craver (2007, 66-69) for similar distinctions.

10 A useful way to spell out the third dimension of universality could be found in Loewers use of "global" (as introduced in the quote in the introduction): "The dynamical laws of classical mechanics are complete and deterministic. Given the state at any time t they determine the state at any other time. The determination is *global* since the position and momentum of any particle at a time t+r is determined only by the global (i.e. the entire) state of that system at time t. That is, to know how any one particle moves at t+x one has to know something at each particle at t. The dynamical laws and a partial description of state at t (except in special cases) do not entail much about the state of the system at other times and, in particular, don't say much about what any particular particle will (was) doing at t+r." (Loewer 2008: 155)

11 A variable X (in the terminology of statistics) is a functional property $X:D\rightarrow ran(X)$ of individuals (or outcomes) $d\in D$ of a domain D, where ran(X) is the set of possible values $x\in ran(X)$ of the variables (for quantitative variables X, ran(X) is the set of rational numbers). For example, temperature is represented by a variable T that has several possible values such as $T=30^{o}$.

ing this challenge is to show how a cp-law that is non-universal$_3$ can avoid Lange's dilemma. To show precisely this will be my aim in sections 4 and 5.

4. Non-Universality for External Circumstances: The Method of quasi-Newtonian Laws

How shall we deal with the third dimension, i.e. the fact that special science laws are sensitive to external factors? Recall that Lange's dilemma only applies to those reconstructions of law statements that are qualified by *only* a cp-clause like '*all disturbing factors are absent*'. But one is *not* committed to this reading of the cp-clause. Is there an alternative reconstruction? I think that this is the case and my positive thesis is: cp-laws are backed up by a methodology which allows to describe (comparatively or quantitatively) the influence of relevant disturbing factors.

I will introduce a methodology in order to argue for my positive thesis: the method of quasi-Newtonian laws. The basic idea of the method of quasi-Newtonian laws is: factors that lead to a counter-instance of the law L are described by other laws. Note that typically laws are not isolated but part of a theory or a model. Some disturbing factor with respect to law *L* is described by law *L** in the same theory or model. Further, it is important to point out that the idea of quasi-Newtonian laws aims at non-epistemic and perfectly objective truth-conditions of cp-laws and it does not merely concern the epistemic acceptability conditions of a cp-law (as, e.g., Pietroski and Rey 1995 argue).

Originally, this key idea of dealing with disturbing factors has been proposed by John Stuart Mill:

> "The disturbing causes have their laws, as the causes which are thereby disturbed have theirs; and from the laws of disturbing causes, the nature and amount of the disturbance may be predicted *a priori*, like the operation of the more general laws which they are said to modify or disturb, but which they might more properly said to be concurrent." (Mill 1836/2008, 50)

For instance, the law of supply states "If the supply of a commodity increases (decreases), then the price decreases (increases)". It is usually added "… *while the demand for this commodity stays the same*" which implies that the law of supply does not hold if the demand increases or decreases. At this point it is crucial to notice that the evolution of the price of a good is not described by a single generalization, i.e. by the law of supply. The evolution of the price also depends on another factor, the demand of a good, described by the law of demand: "if the demand for a commodity increases (decreases) *while the supply remains the same*, then the price increases (decreases)". It is to be emphasized that the equilibrium model of supply and demand also describes what would happen, if demand did not remain the same. In other words, the evolution of the price of a

commodity is described by an equilibrium model, according to which supply and demand can vary independently (Hausman 1992; Mas-Colell, Whinston and Green 1995).

In order to illustrate Mill's original idea how a disturbing factor can also be described by a law, I draw on Tim Maudlin's (2004) concept of a quasi-Newtonian law. In Newton's physics, Newton's First Law describes the inertial behavior of a physical system, i.e. the uniform motion of a physical system when no force acts upon it. Newton's Second Law describes the deviant behavior, i.e. the change of inertial motion if other forces *are* present. Maudlin characterizes the general form of quasi-Newtonian laws by analogy:

> "Let us denominate laws *quasi-Newtonian* if they have this form: There are, on the one hand, *inertial laws* which describe how some entities behave when nothing act on them, and there are *laws of deviation* that specify in what conditions, and in what ways, the behavior will deviate from the inertial behavior." (Maudlin 2004, 431, my italics)

Maudlin (2004, 434) stresses that special science laws are typically quasi-Newtonian. So, let us apply Maudlin's idea to the economic case: The law of demand describes inertial behavior; if the law of supply is integrated in the model, then the whole model describes the deviant behavior (of the price). Thus, the law of demand and supply describing the evolution of the price of a commodity is a quasi-Newtonian law. As Lange observes, the case is analogous concerning the area law (viewed as an inertial law) in island biogeography: It matters how far an island is away form the coast – being very far away might count as a disturbance of the area law. Island bio-geographers describe this disturbing factor (for the area law) by the *distance law*: "*ceteris paribus*, islands farther away from the mainland equilibrate at lower biodiversity levels" (Lange 2002, 419). In Maudlin's terminology, we might call the distance law a law of deviation (with respect to the area law as an inertial law).

Note carefully that there are important disanalogies between Newton's laws and special science laws that are understood in terms of quasi-Newtonian laws: First, in the economic (and the island bio-geographical) case it depends on pragmatic choice which law is dubbed "inertial law" and which "law of deviation". Secondly, Newton's laws and the laws of supply and demand (in an equilibrium model) differ mathematically. Thirdly, in cases of special science laws the deviation laws are (usually) not $\text{universal}_{1\text{-}4}$ as Newton's Second Law is.

Despite these disanalogies, I think that the positive analogy remains intact: Some laws in physics and in the special sciences are quasi-Newtonian because the influence of a disturbing factor on a system describes by a law L can be described (comparatively or quantitatively) by another law L*.

5. Which Disturbing Factors are Important? – A Pragmatic Approach

The method of quasi-Newtonian laws raises a question: Do scientists need to know all the laws of deviation descrinbing the influence of all possible disturbing factors? Are all (possible) disturbing factors equally important? And, if this is not the case: how does one distinguish important disturbing factors from irrelevant ones? Intuitively speaking, it seems quite obvious (and descriptively adequate) that scientists are not interested in all possible disturbing factors. Rather, scientists seem to discriminate relevant and (more or less) irrelevant disturbing factors. I attempt to give a pragmatic answer to the above mentioned questions that builds on ideas by Lange (2000).

According to Lange's core idea, cp-laws are stable (sets of) propositions whose application is *pragmatically* restricted to the purposes of a scientific discipline. As Lange formulates the point with respect to the laws in the special sciences, or inexact sciences as he calls them:

> "A set is stable *for the purpose of an inexact science* if and only if it is invariant under every counterfactual supposition of interest to the science and consistent with the set." (Lange 2002, 416).

Yet, let us not bother with Lange's stability theory of laws here (cf. Lange 2009: chapter 1). Instead let us focus on what he says about disturbing factors. Lange tries to avoid the horns of the Falsity-or-Triviality-Dilemma by treating 'ceteris paribus' as a name for a set **I** of interfering factors. Note that **I** does not list *all* the possible interferences which prevent the occurrence of B in 'cp, all As are Bs', but only those factors that are *relevant* (for a discipline). One can understand Lange in a way that he provides two strategies[12] to determine the members of **I**: (A) the strategy of non-negligibility and (B) the strategy of intended interest of a science. Both strategies try to explicate a methodology that is implicitly used by scientists in a particular discipline (cf. Lange 2000, 170-174).

(A) *The strategy of non-negligibility*: Instead of providing a *complete* list of all interfering factors, scientists merely refer to those interfering factors

> "that arise sufficiently often, and can cause sufficiently great deviations from *G*-hood, that a policy of inferring *F*s to be *G* […] would not be good enough for the relevant purposes" (Lange 2002, 411; Lange 2000, 170f).

For instance, consider the economic law of supply. According to Lange, it may happen that the increase in supply is so small that no decrease in prize results. But it might as well happen that the price does not decrease although the supply

12 Note that to distinguish two "strategies" and to call them "strategies" is my way to reconstruct Lange's approach – Lange does not refer to these ideas in this terminology. According to Lange, these "strategies" are for the most part *implicit* in scientific practice.

increases significantly, because a gigantic comet hitting the planet Earth and destroying all life on its surface disturbs the instantiation of this law. The comet causes sufficiently great deviation from a decrease in the price of a good. Nevertheless comets are negligible for the purposes of economists because their occurrence does not arise sufficiently often to count as interfering factor that is to be explicitly listed in the cp-conditions.

(B) *The strategy of intended interest of a science*: A law may still count as stable if it fails to hold under those counterfactual suppositions that do not fall into the range of the laws intended purpose and application. This point is best illustrated by an example from island biogeography – the area law – provided by Lange (2002, 416f.; Lange 2000, 235f – see introduction).

There are counterfactual suppositions (which describe the occurrence of actual and merely possible disturbing factors) for which the area law is not true. For example, imagine an island where the animals of the species "chicken" exclusively live on chicken farms. Suppose further that on these farms chicken are bred and held under extremely crowded conditions. So, the counterfactual supposition stemming from this example is "chicken on the island in question are bred under extremely crowded, artificial conditions set up by farmers". Obviously, the area law will drastically fail to hold for this case. Moreover, cases of this kind are not far-fetched philosophical thought experiments: they do *not* occur *rarely* in times of cultivated breeding of animals (as required by *the strategy of non-negligibility*). Nevertheless, scientists exclude this kind of exceptions because it conflicts with the intended purpose and application of their discipline (in this case: island biogeography and the area law).

Lange discusses another example of counterfactual suppositions that fail to be consistent with the intended interest and application of island biogeography. In his example, it is counterfactually assumed that the earth lacks a magnetic field. Although the laws of physics appear to be preserved for this counterfactual supposition, it seems that this is not the case for the area law in island biogeography. In other words, assuming drastic changes in gravitational force is a possible disturbing factor for the area law. Does this result indicate that the laws of island biogeography (and, analogously, the laws of other special sciences) fail to avoid Lange's dilemma and fail to be lawlike? Lange offers a pragmatic answer to this question in favor of the stability of special science laws:

> "The area law is not prevented from qualifying as an island-biogeographical law – from belonging to a set that is *stable for the purposes of island biogeography* – by its failure to be preserved under the [...] counterfactual suppositions […]. The supposition concerning Earth's magnetic field falls outside of island biogeography's range of interest. It twiddles with *a parameter that island biogeography takes no notice of*, or at least does *not take it as a variable*." (Lange 2002, 418)

The counterfactual supposition above refers to a parameter or variable that lies 'offstage' (given the intended interests of island bio-geography), as Lange (2000, 232) formulates the issue.

It is illuminating to contrast Lange's strategies of determining a set of disturbing influence **I** (in the light of intended applications of a law) with the use of cp-clauses that does indeed render a statement trivial. Lange (2000, 172; cf. also 2002, 410) illustrates this point by an example of a mere "excuse clause":

> "suppose someone says 'I can run a four-minute mile' but with each failure reveals a proviso that she had not stated earlier: 'except on this track', 'except on sunny Tuesdays in march' and so on. It quickly becomes apparent that this person will not acknowledge having committed herself to any claim by asserting 'I can run a four-minute mile.'"

According to Lange, excuse clauses of this kind differ from cp-clauses as used in the sciences, because the latter refer to strategies of determining disturbing factors (for instance, by describing their influence on the target system by means of quasi-Newtonian laws). Although the relevant set of disturbing factors is not listed explicitly, it is implicit in the scientific practice (and the education and studies) in a particular field of inquiry.

The lesson we can learn from Lange is that, ontologically speaking, all of the disturbing factors are on a par. Yet, from a pragmatic point of view, it seems to be the case that scientists rank the relevance of disturbing factors with respect the aims of the research in their particular fields. Following Lange, two ways to describe this ranking are the *strategy of non-negligibility* and the *strategy of intended interest of a science*. If one considers these strategies to be sensible, then only those disturbing factors that are evaluated as relevant in the light of these strategies have to be described by laws of deviation.

6. Meeting the Challenges: Lange's Dilemma and the Requirement of Relevance

So, does the method of quasi-Newtonian laws help to avoid Lange's dilemma with respect to the third dimension non-universality? Lange's Dilemma addresses a problem of generalizations that are context-sensitive with respect to disturbing factors. Naturally, I claim that our theory of the third dimension of non-universality of special science laws has to deal with Lange's Dilemma: quasi-Newtonian laws are describing the influence of disturbing factors. I will now argue that Lange's Dilemma can be avoided by relying on quasi-Newtonian laws.

Does the *method of quasi-Newtonian laws* avoid Lange's Dilemma? I argue that it does: The *Method of quasi-Newtonian* laws describes the influence of disturbances. It avoids *Falsity* because the occurrence of a disturbance does not render the law L in question false – instead the influence of a disturbing factor is

described by another law L*. It avoids *Triviality* because it is not (exclusively) committed to the fatal expression 'if nothing interferes' – rather quasi-Newtonian laws describe two kinds of situation: undisturbed (i.e. "inertial") behavior and disturbed ("deviant") behavior.

Moreover, the question whether all possible disturbing factors are equally relevant is answered negatively (in section 5). Yet, the ranking of disturbing factors with respect to their relevance is shifted from an ontological terrain to a pragmatic one, i.e. the ranking is relative to the aims of researchers in a particular field.

To sum up, quasi-Newtonian laws and a pragmatic account of relevance of disturbing factors (as presented in section 5) are supplements for a theory of non-universal laws.

7. Conclusion: The Explication of Special Science Laws

I started out by asking how non-universal generalizations in the special science can perform a lawlike function. One important step towards answering this question consists in showing that a theory of non-universal laws can avoid Lange's dilemma. In order to show precisely this, I distinguished four dimensions of non-universality, and I focused only on the third dimension of non-universality in this paper, i.e. the non-universality of special science laws with respect to external, disturbing factors. The non-universality in the third dimension is taken care of by methods that describe the influence of disturbing factors (Mill-Maudlin-view of quasi-Newtonian laws). My explication of a special science law relies, at least, on the following necessary condition: A statement L is a special science law, if L is backed up a methodology to describe disturbing factors, i.e. if L is quasi-Newtonian. I argue that this necessary condition (as part of an explication of special science laws) has an important advantage: it avoids Lange's Dilemma. I dare to conclude that, according to my reconstruction, special science laws are at least good candidates of empirical general statements playing a lawlike role in the special sciences.

References

Armstrong, David: What Is a Law of Nature?. Cambridge University Press, Cambridge, 1983

Braithwaite, Richard: Scientific Explanation. Cambridge University Press, Cambridge, 1959

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

Earman, John, and John Roberts: "Ceteris Paribus, There is no Problem of Provisos". *Synthese,* 118, 1999. S. 439-478

Gadenne, Volker: Philosophie der Psychologie. Verlag Hans Huber, Bern, 2004

Hausman, Daniel: The Separate and Inexact Science of Economics. Cambridge University Press, Cambridge(MA), 1992

Hüttemann, Andreas: "Naturgesetze". In: *Andreas Bartels, and Manfred Stöckler (eds.)*: Wissenschaftstheorie. Mentis, Paderborn, 2007. S. 135-153

Kincaid, Harold: "Are There Laws in the Social Sciences?: Yes". In: *Christopher Hitchcock (Hrsg.):*Contemporary Debates in the Philosophy of Science, Blackwell, Oxford, 2004. S. 168-187

Kincaid, Harold: "Explaining Growth". In: *Kincaid, H., and D. Ross (Hrsg.):* The Oxford Handbook of Economics. Oxford University Press, Oxford, 2009. S. 455-475

Lange, Marc: "Natural Laws and the Problem of Provisos". Erkenntnis, 38, 1993. S. 233-248

Lange, Marc: Natural Laws in Scientific Practice. Oxford University Press, Oxford, 2000

Lange, Marc: "Who's Afraid of Ceteris Paribus Laws? Or: How I learned to stop worrying and love them". In: *J. Earman et al. (Hrsg.):* Ceteris Paribus laws. Erkenntnis, 52 (Special Issue), 2002. S. 407-423

Lange, Marc: Laws and Lawmakers. Oxford University Press, Oxford, 2009

Lewis, David: Counterfactuals. Blackwell, Oxford, 1973

Loewer, Barry: "Why There *Is* Anything Except Physics". In: *Jakob Hohwy and Jesper Kallestrup (Hrsg.):* Being Reduced. New Essays on Reduction, Explanation, and Causation. Oxford University Press, Oxford, 2008. S. 149-163

Mas-Colell, A., M. Whinston, and J. Green: Microeconomic Theory. Oxford University Press, Oxford, 1995

Maudlin, Tim: “Causation, Counterfactuals, and the Third Factor”. In: *Ned Hall et al. (eds.):* Causation and Counterfactuals. MIT Press, Cambridge (MA), 2004. S. 419-444

Mill, John Stuart: “On the Definition and Method of Political Economy”. In: *Daniel Hausman (Hrsg.):* The Philosophy of Economics. An Antholo-gy.Cambridge University Press, New York, Third Edition, 1836/2008. S. 41-58

Mitchell, Sandra: “Dimensions of Scientific Law”. Philosophy of Science, 67, 2000. S. 242-265

Pietroski, Paul and Georges Rey: “When other things aren’t equal: Saving Ce-teris Paribus Laws from Vacuity”. British Journal for the Philosophy of Science, 46, 1995. S. 81-110

Roberts, John: “There are No Laws in the Social Sciences”. In: *Christopher Hitchcock (Hrsg.):* Contemporary Debates in the Philosophy of Science. Blackwell, Oxford, 2004. S. 168-185

Rosenberg, Alexander and Robert McShea: Philosophy of Science. Routledge, London, 2008

Schurz, Gerhard: “Ceteris Paribus Laws: Classification and Deconstruction”. In: *J. Earman et al. (Hrsg.):* Ceteris Paribus laws. Erkenntnis, 52 (Special Issue), 2002. S. 351-372

Tooley, Michael: “The Nature of Laws”. Canadian Journal of Philosophy, 7, 1977. S. 667-98

Weber, Marcel: “Causes without Mechanisms: Experimental Regularities, Physical Laws, and Neuroscientific Explanation.” Philosophy of Science, 75, 2008. S. 995-1007

Woodward, James and Christopher Hitchcock: “Explanatory Generalizations, Part I: A Counterfactual Account”. Nous, 37.1, 2003. S. 1-24

Kreationismus, Bayesianismus und das Abgrenzungsproblem

Gerhard Schurz
schurz@phil-fak.uni-duesseldorf.de
Universität Düsseldorf

1. Anthropisches Prinzip, Kreationismus, und das Abgrenzungsproblem

Wenn auch die moderne Evolutionstheorie zahlreiche Rätsel zur Entstehung des Lebens lösen und den Einfluss des religiösen Kreationismus zurückdrängen konnte, so hat man in jüngerer Zeit herausgefunden, dass die Möglichkeit der Evolution höherer Lebensformen auf unwahrscheinlich fein tarierten Eigenschaften unseres Planeten, Sonnensystems und Universums beruht (s. Ward und Brownlee 2000, Smolin 1997). Darunter fallen Sachverhalte wie z.B. dass unsere Erde in jener sehr engen Temperaturzone um die Sonne kreist, welche flüssiges Wasser ermöglicht, dass ihr Magnetfeld die Rotationsachse stabilisiert und kosmische Strahlung ablenkt, dass Nachbarplaneten wie Jupiter Kometeneinschläge reduzieren, bis dahin dass nur geringfügige Veränderungen der Werte der Naturkonstanten die Materie instabil werden lassen oder zumindest Sternbildung verhindern würden. Die Konsequenzen dieser gut bestätigten Sachverhalte in Bezug auf die ethische Verantwortung des Menschen für seinen Planeten würden eine eigene philosophische Betrachtung verdienen, doch darauf gehe ich hier nicht ein. Mich interessiert hier vielmehr die Frage der Erklärbarkeit dieser Sachverhalte.

Im Zuge des Bestrebens, diese 'Unwahrscheinlichkeit' unserer Welt zu erklären, haben unter dem Namen "anthropisches Prinzip" kreationistische Erklärungsversuche neuen Auftrieb erhalten, sogar innerhalb der Physik und Kosmologie (Barrow und Tipler 1988, Davies 1995). Das anthropischen Prinzip besagt folgendes:

> (AP) Die Parameter unserer Welt sind ist so unwahrscheinlich, wie sie sind, "weil" wir Menschen (bzw. komplexe Lebensformen) darin existieren können.

Offenbar ist das Prinzip mehrdeutig. Man kann das "weil" im schwachen Begründungs- und im starken Erklärungssinn auffassen. Im Begründungssinn ist das AP harmlos, da es im Effekt nur sagt, dass es uns Menschen nicht geben könnte, wären die Parameter unserer Welt nicht so unwahrscheinlich, wie sie

sind.[1] Die angesprochenen neo-kreationistischen Strömungen fassen das "weil" des AP dagegen im Erklärungssinn auf, und zwar im *finalistischen* Erklärungssinn (ein kausaler Erklärungssinn ist unmöglich, da das Erklärte in der Vergangenheit des Erklärenden liegt). Die Parameter unserer Welt sind dieser Interpretation des zufolge so unwahrscheinlich, wie sie sind, *damit* höhere Lebensformen entstehen konnten – denn es gibt einen intelligenten Kreator, welcher die Welt zweckmäßig eingerichtet hat.

Eine Reihe von 'Neo-Kreationisten' haben auf dieser Basis versucht, den Kreationismus als wissenschaftliche Hypothese zu etablieren. Würde das gelingen, so wären damit starke Argumente für Forderungen nach einem gleichberechtigten Unterricht von Evolutionsbiologie und Religion im Schulunterricht gewonnen. Es wurde sogar versucht, kreationistische Erklärungen durch explizite Anwendung gewisser wissenschaftstheoretischer Methoden zu rechtfertigen – nämlich mithilfe bayesianischer Bestätigungsmethoden (Swinburne 1979, Unwin 2005). Wäre dies tatsächlich möglich, so wäre das sogenannte Abgrenzungsproblem – also das Problem einer objektiven Abgrenzung zwischen Wissenschaft und Spekulation – unlösbar. In der Tat wurde nach der durch Thomas Kuhn ausgelösten Wende in der Wissenschaftstheorie die Möglichkeit einer solchen Abgrenzung stark in Zweifel gezogen. Das Abgrenzungsproblem, welches frühere Wissenschaftstheoretiker wie Popper oder Carnap lösen wollten, wurde als obsolet angesehen. Den Wissenschaftstheoretikern ging es damals vorwiegend um die Kritik eines zu engstirnigen Wissenschaftspositivismus, während nur wenige daran dachten, dass eine auf öffentlicher Schulbildung basierende Gesellschaft großen Bedarf besitzt an plausiblen Unterscheidungskriterien zwischen allgemeinverbindlichem Wissen im Gegensatz zu bloßen Spekulationen.

Die meisten heutigen Wissenschaftstheoretiker wenden sich gegen die neo-kreationistischen Versuche, Religion als wissenschaftliche Hypothese zu etablieren. Doch wie könnte nach aller Selbstkritik der Wissenschaftstheorie die hierfür nötige diese Abgrenzung nun vor sich gehen? Kann eine plausible Grenze zwischen Wissenschaft und rationalisierter Religion überhaupt gezogen werden? In diesem Vortrag möchte ich zeigen, dass dies durchaus der Fall ist. Die bayesianische Bestätigungstheorie reicht dafür jedoch nicht aus – sie liefert nur eine notwendige, und keine hinreichende Bedingung für rationale Bestätigung. Die gesuchte Abgrenzung muss auf andere Weise etabliert werden.

1 Es gibt weitere Varianten des APs, auf die hier nicht eingegangen wird (vgl. de.wikipedia.org/wiki/Anthropisches_Prinzip).

2. Empirisch kritisierbarer versus unkritisierbarer Kreationismus

Zunächst müssen zwei Arten kreationistischer Lehrgebäude unterschieden werden:

(1.) Empirisch kritisierbare Kreationismen besitzen empirische Konsequenzen besitzen, an denen sie überprüfbar sind. Darunter fallen die meisten traditionellen Religionen, und sie werden aufgrund ihrer unzutreffenden empirischen Konsequenzen widerlegt oder zumindest wahrscheinlichkeitsmäßig stark geschwächt. Diese Diagnose trifft nicht nur auf strikte *Genesis-Kreationismen* zu, welche falsche historische Faktenbehauptungen implizieren, wie etwa dass unsere Erde erst 60.000 Jahre alt wäre. Er trifft auch auf zahlreiche *Design-Kreationismen* zu, welche in Bezug auf die Genesis einen liberalen Standpunkt vertreten, aber die funktionale Perfektion der gottgeschaffenen Lebewesen hervorheben. Denn die Produkte der Evolution keineswegs funktional perfekt, sondern voller Inperfektheiten. Kein intelligenter Konstrukteur würde beispielsweise auf die Idee kommen, das Skelett von Walflossen mit fünf Fingerknochen zu versehen – erst die evolutionäre Abstammung der Wale von landlebenden Säugetieren liefert hierfür eine plausible Erklärung (s. Ridley 1993, 45). Auch sind die Produkte der Evolution nicht moralisch gut, sondern voller moralischer Grausamkeiten – was übrigens eine Neuauflage des klassischen Theodizeeproblems darstellt.

Mit der Kritik dieser empirisch konsequenzenreichen Kreationismus hat keine der Hauptströmungen der gegenwärtigen Wissenschafts- und Erkenntnistheorie kein Problem, da diese Kreationismen aufgrund ihres Konfliktes mit den Erfahrungstatsachen krisiert werden können.

(2.) Empirisch unkritisierbare Kreationismen: Schwieriger wird es mit den *rationalisierten* Formen des Kreationismus, welche von wissenschaftlich gebildeten Personen so entwickelt werden, dass sie jeglichen Konflikt mit etabliertem Erfahrungswissen vermeiden. Man könnte meinen, dies empirisch unkritisierbaren Kreationismen könnten kritisiert werden, weil sie keine empirischen Konsequenzen hätten und daher nicht empirisch überprüfbar seien – im logischen Sinn dieses Wortes, qua Vergleich mit Beobachtungstatsachen. Doch dies ist nicht der Fall – vielmehr ist es immer möglich, eine solche kreationistische Erklärung mit empirischen Konsequenzen auszustatten.

Beispielsweise ist folgende Minimalformulierung eines Kreationismus in der Tat gehaltleer und unüberprüfbar:

> *(K-Leer)* Wie immer unsere Welt faktisch beschaffend ist, hat sie einen Schöpfer (über den sonst nichts empirisch Gehaltvolles gesagt wird).

Sobald wir aber das Wirken des Kreator in (K-Leer) in Bezug auf bekannte Beschaffenheiten unserer Welt anreichern, erhalten wir daraus rationalisierte Versionen, die durchaus empirisch gehaltvoll ist, also empirische Tatsachen logisch implizieren – wie beispielsweise:

> *(K-Rat)* Unsere Welt hat einen Schöpfer, der bewirkt hat, dass in ihr folgende Tatsachen wahr sind: … (hier folgt eine korrekte Aufzählung aller bis dato bekannten empirischen Tatsachen, z.B. eine Aufzählung aller bekannten Lebewesen, usw.)

Im Gegensatz zu (K-Leer) hat (K-Rat zutreffende empirische Konsequenzen (ebenso Sober 1993, 45-49). Argumentationen, die auf der Linie von (K-Rat) liegen, trifft man in der gegenwärtigen anglosächsischen *intelligent design* Bewegung deren Vertreter behaupten, die Unwahrscheinlichkeiten Welt wären am besten durch einen Schöpfer erklärt, ohne jedoch spezielle Behauptungen über die Kreationsgeschichte oder über den Perfektionsgrad der Biosphäre zu implizieren (s. z.B. Behe 1996, Dembski 1998; zur Kritik s. Sober 2002, 72). Bedeutet dies tatsächlich, dass aus rationalisierte die Kreationismushypothese nun zur wissenschaftlichen Hypothese geworden ist, welche dieselben empirischen Fakten erklären kann wie die Wissenschaft? Intuitiv spüren wir, dass an (K-Rat) immer noch etwas faul ist – aber was könnte das sein?

Das Problem, auf das wir hier stoßen, ist das schon erwähnte *Abgrenzungsproblem*: aufgrund welcher Kriterien lassen sich wissenschaftliche Hypothesen von nichtwissenschaftlichen Spekulationen abgrenzen? In der gegenwärtigen Wissenschaftstheorie besteht Konsens darüber, dass frühere Vorschläge zum Abgrenzungsproblem zu einfach waren. Hier seien nur die zwei wichtigsten herausgegriffen.

Der erste Vorschlag ist das Kriterium der *empirischen Definierbarkeit*.[2] Ihm zufolge sind nur solche Hypothesen wissenschaftlicher Natur, deren nichtlogische Begriffe durch reine Beobachtungsbegriffe definierbar sind. Dieses von klassischen Empiristen und Positivisten vertretene Kriterium ist aber *zu eng*, denn wissenschaftliche Theorien enthalten sogenannte empirisch undefinierbare, sogenannte *theoretische Begriffe* (wie z.B. "magnetische Kraft", "Quantenzustand", etc.) welche *unbeobachtbare* Dinge oder Merkmale bezeichnen (Carnap 1956; Schurz 2006, Kap. 5).

Der zweite Vorschlag, der von Popper und im späteren Wiener Kreis vertreten wurde, verlangt von wissenschaftlichen Hypothesen lediglich, dass sie zumindest im Verbund mit anderen Hypothesen *empirische Konsequenzen* besitzen. Dieses Kriterium ist aber *zu weit*, da, wie wir sahen, auch rein spekulative Hypothesen in *trivialer* Weise zu empirisch gehaltvollen Hypothesen erweiterbar sind (s. Stegmüller 1970, Kap. V). Beispielsweise ist der Satz

---

2 Noch enger ist das Kriterium der empirischen *Verifizierbarkeit*, welches im frühen Wiener vertreten wurde.

"Gott existiert", oder

empirisch gehaltloser und daher unüberprüfbare Hypothesen. Doch wir müssen ihm nur als Konjunktionsglied eine Implikation auf beliebige empirische Tatsache hinzufügen, um daraus einen empirisch gehaltvollen Satz bilden, wie z.B.

"Gott existiert, und wenn Gott existiert, dann ist Gras grün".

Dasselbe Verfahren wurde offenbar auch bei (K-Rat) oben angewandt.

Aufgrund solcher Schwierigkeiten halten viele gegenwärtige Wissenschaftstheoretiker solche Abgrenzungsversuche zwischen wissenschaftlichen versus spekulativen Hypothesen für überholt und nicht zielführend. Eine prominente Strömung, welche diese Auffassung vertritt, ist der Bayesianismus. Ihm zufolge ist die Bestätigung von Hypothesen allein eine Sache ihrer Wahrscheinlichkeit aufgrund gegebener empirischer Evidenzen; darüber hinausgehende Anforderungen an wissenschaftliche Hypothesen sind weder nötig noch sinnvoll. Im folgenden Abschnitt möchte ich zeigen, wie es kommt, dass diese wissenschaftstheoretische Strömung von rationalisierten Kreationisten zur Begründung ihrer Position herangezogen wurde.

3. Bayesianische Rechtfertigung des rationalisierten Kreationismus

Der Bayesianismus betrachtet Wahrscheinlichkeiten als rationale Glaubensgrade. Die bedingte Wahrscheinlichkeit P(E|H) einer (empirischen) Evidenz E gegeben eine Hypothese H mennt man auch das *Likelihood*. Dieses Likehood kann zwar nicht immer aber zumindest oft objektiv bestimmt werden. Beispielsweise ist das Likelihood von E = 'Werfen von Kopf', gegeben H = 'Münzwurfexperiment mit regulärer Münze', aufgrund statistischer Gesetze 1/2. Und das Likelihood von E, gegeben eine Hypothese H die E logisch impliziert, ist aus logischen Gründen 1. Was man nun wissen will, ist natürlich P(H|E), die Wahrscheinlichkeit der Hypothese H gegeben E. Gemäß der berühmten bayesianischen Formel berechnet sich diese aus dem Likelihood und den sogenannten Ausgangswahrscheinlichkeiten wie folgt:

(Bayes1): $P(H|E) = P(E|H) \cdot P(H) / P(E).$

Dabei ist P(H) die Ausgangswahrscheinlichkeit von H, welche das Kernproblem des Bayesianismus ausmacht, denn Glaubensgrade 'vor aller Erfahrung' sind subjektiv und spiegeln meistens nur die eigenen Vorurteile wieder. P(E) ist die Ausgangswahrscheinlichkeit von E – sie wird üblicherweise berechnet als

$$P(E) = \Sigma_{1 \le i \le n} P(E|H_i) \cdot P(H_i),$$

mit $\{H_1,\ldots,H_n\}$ als der Partition von alternativen Hypothesen, welche die fragliche Hypothese H enthält. Auch darin liegt ein Problem, denn eine solche Partiti-

on enthält meist ein "Default-Element" der Form "keine der bisher genannten Hypothesen ist wahr", wobei das Likelihood von E in Bezug auf diese Default-Hypothese gänzlich unbekannt ist – doch dieses Problem sei hier außer Betracht gelassen.

Um dem Problem der Abhängigkeit vom subjektiven Wert der Ausgangswahrscheinlichkeit P(H) zu entgehen, wird im *komparativen* (bayesianischen) Bestätigungsbegriff lediglich die Wahrscheinlichkeitserhöhung von H durch E, P(H|E) > P(H), als Kriterium für die Bestätigung von H durch E angesehen (wobei E als konsistent und H als nichttautologisch angenommen werden).

> *Komparativer bayesianischer Bestätigungsbegriff:* Ein (konsistentes) E bestätigt ein (nichttautologisches) H wenn P(H|E) > P(H).

Unter Voraussetzung des *Normalfalles* $0 < P(E) < 1$ ist leicht beweisbar, dass P(H|E) > P(H) genau dann gilt, wenn P(E|H) > P(E) gilt, bzw. die sogenannte Likelihood-Ratio P(E|H)/P(E) positiv ist. Letzteres gilt jedoch immer, sofern H irgendeine Hypothese ist, die E logisch impliziert. Es gilt also folgender Sachverhalt:

> *(Bayes2):* Jede (nichttautologische) Hypothese H, welche eine empirische Evidenz E mit 0 < P(E) < 1 logisch impliziert, wird durch E im komparativen Sinn bestätigt.

Und das ist die Konsequenz, welche die Vertreter aller Arten von rationalisierter Spekulationen ausschlachten können. Denn gemäß (Bayes2) können offenbar gänzlich abstruse Hypothesen bestätigt werden, sofern sie E nur logisch implizieren (s. auch Schurz 2008a, §7.1). Z.B. bestätigt die Tatsache, dass Gras grün ist, die Hypothese, dass Gott existiert und veranlasst hat, dass Gras grün ist. Dieselbe Tatsache bestätigt aber auch die Hypothese, dass ein Spaghetti-Monster existiert, welches veranlasst hat, dass Gras grün ist – die Spaghetti-Monster-Bewegung ist eine von Physikern initiierte Gegenbewegung zum Kreationismus, welche die Forderung, dass kreationistische Lehren in der Schule unterrichtet werden sollen, ad absurdum treiben will (s. www.venganza.org/aboutr/open-letter). Und dieselbe Tatsache bestätigt auch die Hypothese, dass zwei Spagetti-Monster gemeinsam dies veranlasst haben, oder ein Gott und ein Spagetti-Monster, ein Gott und ein Teufel, oder … usw. , bis hin zur wissenschaftlichen Erklärung der grünen Farbe von Gras. Alle diese Erklärungshypothesen H_i werden gleichermaßen komparativ bestätigt. Wenn sie einen unterschiedlichen konditionalen bayesianischen Glaubensgrad $P(H_i|E)$ besitzen, dann kann dieser gemäß (Bayes1) nur an ihrer unterschiedlichen Ausgangswahrscheinlichkeit $P(H_i)$ liegen, denn $P(E|H_i)$ ist bei allen Hypothesen 1, und auch P(E) ist ein hypothesenunabhängiger Wert.

Bayesianische Wissenschaftstheoretiker sind sich dieser Tatsache bewusst (vgl. Howson/Urbach 1996, 141f). Sie argumentieren, dass wissenschaftliche

Hypothesen eben eine wesentlich höhere Ausgangswahrscheinlichkeit besitzen als religiöse Hypothesen (vgl. Sober 1993, 31f). Aber es erscheint unangemessen, den Unterschied zwischen wissenschaftlichen und spekulativen Hypothesen auf subjektive Vormeinungen zu stützen. Aus religiösen Sicht wird umgekehrt die Kreationismushypothese die höhere Ausgangswahrscheinlichkeit besitzen. Und aus diesem Grund kann die Bayesianische Bestätigungstheorie von Vertretern des Kreationismus benutzt werden, um damit die Bestätigtheit der Schöpferhypothese aufzuzeigen, auf die oben beschriebene Weise. Ein frühes Beispiel hierfür ist Swinburne (1979, ch. 6). In jüngerer Zeit hat Unwin (2005) aufgehend von einer 1.1-Ausgangswahrscheinlichkeit die bedingte Wahrscheinlichkeit, dass Gott existiert, mithilfe der Bayes-Formel auf 67% berechnet.[3]

Die Schwierigkeit, innerhalb dieses Rahmens ein Abgrenzungsargument zu finden, zeigt sich auch in Dawkins sehr agitatorisch geratenem Buch "Der Gotteswahn" (2007). Eines der Dawkinschen Hauptargumente gegen die Existenz Gottes ist nämlich das

Unwahrscheinlichkeitsargument: die Hypothese eines Schöpfergottes, der all dieses Unwahrscheinliche zustande gebracht hat, sei prima facie ihrerseits extrem unwahrscheinlich, wogegen die Annahmen der Evolutionstheorie prima facie wesentlich wahrscheinlicher seien.

Doch Dawkins Argument ist sehr fragwürdig. Da auch die Prämissen der evolutionären Erklärung annehmen müssen, dass unsere Welt die unwahrscheinlichen Parametersetzungen besitzt, auf welche das anthropische Argument rekurriert, ist es intuitiv nicht klar, dass deren Ausgangswahrscheinlichkeit größer ist die der kreationistischen Erklärungsprämissen – ganz abgesehen vom grundsätzlicheren Problem, dass Ausgangswahrscheinlichkeiten immer subjektiv ist, und nicht zu sehen ist, wie auf diesem Wege eine objektiv begründete Abgrenzung möglich sein soll.

Die Tatsache, dass mit der Bayes-Formel völlig abstruse Formeln quasi-bestätigt werden können, scheint eher darauf hinzuweisen, dass die Bayesianische Bestätigungstheorie zu *schwach* ist, um genuine Bestätigung zu erfassen, und genuin bestätigte Hypothesen von bloßer Spekulation abzugrenzen. Nicht dass die Bayes-Formeln falsch wären, im Gegenteil sind sie mathematisch korrekt, und die bayesianische Bedingung einer positiven Likelihood-Ratio ist durchaus eine notwendige Bedingung für Bestätigung – aber keinesfalls eine hinreichende Bedingung. Das Abgrenzungsproblem kann in diesem Rahmen nicht gelöst werden. Im nächsten Abschnitt entwickle ich einen Alternativvorschlag.

3 Woraufhin der Herausgeber der Zeitschrift *Sekptic*, Michael Shermer, eine Gegenrechnung aufstellte und zum Ergebnis von 2% kam.

4. Neue Voraussagen als Abgrenzungskriterium

Was intuitiv am rationalisierten Kreationismus defekt ist, ist offenbar dieses: wie auch immer der empirische Faktenstand aussieht, kann eine solche Erklärung gegeben werden. Die kreationistische Erklärung ist völlig *ex-post*, also im nachhinein zurecht konstruiert, und ein Abgrenzungskriterium sollte diesen Defekt des rationalisierten Kreationismus (K-Rat) ins Zentrum rücken. Der ex-post Charakter einer Hypothese äußert im Fehlen ihrer Fähigkeit, neue Voraussagen zu machen. In der Tat kann der rationalisierter Kreationismus nichts voraussagen, weil kreationistische Erklärungshypothese nichts über die Natur des Schöpfers aussagt, was darüber hinausgeht, dass er die zu erklärenden Fakten bewirkte. Die kreationistische Erklärungshypothese "Gott bewirkte, dass E" lässt sich daher immer nur im nachhinein postulieren, wenn E schon bekannt ist.

Ich nenne dieses Abgrenzungskriterium das *Voraussagekriterium*. Die Grundidee des Voraussagekriteriums ist auch von vielen anderen Wissenschaftlern und Wissenschaftstheoretikern vorgeschlagen worden (z.B. Lakatos 1977; Ladyman and Ross 2007, §2.1.3). Am Voraussagekriterium wurde kritisiert, dass es zu eng sei, weil eine Reihe von Disziplinen, einschließlich der Evolutionstheorie, nur wenig Voraussagen machen. Doch hier liegt ein Missverständnis vor. Denn man versteht den Begriff der Voraussage in diesem Kriterium nicht im zeitlichen, sondern im epistemischen Sinn eines *ex-ante* Argumentes (s. auch Stegmüller 1983, 976). Bei einer Voraussage qua ex-ante Argument wird nicht verlangt, dass sich die Konklusion auf die Zukunft bezieht, sondern lediglich, dass die Prämissen schon vor der Konklusion bekannt waren und die Konklusion erst danach daraus erschlossen wurde. Im Gegensatz dazu ist bei einem ex-post Argument die Konklusion zuerst bekannt, und die Prämissen werden erst nachträglich gefunden bzw. postuliert. Dies eröffnet die Möglichkeit, dass geeignete Prämissen auf die gegebene Konklusion zurechtgeschneidert werden. Ein solches Zurechtscheidern oder Fitten auf die Konklusion ist im Falle eines ex-ante Argumentes, also einer epistemischen Voraussage, dagegen unmöglich.

Je nachdem, ob sich die Konklusion auf die Zukunft oder Vergangenheit bezieht, liegt bei einer epistemischen Voraussage eine zeitliche Voraussage oder aber eine zeitliche Retrodiktion vor. Die Evolutionstheorie macht zwar wenig überprüfbare zeitliche Voraussagen, aber jede Menge Retrodiktionen, die durch gegenwärtige Spuren (geologische Spuren, Fossilien, archäologische Funde etc.) unabhängig empirisch testbar sind (m.a.W., zeitliche Retrodiktionen über vergangene Ereignisse implizieren zeitliche Voraussagen über gegenwärtig zu findende Spuren.) Genau diese *unabhängige Testbarkeit* wird durch die Erfüllung des Voraussagekriterium gewährleistet – die Testbarkeit unabhängig von den Tatsachen, um derer Erklärung willen die Hypothese konstruiert wurde.

In einer Hinsicht muss das Voraussagekriterium jedoch noch verbessert werden. Eine kreationistische ad-hoc Erklärung könnte einfach dadurch neue

Fakten voraussagen, indem sie induktive Zusammenhänge ausnutzt. Z.B. könnte ein Kreationist erklären:

> "Die Sonne geht jeden Tag, auch morgen, auf, weil Gott es so will",

und darauf hinweisen, dass er damit neue Fakten voraussagt, nämlich dass morgen und auch übermorgen (etc.) die Sonne aufgehen wird.

Das ist eine weitere Komplikation unseres Problems. "Gott" ist ja, wie z.B. "Magnetfeld", ein *unbeobachtbarer* bzw. 'theoretischer' Begriff, oder in der Terminologie der Statistik, eine *latente* (nicht-manifeste) Variable. Wir gelangen hier zur altbekannten 'Ockhamschen Problem', wann es denn in der Wissenschaft angemessen ist, zu empirischen Erklärungszwecken unbeobachtbare Entitäten zu postulieren. Ich gebe darauf zwei Antworten, eine einfache und eine kompliziertere. Die einfache Antwort besagt (im Sinne des Ockhamschen 'Rasiermesser'-Kriteriums), dies ist dann angemessen, wenn eine gleichermaßen gute Erklärung nicht auch *ohne* die Annahme von unbeobachtbaren Entitäten hätte gefunden werden können, und zwar mithilfe einer durch einfache induktive Generalisierung gewonnenen Gesetzmäßigkeit zwischen den beobachteten Variablen. Dies ist in obigem Beispiel der Fall, denn das morgige Aufgehen der Sonne ist schon durch die empirische Gesetzeshypothese voraussagbar, derzufolge die Sonne täglich über den Horizont wandert. Damit mit gelangen wir zu unserem ersten Abgrenzungskriterium:

> *(Voraussagekriterium No. 1):* Eine Erklärungshypothese, die unbeobachtbare Entitäten einführt bzw. postuliert, ist nur dann wissenschaftlich legitim, wenn sie potentiell neue Voraussagen impliziert – das sind empirische Konsequenzen, die über das hinausgehen, was aus den von ihr erklärten und schon zuvor (d.h. vor ihrer Konstruktion) bekannten Fakten durch einfache induktive Verallgemeinerung erschließbar ist.

Das Voraussagekriterium No. 1 fordert nur die Implikation von *potentiell* neuen Voraussagen – sein Zutreffen hängt daher nur vom Gehalt der Erklärungshypothese und der empirischer Evidenz zum Zeitpunkt ihrer Konstruktion ab, nicht aber vom pragmatischen Umstand des Noch-Unbekanntseins der neuen Voraussagen. Durch das Bekanntwerden der bislang unbekannten neuen Voraussage hört eine Erklärungshypothese also nicht auf, wissenschaftlich legitim zu sein.[4]

Unter einer induktiven Verallgemeinerung verstehe ich eine Generalisierung von bisher beobachteten Regelmäßigkeiten in die offene Zukunft. Solche Regelmäßigkeiten können freilich sehr komplex sein, weshalb ich mich auf "einfache" induktive Verallgemeinerungen beschränke. Der der Begriff der "einfachen" induktiven Verallgemeinerung ist freilich graduell, und somit ist auch un-

4 Aus diesem Grund ist das Voraussagekriterium auch nicht vom sogenannten Problem der "old evidence" betroffen, welches im Bayesianismus viel diskutiert wurde (s. Howson und Urbach 1992, 403ff).

ser erstes Voraussagekriterium graduell. Dennoch reicht es für alle praktischen Abgrenzungszwecke aus, denn eine Hypothese, die Voraussagen macht, welche weder zuvor bekannt waren noch durch simple Induktion aus bekannten Beobachtungen gewonnen werden können, sollte ernst genommen werden, auch wenn sie neuartige spekulative Annahmen bzw. unbeobachtbare Entitäten postuliert. Selbst religiöse Hypothesen, würden sie solche Voraussagen machen, sollten vor ernstzunehmenden Überprüfungsversuchen nicht ausgeschlossen werden. Allerdings kenne ich keine religiösen Glaubenssysteme, die solche Voraussagen machen und bislang noch nicht falsifiziert worden sind.

Man kann noch stringenter sagen, wann und warum es wissenschaftlich angemessen ist, zur Erklärung empirischer Phänomene latente Variablen zu benutzen, sofern man schwache kausalitätstheoretische Annahmen macht – nämlich die sogenannten *Reichenbach-Bedingungen* der Kausalität. Diesen Annahmen zufolge ist jede Korrelation zwischen zwei oder mehreren (empirischen bzw. manifesten) Variablen entweder auf gerichtete Kausalbeziehungen zwischen den Variablen zurückzuführen, oder aber darauf, dass diese Variablen eine versteckte Variable als gemeinsame Ursache besitzen. Wann immer also die Korrelationen zwischen einer Menge von empirischen Variablen nicht durch die Annahme von Ursache-Wirkungs-Beziehungen zwischen diesen Variablen erklärbar ist, ist die Annahme von latenten Variablen als gemeinsame Ursachen berechtigt. Dies ist nun insbesondere dann der Fall, wenn es sich bei den empirischen Variablen um *Dispositionen* handelt. Dispositionen von Objekten bestehen darin, dass sie unter gewissen Umständen spezifische Wirkungen hervorbringen. Dispositionen in diesem Verständnis sind also funktionale Merkmale höherer Stufe, die Regelmäßigkeiten ausdrücken und daher selbst nicht als Ursache für andere Dispositionen in Frage kommen (s. auch Prior et al. 1982).[5] In Schurz (2008a, §7.2, 2008b) habe ich zu zeigen versucht, dass eine Vielzahl von theoretischen Begriffen der Wissenschaft (und vielleicht sogar alle) als gemeinsame Ursachen von korrelierten Dispositionen eingeführt wurden. Mein zweiter Explikationsvorschlag lautet also:

> *(Voraussagekriterium No. 2):* Eine Erklärungshypothese, die unbeobachtbare Entitäten einführt bzw. postuliert, ist wissenschaftlich legitim, wenn sie als gemeinsame Ursache von korrelierten Dispositionen fungiert.

Wie in Schurz (2009) zu zeigen versucht wird, impliziert eine Erklärungshypothese, die als eine solche gemeinsame Ursache fungiert, immer potentiell neue Voraussagen, die über simple Induktion hinausgehen, denn mithilfe gemeinsamer theoretischer Ursachen kann man vom Verhalten von Objekten in einem Anwendungsbereich auf deren Verhalten in einem ganz anderen Anwendungs-

5 Es gibt auch andere Konzeptionen von "Disposition" in denen Dispositionen kausal wirksam sein können – dabei werden Dispositonen nicht mit Regelmäßigkeiten identifiziert (vgl. Mumford 1998).

bereich schließen – beispielsweise vom Gewicht von Körpern auf der Erde auf ihre Verhalten in einer Umlaufbahn, usw.

Das Voraussagekriterium 2 verwendet den Ursachebegriff im probabilistischen Sinn, und schließt somit nicht aus, dass es neben der die fraglichen gemeinsamen Ursache auch noch andere gibt, bzw. dass die fragliche gemeinsame Ursache Bestandteil eines ganzen Netzes von latenten und kausal verbundenen Variablen ist. Davon abgesehen habe ich vorsichtshalber das Voraussagekriterium 2 nur als *hinreichende*, aber nicht als notwendige Bedingung eingeführt, da es auch noch andere Fälle geben mag, in denen die Einführung theoretischer Begriffe wissenschaftlich legitim ist, obwohl mir bislang kein solcher Fall bekannt ist. Dagegen habe ich das Voraussagekriterium 1 vorsichtshalber nur als *notwendige* Bedingung formuliert, die möglicherweise um weitere Bedingungen zu verstärken ist.

5. Lösung des Abgrenzungsproblems innerhalb des Bayesianismus?

Da der Bayesianismus ein methodologisches Rahmenwerk bildet, das auch viele Vorzüge besitzt, wäre es wünschenswert, wenn sich die Inadäquatheit von ex-post Erklärungen auch *innerhalb* des bayesianischen Ansatzes aufzeige ließe. Hierfür würde sich prima facie folgender Weg anbieten. Der *ex-post* Charakter einer Hypothese geht ja wie erläutert Hand in Hand damit, dass man auf diese Weise *beliebige* alternative Erklärungshypothesen konstruieren könnte, von Göttern bis zu Spaghetti-Monstern usw. Dies bedeutet, dass es, wenn man spekulative Hypothesen zulässt, immer *unendlich viele gleichermaßen in Frage kommende* Alternativhypothesen gibt. Das impliziert aber, dass man sämtlichen solchen Hypothesen die Ausgangswahrscheinlichkeit *null* geben muss. Und daraus folgt wiederum gemäß (Bayes1), dass völlig unabhängig vom Likelihood P(E|H) der resultierende Wert von P(H|E) ebenfalls null beträgt, sofern nur P(E) positiv ist:

$$P(H|E) = P(E|H)\cdot P(H)/P(E) = 0/P(E) = 0.$$

Selbst wenn man argumentiert, dass unser Explanandum E, die lebensermöglichenden Eigenschaften unserer Welt, so extrem unwahrscheinlich seien, dass wir P(E) ebenfalls als null ansetzen müssen, erhalten wir keinen positiven Wert von P(H|E), sondern lediglich den Wert 0/0 = unbestimmt.

In anderen Worten besagt diese Argumentation, dass die Anwendung des Bayesianismus auf spekulative Hypothesen aufgrund deren unendlicher Vielfalt ins Leere geht. Nur wenn man Hypothesen betrachtet, für die man Gründe besitzt, eine Ausgangswahrscheinlichkeit größer Null anzunehmen, kann man mithilfe des bayesianischen Kriteriums überhaupt zu sinnvollen Bestätigungsaussa-

gen gelangen. Und solche Gründe besitzt man nur dann, wenn die Hypothese keine reine ex-post Hypothese ist, sondern das im obigen Abschnitt erläuterte Voraussagekriterium erfüllt.

Das Problem dieser Argumentation liegt darin, dass man nur schwer begründen kann, warum es zu einem gegebenes Explanandum nicht eventuell auch unendlich viele alternative wissenschaftliche Erklärungshypothesen geben könnte, die allesamt über simple Induktion hinausgehende neue Voraussagen generieren und somit das Voraussagekriterium 1 erfüllen. Bedenkt man, dass die postulierten gemeinsamen Ursachen in Voraussagekriterium 2 Bestandteile eines Netzes von latenten Ursachen sein können, so könnten diese unendlich vielen Erklärungshypothesen eventuell sogar allesamt auch das Voraussagekriterium 2 erfüllen. Nachdem alle diese unendlich vielen wissenschaftlich legitimen Erklärungshypothesen prima facie gleichberechtigt sind, müssten man auch ihnen allen die Ausgangswahrscheinlichkeit null zuweisen, und wir landen beim selben Problem wie oben bei den unendlich vielen spekulativen Hypothesen. Wir haben damit also eher ein weiteres Problem des Bayesianismus aufgespürt, als eine bayesianische Lösung des Abgrenzungsproblems gegeben.

Die Überlegung, dass eine Erklärungshypothese es überhaupt nur dann wert ist, einer Überprüfung unterzogen zu werden, wenn sie das Voraussagekriterium erfüllt, weil sie nur *bestätigungsfähig* ist, bleibt nach wie vor richtig. Es ist aber fraglich, ob diese Überlegung auf die Betrachtung von Ausgangswahrscheinlichkeitsverteilungen zurückgeführt werden kann. Diese abschließende Überlegung stützt somit die Diagnose, dass der Bayesianismus nur ein notwendiges, aber kein hinreichendes Kriterium für die Bestätigung von Hypothesen darstellt.

Literaturverzeichnis

Barrow, J.D.: Tipler, F.: The Anthropic Cosmological Principle. Oxford University Press, Oxford, 1988

Behe, M.: Darwin's Black Box, Free Press, New York, 1996

Carnap, R.: "The Methodological Character of Theoretical Concepts". In: *Feigl, H./ Scriven, M. (Hrsg.):* Minnesota Studies in the Philosophy of Science, Vol. I. Univ. of Minnesota Press, Minneapolis, 1956. S. 38-76

Davies, P.: Der Plan Gottes. The mind of God. 1992. Insel, Frankfurt/M., 1995

Dawkins, R.: Der Gotteswahn. Ullstein, Berlin, 2007

Dembski, W.: The Design Inference. Cambridge Univ. Press, Cambridge, 1998

Howson, C./ Urbach, P.: Scientific Reasoning: The Bayesian Approach. Open Court, Chicago, 2. Aufl., 1996

Ladyman, J./Ross, D.: Every Thing Must Go. Metaphysics Naturalized. Oxford University Press, Oxford, 2007. (with D. Spurrett and J. Collier.)

Lakatos, I.: "Science and Pseudoscience". In: *Lakatos, I.:* Philosophical Papers, Vol. 1. Cambridge Univ. Press, Cambridge, 1977. S. 1-7

Mumford, S.: Dispositions. Oxford Univ. Press, Oxford, 1998

Prior, E.W./Pargetter, R./Jackson, F.: "Three Theses about Dispositions". American Philosophical Quarterly, 19, 1982. S. 251-7

Ridley, M.: Evolution. Blackwell Scientific Publications, Oxford, 1993

Schurz, G.: Einführung in die Wissenschaftstheorie. WBG, Darmstadt, 2. Aufl., 2008, 2006

Schurz, G.: "Patterns of Abduktion". Synthese, 164, 2008a. S. 201-234

Schurz, G.: "Common Cause Abduction and the Formation of Theoretical Concepts". In: *Dégremont, C./Keiff, L./ Rückert H.* (Hrsg.): Dialogues, Logics, and Other Strange Things. Essays in Honour of Shahid Rahman. College Publications, London, 2008b. S. 337-364

Schurz, G.: "When Empirical Success Implies Theoretical Reference: a Structural Correspondence Theorem". British Journal for the Philosophy of Science, 60, 2009. S. 101-133

Smolin, L.: The Life of the Cosmos. Oxord Univ. Press, New York, 1997

Sober, E.: Philosophy of Biology. Boulder, Westview Press, 1993

Sober, E.: "Intelligent Design and Probability Reasoning". International Journal for Philosophy of Religion, 52, 2002. S. 65-80

Stegmüller, W.: Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band I: Erklärung - Begründung - Kausalität. Springer, Berlin, zweite verbesserte und erweiterte Auflage, 1983

Swinburne, R.: The Existence of God. Clarendon Press, Oxford, revised 2nd ed. 2004, 1979

Unwin, S.T.: Die Wahrscheinlichkeit der Existenz Gottes. Mit einer einfachen Formel auf der Spur der letzten Wahrheit. discorsi, Hamburg, 2005

Ward, P./Brownlee, D.: Rare Earth. Springer, New York, 2000 (dt. bei Spektrum)

Grundzüge einer Zustimmungslogik

Werner Stelzner
w.stelzner@yahoo.de
Institut für Philosophie der Universität Bremen

Abstract/Zusammenfassung

In the explication of the logical properties of the concept of assent and the corresponding intuitive semantics for this term the following three principles are realized:

1. It is expected that also inconsistent assents can be performed and that the concept of assent in this case is not trivial (paraconsistency of assent).
2. It is assumed that assent is performed under the condition of limited logical resources (or limited logical skills) of epistemic subjects. By agreeing to certain sentences the epistemic subject does not assent to arbitrary logical consequences of these sentences. It also does not unconditionally assent arbitrary logically valid propositions. Therefore, no logical omniscience of the assenting subject is assumed.
3. Despite these limitations logically relevant relationships between assents can be handled. This is achieved by a typification of the epistemic subjects according to their abilities to draw logical conclusions. The corresponding conclusions are completed stepwise according to simple logical rules.

In the paper three aspects of restrictions of logical resources of epistemic subjects are expressed:

1. The relationship of deducibility, defined for subjects of type P^1, is closed under the system of tautological entailments, but not complete relative to this system and therefore such subjects are resource limited concerning this system.
2. Another type of resource limitation is connected with the fact that an epistemic subject of type P^1 or type P^2 can carry out only a certain number of derivation steps.
3. Immediate derivation rules (which can be completed in one step) for epistemic subjects of type P^1 are resource limited in the sense, that in these rules occur no more than two monadic and one two-place connectives.

It is shown, that the system of belief logic introduced by Lévesque (2000) as a resource limited system, in none of the three mentioned aspects happens to be resource limited.

Bei der Bestimmung der logischen Eigenschaften des Zustimmungsbegriffs und der damit verbundenen Entwicklung intuitiver Semantiken für diesen Begriff werden drei Prinzipien realisiert:

1. Es wird davon ausgegangen, dass auch widersprüchliche Zustimmungen vollzogen werden können und dass der Zustimmungsbegriff auch in diesem Falle nicht trivial wird (Parakonsistenz der Zustimmung).
2. Es wird angenommen, dass die Zustimmung unter der Bedingung beschränkter logischer Ressourcen (bzw. beschränkter logischer Fähigkeiten) der epistemischen Subjekte erfolgt. Mit der Zustimmung zu bestimmten Sätzen ist nicht die Zustimmung zu jeder logischen Folge dieser Sätze verbunden. Auch wird logisch gültigen Sätzen nicht bedingungslos zugestimmt. Es wird also keine logische Allwissenheit des Zustimmenden vorausgesetzt.
3. Trotz dieser Beschränkungen sollen logisch relevante Beziehungen zwischen Zustimmungen behandelt werden können. Das wird dadurch erreicht, dass eine Typisierung der epistemischen Subjekte nach ihren Fähigkeiten, logische Schlüsse zu ziehen, vorgenommen wird. Die entsprechenden Schlüsse werden dabei schrittweise entsprechend einfachster logischer Regeln vollzogen.

Dabei werden drei Aspekte der Beschränkung der logischen Ressourcen von epistemischen Subjekten ausgedrückt:

1. Die für Subjekte der Art P^1 definierten Ableitbarkeitsbeziehungen sind unter den tautological entailments geschlossen, aber nicht vollständig bezüglich des Systems der tautological entailments und in diesem Sinne ressourcenbeschränkt.
2. Eine andere Art der Ressourcenbeschränkung ist damit verbunden, dass ein epistemisches Subjekt der Art P^1 oder der Art P^2 nur eine bestimmte Anzahl von Ableitungsschritten in Frage-Antwortsequenzen durchlaufen kann.
3. sind unmittelbare Ableitungsregeln (für solche Ableitungen, die in einem Schritt vollzogen werden können) für epistemische Subjekte der Art P^1 in dem Sinne ressourcenbeschränkt, dass in diesen Regeln höchstens zwei einstellige und eine zweistellige Verbindung vorkommen können.

Es wird gezeigt, dass das von Levesque (2000) als ressourcenbeschränktes System der Glaubenslogik vorgestellte System in keinem der drei genannten Aspekte ressourcenbeschränkt ist.

1. Zum Zustimmungsbegriff

Wenn wir uns der logischen Analyse der Zustimmung zuwenden, ist der Unterschied zwischen innerer Zustimmung und äußerer Zustimmung zu beachten. Während bei der inneren Zustimmung das epistemische Subjekt einem Satz innerlich seine Zustimmung gibt, was diesen Begriff in die Nähe des Glaubens rückt, wird bei der äußeren Zustimmung öffentlich und explizit einem Satz zugestimmt, wie dies z. B. in Behauptungen als spezifischen Zustimmungshandlungen geschieht.[1]

In einer Serie von Arbeiten hat R. Marcus gegen die sprachzentrierte Auffassung des Glaubens als spezifischer Form der Zustimmung argumentiert, nach der Glaube eine Relation zu Sätzen oder linguistischen Entitäten ist.[2]. Engel erörtert dagegen eine Zustimmungstheorie des Glaubens, nach der zu glauben, dass G gilt, bedeutet, zur Zustimmung zu einer internen Repräsentation der Bedeutung, dass G, disponiert zu sein. Dabei ist Zustimmung immer Zustimmung zu einem Satz oder zu einer satzähnlichen Repräsentation.[3]

Wir wollen unserer Analyse auch für die innere Zustimmung einen Begriff der Zustimmung zugrunde legen, der zumindest für das jeweilige epistemische Subjekt, dessen Zustimmungen infrage stehen, effektiv ist. Dazu bietet sich der Begriff der expliziten inneren Zustimmung (der abgeschwächt „aktivistische" Begriff der inneren Zustimmung) an, dessen Einführung unmittelbar sichert,

1 Zum Verhältnis zwischen linguistischer und mentaler Zustimmung vgl. Shoemaker (1996, 78), Moltmann 2003 und Proust 2001.

2 Vgl. Marcus 1981, 1985, 1986, 1993, 1995.

3 Vgl. Engel 1999, 212. Mitunter wird in diesem Fall auch von einer „judgmental" Konzeption des Glaubens gesprochen (vgl. Pettit 1998). Zur Akzeptation vgl. auch Stalnaker 1984, Lehrer 1990, Cohen 1992, Bratman 1992. Wir betrachten diese judgmental Konzeption als unmittelbar für die Zustimmung zutreffend, die einen linguistischen Akt darstellt.

dass das epistemische Subjekt unproblematisch entscheiden kann, ob es einem bestimmten Satz explizit innerlich zustimmt oder nicht.

2. Eine intuitive Semantik für die Zustimmungslogik

Bei der Bestimmung der logischen Eigenschaften dieses Zustimmungsbegriffs und der damit verbundenen Entwicklung intuitiver Semantiken für diesen Begriff sollen drei Prinzipien realisiert werden:

1. Es ist davon auszugehen, dass auch widersprüchliche Zustimmungen vollzogen werden können und dass der Zustimmungsbegriff auch in diesem Falle nicht trivial wird, d. h. nicht Beliebigem zugestimmt wird (Parakonsistenz der Zustimmung).
2. Es wird angenommen, dass die Zustimmung unter der Bedingung beschränkter logischer Ressourcen (bzw. beschränkter logischer Fähigkeiten) der epistemischen Subjekte erfolgt. Zwar wird angenommen, dass die Regeln für die logische Ableitung von Zustimmungen aus gegebenen expliziten Zustimmungen unter der klassischen Folgebeziehung geschlossen sind, es kann aber nicht angenommen werden, dass diese Regeln bezüglich der klassischen Folgebeziehung und davon abweichender nichtklassischer Folgebeziehungen, wie z. B. den tautological entailments, vollständig sind: Mit der Zustimmung zu bestimmten Sätzen ist nicht die Zustimmung zu jeder logischen Folge dieser Sätze verbunden. Auch wird logisch gültigen Sätzen nicht bedingungslos zugestimmt. Es wird also keine logische Allwissenheit des Zustimmenden vorausgesetzt. Das epistemische Subjekt bildet seine Zustimmungen unter der Bedingung der Ressourcenbeschränktheit seiner logischen Kapazitäten.
3. Trotz dieser Beschränkungen sollen logisch relevante Beziehungen zwischen Zustimmungen behandelt werden können.

Diesen Zielstellungen entsprechend werden die folgenden semantischen Festlegungen gewählt:

1. Wir führen zwei Arten von Welten ein:

1.1 Eine nichtleere Menge W_r von Reality-Welten, die klassisch bestimmt sind.

1.2 Eine nichtleere Menge W_e von epistemischen Welten (Set-up-Welten), mit denen Zustimmungszuordnungen bestimmt werden.

2. Neben der zweistelligen Erreichbarkeitsrelation R, die zwischen Reality-Welten definiert ist, wird eine dreistellige Relation wiRxwk eingeführt, eine epistemische Erreichbarkeitsrelation, die neben Reality-Welten auch für epistemische Welten definiert ist. Mit dem Index i in R_i wird auf das jeweilige epistemische Subjekt verwiesen, dessen Zustimmungen untersucht werden sollen. $w_iR_xw_k$ wird also gelesen: „Die Welt w_2 ist für das epistemische Subjekt x von der Welt w_1 aus erreichbar".
3. Die Interpretationsfunktion *v* stellt sich für Welten aus W_r und aus W_e unterschiedlich dar:

3.1 Für Für Reality-Welten (w, $w_1 \in W_r$) gelten die üblichen klassischen Interpretationsregeln:

$\nu(G, w) \in \{t, f\}$
$\nu(\sim G, w) = t$ gdw. $\nu(G) = f$
$\nu(G \wedge H, w) = t$ gdw. $\nu(G, w) = t$ und $\nu(H, w) = t$
$\nu(G \vee H, w) = f$ gdw. $\nu(G, w) = f$ und $\nu(H, w) = f$
$\nu(G \supset H, w) = f$ gdw. $\nu(g, w) = t$ und $\nu(H, w) = f$
$\nu(\Box G, w) = t$ gdw. Für alle w_1: Wenn wRw_1, so $\nu(G, w_1) = t$

Wird angenommen, dass R eine Äquivalenzrelation ist, so erhalten wir für die Reality-Welten eine adäquate Semantik für S5.

3.2 Für Set-up-Welten ($w \in W_e$) wird nur die folgende Interpretationsregel für beliebige Ausdrücke G angenommen.

$\nu(G, w) \in \{t, f\}$

Die Wertezuordnungen in epistemischen Welten werden für G und ~G unabhängig voneinander vorgenommen.

3.3 Die Interpretation des Zustimmungsbegriffes erfolgt sowohl für Reality-Welten als auch für Set-up-Welten folgendermaßen:

$\nu(A(i, G), w) = t$ gdw. Für alle Welten w_1 gilt:
Wenn $w_1 \in W_e$ und wR_iw_1, so $\nu(G, w_1) = t$

Hier ist also durchaus möglich, dass erfüllt ist $A(x, p) \wedge A(x, \sim p)$
Und es gilt logisch weder $A(x, p) \vee A(x, \sim p)$ noch $A(x, p) \supset \sim A(x, \sim p)$.

Damit ist Punkt 1 unserer Ausgangsforderungen an die aufzubauende Semantik erfüllt.

Es gelten aber auch keine anderen Beziehungen, die sich auf die innere Logik von Zustimmungen beziehen. Damit ist auch Punkt 2 dieser Forderungen erfüllt, wenn auch in extremer Weise:
Für die Zustimmung gilt keine logische Allwissenheit, d. h. weder die Gödelregel

Wenn $\sim G$, so $\sim A(x, G)$,

noch die Vollständigkeit der Folgebeziehungen zwischen Zustimmungen bezüglich der logischen Folgebeziehung

Wenn $G \sim H$, so $A(x, G) \sim A(x, H)$

gelten.

Die innere Logik der Zustimmung ist leer, es gelten keinerlei Prinzipien der Art $A(x, G) \sim A(x, H)$, wobei G und H unterschiedliche Ausdrücke sind. Es gelten aber auch keine Ausschlussprinzipien für Zustimmungen der Art $A(x, G) \sim\sim A(x, H)$, insbesondere gilt nicht (wie bereits oben angemerkt) $A(x, G) \sim\sim A(x, \sim G)$. Die Beziehungen zwischen unterschiedlichen Zustimmungen sind also weder positiv noch negativ logisch reglementiert.

Logische Reglementierungen der Zustimmung setzen spezifische Typen von Zustimmenden voraus, deren Existenz nicht logisch postuliert werden kann, sondern deren Annahme, auch wenn sie sich auf die Gültigkeit einer inneren Logik von Zustimmungen bezieht, empirisch bestimmt ist:

Das kann durch die Annahme spezieller Eigenschaften der epistemischen Erreichbarkeitsrelationen[4] und zum anderen durch die Angabe von Interpretationsregeln für die innere Verwendung von logischen Junktoren in Zustimmungskontexten geschehen. Solche zusätzlichen Interpretationsregeln sind allerdings für epistemische Subjekte keine logischen Regeln, obwohl sie sich auf logische Beziehungen richten, sondern sie gelten nur bedingt unter der Voraussetzung bestimmter Sprechertypen, die über bestimmte elementare oder weiter ausgebildete logische Fähigkeiten und Ressourcen verfügen. Unter expliziter Einbeziehung solcher Sprechertypen wird dann auch Forderung 3 erfüllt und es werden logische Beziehungen zwischen Zustimmungen unter der Voraussetzung bestimmter Sprechertypen ausgedrückt, die über entsprechende logische Kapazitäten verfügen.

3. Ressourcenbeschränkte logische Sprechertypen

Bisher wurden keinerlei Regeln für die logische Normierung innerer Zustimmungsakte angegeben. Die so charakterisierten Sprechertypen können als logisch unbestimmt bezeichnet werden. Für diese Sprechertypen werden keine empirischen Annahmen über ihre logischen Fähigkeiten und Ressourcen gemacht. Damit ist eine extreme Art logischer Ressourcenbeschränktheit expliziert.

Weniger starke Arten von Ressourcenbeschränktheit werden im Folgenden dargestellt. In der positiven Variante verfügen die diesen Formen der Ressourcenbeschränktheit entsprechenden Sprecher über bestimmte Fähigkeiten zur logischen Normierung ihrer Zustimmungsakte. Wir werden drei Arten solcher Sprechertypen explizieren: Primitiv-logische Sprechertypen, die nur einfachste logische Operationen vollziehen können, wobei dieser Vollzug schrittweise in Form von Frage-Antwort-Sequenzen vor sich geht. Über mehrere solcher Schritte können dann relativ komplizierte logische Folgen erarbeitet werden, ohne dass eine logische Vollständigkeit bezüglich der klassischen Logik bzw. des Systems der tautological entailments bei unbegrenzter Schrittmenge erreicht werden würde.

Eine zweite Form der logischen Normierung von Sprechern liegt vor, wenn die elementaren Regeln der Zustimmung so gefasst sind, dass über unbeschränk-

4 So wird durch die Reflexivität der Ri gesichert, dass $A(i, A(i, p)) \supset A(i, p)$ gilt (aber nicht $A(i, p) \supset p$) oder durch Transitivität von Ri erhalten wir $A(i, p) \supset A(i, A(i, p))$.

te Frage-Antwort-Sequenzen eine Vollständigkeit bezüglich des Systems der tautological entailments erreicht werden kann.[5]

In den genannten positiv ressourcenbeschränkten Ansätzen wird nicht die Widerspruchsfreiheit der Zustimmung vorausgesetzt. Hier ergibt sich eine weitere Möglichkeit zu einer verstärkten logischen Normierung indem angenommen wird, der in dieser Weise negativ normiert zustimmende Sprecher stimme keinen widersprüchlichen Sätzen zu. Wird diese Annahme mit der Annahme der Vollständigkeit im Sinne der tautological entailments kombiniert, erhalten wir einen Sprecher, der klassisch schrittweise vollständig ist.

Schließlich können ressourcenunbeschränkte Sprecher betrachtet werden, die bezüglich eines bestimmten logischen Systems unmittelbar beliebigen logischen Folgen aus zugestimmten Sätzen zustimmen. Das von Levesque dargestellte System ist in diesem Sinne bezüglich des Systems der tautological entailments ressourcenunbeschränkt, bleibt aber bezüglich der klassischen Logik ressourcenbeschränkt.

Die Umsetzung dieser Überlegungen in Weltensemantiken erfordert, dass aus der Menge der epistemischen Welten W_e Mengen von speziellen logisch strukturierten Welten hervorgehoben werden, mit denen die logischen Fähigkeiten ressourcenbeschränkter bzw. ressourcenunbeschränkter Sprecher expliziert werden. Wir wollen vorläufig folgende logisch strukturierte Arten epistemischer Welten hervorheben:

P^1: primitiv-logisch strukturierte Welten
P^2: tautologisch strukturierte Welten

(Wenn $G \rightarrow H$ ein gültiges tautological entailment ist, so gilt, wenn es eine Schrittzahl gibt, in der die Zustimmung zu G erreichbar ist, so gibt es auch eine Schrittzahl, in der die Zustimmung zu H erreichbar ist.)

P^3: widerspruchsfreie epistemische Welten
(Wenn die Zustimmung zu G in r Schritten erreichbar ist, so ist die Zustimmung zu ~G in einer beliebigen Anzahl von Schritten nicht erreichbar.)

P^4: klassische epistemische Welten
(Wenn H aus G klassisch ableitbar ist und G zugestimmt wird, so gilt, wenn es eine Schrittzahl gibt, in der die Zustimmung zu G erreichbar ist, so gibt es auch eine Schrittzahl, in der die Zustimmung zu H erreichbar ist.)

Für alle P^i (i = 1, 2, 3, 4) gilt: $P^i \subseteq W_e$.

Grundlegend für die Bestimmung der ressourcenbeschränkten Welten ist der Begriff der Ableitbarkeit eines Ausdrucks G in n Schritten in einer bestimmten Welt w:

5 Eine solche Vollständigkeit wird bei der Behandlung ressourcenbeschränkten Glaubens bei Levesque (Levesque/Lakemeyer 2000) angenommen, wobei aber dort die Vollständigkeit bezüglich der tautological entailments unmittelbar gegeben ist, also nicht lediglich durch eine Abfolge von Frage-Antwort-Sequenzen erreichbar wird.

$v(G, r, w) \in \{t, f\}$

$v(G, r, w) = t$ gdw.	G ist aus den in w explizit enthaltenen Sätzen in r Schritten ableitbar.

Für die üblicherweise (und auch oben) ohne Schrittzahlparameter angegebenen Wertezuordnungen ergeben sich daraus drei unterschiedliche Interpretationsmöglichkeiten:

a)	$v(G, w) = t$ gdw.	$v(G, 0, w)=t$, d. h., G ist in 0 Schritten ableitbar bzw. G ist explizit in w enthalten.
b)	$v(G, w) = t$ gdw.	$v(G, 1, w)=t$, d. h., G ist in einem Schritt ableitbar bzw. G ist unmittelbar in einem Ableitungsschritt in w gewinnbar.
c)	$v(G, w) = t$ gdw.	$\exists r\, v(G, r, w)=t$, d. h., G ist in w ableitbar bzw. G ist implizit in w enthalten.

Dabei ist auch die unter c) genannte Interpretation durchaus mit dem Ausdruck einer Ressourcenbeschränktheit vereinbar, denn von den vorhandenen Ressourcen (auch zeitlichen) hängt es ab, ob zu G führende Ableitungsschritte tatsächlich abgearbeitet werden können.

4. Primitiv-logische Zustimmungswelten

Primitiv-logische Zustimmungswelten können unterschiedlich logisch strukturiert sein. Insofern gibt es durchaus eine Vielzahl primitiv-logischer Zustimmungswelten, die unterschiedlichen Intuitionen primitiv-logischer Strukturierung epistemischer Welten und einer dementsprechenden Ressourcenbeschränktheit entsprechen. Wir wollen hier eine Art solcher primitiv-logischer Welten exemplarisch definieren, ohne den Anspruch zu erheben, damit die Vielfalt möglicher primitiv-logischer Welten zu erschöpfen.[6]

6 Zum Beispiel entspricht die hier gegebene Explikation primitiv-logischer Welten nicht der Intuition der strengen logischen Folgebeziehung Sinowjews bzw. der analytischen Implikation Parry/Dunns, nach denen in den abgeleiteten Ausdrücken kein begriffliches Material (hier: keine Aussagenvariable) vorkommen darf, das nicht in den Prämissen vorkommt. Damit wäre die unten angenommene Alternativeinführung ausgeschlossen, die klassisch und im System der tautological entailments gilt. Vgl. Sinowjew 1970, Parry 1933, 1989, Dunn 1970.

IR1. Wenn $w \in P_1$ (w ist eine primitiv-logisch strukturierte Welt), so gilt

$v(G, r, w) = t \Rightarrow v(G, r+1, w) = t$
$v(G, r, w) = t \Rightarrow v(\sim\sim G, r+1, w) = t$
$v(\sim\sim G, r, w) = t \Rightarrow v(G, r+1, w) = t$
$v(G, r, w) = t \ \& \ v(H, s, w) = t \Rightarrow v(G \wedge H, r+s+1, w) = t$
$v(G \wedge H, r, w) = t \Rightarrow v(G, r+1, w) \ \& \ v(H, r+1, w) = t$
$v(G, r, w) = t \ \mathrm{I} \ v(H, r, w) = t \Rightarrow v(G \vee H, r+1, w) = t$
$v(G \vee H, r, w) = t \Rightarrow v(G, r+1, w) = t \ \mathrm{I} \ v(H, r+1, w) = t$
$v(\sim G, r, w) = t \ \mathrm{I} \ v(H, r, w) = t \Rightarrow v(G \supset H, r+1, w) = t$
$v(G \supset H, r, w) = t \Rightarrow v(\sim G, r, w) = t \ \mathrm{I} \ v(H, r+s+1, w) = t$

Um eine dieser Semantik angemessene Syntax der Zustimmung aufzubauen, muss die Relativierung auf den Ableitbarkeitsparameter auch im Zustimmungsprädikat ausgedrückt werden:

A(x, G, r) heißt dementsprechend „x kann in r Frage-Antwort-Schritten zur Zustimmung zu G geführt werden“. Die Interpretationsregel für A(x, G, r) wird folgendermaßen bestimmt:

IR2. $v(A(x, G, r), w) = t$ gdw. $\forall w_1(w_1 \in W_S \ \& \ wR_x w_1 \Rightarrow v(G, r, w_1) = t)$

Neben der so bestimmten Zustimmung tritt eine schwächere Form der Erreichbarkeit der Zustimmung in r Schritten auf, die dann besteht, wenn der Satz, dem zugestimmt wird, in mindestens einer erreichbaren Zustimmungswelt in r Ableitungsschritten gewinnbar ist ($A^S(x, G, r)$). G ist also in mindestens einer Zustimmungswelt in r Schritten gewinnbar. Vom epistemischen Subjekt wird die Möglichkeit des schwach zugestimmten Satzes eingeräumt. Die entsprechende semantische Regel lautet:

IR3. $v(A^S(x, G, r), w) = t$ gdw. $\exists w_1(w_1 \in W_S \ \& \ wR_x w_1 \ \& \ v(G, r, w_1) = t)$

Setzt man die Seriellität ($\forall w \exists w_1(wRw_1)$) der Erreichbarkeitsrelationen R_x voraus, so gilt

$$\sim A(x, p, r) \supset A^S(x, p, r)$$

Die Seriellität von R_x führt aber im Gegensatz zu alethischen Modallogiken, bzw. solchen Modallogiken in deren Semantik nur widerspruchsfreie erreichbare Welten angenommen werden, nicht dazu, dass aus dem Vorliegen der Gewinnbarkeit einer Zustimmung zu G in r Schritten darauf geschlossen werden kann, dass eine Zustimmung zu ~G in r oder einer anderen Anzahl von Schritten nicht gewinnbar ist. Das heißt, es gilt nicht

$$A(x, p, r) \supset \sim A(x, \sim p, r).$$

Und es gilt auch nicht die Interdefinierbarkeit von A und A^S in der Form

$$A(x, G, r) \equiv \sim A^S(x, \sim G, r),$$

da die epistemischen Zustimmungswelten weder widerspruchsfrei noch vollständig sind.

Außerdem muss die Beschränkung des ressourcenbeschränkten epistemischen Subjekts auf bestimmte logisch strukturierte epistemische Welten syntaktisch ausgedrückt werden. Das kann einmal durch die Einführung mehrsortiger Variablen für epistemische Subjekte geschehen. Zum anderen – und das ist der hier beschrittene Weg – können Zusatzprädikate eingeführt werden, mit denen ausgedrückt wird, dass das epistemische Subjekt nur auf epistemische Welten relativiert ist, die bestimmte logische Strukturierungskriterien erfüllen. Für den primitiv-logisch Zustimmenden kann das folgendermaßen geschehen:
$P^1(x)$ soll für „x ist primitiv-logisch Zustimmender“ stehen, und es ist semantisch folgendermaßen zu bestimmen:

IR4. $v(P^1(x), w) = t$ gdw. $\forall w_1(wR_x w_1 \Rightarrow w_1 \in P^1)$

Analog würde die Sortenzugehörigkeit anderer ressourcenbeschränkter epistemischer Subjekte über entsprechende Zusatzprädikate ausgedrückt.

In der gegebenen Semantik ist der folgende Ausdruck allgemeingültig, der zugleich als Axiom zum alethischen System hinzugefügt ein vollständiges und widerspruchsfreies Axiomensystem der Zustimmungslogik liefert, dessen Axiome so bestimmt sind, dass die abgeleiteten Zustimmungen ressourcenbeschränkt in einem Schritt aus den vorausgesetzten Zustimmungen gewonnen werden können:

$$
\begin{aligned}
\sim \quad & P^1(x) \supset \\
& (A(x, p, r) \supset A(x, \sim\sim p, r+1)) \wedge \\
& (A(x, \sim\sim p, r) \supset A(x, p, r+1)) \wedge \\
& (A(x, p, r) \wedge A(x, q, s) \supset A(x, p \wedge q, r+s+1)) \wedge \\
& (A(x, p \wedge q, r) \supset A(x, p, r+1) \wedge A(x, q, r+1)) \wedge \\
& (A(x, p, r) \vee A(x, q, r) \supset A(x, p \vee q, r+1)) \wedge \\
& (A(x, p \vee q, r) \wedge \sim A^S(x, p, r) \supset A(x, q, r+1)) \wedge \\
& (A(x, \sim p, r) \vee A(x, q, r) \supset A(x, p \supset q, r+1)) \wedge \\
& (A(x, p \supset q, r) \wedge \sim A^S(x, \sim p, r) \supset A(x, q, r+1)))
\end{aligned}
$$

5. Primitiv-logische Zustimmungswelten und Widersprüche

Obwohl sowohl

$$\sim \quad P^1(x) \supset (A(x, p \vee q, r) \wedge \sim A^S(x, p, r) \supset A(x, q, r+1))$$

als auch

$$\sim \quad P^1(x) \supset (A(x, p \supset q, r) \wedge \sim A^S(x, \sim p, r) \supset A(x, q, r+1))$$

gelten,

gelten (wie für Teilsysteme der tautological entailments nicht anders zu erwarten) weder der disjunktive Syllogismus

$$P^1(x) \supset (A(x, G \vee H, r) \wedge A(x, \sim G, s) \supset A(x, H, r+s+1))$$

noch der modus ponens

$$P^1(x) \supset (A(x, G \supset H, r) \wedge A(x, G, s) \supset A(x, H, r+s+1))$$

bezüglich der inneren Logik von Behauptungen für den primitiv-logisch Zustimmenden.

Diese Beziehungen würden für einen primitiv-logisch Zustimmenden x nur gelten, wenn er in folgendem Sinne nicht widersprüchlich zustimmt, d. h., wenn für x gilt, $A(x, G, r) \supset \sim A^s(x, \sim G, r)$ und $A(x, \sim G, r) \supset \sim A^s(x, G, r)$. Unter diesen Voraussetzungen würden dann der modus ponens und der disjunktive Syllogismus gelten.

Dieses Ergebnis kann verallgemeinert und die strenge Widespruchsfreiheit in der Semantik ausgedrückt werden, indem die primitiv-logischen Zustimmungswelten als global widerspruchsfreie Zustimmungswelten spezifiziert werden, die folgendermaßen bestimmt werden:

IR5. Wenn $w \in P_3$ so gilt $\nu(G, r, w) = t \Rightarrow \nu(\sim G, s, w) = f$

Damit gilt:

$$\sim \quad P^3(x) \supset (A(x, G, r) \supset \sim A^s(x, \sim G, s))$$

und natürlich auch das schwächere

$$\sim \quad P^3(x) \supset (A(x, G, r) \supset \sim A(x, \sim G, s)).$$

Für global widerspruchsfreie primitiv-logische Sprecher gelten dann sowohl der modus ponens als auch die Beseitigungsregel für die Alternative:

$$\sim \quad P^1(x) \wedge P^3(x) \supset ((A(x, G \supset H, r) \wedge A(x, G, s) \supset A(x, H, r+s+1)$$
$$\sim \quad P^1(x) \wedge P^3(x) \supset ((A(x, G \vee H, r) \wedge A(x, \sim G, s) \supset A(x, H, r+s+1)),$$

also zwei klassisch, aber nicht tautologisch gültige Prinzipien.

6. Distributivgesetze und primitivlogische Welten

Im System der tautological entailments gelten folgende Distributivgesetze:

$$(p \vee q) \wedge (r \vee s) \rightarrow p \wedge r \vee p \wedge s \vee q \wedge r \vee q \wedge s$$
$$(p \wedge q) \vee s \rightarrow (p \vee s) \wedge (q \vee s)$$

Für den primitiv-logisch Zustimmenden gelten diese Beziehungen nicht für den auf einen Schritt ressourcenbeschränkten Übergang vom Antezedent zum Konsequent, d. h., es gilt nicht:

$$P^1(x) \supset (A(x, (p \vee q) \wedge (p_1 \vee q_1), r) \supset A(x, p \wedge p_1 \vee p \wedge q_1 \vee q \wedge p_1 \vee q \wedge q_1, r+1))$$

Analog gilt auch nicht

$$P^1(x) \supset A(x, (p \wedge q) \vee p_1, r) \supset A(x, (p \vee p_1) \wedge (q \vee p_1), r+1))$$

Allerdings gelten folgende Abschwächungen:

$$\sim P^1(x) \supset (A(x, (p \vee q) \wedge (p_1 \vee q_1), r) \supset \exists u A(x, p \wedge p_1 \vee p \wedge q_1 \vee q \wedge p_1 \vee q \wedge q_1, u))$$

und

$$\sim P^1(x) \supset A(x, (p \wedge q) \vee p_1, r) \supset \exists u A(x, (p \vee p_1) \wedge (q \vee p_1), u))$$

Auf konkrete Schrittzahlen bezogen gilt:

$$\sim P^1(x) \supset A(x, (p \wedge q) \vee p_1, r) \supset A(x, (p \vee p_1) \wedge (q \vee p_1), 2r+7))$$

Um das zu demonstrieren:

Angenommen $x \in P^1$ und $v((p \wedge q) \vee p_1, r, w) = t$.

Dann gilt:

(*) $v(p \wedge q, r+1, w) = t$ I $v(p_1, r+1, w) = t$.

Verfolgen wir die linke Seite der Alternative (*). Dann gilt

$v(p, r+2, w) = t$ & $v(q, r+2, w) = t$.

Daraus erhalten wir

$v(p \vee p_1, r+3, w) = t$ & $v(q \vee p_1, r+3, w) = t$.

Hieraus:

$v((p \vee p_1) \wedge (q \vee p_1), (r+3)+(r+3)+1, w) = t$, also
$v((p \vee p_1) \wedge (q \vee p_1), 2r+7, w) = t$.

Für die rechte Seite der Alternative (*) erhalten wir

$v(p \vee p_1, r+2, w) = t$ & $v(q \vee p_1, r+2, w) = t$. Daraus
$v((p \vee p_1) \wedge (q \vee p_1), 2r+5, w) = t$ und daraus
$v((p \vee p_1) \wedge (q \vee p_1), 2r+7, w) = t$.

Da dies für beide Seiten der Alternative (*) gilt, folgt dies also auch aus der Voraussetzung selbst. Folglich ist allgemeingültig:

$$\sim P^1(x) \supset A(x, (p \wedge q) \vee p_1, r) \supset A(x, (p \vee p_1) \wedge (q \vee p_1), 2r+7))$$

Unter Ressourcenbeschränkung ist damit jedoch nicht gesichert, dass von der den Antezedent betreffenden Zustimmung aus, die den Konsequent betreffende Zustimmung tatsächlich erreicht werden kann. Es kann also der Fall eintreten, dass dem Antezedent tatsächlich zugestimmt wird, während die Zustimmung zum Konsequent unter den gegebenen Ressourcenbeschränkungen nicht erreicht werden kann, da selbst bei expliziter Zustimmung zu $(p \wedge q) \vee p_1$ (wenn also der

Fall ist A(x, $(p \wedge q) \vee p_1$, 0)) immerhin 7 Schritte notwendig sind, um die Zustimmung zu $(p \vee p_1) \wedge (q \vee p_1)$ zu sichern.

Konkret kann die Ressourcenbeschränkung über die Erreichbarkeitsdifferenz zwischen Antezedent und Konsequent bestimmt werden. Die untere Grenze (und maximale Ressourcenbeschränkung) wird dabei durch 0 gesetzt, wobei dann keinerlei aus expliziten Zustimmungen erreichbare implizite Zustimmungen auftreten. Der Zustimmungsraum bleibt also auf die expliziten Zustimmungen beschränkt.

Eine strenge Ressourcenbeschränkung liegt vor, wenn die Erreichbarkeitsdifferenz zwischen Erreichbarkeit der Zustimmung zum Antezedent und der Erreichbarkeit der Zustimmung zum Konsequent auf 1 gesetzt wird, wodurch also nur unmittelbare erreichbare (in einem Argumentationsschritt gewinnbare) Ableitungen sicher zu einer neuen Zustimmung geführt werden können. Der entsprechende Zustimmungsraum wird durch die oben angegebene Axiomatisierung für die 1-stufig erreichbare Zustimmung beschrieben.

Über Zwischenformen der Ressourcenbeschränktheit auf n Ableitbarkeitsschritte (n>1) kann schließlich zur Aufgabe der Ressourcenbeschränkung übergegangen werden, indem zugelassen wird, dass die Erreichbarkeitsdifferenz zwischen Antezedent und Konsequent beliebig groß sein kann. Das führt zu Systemen logischer Allwissenheit bezüglich einer zugrunde gelegten Semantik, wie das bei Levesque bezüglich des Systems der tautological entailments geschieht. Der von Levesque erhobene Anspruch, ein System ressourcenbeschränkten Glaubens geliefert zu haben, wird also nicht erfüllt, denn bezüglich des Systems der tautological entailments ist das von Levesque geliefert System absolut ressourcenunbeschränkt, obwohl es nicht vollständig bezüglich der klassischen Logik ist.

7. Tautologisch-Entailment strukturierte Welten

Die im gegebenen System explizierte Folgebeziehung kann in folgendem Sinne für primitiv-logisch Behauptende als Teilsystem der tautologischen entailments aufgefasst werden:

> Wenn gilt $\sim P^1(x) \supset (\exists r A(x, G, r) \supset \exists r A(x, H, r))$,
> so ist $G \rightarrow H$ ein gültiges tautological entailment.

Die Umkehrung dieses Satzes gilt allerdings nicht. So ist z. B.

$$\sim(p \wedge q) \rightarrow \sim p \vee \sim q$$

ein gültiges tautological entailment, es gilt aber nicht

$$P^1(x) \supset (\exists r A(x, \sim(p \wedge q), r) \supset \exists r A(x, \sim p \vee \sim q, r)).$$

Wird neben primitiver Logizität auch die Widerspruchsfreiheit der Set-up-Welten (und damit die Widerspruchsfreiheit des Zustimmenden) vorausgesetzt, so erhalten wir eine Folgebeziehung für die Zustimmung, die unter der klassischen Folgebeziehung geschlossen ist, nicht aber unter den tautological entailments. Es gilt zwar

Wenn gilt $\sim P^1(x) \wedge P^3(x) \supset (\exists rA(x, G, r) \supset \exists rA(x, H, r))$,
so ist $G \sim H$ eine klassisch gültige Folgebeziehung.

Aber es gilt nicht mehr

Wenn gilt $\sim P^1(x) \wedge P^3(x) \supset (\exists rA(x, G, r) \supset \exists rA(x, H, r))$,
so ist $G \rightarrow H$ ein gültiges tautological entailment.

Bespiele dafür sind die Gültigkeit des modus ponens und der Alternativbeseitigung unter Voraussetzung von $P^1(x) \wedge P^3(x)$ (d. h. $\sim P^1(x) \wedge P^3(x) \supset (\exists rA(x, (p \supset q) \wedge p, r) \supset \exists rA(x, q, r)))$, obwohl $(p \supset q) \wedge p \rightarrow q$ kein gültiges tautological entailment ist.

Ein bezüglich der tautological entailments adäquates System der Glaubenslogik hat Levesque entwickelt, wobei dort allerdings die Ressourcenbeschränktheit des Glaubens nicht berücksichtigt wurde. Dieses System kann zu einem klassisch adäquaten System erweitert werden, in dem klassische Set-up-Welten vorausgesetzt werden, d. h., solche, die widerspruchsfrei und vollständig sind.

Eine adäquate Semantik für eine zum System der tautological entailments adäquaten Zustimmungsfolgebeziehung kann folgendermaßen gegeben werden:

Der Wert von G bezüglich der Schrittzahl r in der Welt w ([G, r, w]) wird folgendermaßen bestimmt:

$[G, r, w] \subseteq \{(G, r, w), (\sim G, r, w)\}$

Jetzt wird folgende Interpretationsregel für tautologisch strukturierte Welten angenommen:

IR6. Wenn $w \in P_2$ so gilt

1. $v(G, r, w) = t \Leftrightarrow (G, r, w) \in [G, r, w]$
   $v(G, r, w) = f \Leftrightarrow (\sim G, r, w) \in [G, r, w]$

Hier ist es möglich, dass sowohl $v(G, r, w) = t$ als auch $v(G, r, w) = f$ wahr sind, oder auch nur das erste, nur das zweite oder keines von beiden. Es gilt also weder $(v(G, r, w) = t$ I $v(G, r, w) = f)$ noch $\neg((v(G, r, w) = t \,\&\, v(G, r, w) = f)$.

2. $v(\sim G, r, w) = t \Leftrightarrow v(G, r, w) = f$
   $v(\sim G, r, w) = f \Leftrightarrow v(G, r, w) = t$
3. $v(G, r, w) = t \;\&\; v(H, s, w) = t \Rightarrow v(G \wedge H, r+s+1, w) = t$
   $v(G \wedge H, r, w) = t \Rightarrow v(G, r+1, w)=t \;\&\; v(H, r+1, w)=t$
   $v(G, r, w) = f \;\&\; v(H, s, w) = f \Rightarrow v(G \wedge H, r+s+1, w) = f$
   $v(G \wedge H, r, w) = f \Rightarrow v(G, r+1, w)=f$ I $v(H, r+1, w)=f$

4. $v(G, r, w) = t \text{ I } v(H, r, w) = t \Rightarrow v(G \vee H, r+1, w)$
   $v(G \vee H, r, w) = t \Rightarrow v(G, r, w) = t \text{ I } v(H, r, w) = t$
   $v(G, r, w) = f \;\&\; v(H, s, w) = f \Rightarrow v(G \vee H, r+s+1, w)=f$
   $v(G \vee H, r, w) = f \Rightarrow v(G, r+1, w) = f \;\&\; v(H, r+1, w) = f$
5. $v(G, r, w) = f \text{ I } v(H, r, w) = t \Rightarrow v(G \supset H, r+1, w) = t$
   $v(G \supset H, r, w) = t \Rightarrow v(\sim G, r+1, w) = t \text{ I } v(H, r+1, w) = t$
   $v(G, r, w) = t \;\&\; v(H, s, w) = f \Rightarrow v(G \supset H, r+s+1, w) = f$
   $v(G \supset H, r, w) = f \Rightarrow v(\sim G, r+1, w) = f \;\&\; v(H, r+1, w) = f$

Weiter gilt

IR7. $v(A(x, G, r), w)=t$ gdw. $\forall w_1(w_1 \in W_S \;\&\; wR_x w_1 \Rightarrow v(G, r, w_1)=t)$
$v(A(x, G, r), w)=f$ gdw. $\neg (v(A(x, G, r), w)=t)$

Unter Voraussetzung tautologisch Zustimmender ist die dieser Semantik entsprechende Zustimmungsfolgebeziehung mit der durch die tautological entailments explizierten Folgebeziehung äquivalent, d. h., widerspruchsfrei und vollständig. D. h. es gilt:

Der Ausdruck $G \rightarrow H$ ein genau dann ein gültiges tautological entailment, wenn $\sim P^2(x) \supset (\exists r A(x, G, r) \supset \exists r A(x, H, r))$, wenn also gilt, dass wenn es eine Schrittzahl gibt, in der die Zustimmung zu G erreichbar ist, so gibt es auch eine Schrittzahl, in der die Zustimmung zu H erreichbar ist.

Um eine zur klassischen Folgebeziehung adäquate Zustimmungsfolgebeziehung zu erhalten, kann zu den obigen Interpretationsregeln folgende Festlegung hinzugefügt werden:

IR8. Wenn $w \in P_4$, dann gilt
$\exists r\, v(G, r, w)=t \Leftrightarrow \neg \exists r\, v(\sim G, r, w)=t$

In den Ableitungsmengen der klassischen Welten ist also entweder G oder nicht ~G enthalten, nicht beide zugleich (Widerspruchsfreiheit) und mindestens eines von beiden (Vollständigkeit). Wenn gilt $P^2(x) \wedge P^4(x)$ (also lediglich klassische und tautologische Set-up-Welten für x zulässig sind), so ist die Zustimmungsfolgebeziehung von x klassisch bestimmt.
In unserer obigen Darstellung haben wir drei Aspekte der Beschränkung der logischen Ressourcen von epistemischen Subjekten ausgedrückt:

1. Die für Subjekte der Art P^1 definierten Ableitbarkeitsbeziehungen sind unter den tautological entailments geschlossen, aber nicht vollständig bezüglich der tautological entailments und in diesem Sinne ressourcenbeschränkt bezüglich des Systems der tautological entailments.
2. Eine andere Art der Ressourcenbeschränkung ist damit verbunden, dass ein epistemisches Subjekt der Art P^1 oder der Art P^2 nur eine bestimmte Anzahl von Ableitungsschritten in Frage-Antwortsequenzen durchlaufen kann und damit selbst dann, wenn für dieses epistemische Subjekt gilt $\exists r A(x, G, r) \supset \exists r A(x, H, r)$, x selbst dann wenn x zur expliziten Zustimmung zu G geführt werden kann, x nicht zur expliziten Zustim-

mung zu H geführt werden kann, da die dazu notwendigen Schritte die Menge der von x tatsächlich vollziehbaren Schritte überschreitet.

3. Sind unmittelbare Ableitungsregeln (für solche Ableitungen, die in einem Schritt vollzogen werden können) für epistemische Subjekte der Art P^1 bezüglich dieser Subjekte in dem Sinne ressourcenbeschränkt, dass in diesen Regeln zur 1-Schritt-Ableitung höchstens zwei einstellige Verbindungen und eine zweistellige Verbindung vorkommen können.

Das von Levesque als ressourcenbeschränktes System der Glaubenslogik vorgestellte System ist in keinem der drei genannten Aspekte ressourcenbeschränkt. Es ist bezüglich des Systems der tautological entailments vollständig und widerspruchsfrei und lediglich bezüglich der klassischen Folgebeziehung in dem Sinne ressourcenbeschränkt, dass keine klassisch gültige Ableitung im System gültig ist, die nicht tautologisch gültig ist.

Literaturverzeichnis

Anderson, Alan Ross/Belnap Jr., Nuel D.: Entailment, vol. 1. Princeton University Press, Princeton, 1975

Anderson, Alan Ross/Belnap Jr., Nuel D./Dunn, Jon Michael: Entailment, vol. 2. Princeton, 1992

Dunn, Jon Michael: "A Modification of Parry's Analytic Implication". Notre Dame Journal of Formal Logic, 13, 1972. S. 195–205

Engel, Pascal: Dispositional Belief, Assent, and Acceptance. Dialectica, 53, 1999. S. 211–226.

Frege, Gottlob: Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. L. Nebert, Halle a. S., 1879

Frege, Gottlob: Grundgesetze der Arithmetik, Bd. I und II. H. Pohle, Jena, 1893/1903

Gabriel, Gottfried: Zustimmung. II. „Neuzeit". In: *J. Ritter/K. Gründer/G. Gabriel (Hrsg.):* Historisches Wörterbuch der Philosophie, Bd. 12. Schwabe-Verlag, Basel, 2004. S. 1466–1470

Levesque, Hector J./Lakemeyer, Gerhard: The Logic of Knowledge Bases. MIT, Cambridge/London, 2000

Lorenz, Kuno: „Dialogspiele als semantische Grundlage von Logikkalkülen". Archiv für mathematische Logik und Grundlagenforschung, 11, 1967/68

Lorenzen, Paul: Formale Logik. De Gruyter, Berlin, 1967

Lorenzen, Paul: „Szientismus versus Dialektik“. In: Rehabilitierung der praktischen Philosophie, Bd. 2: Rezeption, Argumentation, Diskussion. Verlag Rombach, Freiburg i. Br., 1972. S. 335–351

Lorenzen, Paul: „Rationale Grammatik“. In: *C. F. Gethmann (Hrsg.):* Theorie des wissenschaftlichen Argumentierens. Suhrkamp, Frankfurt a. M., 1980. S. 73–94

Lorenzen, Paul/Lorenz, Kuno: Dialogische Logik. Wissenschaftliche Buchgesellschaft, Darmstadt, 1978

Marcus, Ruth Barcan: “A Proposed Solution to a Puzzle about Belief”. Midwest Studies in Philosophy, IV, 1981. S. 501-510

Marcus, Ruth Barcan: “Rationality and Believing the Impossible”. The Journal of Philosophy, 80, 1985. S. 321–338

Marcus, Ruth Barcan: “Some Revisionary Proposals about Belief and Believing”. Philosophy and Phenomenological Research, 50, 1986. S. 133–153

Marcus, Ruth Barcan: Modalities. Oxford University Press, New York, 1993

Marcus, Ruth Barcan: “The Anti-naturalism of some Language-centered Accounts of Belief”. Dialectica, 49, 1995. S. 113–129

Moltmann, Frederike: “Propositional Attitudes without Propositions”. Synthese, 135, 2003. S- 77–118

Parry, William T.: „Ein Axiomensystem für eine neue Art von Implikation (analytische Implikation)“. Ergebnisse eines Mathematischen Kolloquiums, 4, 1933, 5–6. In: *K. Berka/L. Kreiser (Hrsg.):* Logik-Texte. Akademie-Verlag, Berlin, 3. Aufl., 1983. S. 163–164

Parry, William T.: „Analytic Implication. Its History, Justification and Varieties“. In: *J. Norman/R. Sylvan (Hrsg.):* Directions in Relevant Logic. Kluwer, Dordrecht, 1989. S. 101–118

Pettit, Dean: “Practical Belief and Philosophical Theory”. *The Australasian Journal of Philosophy*, 76, 1998. S. 15–33

Piel, Wilhelm: „Zur formalen Pragmatik konstativer Performatoren“. In: *C. F. Gethmann (Hrsg.):* Theorie des wissenschaftlichen Argumentierens. Suhrkamp, Frankfurt/M., 1980. S. 165–189

Proust, Joëlle: „A Plea for Mental Acts“. Synthese, 129, 2001. S. 105–128

Shoemaker, Sydney: “Moore’s Paradox and Self-Knowledge”. In: The First-Person Perspective and Other Essays. Cambridge University Press, Cambridge, 1996. S. 74–93.

Sinowjew, Alexander A.: Komplexe Logik. Verlag der Wissenschaften, Berlin, 1970

2 Erkenntnistheorie

Cartesian Arguments for Scepticism - Their Interrelations and Presuppositions[1]

Jochen Briesen
briesen@semuin.de
Humboldt-Universität Berlin

Abstract/Zusammenfassung

Cartesian Arguments make essential use of so-called sceptical hypotheses in order to draw their sceptical conclusion, that we know (almost) nothing about the external world. There are two promising ways to argue for this conclusion via highlighting a sceptical hypothesis: The closure and the underdetermination argument. With respect to these arguments Anthony Brueckner (1994, 2005) defended two claims. First, he claims that the closure argument rests on the underdetermination argument. If Brueckner were right, most of the contemporary discussions of scepticism would be focused on a superfluous argument. Second, he claims that both arguments presuppose infallibilism. If this claim were true, fallibilists would be right in not taking the problems posed by these two arguments seriously. As many epistemologists are sympathetic to fallibilism, this would be a very interesting result. However, in this paper I will argue that Brueckner's claims are wrong: The closure and the underdetermination argument are not as closely related as he assumes and neither rests on infallibilism. Thus even a fallibilist should take these Cartesian arguments to raise serious problems that must be dealt with somehow.

Cartesianische Argumente für den Skeptizismus zeichnen sich durch ihren Gebrauch skeptischer Hypothesen aus. Zwei vielversprechende Argumente dieser Art lassen sich unterscheiden: das Geschlossenheits- und das Unterbestimmtheitsargument. In Bezug auf diese Argumente verteidigt Anthony Brueckner (1994, 2005) zwei Thesen. Er behauptet erstens, dass das Geschlossenheitsargument auf dem Unterbestimmtheitsargument beruht. Wenn Brueckner Recht behielte, so wären die meisten gegenwärtigen Diskussionen des Skeptizismus auf ein letztlich überflüssiges Argument fokussiert. Zweitens behauptet er, dass beide Argumente den Infallibilismus voraussetzen. Wenn diese Behauptung korrekt wäre, so gäbe es keinen Grund für Fallibilisten, Cartesianische Argumente ernst zu nehmen. Weil die meisten gegenwärtigen Erkenntnistheoretiker erklärte Fallibilisten sind, wäre dies sicher ein interessantes Ergebnis. Allerdings werde ich in diesem Aufsatz dafür argumentieren, dass Brueckners Thesen falsch sind: Der Zusammenhang des Geschlossenheits- und des Unterbestimmtheitsarguments ist nicht so eng wie er vermutet und keines der Argumente basiert auf dem Infallibilismus. Auch Fallibilisten müssen daher akzeptieren, dass die beiden Cartesianischen Argumente ernste Probleme darstellen, die einer echten Lösung bedürfen.

1 This is a slighty modified version of my paper 'Reconsidering Closure, Underdetermination, and Infallibilism', in: *Grazer Philosophische Studien* 80 (2010), pp.221-234.

1 Introduction

Some sceptical arguments make essential use of so-called sceptical hypotheses in order to draw their sceptical conclusion, that we know (almost) nothing about the external world. An argument makes essential use of a sceptical hypothesis, if it is possible to block the argument by ruling out the hypothesis or by knowing that the hypothesis in question is false. I will call arguments of this kind "Cartesian arguments" and the resulting form of scepticism "Cartesian scepticism". Cartesian arguments can vary along the following parameters:

(1) *Target.* A Cartesian argument can be directed against knowledge or justification.
(2) *Depth.* A Cartesian argument can be directed against first- or higher-order knowledge/justification.
(3) *Scope.* A Cartesian argument can be directed against our knowledge/justification in general or against a certain sufficiently wide range of knowledge/justification (e.g. knowledge/justification of the external world).
(4) *Sceptical Hypothesis.* Cartesian arguments can vary along the choice of the sceptical hypothesis.
(5) *Argumentative Structure.* Cartesian arguments can vary in argumentative structure.

In this paper I am interested in the argumentative structure of Cartesian arguments. Therefore, I will fix the parameters (1)-(4). That means I will focus on Cartesian arguments which are: (1) knowledge-specific (target), (2) restricted to first-order-knowledge (depth), (3) directed against our knowledge of the external world (scope), and (4) concerned with a sceptical hypothesis that is incompatible with our actual beliefs about the external world.
With these fixations in place, Cartesian arguments only vary along (5): their argumentative structure. There are two promising ways to argue for a sceptical conclusion via highlighting a sceptical hypothesis. One route to Cartesian scepticism is based on the closure and the other on the underdetermination principle. These two principles lead to different patterns of sceptical arguments: the closure and the underdetermination argument. The central questions of this paper are: How are the two standard patterns of Cartesian arguments interrelated? Does one of the arguments depend on the other? Are there different presuppositions in place for the closure- and the underdetermination-argument?

I will tackle these questions via a discussion of two claims repeatedly defended by Anthony Brueckner:

Claim A. The closure argument presupposes the unterdetermination argument: In order to motivate one of the premises of the closure argument the sceptic has to refer to the underdetermination argument. As a consequence, the closure argument is superfluous in motivating Cartesian scepticism (Brueckner 1994, 830-833; 2005, 388-390).

Claim B. The underdetermination argument and with it the closure argument are based on infallibilism: In order to motivate the premises of the arguments the sceptic has to refer to an infallibility principle. As a conse-quence, a fallibilist does not have to take these Cartesian arguments to raise a serious challenge (Brueckner 2005, 389-390).

Claim A is interesting because most of the contemporary discussions of scepticism are focused on the closure argument. If claim A were true, the closure argument would persuppose the underdetermination argument and most of the discussions would be concerned with a superfluous argument. And claim B is interesting because most epistemologists are fallibilists. So if claim B were true and both Cartesian arguments would presuppose infallibilism, for most epistemologists Cartesian scepticism would not pose an interesting problem at all. Cartesian arguments for scepticism are philosophically interesting because their conclusion is very implausible but their premises are at least *prima facie* very plausible. The challenge is to locate the mistake in these arguments. But if claim B were true and sceptical arguments relied on an infallibility principle as a premise, then – at least for fallibilists – it would be very easy to locate the mistake in these arguments.

However, I will argue that Brueckner's claims A and B are false. The closure argument does not depend on the underdetermination argument and neither the closure nor the underdetermination argument presupposes infallibilism. Before I give a short overview of the paper, let me make one further remark. This paper is not concerned with the logical relation of the two Cartesian arguments, which has been discussed by Stewart Cohen (1998) and Duncan Pritchard (2005). Both authors agree that the closure principle for *justification* entails the underdetermination principle but they also argue convincingly, contra Anthony Brueckner (1994), that the two principles are not equivalent. From a logical point of view we are thus faced with two distinct epistemic principles and therefore with two different Cartesian arguments that employ these principles, respectively. However, it remains an interesting question, whether these arguments are interrelated in another sense: Is Brueckner right in thinking that the two Cartesian arguments depend on each other and on infallibilism in order to get their premises motivated?

The following discussion of Brueckner's claims A and B is organized as follows. In section 2, I will introduce the closure and the underdetermination argument as well as the principle of infallibilism. In section 3 and 4, I will critically discuss Brueckner's claims A and B and explain why his reasoning is defective. Finally, I will give a short summary in section 5.

2 Introducing the Arguments and Infallibilism

In the following arguments *S* stands for an epistemic subject, *p* stands for a proposition concerning the external world and *SH* stands for a sceptical hypothesis that is incompatible with *p*. Let *SH* be the proposition *that we are all*

brains-in-vats (placed on a deserted planet) with induced hallucinated experiences indistinguishable from our actual experience.[2]

Closure Argument

(c1) If *S* knows *p*, then *S* knows ¬*SH*.
(c2) *S* does not know ¬*SH*.
(c3) Hence, *S* does not know *p*.

This argument is called "closure-argument" because premise (c1) is based on the principle that knowledge is closed under known entailment:

(CP) For all *S*, *p*, *q*, if *S* knows *p* and *S* knows that *p* entails *q*, then *S* knows *q*.[3]

By accepting (CP) premise (c1) can easily be justified because (c1) can be seen as an abbreviated instance of (CP). For the rest of the paper I accept (CP), which I take to be a very plausible principle. In our context premise (c2) is more interesting because it is this premise which Brueckner holds to rest on the underdetermination argument.

What does the underdetermination argument look like? The standard version of the argument is this (see e.g. Brueckner 1994; Pritchard 2005):

Underdetermination Argument

(u1) If *S's* evidence for believing *p* does not favour *p* over *SH*, then *S*'s evidence does not justify *S* in believing *p*.
(u2) *S's* evidence for believing *p* does not favour *p* over *SH*.
(u3) Therefore, *S* lacks justification for believing *p*.
(u4) Hence, *S* does not know *p*.

This line of thought is called "underdetermination argument" because premise (u1) rests on the underdetermination principle:

(UP) For all *S*, *p*, *q*, if *S's* evidence for believing *p* does not favour *p* over some incompatible hypothesis *q*, then *S's* evidence does not justify *S* in believing *p*.

What is *favouring*? Following Brueckner I will understand the term "favouring" in the following way. If my evidence favours *p* over *q*, then *p* has some epistemic credit which *q* lacks. In other words: If my evidence favours *p* over *q*, then it is more reasonable for me to believe *p* than *q* (Brueckner 2005, 389).

---

2 If you think that the brain-in-a-vat scenario is inconsistent with plausible views of semantic externalism, replace the standard brain-in-a-vat scenario by the recent-envatment scenario and restrict the scope of the sceptical argument to knowledge of our current environment.

3 John Hawthorne (2003) argues that only a reformulated and weaker version of the closure principle can be defended successfully. If you think this is true, adjust the sceptical premises (c1) and (c2) to this reformulation. It will not undermine the considerations in this paper.

(UP) is a very plausible principle as well. If my evidence cannot favour *p* over *q*, then it seems that my evidence can give me at most justification for the belief in the disjunction *p or q*. Why should I rationally prefer any one of the disjuncts in this case?

But the argument depends on two more principles. In order to infer (u3) from (u1) and (u2), we have to assume that all justification is evidential, that a lack of appropriate evidence entails a lack of justification – there is no such thing as warrant for nothing or justification by default. Therefore, everybody who endorses the underdetermination argument is committed to the following evidence principle of justification:

(EP) For all *S*, *p*, if *S* has justification for believing *p*, then the belief in *p* is justified by evidence.

And to infer (u4) from (u3), we have to assume that justification is a necessary condition for knowledge, that a lack of justification for the belief that *p* entails a lack of knowledge that *p*. Therefore, everybody who endorses the underdetermination argument is also committed to the following justification principle of knowledge:

(JP) For all *S*, *p*, if *S* knows *p*, then *S* has justification for believing *p*.

By adding two premises to the underdetermination argument that are based on these two principles we would get the full version of the argument. But for simplicity's sake I will focus on the abbreviated version. Even though the two mentioned principles are very controversial, I will accept them for the rest of the paper. With all the mentioned principles in place, the only problematic premise of the underdetermination argument is (u2). It is (u2) which Brueckner holds to rely on infallibilism.

What is infallibilism? Following Brueckner, I will focus on infallibilism with regard to justification. This form of infallibilism holds the following infallibility principle:

(INF) For all *S*, *p*, if *S* has justification for believing *p* in virtue of having evidence *E*, then *E* guarentees *p*.

In most of the literature the infallibility principle is formulated with the term "entail" rather than with the term "guarantee". But like Brueckner I try to keep the discussion concerning the relation between Cartesian arguments and infallibilism indifferent with regard to the question whether our evidence is propositional in structure or not. And since an entailment relation can only obtain between propositions or propositional states, I formulated the infallibility principle with the term "guarantee" instead of the term "entail". Whereby "*S's* evidence *E* guarantees *p*" is equivalent to "*S's* evidence *E* entails *p or* the proposition that *S* has evidence *E* entails *p*". So the infallibility principle (INF) is a combination of Brueckner's principles (JEP) and (JEP*) (see Brueckner 2005, 384-386).

Combining the two principles in one principle only simplifies the discussion a little – nothing substantial hinges on this combination.[4]

In contrast to the closure and underdetermination principle, the infallibility principle (INF) is very implausible – it is incredibly strong and it would make inductive inferences completely unreasonable. With a strong principle like this in place it is easy to argue for a sceptical conclusion. The following sceptical argument is called infallibility argument:

Infallibility Argument

(i1) If *S* has justification for believing that *p* in virtue of having evidence *E*, then *E* guarentees that *p*.
(i2) *S's* evidence *E* for believing that *p* does not guarantee *p*.
(i3) Therfore *S* lacks justification for believing that *p*.
(i4) Hence, *S* does not know that *p*.

Obviously the infallibility argument just as much rests on (EP) and (JP) as the underdetermination argument does. But in contrast to the underdetermination argument, the infallibility argument also rests on the very implausible infallibility principle (INF). It is thus a bad sceptical argument, which for most epistemologists does not even pose a serious challenge. At least for a fallibilist, who rejects (INF), it is very easy to block the infallibility argument.

3 Brueckner's Claim A

Brueckner's claim A is that the closure argument presupposes the unterdetermination argument: In order to motivate one of the premises of the closure argument the sceptic has to refer to the underdetermination argument. Why should we think that claim A is true?

Brueckner argues that the sceptic cannot defend (c2) by referring to Robert Nozick's (1981) sensitivity account of knowledge. Therefore the sceptic has to justify (c2) by referring to the underdetermination argument (Brueckner 2005, 388; 1994, 828-830). Just replace *p* by ¬*SH* in the underdetermination argument to see how this might be done.

Why does Brueckner think that (c2) cannot be established via the sensitivity account? Nozick's sensitivity account appeals to the following *tracking condition*:

4 Another difference in terminology has to be noted. Brueckner reserves the term "evidential justification" for justification by propositionally structured evidence (believed evidential propositions) (see Brueckner 2005, 386). This is why he refers to non-propositional evidence with "non-*evidential* justifier". I will use the term "evidence" and "evidential justification" in a more general way, covering both propositional and non-propositional justifiers. Again, nothing substantial hinges on this terminological difference.

(N) If *S* knows that *p*, then: if *p* were false, *S* would not mistakenly believe that *p*.

My belief that ¬*SH* clearly does not satisfy (N). So, given (N), I do not know ¬*SH* (see (c2)). But Brueckner takes this strategy for establishing (c2) to be useless to the sceptic, because by appealing to the sensitivity account the sceptic would undermine her justification for premise (c1), which rests on the closure principle (CP). After all, (CP) is false if (N) is true (Brueckner 2005, 388). But this thought is only correct as long as we think sensitivity to be *necessary and* – possibly together with other conditions uncontroversially met by the beliefs in question – *sufficient* for knowledge.[5] However, I see no reason why a sceptic should be committed to that. Thus a sceptic who takes sensitivity to be a necessary but not sufficient condition for knowledge might very well establish (c2) via the sensitivity principle (N) without being incoherent.

Maybe there is a better reason why the sceptic should not motivate (c2) via sensitivity: The sensitivity account of knowledge is simply wrong – as many counterexamples show, sensitivity is neither necessary nor sufficient for knowledge. However, even if we accepted that the sceptic cannot defend (c2) via (N), we would not be forced to accept Brueckner's conclusion that the sceptic has to appeal to the underdetermination argument in order to motivate (c2). There might be other ways for the sceptic to defend the second premise of the closure argument. Here is a way to justify premise (c2) that is independent of the underdetermination principle and *prima facie* appealing.

Assume that justification is necessary for knowledge (see the justification principle (JP)) and that basing a belief on evidence is necessary for justification (see the evidence principle (EP)). Now the sceptic can defend (c2) by showing that ¬*SH* cannot be justified, neither by empirical nor non-empirical evidence.

Why is ¬*SH* not justifiable by non-empirical evidence? The proposition ¬*SH* concerns the position of an epistemic subject in the world. And to justify such a proposition you need empirical evidence – you have to take a look at the world to locate your position in it. Thus we might say: ¬*SH* cannot be justified by non-empirical evidence and therefore we cannot justify ¬*SH* a priori.

Why is ¬*SH* not justifiable by empirical evidence? We acquire empirical evidence as a result of an empirical procedure. And this kind of evidence cannot rationally be regarded as any stronger than one's independent reason for supposing that the procedure in question has been executed properly. Therefore, the evidence for the proposition ¬*SH, that I am not a brain-in-a-vat with hallucinated experiences*, cannot rationally be regarded as stronger than my independent reason for thinking that the relevant procedure has been executed properly, hence that it has been executed at all and not just hallucinated to be executed. But that means that justifying ¬*SH* by empirical evidence already requires that I

5 That in effect is Nozick's view, which is known as the "sensitivity account of knowledge" (Nozick 1981, 167-188).

have justification for the exact same proposition (see Wright 2004, 168). Thus we might say: ¬*SH* cannot be justified by empirical evidence (for the first time) and therefore we cannot justify ¬*SH* a posteriori.

The considerations that make (c2) plausible can be summed up thus:[6]

(1) All evidence is either empirical or non-empirical.
(2) *S* is not justified in believing ¬*SH* by non-empirical evidence.
(3) *S* is not justified in believing ¬*SH* by empirical evidence.
(4) If *S* is justified in believing ¬*SH*, then ¬*SH* is justified by evidence. [based on (EP)]
(5) Therefore, *S* is not justified in believing ¬*SH*. [from (1), (2), (3), (4)]
(6) If *S* knows ¬*SH*, then *S* is justified in believing ¬*SH*. [based on (JP)]
(c2) Hence, *S* does not know ¬*SH*. [from (5), (6)]

Even though this line of thought is *prima facie* appealing, it is surely not uncontroversial.[7] But I did not want to suggest that we should endorse this reasoning. The crucial point is that we have given an argument to defend (c2), which (i) does not refer to the underdetermination principle, (ii) is not utterly implausible and (iii) has been used by some philosophers to argue for the sceptical premise (c2). Thus we can conclude: Brueckner's claim A is false and the closure argument is not superfluous in motivating Cartesian scepticism.

Notice that the given reasoning in favour of (c2) forces the sceptic to accept the two additional principles on which the underdetermination argument were based: the justification principle (JP) and the evidence principle (EP).

4 Brueckner's Claim B

Brueckner's Claim B is that the underdetermination argument and with it the closure argument are based on infallibilism: In order to motivate the premises of the arguments the sceptic has to refer to an infallibility principle. Brueckner motivates claim B by showing that the underdetermination argument, when used to argue for the second premise (c2) of the closure argument, presupposes infallibilism. And via claim A he infers that the closure argument in general is based on infallibilism. If this reasoning were correct, then the sceptic would

6 See Weatherson (2007) for a related argument.

7 Some semantic externalists and defenders of transcendental arguments of other kind would not accept (2). Some epistemic externalists with regard to justification and dogmatists would not accept (3). Other epistemic externalists and defenders of the view that there is something like *warrant for nothing* or *justification by default* would not accept (4). And epistemic externalists with regard to knowledge would not accept (6). The fact that all the important antisceptical strategies in the literature are so easily mapped on the argument speaks in favour of my view, that something like this argument lies at the heart of the sceptics motivation of (c2).

need infallibilism to motivate her sceptical line of thought. And since fallibilism about knowledge and justification is a widely held view in epistemology, many philosophers were right in not taking Cartesian scepticism seriously. But we have already seen that claim A is false, and this is the reason why the second step of Brueckner's consideration is defective. The closure argument is not based on infallibilism because our motivation of (c2) was neither based on the underdetermination nor on the infallibility principle.

Before we turn to the relation of the underdetermination argument and infallibilism let me make one further remark. It might be true that the closure argument along with our motivation for (c2) entails the infallibility principle (INF). But of course this does not mean that a fallibilist is able to sidestep sceptical worries. The fact that the sceptical premises can be used to argue for implausible theses was clear all along. After all, the sceptical premises entail that we know almost nothing about the external world. That the *prima facie* plausible premises of the closure argument also entail the very implausible infallibilty principle should be even more worriesome for the fallibilist – it is surely not a free ticket to ignore sceptical arguments. As long as the sceptic does not use the infallibility principle in order to motivate her premises, the closure argument raises a serious problem even (or especially) for fallibilists.

But what about the underdetermination argument and its relation to infallibilism? As mentioned before, Brueckner only shows that the underdetermination argument relies on infallibilism, when used as a motivation for (c2) (Brueckner 2005, 388-390). But his line of thought can easily be reconstructed to argue that the underdetermination argument in general is based on infallibilism.

Underdetermination Argument

(u1) If *S's* evidence for believing *p* does not favour *p* over *SH*, then *S's* evidence does not justify *S* in believing *p*. [based on (UP)]
(u2) *S's* evidence for believing *p* does not favour *p* over *SH*.
(u3) Therefore, *S* lacks justification for believing *p*. [from (u1), (u2), (EP)]
(u4) Hence, *S* does not know *p*. [from (u3), (JP)]

The crucial premise in our context is (u2). Brueckner tries to argue that in order to establish (u2), we finally have to refer to the infallibility principle. The first step in the envisaged motivation for (u2) is what Brueckner calls the "sameness of evidence lemma":

(SEL) One has exactly the same evidence in the good case and in the bad case. [8]

8 Brueckner eventually speaks of the "sameness of *justifier* lemma" (SJL), because he reserves the term " evidence" for justifiers with propositional content. I will use the term "evidence" to cover both propositional and non-propositional justifiers (see fn:4).

The good case is that in which p is true and *SH* is false and the bad case is that in which *SH* is true and *p* false. Brueckner accepts that it is reasonable to think that (u2) follows from (SEL) (Brueckner 2005, 389).[9] As far as I can see (SEL) is plausible. As long as we think of our evidence for *p* as our perceptual evidence, and do not take a disjunctivist view on perception or Timothy Williamson's (2000) view on evidence, (SEL) seems to follow from the description of *SH* alone.[10] For example, take our perceptual evidence to be our seemings. Then my evidence for the belief that there is a table in front of me is that it seems to me that there is a table in front of me. And of course it would seem to me that there is a table in front of me if *SH* were realized – this is how the sceptical hypothesis is designed. So (SEL) follows from the description of *SH* along with plausible views on perceptual evidence.

But Brueckner thinks that in espousing (SEL), the sceptic is calling attention to the (alleged) fact that it is possible that my evidence *E* for *p* should be present when *SH* is true – in other words, the proposition that I have evidence *E* for *p* is consistent with *SH* and, concomitantly, with the denial of *p* – in other words, my evidence *E* for *p* fails to guarentee *p*. Thus in espousing (SEL) the sceptic is calling attention to the fact that my evidence *E* does not guarentee *p* (Brueckner 2005, 390). Brueckner takes the upshot of this observation to be: The sceptic's strategy of using (UP), (u2) and (SEL) to argue for (u3) (that *S* lacks justification for *p*) is simply a use of the infallibility principle (INF) in arguing for (u3) (Brueckner 2005, 390).[11]

---

9 Note that Brueckner refers to the second premise of the underdetermination argument (u2) with "~F".

10 Of course other accounts of perceptual evidence that would be incompatible with (SEL) are logically possible. But the most plausible and recently defended views of this sort are disjunctivism and Williamson's view on evidence. On a disjunctivist view the intrinsic nature of a perceptual state is determined by the perceived object. This is why a disjunctivist thinks that the perceptual states that I have when perceiving a fish and when merely halluzinating a fish are not tokens of a single perceptual-state-type, even though we cannot tell the difference from a first persons perspective. Therefore they think that our perception in the good p-case and in the bad SH-case are very different and hence that we cannot have the same evidence in the two cases. Timothy Williamson (2000) on the other hand holds the view that the body of our evidence is the body of our knowledge. As long as you think that the body of knowledge of a peson in a situation where p is realized differs from the body of knowledge of a person in a situation where SH is realized, it follows that we cannot have the same evidence in the good p-case and in the bad SH-case. But of course both of the mentioned views are very controversial.

11 Here is a quote of the relevant passage: "In espousing SJL, the skeptic is calling attention to the (alleged) fact that it is possible that my putative experiential justifier for ~SK should be present when SK is true. In other words, the proposition that I have the putative experiential justifier for ~SK is consistent with SK and, concomitantly, with the denial of ~SK. In other words, the proposition that I have the putative experiential justifier for ~SK fails to entail ~SK. Sound familiar? The upshot is that the skeptic's strategy of using SJL, ~F,

It is not easy to understand Brueckner's line of thought. I do not really see why Brueckner's reconstruction of the sceptic's strategy would amount simply to a use of the infallibility principle (INF). The principle (INF) just did not arise in his reconstruction. But instead of speculating what exactly Brueckner has in mind, I will do the following. According to Brueckner, the sceptic's reliance on infallibilism is somehow due to the motivation of (u2) via (SEL) and the interrelation of (UP) and (SEL). So in order to sidestep Brueckner's worries altogether, we should find a way to motivate the crucial second premise (u2) of the underdetermination argument without referring to the sameness of evidence lemma (SEL). I will present two promising principles that can be used for that purpose.

The first principle is the explanation principle:

(EXP) For all *S*, *p*, *q*, if *q* causally explains *S's* evidence *E* (at least as good as *p* does), then *E* cannot favour *p* over the incompatible alternative *q*.

The explanation principle (EXP) is at least *prima facie* very plausible. The principle is much weaker than (INF) and it allows that we can evidentially favour some possibilities over others by induction. But it is strong enough to establish (u2):

(1*) If *SH* causally explains *S's* evidence *E* (at least as good as *p* does), then *E* cannot favour *p* over *SH*. [based on (EXP)]

(2*) *SH* causally explains *E* (at least as good as *p* does).

(u2) Therefore, *S's* evidence cannot favour *p* over *SH*.

Of course assumption (2*) is controversial. But as many failed antisceptical attempts based on the inference to the best explanation illustrate, it is very hard to argue for the view that the explanation of our experience by the sceptical hypotheses is worse than our standard explanation.[12]

The second principle I want to present is the entailment principle:

(ENT) For all *S*, *p*, *q*, if *q* entails the proposition that *S* has evidence *E*, whereas the incompatible alternative *p* does not entail the proposition that *S* has evidence *E*, then *E* cannot favour *p* over *q*.

In my view principle (ENT) seems even more plausible than (EXP). Let *S's* evidence *E* consist in *S's* many perceptual experiences of white swans, and let this evidence *E* be the only evidence she has got for the beliefs in the following

---

and UP to establish his premise 2 is simply a use of the entailment principle JEP* to show a lack of justification for ~SK." (Brueckner 2005, 390) For an explanation of the terminological differences see fn. 4, 8, 9, and keep in mind that I reconstructed Brue-ckner's thought to argue that the underdetermination argument in general is based on infallibilism.

12 For an interesting antisceptical attempt to block the sceptical argument via an inference to the best explanation see Vogel (1990).

propositions. Let *w* stand for the proposition *that most swans are white* and *b* for the proposition *that most swans are black*. Neither *w* nor *b* entail the proposition that *S* has evidence *E*. Therefore, the principle (ENT) allows that *S's* evidence can favour *w* over the incompatible alternative *b*. Thus the principle is weak enough to allow that we evidentially favour some beliefs over others by induction. Now let *b** stand for the proposition *that most swans are black but S only perceived white ones.* It is intuitively correct to think that *S's* evidence *E*, which solely consists in her perception of white swans, cannot favour *w* over *b**. And this is exactly what principle (ENT) predicts: *b** entails that *S* has evidence *E*, whereas *w* does not, hence *S's* evidence *E* cannot favour *w* over the incompatible alternative *b**.[13] Now, by accepting (ENT) it is easy to establish (u2):

(1**) If *SH* entails the proposition that *S* has evidence *E*, whereas *p* does not entail the proposition that *S* has evidence *E*, then *E* cannot favour *p* over *SH*. [based on (ENT)]

(2**) *SH* entails the proposition that *S* has evidence *E* and *p* does not.

(u2) Therefore, *S's* evidence cannot favour *p* over *SH*.

Of course (2**) is controversial. But again, as long as we think of our evidence *E* as our perceptual evidence and do not take a disjunctivist view on perception or Williamson's view on evidence, (2**) looks very plausible as well (see fn. 9).

Note that I do not mean to suggest that we should endorse the arguments that establish (u2) via (EXP) and (ENT). Even though I take the arguments to be at least *prima facie* very plausible, both arguments are surely controversial. The crucial point is that we have given two arguments to establish (u2) with the follwing characteristics. First, the arguments are at least as plausible as Brueckner's motivation for (u2) via (SEL). Second, the arguments are not referring to the sameness of evidence lemma (SEL), which in Brueckner's view is somehow responsible for the sceptic's dependence on infallibilism in motivating (u2). Third, the arguments in favour of (u2) we have presented are not relying on the infallibility principle – the principles (EXP) and (ENT) are weaker than an analogous infallibility principle and thus more plausible, but they are strong enough to motivate the second premise of the underdeter-mination argument (u2). Thus we can conclude: Brueckner's claim B is false – neither the closure nor the underdetermination argument is based on infalli-bilism. A fallibilist should accept that these arguments raise a serious a chall-enge that must be answered somehow.

Again, it might be true that the underdetermination argument together with our motivation of (u2) entail the infallibility principle (INF). But as with regard to the closure argument, this is surely not a free ticket for fallibilists to ignore the underdetermination argument. As long as the sceptic does not refer to (INF)

13 Remember that we only allowed *S's* perceptual experiences of white swans as *S's* evidence for *w* – no background assumptions whatsoever are in place as additional evidence.

in establishing her premises, the underdetermination argument raises a challenge even (or especially) for fallibilists.

5 Summary

Brueckner's claim A is that the closure argument rests on the underdetermination argument: In order to motivate the second premise (c2) of the closure argument the sceptic has to refer to the underdetermination argument. I have shown that claim A is false. I have given a plausible argument in favour of (c2) without referring to the underdetermination principle. Thus the closure argument is not superfluous in motivating Cartesian scepticism.

Brueckner's claim B is that the underdetermination argument and with it the closure argument rest on infallibilism: In order to establish their premises the sceptic has to refer to the infallibility principle (INF). My answer to claim A already shows that the second part of claim B is not true – the closure argument does not rest on infallibilism because our motivation of (c2) is neither relying on the underdetermination nor on the infallibility principle.

What about the underdeterminaton argument? A presupposition of Brueckner's reasoning why the underdetermination argument eventually rests on infallibilism is that the scpetic has to refer to the sameness of evidence lemma (SEL) in order to motivate the second premise of the underdetermination argument (u2). I have presented two ways to motivate (u2) without referring to (SEL) or the infallibility principle (INF). Therefore, the underdetermination argument does not rest on infallibilism, and even a fallibilist has to take the arguments to raise a serious challenge that must be answered somehow.[14]

References

Briesen, J.: "Reconsidering Closure, Underdetermination, and Infallibilism". Grazer Philosophische Studien, i.V.

Brueckner, A.: "The Structure of the Skeptical Argument". Philosophy and Phenomenological Research, 54, 1994. S.827-834

Brueckner, A.: "Fallibilism, Underdetermination, and Skepticism". Philosophy and Phenomenological Research, 71, 2005. S. 384-391

Cohen, S.: "Two Kinds of Skeptical Arguments", Philosophy and Phenomenological Research, 58, 1998. S. 143-159

14 Thanks to Anthony Brueckner, Martin Smith, Brian Weatherson, and Elia Zardini for helpful comments and discussions on the material.

Hawthorne, J.: "The Case for Closure". In: *Steup, M./Sosa, E.* (eds.): Contemporary Debates in Epistemology, Blackwell Publishing, Oxford, 2003

Nozick, R.: Philosophical Explanations. Harvard University Press, Cambridge, 1981

Pritchard, D.: "The Structure of Sceptical Arguments", The Philosophical Quarterly, 55, 2005. S. 37-52

Stroud, B.: The Significance of Philosophical Scepticism, Oxford University Press, Oxford, 1984

Vogel, J.: "Cartesian Skepticism and Inference to the best Explanation", Journal of Philosophy, 87, 1990. S. 658-666

Weatherson, B.: "The Bayesian and the Dogmatist", Proceedings of the Aristotelian Society, 107, 2007. S. 169-185

Williamson, T.: Knowledge and its Limits, Oxford University Press, Oxford, 2000

Wright, C.: "Warrant for Nothing: Notes on Epistemic Entitlement", Proceedings of the Aristotelian Society, Supplementary Volume 78, 2004. S.167–212

Four Ways to Gettierize

Wolfgang Freitag
Wolfgang.Freitag@uni-konstanz.de
University of Konstanz

Abstract/Zusammenfassung

Gettier problems arise when it is only by luck that a justified belief is true. Such a belief, though true in fact, might equally well have been false. Therefore, it is generally ruled out to be knowledge. Ruling out the Gettier problems means ruling out luck in a justified true belief. There are two ways in which *truth* might be a matter of luck, both exemplified by prominent Gettier examples. One is specific to the inferential cases, the other is specific to the noninferential cases of (putative) knowledge. My aim is to show that the Gettier problem is of a more general sort, not restricted to (two kinds of) accidental *truth*. Where there is a true belief, there is not only truth; there is also belief. As the truth may be due to pure luck, so may the belief. From this it is but a short step to recognizing that in the inferential case there are two further ways in which a justified true belief may fail to be knowledge: it is not knowledge if it is due to pure luck that the epistemic subject has a certain inferential belief, and it is not knowledge if the procedure by which the subject arrives at her inferential belief is only accidentally valid. My task is to depict the four different ways of Gettierizing a *prima facie* legitimate claim to knowledge. But to show what can go wrong is to show what needs to be right. Pointing out the four kinds of Gettier problem will also be a means to attain a further goal: that of revealing the structure of knowledge, and hence the structural constraints on any account thereof.

Gettier problems arise when it is only by luck that a justified belief is true. Such a belief, though true in fact, could equally well have been false. Therefore, it is generally ruled out to be knowledge. Ruling out Gettier-problem cases means ruling out luck in a justified true belief. To rule out luck is to make knowledge a kind of necessity and hence to give a modal account of knowledge. My primary task here is not to *solve* the Gettier problem, and therefore I am not so much concerned with the kind of necessity involved in knowledge. This paper rather *uses* the Gettier problem to determine the structure of knowledge.

It will be argued that there is not only one way to Gettierize a justified true belief, but that there are *four* such ways. There are two ways in which *truth* might be a matter of luck, one specific to the noninferential case, the other to the inferential case. These two forms of alethic Gettierization categorize existing Gettier examples. My main thesis is that the Gettier problem is of a more general sort, not restricted to (the two kinds of) accidental *truth*. Where there is a true belief, there is not only truth, there is also belief. As the truth may be due to pure luck, so may the belief. From this it is but a short step to recognizing that in

the inferential case there are two further ways in which a justified true belief may fail to be knowledge: it is not knowledge if it is due to pure luck that the epistemic subject has a certain noninferential belief, and it is not knowledge if the procedure by which the subject arrives at her noninferential belief is only accidentally successful.

My task is to depict the four different ways of Gettierizing a *prima facie* legitimate claim to knowledge of contingent propositions.[1] But to show what can go wrong is also to show what needs to be right. Pointing out the four ways in which the Gettier problem may arise will also be a means to attain my ultimate goal: that of revealing the structure of knowledge, and hence the structural constraints on any account thereof.

The paper has four sections. In the first, I discuss the Gettier problem as the central problem for a theory of knowledge and briefly present my own favourite strategy of handling it. The second section concerns the two *alethic* Gettierizations, i.e., those cases in which a belief is *true* only by luck. Section 3 presents the two forms of *doxastic* Gettierization. Section 4 provides a definition of (inferential) knowledge based on the previous considerations.

1 The Gettier problem and the modal account of knowledge

Historically, the Gettier problem was directed against the tripartite definition of knowledge as justified true belief, hence against

(JTB) K(*S*, *p*) if and only if
  (1) *B(S, p)*,
  (2) *p* is true, and
  (3) *S*'s belief that *p* is justified, i.e., *S* has sufficient grounds to believe that *p*.[2]

Gettier (1963) argued in his paper that there are cases of justified true belief which are not knowledge, and that therefore the three conditions are not jointly sufficient for knowledge. Gettier argues by way of example. He describes two possible cases in which conditions (1) to (3) are fulfilled, but which we would not count as knowledge. I will introduce an example that is similar to the ones given by Gettier, with the purpose of laying bare the essence of the Gettier problem. Let me start with a basic scenario, subsequently to be modified in different ways:

1 In general, my discussion is restricted to (putative) knowledge of *contingent* propositions. An account of knowledge of necessary propositions poses problems of its own and will therefore not be attempted here.

2 Here and later, 'K(*S*, *p*)' and 'B(*S*, *p*)' stand for '*S* knows that *p*' and '*S* believes that *p*' respectively.

BASIC: The chemist Che carries out the litmus test with a certain sample of a fluid. The litmus paper turns red (*q*) and Che has the belief that the litmus paper turns red; Che believes that *q*. Che also believes that the tested fluid is acid (*p*). The proposition *p* is indeed true.

BASIC represents a case of a justified true belief and appears to describe a case of knowledge. But consider the following specification:

Addition 1: Yet Che's belief that the litmus paper turns red is due to a hallucination, which by luck occurs while the litmus paper indeed turns red.[3]

Addition 1 to BASIC results in what I refer to as *CASE 1*. (In general, BASIC+Addition X will be referred to as CASE X.) In CASE 1, Che does not *know* that the tested fluid is acid. Che's belief that the fluid is acid is true, but it could equally well have been false. The accident of bad luck – Che's belief is based on a hallucination – is balanced out by an accident of good luck – the hallucination is veridical and the resulting belief is true. Because of the luckiness of the truth of the belief on which the target belief is based, Che's belief that the fluid is acid is not knowledge.[4]

Gettier has shown that knowledge must not be based on lucky truth. Only belief which is non-accidentally true (in the right way) can be knowledge.[5] But to avoid any form of accidentality, the truth must be necessary in some sense. Knowledge is not only factive; knowledge is, to coin a phrase, *necessitative*. Modal epistemology takes knowledge to be based on some kind of modality. In the noninferential case, *S*'s belief that *p* (B(*S*, *p*)) is knowledge if and only if (roughly) 'B(*S*, *p*) → *p*' (the *K-conditional*) is true in all relevant possible worlds. The important task is then to determine the set of relevant possible worlds, the *K-set* $\mathfrak{K}$. Let $\mathfrak{a}$ be the actual world. I take it that a (preliminary) account of noninferential knowledge has the form

3 A structurally similar example is found already in Lehrer and Paxson 1969, pp. 234–235.

4 The core of the Gettier problem is truth by luck, as described in the main text. That the problem can be used, and usually has been used, to undermine the tripartite definition is owed to a conception of justification according to which it is not truth-entailing. As Gettier explains, "in that sense of 'justified' in which *S*'s being justified in believing that *P* is a necessary condition of *S*'s knowing that *P*, it is possible for a person to be justified in believing a proposition that is in fact false" (1963, p. 121). I contend that Gettier's assumption is very plausible for most notions of justification, but that this should not blind us to the fact that the Gettier examples do not unconditionally show that (JTB) is inadequate. If the assumption that justification is not truth-entailing is denied, one is free to escape Gettier's conclusion that (JTB) is an inadequate analysis of knowledge.

5 The analysis of the Gettier examples as being based on luck has in its essentials already been proposed by Kirkham 1984, Zagzebski 1994, Heller 1999 and Pritchard 2005. As early as 1968, Peter Unger suggested that knowledge is belief which is not at all accidentally true.

(MBT) K(S, p) iff
(1) $\mathfrak{a} \models B(S, p)$
(G1) $\forall x \in \mathfrak{K}: x \models B(S, p) \rightarrow p$
(G2) $\mathfrak{a} \in \mathfrak{K}$.

Note that no separate truth condition is needed. That p is true is entailed by (1), (G1), and (G2). If the conjunction of (G1) and (G2) is fulfilled, I will say that B(S, p) *guarantees p with respect to* $\mathfrak{K}$. (In general, I will say that *A guarantees B with respect to* $\mathfrak{K}$, or *guarantees*$_{\mathfrak{K}}$, if and only if (i) $\forall x \in \mathfrak{K}: x \models A \rightarrow B$, and (ii) $\mathfrak{a} \in \mathfrak{K}$. Conditions (i) and (ii) constitute the first and second guarantee condition, respectively.) If context excludes possible misunderstandings, I speak of guarantee *simpliciter*. Given these stipulations, we can give the following definition of noninferential knowledge:

(D1) A belief is noninferential knowledge if and only if it guarantees$_{\mathfrak{K}}$ its truth.

This definitional scheme covers a wide range of possible modal analyses, given a suitable determination of the K-set. Two extreme possibilities suggest themselves. (i) The K-set comprises *all* possible worlds. This is the most straightforward way of fleshing out the idea that knowledge is not only factive but necessitative; it unconditionally necessitates the truth of its content. Such a determination would rule out Gettier cases of the sort described as cases of knowledge, but obviously such an account is not tenable, because it would make knowledge of contingent propositions impossible.[6] To avoid scepticism, the K-set must be a proper subset of the set of all possible worlds. (ii) The other extreme would be to claim that $\mathfrak{K}=\{\mathfrak{a}\}$, i.e., that the K-set comprises only the actual world. In this way knowledge is not a rare commodity anymore, but rather too abundant. Any true belief would qualify as knowledge, a position which I consider to be (*pace* Crispin Sartwell (1991, 1992)) untenable.

More plausible approaches are sensitivity accounts (Nozick 1981; cf. Dretske 1971) and safety theories (Pritchard 2005, Sosa 2000: 14, and 2007, and Williamson 2000[7]). There are important differences between the two, but these need not interest us here. Roughly speaking, these approaches understand the K-set to be the set of those possible worlds that are closest or nearest to the actual world. Let 'R' denote this relation and let $\mathfrak{K}_{R\mathfrak{a}}=\{x \mid R(x, \mathfrak{a})\}$. Safety and sensitivity conditions propose, very roughly, the following definition:

6 Or rather it would mean that knowledge is impossible for *almost* all contingent propositions. Knowledge of propositions which are the conditions of the possibility of B(S, p) would still be possible.

7 Let me add that Williamson is not, of course, concerned with an *analysis* of knowledge, since he considers an analysis hopeless. Nevertheless, he provides something like a safety condition.

($MBT_{R\alpha}$) K(*S*, *p*) iff
(1) $\alpha \vDash B(S, p)$
($G1_{R\alpha}$) $\forall x \in \mathfrak{K}_{R\alpha}: x \vDash B(S, p) \rightarrow p$
($G2_{R\alpha}$) $\alpha \in \mathfrak{K}_{R\alpha}$.

What unites the safety and the sensitivity accounts is that the set of relevant possible worlds is determined by way of closeness to the actual world.[8] Both are based on the idea that the question of whether the true belief that *p* is knowledge is decided by reference to the truth of the material conditional 'B(*S*, *p*)→*p*' in possible worlds which are R-related to the actual world. Since the K-set is determined as the set of those worlds that stand in relation R to α, and, we may assume, the relation R is reflexive, the actual world α is automatically an element of $\mathfrak{K}_{R\alpha}$. Condition ($G2_{R\alpha}$) is therefore true by the very definition of the relevant set.

The accounts discussed so far try to cope with the Gettier problem by reference to the first guarantee condition, since the second is automatically fulfilled. Gettier cases are ruled out to be knowledge by condition ($G1_{R\alpha}$); here is a close, non-actual possible world, in which Che has the false belief that the litmus paper turns red. But why should we think that the Gettier problem arises from a fault in *other* possible worlds? Isn't something wrong with the *actual* world, e.g., as suggested by the fact that the litmus paper is so manipulated that it does not function as it is expected to function? Perhaps we should determine the K-set in such a way that the Gettier effect arises from the fact that the actual world is not an element of it. The closeness-accounts determine the K-set by way of closeness to the actual world. To determine an alternative, let us reflect on CASE 1. That Che's belief that *q* is not based on proper perception but on hallucination is the source of the trouble. This comes as quite a surprise: the expected conditions, the standard or normal conditions, are not present. As a first stab, let's therefore assume that the K-set is the set of normal possible worlds.[9] There would be a lot to say about the notion of normal conditions involved, but for reasons of space, I will refine the basic idea in only one dimension here. I suggest that the world need not be normal with respect to all of its aspects; it suffices that the actual world be normal in all *relevant* respects. What is relevant depends on the circumstances, in particular also on the *relata* of the guarantee

8 They differ, of course, in the exact content of the R-relation.

9 Normality as the background condition for the possibility of knowledge has been suggested in several places in the epistemological discussion. To cite only a few of these, we find them in Stine 1976, Heller 1999 and Sosa 2007. As far as I know, there is no well-worked-out theory of normal conditions as relevant for knowledge. Sometimes it is suggested that *ceteris paribus* clauses refer to normal conditions, but to provide an analysis in terms of *ceteris paribus* clauses would get us only from the frying pan into the fire.

condition.[10] This leads to the following definition of noninferential knowledge. Let $\mathfrak{K}_{\mathfrak{N}1}$ be the set of relevantly normal worlds:

(MBT$_{\mathfrak{N}1}$) K(S, p) iff
(1) $@\models$B(S, p)
(G1$_{\mathfrak{N}1}$) $\forall x \in \mathfrak{K}_{\mathfrak{N}1}$: $x \models$B(S, p) $\rightarrow p$
(G2$_{\mathfrak{N}1}$) $@ \in \mathfrak{K}_{\mathfrak{N}1}$.

Given (G1$_{\mathfrak{N}1}$) and (G2$_{\mathfrak{N}1}$), B(S, p) guarantees p with respect to $\mathfrak{K}_{\mathfrak{N}1}$.

2 Two types of *alethic* Gettierization

Endowed with this definition of noninferential knowledge, we can explain what has gone wrong in CASE 1. Che's belief that the litmus paper has turned red necessitates that the litmus paper has turned red, assuming relevantly normal circumstances. The first guarantee$_{\mathfrak{N}1}$ condition is fulfilled, whence the whole scenario possibly represents a case of knowledge in BASIC. Where things go wrong is when Che's basic belief is not due to proper perception but due to a hallucination; the conditions here are not normal in the relevant way. Therefore the second guarantee condition (G2$_{\mathfrak{N}1}$) is violated. The specific problem concerns the relation between the belief and the truth of its content. Because of this, I will speak of the Gettierization of the *belief–truth* (*B–T*) relation. Many of the known Gettier examples are of the B–T type. Russell's clock case is of this type, at least under the standard interpretations, in which the belief that it is 3 pm is a basic belief, i.e., a belief not based on another belief. Another is Goldman's example of the bogus barns, which roughly runs as follows:[11]

> Barney travels through a landscape which is replete with what look like barns but are, unbeknownst to Barney, in fact only barn façades, perhaps erected for a movie to be shot at this location. Suppose that amongst these many bogus barns, there is one real barn. Barney, looking at the real barn, correctly believes that this thing is a barn.

Conditions are such that Barney's belief could equally well have been false. It is true only by luck. Assuming that Barney's barn belief is a perceptual belief not resting on any other belief, we must understand this as a case of B–T Gettierization.[12]

B–T Gettierization is not the only way in which the truth of p may be a matter of mere luck. I will now discuss inferential knowledge, and this allows for three more ways of Gettierization. In all of the cases cited, the actual world is

10 There is also room for attributer contextualism here, a possibility not explored in this essay.
11 See Goldman 1976. Yet another example would be Chisholm's sheep example.
12 Many accounts of knowledge, e.g., those of Dretske (1971) and Nozick (1981), can be seen to be responses to Gettier problems of the B–T type, primarily.

not normal in the relevant respect demanded by the pertinent K-set; the second guarantee condition is never fulfilled. Thus, (MBT), though providing the right scheme for noninferential knowledge, must be developed further to also cover inferential knowledge. It should be noted that the relevant normality required is different in each of the four cases of Gettierization; the set of relevantly normal worlds will be different in each case. Note also that in what is to follow, my specific account of the K-set in terms of relevantly normal conditions, and hence my specific account of the fault in the Gettier cases, is not needed to see these forms of Gettierizations. The different types of Gettier problems are independent of the exact modal account of the necessity involved. I use my own account of guarantee because I find it intuitive and for the sake of definiteness, but other modal accounts would do the job equally well.

To see the second type of alethic Gettierization, consider the following addition to BASIC:

> Addition 2: Yet the litmus paper is covered with a coat that turns red when put in any liquid.

Suppose that Che does not by hallucination arrive at the belief that the litmus paper turned red, but by proper perception, and suppose that all other relevant circumstances are normal. Then there is B–T guarantee. Still the resulting scenario does not describe a case of knowledge. CASE 2 is so constructed that the litmus paper's turning red does not guarantee the acidity of the fluid (with respect to the set of normal possible worlds relevant in this case, $\mathfrak{K}_{\mathfrak{N}2}$). The conditional 'The litmus paper turns red (q) → the fluid is acid (p)' is true in all relevant normal possible worlds, but again the actual world is not normal. The second guarantee condition for q and p is violated and therefore the guarantee relation between the truth of q and the truth of p is undermined. This Gettierization concerns the relation between the *truths* of the belief contents, whence I will speak of CASE 2 as representing the *truth–truth* (*T–T*) case. CASE 2 is a T–T-Gettierized version of BASIC. In order to obtain inferential knowledge, one must have T–T guarantee with respect to the set of normal worlds relevant to the case.

Gettier's original examples all represent T–T Gettierizations. Consider only his first scenario:

> Smith and Jones have applied for a certain job and Smith has strong evidence for the following conjunctive proposition (assume that Smith remembers that the president of the company assured him that Jones would get the job and further that he, Smith, had counted the coins in Jones's pocket just now):
>
> (p) Jones is the man who will get the job, and Jones has ten coins in his pocket.

Proposition *p* entails:

(*p′*) The man who will get the job has ten coins in his pocket.
Assume that Smith sees the entailment from *p* to *p′* and accepts *p′* on the grounds of *p*. But assume, further, that unknown to Smith, he himself, not Jones, will get the job and he himself has, as a matter of pure accident, ten coins in his pocket. Proposition *p′* is true. (Adapted from Gettier 1963, p. 122)

Smith has three beliefs which are directly relevant here. The first belief is that there is good testimonial/perceptual evidence for the proposition *p*, the second belief is that *p*, the third is that *p′*. The belief that there is testimonial/perceptual evidence for *p* is described as being 'remembered'. Therefore we safely assume it to constitute knowledge. Proposition *p′* follows from *p* deductively, thus there is no possibility of luck stepping in the relation between the truth of *p* and the truth of *p′*. Consider, however, the relation between the proposition *There is good testimonial/perceptual evidence for p*, and *p* itself. Under normal conditions, the president's testimony that Jones will get the desired job entails that Jones does get the desired job – otherwise no knowledge could ever arise from hearing such news from the president's mouth. The T–T conditional important here, '(The president said that *Jones gets the job*) → *Jones gets the job*', is true in all normal worlds relevant to the case. But the world is not normal, as indicated by the fact that Jones does not get the job. The second guarantee condition for the crucial conditional is violated: though *p′* is true (a possibility engendered by *p′* being a logical weakening of the false *p*), it is true only by luck. Smith's belief that the man who will get the job has ten coins in his pockets is therefore not knowledge. The relation of guarantee Gettierized does not concern a belief and the truth of its contents, but rather the truth of one believed proposition (*There is good testimonial/perceptual evidence for p*) and the truth of another believed proposition (*p*). Since the second guarantee condition between two truths is undermined by this example, we are here concerned with T–T Gettierization. Analogous considerations apply to Gettier's second example.[13] Existing Gettier-examples may be classified as being of the T–T type or of the B–T type. Both types offer different ways in which the *truth* of *p* is not guaranteed by the

13 In Gettier's original case, the fact that the second guarantee condition is violated is not made explicit. It is only inferable from the fact that *p* – necessitated to be true under normal conditions – is not true. (It also remains undetermined what exactly is wrong with the actual world.) Because of this indirect way of pointing at the violation of the second guarantee condition, Gettier needs to involve a false intermediate belief. The same result could have been achieved by letting *p* be true, but making explicit that circumstances are relevantly abnormal, e.g., by adding 'the president sometimes mixes up the names of Jones and Smith', etc. In this case it would be clear that the fact that there is good testimonial/perceptual evidence for *p* does not guarantee the truth of *p*. Thus, the Gettier problem does not need to involve a false intermediate belief at all. It suffices to show that the situation is such that the belief in question is only true by luck and not by guarantee.

belief that *q*. The different ways arise from the difference between inferential and noninferential knowledge. In the noninferential case, there can only be B–T Gettierization; the inferential case allows for both B–T and T–T Gettierization.

3 Two ways of *doxastic* Gettierization

Alethic Gettierization, namely the failure of guaranteeing truth, has been in exclusive focus in epistemology. I will now provide examples intended to show that there are Gettier examples which do not concern the *truth* of a proposition. Suppose that no T–T and no B–T Gettierizations of BASIC obtain. Neither is Che the victim of a hallucination, nor is anything wrong with the litmus paper or with the chemical laws. Still the belief that *p* need not be knowledge. Consider CASE 3 resulting from BASIC with

> Addition 3: Che believes that the fluid is acid, not because he believes the litmus paper to be red, but because his astrologer has told him so.

In this case, Che's belief that *p* is justified and true, as usual.[14] Yet, again, it is not knowledge. Obviously, something has gone wrong with Che's *motivation* for believing. We have expected that Che holds the belief that the fluid is acid based on his belief that the litmus paper has turned red. But as it turns out, Che would have had the former belief even if he had not had the latter. The *belief* that *p* does not guarantee the *belief* that *q*. Under normal conditions, Che's belief that the fluid is acid would guarantee that he possesses a belief such that this further belief has B–T and T–T guarantee. However, conditions are abnormal. The belief that the fluid is acid is *not* based on another belief with the required types of guarantee, and therefore it is not knowledge. CASE 3 constitutes a *belief–belief* (*B–B*) Gettierization.

The most natural way of interpreting this case is to say that Che does not possess an adequate reason. That is not to say that Che does not have any reason at all. The reason for his belief that the fluid is acid is his belief in the astrologer's prediction. Let us call such a belief related to the target belief by B–B guarantee a *doxastic* reason. The belief that the astrologer made this prediction is Che's doxastic reason for believing that the fluid is acid; the belief that the litmus paper turned red is no doxastic reason. Yet only the latter, not the former, is a good

14 I assume in this example that a belief that *p* may be justified by a belief that *q*, without the believing that *p* strictly implying the believing that *q*. This requirement is analogous to the requirement necessary for T–T Gettierization, namely that a belief that *p* may be justified by a belief that *q*, without *q* strictly implying *p*. Whether my assumption is correct or not does not really matter. My concern is not with justification and its possible role in an analysis of knowledge, but the different ways in which epistemic luck undermines would-be knowledge.

or a *proper* reason. Only the belief that the litmus paper has turned red has B–T guarantee and confers T–T guarantee on the belief that the fluid is acid. The belief that the astrologer made the acidity prediction, though presumably endowed with B–T guarantee (we may assume that Che is not the victim of an auditory hallucination), nevertheless does not confer T–T guarantee. The proposition that the astrologer says so does not guarantee that the fluid really is acid. The prolem with CASE 3 is therefore not the lack of a doxastic reason, but the lack of a *proper* doxastic reason. Che's doxastic reason is not a proper reason, since it is unable to carry the epistemic burden placed on it; and the proper reason is not a doxastic reason, since it is not what grounds Che's target belief. If Che's doxastic reason were a proper reason, i.e., if the astrologer's saying so guaranteed that the fluid is acid, Che's belief would be knowledge. Conversely, if Che's proper reason were also a doxastic reason, i.e., if he based his belief that the fluid is acid on the belief that the litmus paper has turned red, he would know too that the fluid is acid. In general, if Che based the belief that *p* on a proper reason, Che's belief that *p* would be knowledge.

The B–B case brings the subject into play as a subject whose doxastic interrelations are relevant for the question of whether knowledge is present. Inferential knowledge demands what we might call *doxastic inference* on behalf of the subject. In proposing that B–B guarantee can constitute such a relation of inference, I must mention two things: Firstly, the inference need not be the content of a conscious process. As already Goldman (1967, pp. 360–361) says, "to say that *S* knows *p* by 'inference' does not entail that *S* went through an explicit, conscious process of reasoning."[15] Secondly, and perhaps more importantly, 'inference' as used in connection with inferential knowledge should not be taken to refer to the *genesis*, the coming about, of a belief. With 'inference' I wish to be understood as meaning something like 'legitimate inference', or 'epistemic support' of belief. A belief is inferred in this sense if it is held *on the basis of* another belief, a relation which is independent of actual aetiology.[16] 'Inference' is therefore meant to refer to a dependence relation, perhaps a relation of something like causal or justification sustenance.[17]

---

15 Goldman uses different ways of specifying these relations. In Goldman 1967 he at least tentatively proposes an explication in terms of causality, while later he favours reliable inference procedures, e.g., in Goldman 1986. Whether causation or reliable inference are suitable for evading the Gettier problem depends on their filling out the demands made explicit in the notion of guarantee.

16 Need it be based on another *belief*? – This is an interesting question when it comes to the ultimate basis of knowledge and justification. I have the tendency, but only a tendency, to think that also non-doxastic states – appearings and seemings – might fill that role.

17 I have also the hope that the B–B guarantee provides an alternative understanding of what is internalist in inferential knowledge, namely a subjective representation of the external guarantee relation. Let this short remark suffice as a signpost in the direction of a huge topic which I cannot go into within the confines of the present essay.

Why are there three possible targets for Gettierization? In the noninferential case, there is but one thing relevant; namely whether *S*'s holding *p* to be true guarantees *p*'s being true, i.e., whether B–T guarantee obtains. For the inferential case, we get *two* more cases, the classical T–T case and the new B–B case. Why do we get these two additional cases? A true belief is a believed proposition that is true. Yet a believed proposition that is true is a true proposition that is believed. Since both truth and belief are relevant for knowledge, the inference relation splits into two different relations; the inferred proposition might be only accidentally true, given the truth of its basis, and it may be only accidentally accompanied by a proper reason. Abbreviating '*S*'s believing that *x*' with 'B(*S*, *x*)', 'the truth of *x*' with 'T(*x*)', and representing the guarantee relation with the arrow symbol, we obtain the following suggestive picture for the case of inferential knowledge:[18]

$$\begin{array}{ccc} B(S, p) & & T(p) \\ \downarrow (B\text{–}B) & & \uparrow (T\text{–}T) \\ B(S, q) & \rightarrow & T(q) \\ & (B\text{–}T) & \end{array}$$

At this point it may appear that our analysis is complete, that all eventualities have been provided for, that there is nothing else that could stop us from giving a complete analysis of knowledge. But, unfortunately, this is not so. Goldman, again, provides a hint at what is still missing:

> Suppose our friend Humperdink has attended a series of talks on mathematics by Elmer Fraud. These talks are not under the auspices of any certified educational institution, and Humperdink has been warned that Fraud has no credentials in mathematics. Humperdink hears Fraud enunciate numerous principles and algorithms, almost all of them defective. Nonetheless, being a complete novice – and a gullible one at that – Humperdink blindly accepts and applies them all. In one case, however, Fraud happens to teach a perfectly correct algorithm. Humperdink internalizes this one along with the others, and applies it to a relevant class of problems. In using this algorithm to solve a problem, Humperdink gets the answer right and forms a true belief in the answer. This belief is the result of a reliable process, namely, the algorithm. […] Clearly, though, Humperdink should not be credited with knowledge. (Goldman 1986, pp. 51–52)

Goldman (1986, p. 52) concludes from this case that "[n]ot only must the belief result from a reliable process, or method, the process or method used must have been acquired (or sustained) by a suitable second-order process". As usual, Goldman emphasizes the genetic aspect of knowledge, an aspect which I hold to be rather insignificant. Nevertheless, we may make Goldman's example pertinent to our discussion if we replace the genetic reading by a guarantee reading.

18 This scheme presents the simplest model of inferential knowledge by assuming only two beliefs to be involved. Complications stemming from the fact that there is also knowledge involving 'mediating' beliefs or other 'basic' beliefs can be ignored for the purposes of our discussion.

The example can be understood as one in which the three ways to Gettierize discussed so far are avoided. To properly understand the systematic impact of Goldman's line of thought to the question of Gettierization, let us discuss a final variation of BASIC, a form displaying *second-order* Gettierization:

> Addition 4: Che has the belief that *p* because of the belief that *q*, but that this is so is not owed to chemical training of any sort. It is due to the wisdom obtained from his eccentric granny, who – unknowledgeable of any chemistry – taught Che this particular whim together with her ideas about weather change and angelic potency.

Again, Che would not be said to know that the fluid is acid. Che has the belief that *p* because of the belief that *q*, but the 'because' is due to his granny's (unreliable) teaching and not to chemical knowledge. By the same training, Che would also derive the belief, say, that exactly two angels can dance on the head of a pin from the belief that angels only dance the waltz. Given this particular instance of B–B guarantee$_{\mathfrak{KM}3}$ in CASE 4, T–T guarantee$_{\mathfrak{KM}2}$, though obtaining, could equally well have been missing. Under normal conditions relevant to the case, the fact that Che bases a chemical belief on another belief necessitates that the truth of the latter guarantees the truth of the former. Again, however, conditions are not normal, since Che, though a chemist, is a rather poor one. That a T–T guarantee corresponds to Che's B–B guarantee is merely accidental, a matter of pure luck. This last way of Gettierizing, call it *BB–TT*, can be understood as the inferential complement to the noninferential B–T Gettierization: not only must the psychological state of belief reflect the facts, the doxastic inference must furthermore reflect the guarantee relation present in the world. (Observe that in both cases the direction of the guarantee relation supports our realistic inclinations by making sure that the subjective side aligns with the world, i.e., that the propositional attitude and the B–B guarantee do guarantee their worldly counterparts.) BB–TT guarantee (with respect to the set of normal possible worlds relevant in this case, $\mathfrak{K}_{\mathfrak{M}4}$) makes a subject's doxastic inference epistemically legitimate. In this way it prevents that some 'weird' case of doxastic inference, correct only by luck, can lead to knowledge.

4 The structure of inferential knowledge

The four ways of Gettierizing indicate possible problems with putative cases of knowledge, and reveal, by the same token, the relations necessary for knowledge. Knowledge is based on four different guarantee relations, all of which must be explained by a solution to the Gettier problem. We are now in a position

to propose a refined account of the individually necessary and, I suppose, jointly sufficient conditions for inferential knowledge.[19]

(D2) *S*'s belief that *p* is a case of knowledge if and only if there is a *q* such that
(i) <*S* believes that *p*> guarantees$_{\mathfrak{RM}3}$ that <*S* believes that *q*>, and
(ii) <*S* believes that *q*> guarantees$_{\mathfrak{RM}1}$ <*q* is true>,
(iii) the truth of *q* guarantees$_{\mathfrak{RM}2}$ the truth of *p*,
(iv) condition (i) guarantees$_{\mathfrak{RM}4}$ condition (iii).[20]

The four conditions are meant to each rule out one type of Gettierization. Condition (i) is directed against B–B Gettierization, (ii) eliminates B–T Gettierization and (iii) prevents T–T Gettierization. Condition (iv), finally, blocks BB–TT Gettierization.

Note that there is no need for separately stating that *S* has a belief that *q*, or that *p* is true. Given the supposition that *S* believes that *p*, the former is entailed by (i). The latter is entailed by (ii) and (iii) and B(*S*, *q*). Of course, it is not a defect of (D2) that neither the belief- nor the truth-conditions need to be separately stated; it is a mere consequence of the constraints derived from the Gettier problem: this type of problem occurs only if there is a lack of guarantee, which allows for luck to step in. In general, definitions of knowledge in which the truth condition needs to be stated separately are inadequate. Knowledge is necessitative. The factivity of knowledge flows from its necessitativity.

(D2) allows for *p* and *q* to be identical. If they are, conditions (i) and (iii) are vacuously true and, therefore, (iv) also obtains. Thus, if *p* is identical to *q*, (D2) reduces to

(D3) *S*'s belief that *p* is a case of knowledge if and only if
(ii) <*S* believes that *p*> guarantees <*p* is true>.

(D3) is the definition of basic knowledge. Noninferential knowledge can therefore be understood as constituting a degenerate case of inferential knowledge. As I said, noninferential knowledge has often been the focus of modal epistemology, because it provides the basic problem; namely that of determining the

19 I have no proof that this picture is complete. I merely assume there are no other relevant relations than the four identified, until the opposite has been shown. Note parenthetically that, of course, there are more ways in which luck may play a role in the generation of putative knowledge. The detective might only by luck find the decisive piece of evidence, and thus only accidentally acquire the crucial belief. Or, more generally, a subject may have been in the right place at the right time only through a mere accident, etc. These kinds of luck do not concern epistemic luck, luck which is important for the question of knowledge. (For more on the distinction between epistemically relevant and irrelevant luck, see Pritchard 2005.)

20 Note that, given (i) and (iv), (iii) follows. Its explicit mention would therefore not be needed.

right K-set. I hope to have shown that also inferential knowledge receives a modal analysis.

References

Dretske, Fred: "Conclusive Reasons". Australasian Journal of Philosophy, 49, 1971. S. 1–22. In: Dretske, Fred: Perception, Knowledge and Belief. Cambridge University Press, Cambridge, 2000. S. 3–29

Gettier, Edmund: "Is Justified True Belief Knowledge?". Analysis, 23, 1963. S. 121–123

Goldman, Alvin: "A Causal Theory of Knowing". Journal of Philosophy, 64, 1967. S. 357–372

Goldman, Alvin: "Discrimination and Perceptual Knowledge". Journal of Philosophy, 73, 1976. S. 771–791

Goldman, Alvin: Epistemology and Cognition. Harvard University Press, Cambridge (MA)/London, 1986

Heller, Mark: "The Proper Role of Contextualism in an Anti-Luck Epistemology". Philosophical Perspectives, 13, 1999. S. 115–129

Kirkham, Richard L.: "Does the Gettier Problem Rest on a Mistake?". Mind, 93, 1984. S. 501–513

Lehrer, Keith/Paxson, Thomas: "Knowledge: Undefeated Justified True Belief". Journal of Philosophy, 66, 1969. S. 225–237

Nozick, Robert: Philosophical Explanations. Clarendon Press, Oxford, 1981

Pritchard, Duncan: Epistemic Luck. Oxford University Press, Oxford, 2005

Sartwell, Crispin: "Knowledge is Merely True Belief". American Philosophical Quarterly, 28, 1991. S. 157–165

Sartwell, Crispin: "Why Knowledge is Merely True Belief". Journal of Philosophy, 89, 1992. S. 167–180

Sosa, Ernest: "Scepticism and Contextualism". Philosophical Issues, 10, 2000. S. 1–18

Sosa, Ernest: A Virtue Epistemology. Apt Belief and Reflective Knowledge. Clarendon Press, Oxford, 2007

Unger, Peter: "An Analysis of Factual Knowledge". Journal of Philosophy, 65, 1968. S. 157–170

Williams, Michael: Unnatural Doubts. Epistemological Realism and the Basis of Scepticism. Blackwell, Oxford, 1991
Williamson, Timothy: Knowledge and Its Limits. Oxford University Press, Oxford, 2000

Zagzebski, Linda: "The Inescapability of Gettier Problems". Philosophical Quarterly, 44, 1994. S. 65–73

Die Natur und der epistemische Status von philosophischen Intuitionen

Martin Grajner
martin.grajner@gmx.de
Institut für Philosophie, Technische Universität Dresden

Abstract/Zusammenfassung

In this paper, I try to sketch an account of intuitions according to which intuitions are seemings. My paper consists of four sections. I the first section, I give an overview of the conceptions of intuitions that are endorsed in the literature and examine their interrelations. In the second section of the paper, I present an argument in favour of the view that intuitions are seemings. I argue that this view is backed up by certain linguistic data and that there exist linguistic data against rival views, according to which intuitions can be reduced to doxastic states or dispositions to doxastic states. In the third section of the paper, I try to motivate the claim that the content of an intuition is not modal. In the last section of the paper, I discuss whether some intuitions are a priori. First of all, I examine an argument due to Alvin Goldman that intuitions do not confer a priori warrant. I try to argue that Goldmans argument does not establish that intuitive warrant is not a priori. In the remainder of the section, I try to develop a positive view on what the apriority of an intuition might depend. Due to the fact that I do not endorse the view that rational intuitions are characterized by their modal content, the apriority of an intuition cannot be a matter of their content. I argue that the fact whether an intuition is a priori is rather a matter of their etiology.

In diesem Aufsatz will ich für ein Modell von Intuitionen plädieren, wonach Intuitionen Einstellungen oder Zustände eigener Art sind, die sich nicht auf doxastische Zustände oder Dispositionen zu doxastischen Zuständen reduzieren lassen. Mein Aufsatz besteht aus vier Abschnitten. Im ersten Abschnitt gebe ich zunächst einen Überblick über die unterschiedlichen Modelle von Intuitionen, die in der Literatur vertreten werden, und versuche kurz ihre Beziehungen zueinander zu klären. Im zweiten Abschnitt des Aufsatzes entwickele ich ein Argument für die von mir favorisierte Konzeption von Intuitionen. Es besteht darin, dass für das von mir favorisierte Konzept gewisse natursprachliche Daten existieren und dass alternative Modelle von Intuitionen nicht durch derartige Daten gestützt werden. Im dritten Abschnitt des Aufsatzes gehe ich auf die Frage ein, wie man Intuitionen näher charakterisieren sollte. Ich argumentiere contra Bealer, dass die Annahme nicht plausibel ist, dass der Gehalt von rationalen Intuitionen essentieller Weise modal ist. Im letzten Abschnitt des Aufsatzes gehe ich noch auf die Frage ein, ob manche Intuitionen a priori sind. Ich diskutiere zuerst ein Argument von Alvin Goldman, wonach Intuitionen keine erfahrungsunabhängige Rechtfertigung erzeugen. Ich versuche zu zeigen, dass Goldmans Argument nicht überzeugend ist. In dem restlichen Teil dieses Anschnittes versuche ich dann die Frage zu behandeln, woran die Apriorität einer Intuition liegen könnte. Da ich in diesem Aufsatz nicht die Auffassung akzeptiert habe, dass philosophische oder rationale Intuitionen essentieller Weise einen modalen Gehalt besitzen, kann die Erfahrungsunabhängigkeit einer Intuition nicht an ihrem Gehalt liegen. Ich

versuche demgegenüber für die Auffassung zu plädieren, dass die Apriorität einer Intuition an der Ätiologie liegt

Was die Philosophie von anderen wissenschaftlichen Disziplinen unterscheidet, ist der Rückgriff auf Intuitionen. Während in den Naturwissenschaften Intuitionen keine evidentielle Rolle spielen, sondern eher Beobachtungen, Laboruntersuchungen oder Experimente, scheint man in der Philosophie ohne Intuitionen nicht auskommen zu können. Obwohl der Rückgriff auf Intuitionen in der Philosophie sich bereits in der Antike beobachten lässt und in der Gegenwartsphilosophie sehr weit verbreitet ist, herrscht doch relative Uneinigkeit darüber, was Intuitionen genau sind. Sind Intuitionen Überzeugungen? Sind sie Urteile? Oder sind sie gar Zustände eigener Art? Die Frage danach, was Intuitionen sind, ist nicht nur für sich betrachtet von Interesse, sondern ist auch für die Frage relevant, ob man Intuitionen als verläßlich ansehen sollte oder nicht.[1]

In diesem Aufsatz möchte ich auf diese ganz grundlegende Frage eingehen und für ein Modell von Intuitionen plädieren, das Intuitionen als Einstellungen sui generis versteht, die sich nicht auf Überzeugungen, Urteile oder Dispositionen zu Überzeugungen reduzieren lassen. In der gegenwärtigen Diskussion wurde ein derartiges Modell vor allem von George Bealer populär gemacht.[2] In diesem Aufsatz werde ich zu zeigen versuchen, dass ein derartiges Modell durch gewisse natursprachliche Daten gestützt wird. Ich werde ferner zu zeigen versuchen, dass für alternative Konzeptionen, wonach Intuitionen auf andere doxastische Einstellungen oder auf Dispositionen zu doxastischen Einstellungen reduziert werden können, keine derartigen Daten existieren und dass diese Modelle entsprechend die Wahrheitsbedingungen von Sätzen, in denen das Prädikat „hat die Intuition, dass ...“ auftaucht, nicht korrekt bestimmen. Mein Aufsatz ist folgendermaßen aufgebaut. Im ersten Abschnitt werde ich zunächst unterschiedliche Konzepte von Intuitionen vorstellen, die in der Literatur vertreten werden. Ich werde dabei kurz darauf hinweisen, wie sich diese einzelnen Konzeptionen zueinander verhalten. Im zweiten Abschnitt werde ich die vorgestellten Konzepte diskutieren und mein Argument für das von mir favorisierte Modell entwickeln. Im dritten Abschnitt werde ich kurz darauf eingehen, wie man Intuitionen, wie sie von mir favorisiert werden, näher charakterisieren sollte. Im vierten und letzten Abschnitt werde ich auf die Frage eingehen, ob das von mir favorisierte Konzept nach sich zieht, dass zumindest manche Intuitionen so etwas wie

1 Die Vertreter der sog. experimentellen Philosophie gehen meistens von einem doxastischen Konzept von Intuitionen aus (Siehe etwa Weinberg, Nichols und Stich (2001)). Manche Autoren argumentieren dahingehend, dass ihr favorisiertes Konzept von Intuitionen nicht durch die empirischen Studien berührt wird (etwa Ludwig (2007)).

2 Siehe Bealer (2000). Es gibt zahlreiche Autoren, die sich Bealer angeschlossen haben. Etwa Weatherson (2003) oder Grundmann (2007).

erfahrungsunabhängige Gründe für philosophische Behauptungen darstellen. Ich werde zu argumentieren versuchen, dass es in der Tat so ist, dass manche Intuitionen a priori sind.

Wenn man die Frage beantworten möchte, was Intuitionen sind, muß man, ähnlich wie bei der Analyse anderer philosophischer Begriffe, die rechte Seite des folgenden Bikonditionals vervollständigen:

(I) S hat die Intuition, dass p $\leftrightarrow$ ...

Mit diesem Analysekonzept würde man einzeln notwendige und gemeinsam hinreichende Bedingungen dafür angeben, wann eine gegebene Person über eine Intuition verfügt. Man könnte die Frage nach dem, was Intuitionen sind, auch etwas spezifischer formulieren. Wenn man sich nämlich die Frage stellt, was Intuition sind, muss man zunächst näher bestimmen, welche *Art von Einstellung* eine Intuition gegenüber einem bestimmten Gehalt p ist. Man würde entsprechend zu dem folgenden Analyseschema gelangen:

(I*) S hat die Intuition, dass p $\leftrightarrow$ S hat die *Einstellung* φ gegenüber dem Gehalt, dass p.

Ich werde in der folgenden Diskussion von dem Analyseschema (I*) ausgehen. Es wird mir im Folgenden also primär darum gehen zu bestimmen, welche Art von Einstellung eine Intuition gegenüber einem Gehalt p ist. Wie die zusätzlichen, hinreichenden Bedingungen aussehen, damit tatsächlich eine Intuition vorliegt und nicht eine andere Einstellung der gleichen Art, die aber keine Intuition darstellt, soll in der Diskussion ausgeklammert werden.[3]

Welche Konzepte von Intuitionen gibt es nun in der Literatur? Eine prominente Konzeption von Intuitionen besteht darin, Intuitionen auf Überzeugungen zurückzuführen. Ein Autor, der einen derartigen Ansatz vertreten hat, ist David Lewis. Lewis schreibt:

> It might be otherwise, as if some philosophers seem to think, we had a sharp line between ‚linguistic intuition,' which must be taken as unchallengeable evidence, and philosophical theory, which must at all costs fit this evidence. If that were so, conclusive refutations would be dismantingly abundant. But whatever may be said for foundationalism in other subjects, this foundationalist theory of philosophical knowledge seems ill-founded in the extreme. Our "intuitions" are simply opinions; our philosophical theories are the same. Some are commonsensical, some sophisticated; some are particular, some are general; some are more firmly held, some less. But they are all opinions, and a reasonable goal for a philosopher is to bring them into equilibrium.[4]

Obwohl Lewis in diesem Zitat ganz allgemeine Bemerkungen zur Methode der Philosophie macht, legt er sich in diesem Zitat auf eine ganz spezifische Kon-

3 Ich werde im Verlauf des Textes auch das Bikonditional verwenden, obwohl man für den hier anvisierten Zweck auch ein Konditional verwenden könnte.

4 Lewis (1983), S. x.

zeption von Intuitionen fest, wonach Intuitionen nichts anderes als Überzeugungen sind:

(IÜ) S hat die Intuition, dass p ↔ S ist von p überzeugt.

Gegen diese Idee scheint natürlich auf den ersten Blick zu sprechen, dass nicht jede Überzeugung eine Intuition ist und man müsste (IÜ) entsprechend um zusätzliche Bedingungen oder Merkmale ergänzen, die eine Intuition von einer herkömmlichen Überzeugung unterscheiden. Im Folgenden will ich, wie eben bereits erwähnt, diese Fragestellung ausklammern.

Neben diesem Konzept von Intuitionen gibt es auch noch die Auffassung, dass Intuitionen Urteile sind. Kirk Ludwig ist ein Autor, der einen derartigen Ansatz vertritt.[5] Er identifiziert Intuitionen mit Urteilen, die alleine aus begrifflicher Kompetenz heraus gefällt werden, identifiziert. Ludwig schreibt:

> For terminological clarity, I will use „intuition" to mean an occurrent judgement formed solely on the basis of conceptual competence in the concepts involved in response to a question about a scenario, or simply an occurrent judgement formed solely on the basis of competence in the concepts involved in it (in response, we might say, to the null scenario).[6]

Ludwig vertritt offenbar die folgende Position:

(IU) S hat die Intuition, dass p ↔ S fällt alleine aus begrifflicher Kompetenz das Urteil, dass p (in Bezug auf ein vorgestelltes Szenario).

Zwischen diesem und dem ersten angeführten Konzept von Intuitionen besteht ein sehr enger Zusammenhang. Denn Urteile sind Einstellungen, die eine entsprechende Überzeugung nach sich ziehen bzw. voraussetzen. Man vergegenwärtige sich die folgende Aussage:[7]

(1) *Ich urteile, dass p, aber ich glaube nicht, dass p.

Die Aussage (1) ist sinnlos, was nahelegt, dass Urteile Glaubenszustände gegenüber demselben Gehalt beinhalten.

Ein weiteres Konzept von Intuitionen, das in der Gegenwart von manchen Autoren vertreten wird, besteht darin, Intuitionen mit Dispositionen zu Überzeugungen zu identifizieren[8]:

---

5 Goldman (2007) scheint auch einen derartigen Ansatz zu bevorzugen.

6 Ludwig (2007), S. 135.

7 Moulineux und Earlenbaugh nennen das (1) angewandte Testverfahren den Konjunktionstest (*conjunction-test)* (siehe Earlenbaugh und Moulineux (2009), S. xx). Dieser Test weist nach, dass in jedem Fall, wenn eine Einstellung □ gegenüber einem Gehalt, dass p vorliegt, auch eine andere Einstellung □ vorliegt, sofern eine Konjunktion mit den beiden Einstellungen akzeptabel ist.

8 Siehe Earlenbaugh und Moulyneux (2009).

(ID) S hat die Intuition, dass p ↔ S ist disponiert zu glauben, dass p.

Dieses Konzept ist gegenüber den ersten beiden angeführten Konzepten schwächer, als es weder nach sich zieht, dass man, sofern man eine Intuition hat, eine Überzeugung gegenüber dem Gehalt p besitzt noch dass man ein Urteil mit dem Gehalt p fällt. Die meisten Autoren, die ein derartiges Konzept vertreten, sehen dies auch als einen Vorteil gegenüber diesem Konzept an.

Schließlich gibt es noch die Konzeption, wonach Intuitionen Einstellungen eigener Art sind. George Bealer formuliert dies folgendermaßen:

> For you to have an intuition that A is just for you to seem that A. Here ‚seems' is understood, not as a cautionary or „hedging" term, but in its use as a kind of term for a genuine kind of conscious episode. For example, when you first consider one of de Morgan's laws, often it neither seems to be true nor seems to be false; after a moments reflection, however, something new happens: suddenly it just seems true. Of course, this kind of seeming is intellectual, not sensory or introspective (or imaginative). For this reason, intuitions are counted as „data of reason" not „data of experience."[9]

Bealer vertritt, mit gewissen an anderen Stellen noch näher ausgeführten Einschränkungen hinsichtlich dessen, wie man philosophische oder rationale Intuitionen charakterisieren sollte, offenbar die folgende Analyse:

(IS) S hat die Intuition, dass p ↔ S erscheint es so, dass p.

Bei einem derartigen Konzept von Intuitionen ist man nicht darauf festgelegt, dass man disponiert ist, den Gehalt der Intuition zu glauben. Trivialerweise ist man auch nicht darauf festegelegt, dass man den Gehalt der Intuition tatsächlich glaubt oder dass man hinsichtlich des Gehalts ein Urteil fällt. Aufgrund dieser Merkmale ist das von Bealer vorgeschlagene Modell von Intuitionen das schwächste der angeführten Modelle bezüglich dessen, welche anderen doxastischen Einstellungen durch das Vorliegen einer Intuition nach sich gezogen werden.

1. Diskussion

Welches dieser Konzepte ist nun plausibel? Ich möchte die hier angeführten Konzepte nun – wie bereits erwähnt – dahingehend diskutieren, ob sie tatsächlich die Wahrheitsbedingungen von Aussagen, in denen das Prädikat „hat eine Intuition, dass ..." auftaucht, korrekt bestimmen. Ich werde zu zeigen versuchen, dass es gewisse natursprachliche Daten gibt, die nahelegen, dass die ersten drei Konzepte unplausibel sind und dass für die vierte oben angeführte Konzeption Daten existieren, die sie bestätigen.

9 Bealer (2000), S. 3

Gegen das erste Konzept von Intuitionen scheint man einwenden zu können, dass die folgende Aussage sinnvoll ist:

(1) Ich habe die Intuition, dass p, aber ich glaube nicht, dass p.

Es ist nicht schwierig, Beispiele zu finden, in denen man einer Person zwar wahrheitsgemäß zuschreibt, eine Intuition gegenüber einem Gehalt zu besitzen, aber nicht die entsprechende Überzeugung. Man denke etwa an Intuitionen, die man hinsichtlich Überzeugungen (oder Prinzipien) hat, die zu Paradoxien führen – wie etwa dass es für jede Eigenschaft eine Menge von Objekten gibt, die diese Eigenschaft besitzen ist oder dass ein Sandhaufen ein Haufen bleibt, wenn ein Sandkorn entfernt wird. Wir wissen natürlich, dass die beiden Überzeugungen (oder Prinzipien) falsch sind, weil sie zu Paradoxien führen und glauben sie entsprechend nicht. Dennoch sind die beiden Prinzipien intuitiv und scheinen auch auf den ersten Blick wahr zu sein. (Es gibt selbstverständlich noch weitere Beispiele dieser Art, wie etwa Intuitionen hinsichtlich Wahrscheinlichkeitsverteilungen oder der Mächtigkeit von Mengen).

Dass man den Gehalt, über den man eine Intuition hat, nicht zwingender Weise glaubt, zieht nach sich, dass der Umstand, ob man den Gehalt einer Intuition tatsächlich glaubt, nicht die Wahrheitsbedingungen einer Aussage, in der das Prädikat „hat eine Intuition“ auftaucht, beeinflußt. Wahrscheinlich wird durch die Verwendung des Prädikats „hat die Intuition, dass ...“ konversational implikiert, dass man den Gehalt tatsächlich auch glaubt. Gemäß den von Grice vorgeschlagenen Merkmalen für konversationale Implikaturen würde der Umstand, dass man den Gehalt einer Intuition auch tatsächlich glaubt, das sog. Stornierbarkeitskriterium erfüllen – wie die Aussage (1) belegt. Da konversationale Implikaturen nicht die Wahrheitsbedingungen einer Aussage beeinflussen, würde dies erklären, weshalb es keine notwendige Bedingung für das Vorliegen einer Intuition ist, dass man den Gehalt tatsächlich glaubt. Dennoch würde die Auffassung, dass Intuitionen keine Überzeugungen sind, nicht dem Phänomen widersprechen, dass wir in den meisten Fällen (insbesondere beim Rückgriff auf Intuitionen in der Philosophie) den Gehalt, über den wir eine Intuition besitzen, auch tatsächlich glauben.[10] Wenn wir die Aussage (2)

(2) Das Fenster ist offen.

verwenden, würden wir in den meisten Fällen dem Hörer zu verstehen geben, dass es uns stört, dass das Fenster offen ist. Doch die Aussage (2) bleibt auch dann wahr, wenn man dies nicht dem Hörer mitteilen will (da der implikierte Gehalt, dass es den Sprecher stört, dass das Fenster offen ist, nicht die Wahrheitsbedingungen von (2) beeinflußt). Ähnlich würden wir, wenn wir uns selbst zuschreiben, eine Intuition zu besitzen, damit implikieren, dass wir den Gehalt

10 Gerhard Ernst hat mich in der Diskussion auf diesen Punkt gebracht.

der Intuition auch tatsächlich glauben. Allerdings ist der Umstand, ob wir den Gehalt der Intuition tatsächlich glauben oder nicht glauben, wie bereits erwähnt, nicht dafür relevant, ob wir überhaupt eine Intuition besitzen. Intuitionen scheinen deshalb auch keine Überzeugungen zu sein.

Ein ähnliches Argument läßt sich gegen die Position vorlegen, wonach Intuitionen Urteile sind. Auch die folgende Aussage ist sinnvoll:

(3) Ich habe die Intuition, dass p, aber ich urteile nicht, dass p.

Ob man gewillt ist, über den Gehalt einer Intuition ein Urteil zu fällen, scheint ebenfalls nicht die Wahrheitsbedingungen einer Aussage zu beeinflussen, in der der Ausdruck „hat die Intuition, dass ...“ auftaucht. Auch hier wäre die Deutung naheliegend, dass man in bestimmten Situationen durch die Verwendung des Intuitionenprädikats konversational implikiert, dass man hinsichtlich des Gehalts auch ein Urteil fällt.

Schließlich kann man auch gegen die Auffassung, dass Intuitionen Dispositionen zu Urteilen sind, Beispiele dieser Art vorlegen.

(4) Ich habe die Intuition, dass p, aber ich bin nicht disponiert zu glauben, dass p.

Auch die Aussage (4) scheint sinnvoll zu sein, woraus daraus abgeleitet werden kann, dass Intuitionen keine Dispositionen zu Überzeugungen sind.[11]

Wenn man die eben geführte Diskussion zusammenfasst, legen die angeführten Daten nahe, dass eine Intuition eben keine doxastische Einstellung gegenüber einem Gehalt ist.[12] Andere Arten von Einstellungen, die sich nicht als doxastische Einstellungen verstehen oder auf sie reduzieren lassen, sind etwa die Einstellungen, sich etwas zu wünschen, zu hoffen, zu sehen, zu hören oder etwas mitgeteilt zu bekommen. Auch bei dem Vorliegen derartiger Einstellungen kann man konsistent verneinen, dass ein Zustand des Überzeugtseins (oder der Disposition, überzeugt zu sein) vorliegt. Denn auch die folgenden Aussagen sind sinnvoll[13]:

(ND 1) Ich wünsche mir, dass p, aber ich glaube nicht, dass p.
(ND 2) Ich hoffe, dass p, aber ich glaube nicht, dass p.
(ND 3) Ich sehe, dass p, aber ich glaube nicht, dass p.
(ND 4) Ich höre, dass p, aber ich glaube nicht, dass p.
(ND 5) Jemand teilt mir mit, dass p, aber ich glaube nicht, dass p.

Die Prädikate „wünschen“, „hoffen“, „sehen“, „hören“ und „mitteilen“ sind, wie die Aussagen (ND 1) bis (ND 5) belegen, keine doxastischen Einstellungen.

11 Moulineux und Earlenbaugh versuchen diesen Einwand zu umgehen. Ich kann leider hier darauf nicht eingehen. Siehe Moulineux und Earlenbaugh (2009).

12 Ich komme gleich darauf zu sprechen, dass man gegenüber letzterem durchaus eine ganz spezfische doxastische Einstellung finden kann, die im Fall einer Intuition vorliegt, die aber nicht mit den Einstellungen zusammenfällt, die hier bisher diskutiert wurden.

13 Die Beispiele sind angelehnt an Moulineux und Earlenbaugh (2009).

Obwohl die ersten beiden Arten von Einstellungen, also etwas zu wünschen und etwas zu hoffen, keine epistemische Relevanz besitzen, sind die letzten drei Arten von Einstellungen durchaus epistemisch relevant oder wertvoll.

Doch was sind Intuitionen dann? Betrachten wir nun mal eine an Bealer angelehnte Konzeption von Intuitionen, wonach Intuitionen keine doxastische Einstellungen sind, sondern eigenständige Einstellungen gegenüber einem Gehalt. Die folgende Aussage scheint nicht sinnvoll zu sein:

(4)*Ich habe die Intuition, dass p, aber es erscheint mir nicht so, dass p.

Dass der Gehalt einer Intuition auf eine gewisse Weise erscheint, kann nun im Vergleich zu den anderen oben angeführten Konzepten und die in ihnen enthaltenen konstitutiven Merkmalen für Intuitionen nicht storniert werden. Daraus lässt sich ableiten, dass der Umstand, ob mir ein Gehalt auf eine gewisse Weise erscheint, zunächst keine konversationale Implikatur darstellt. Die Deutung scheint nahe liegend zu sein, dass der Umstand, wie mir der Gehalt einer Intuition erscheint, eher eine notwendige Bedingung darstellt, damit überhaupt eine Intuition vorliegt. Dass der Gehalt einer Intuition als wahr erscheint, scheint entsprechend die Wahrheitsbedingungen einer Intuitionenzuschreibung zu beeinflussen. Entsprechend würde die Aussage (4) die auf Bealer zurückgehende Konzeption von Intuitionen bestätigen.

Wenn man sich andere epistemisch relevante Einstellungen oder Zustände wie etwas zu sehen, zu hören oder etwas mitgeteilt zu bekommen, im Lichte von (4) ansieht, wird man erkennen können, dass auch hier die folgenden Aussagen nicht sinnvoll sind:

(5) *Ich sehe, dass p, aber es erscheint mir nicht so, dass p.
(6) *Ich höre, dass p, aber es erscheint mir nicht so, dass p.
(7) *Jemand teilt mir mit, dass p, aber es erscheint mir nicht so, dass p.

Auch diese Zustände sind Zustände, in denen jemandem ein Gehalt auf eine gewisse Weise „erscheint“. Wie sich diese Zustände letztlich von Intuitionen abgrenzen lassen, soll hier nicht weiter diskutiert werden.

2. Wie sind Intuitionen näher zu charakterisieren?

Da Intuitionen nun keine doxastischen Einstellungen gegenüber Gehalten sind, stellt sich noch die Frage, wie man Intuitionen näher charakterisieren sollte. Ich möchte hier nur auf eine Frage eingehen, nämlich welchen Gehalt (bzw. welche Phänomenologie) Intuitionen besitzen. Bealer hat als ein Charakteristikum für rationale oder philosophische Intuitionen eingeführt, dass einem Subjekt die Gehalte derartiger Intuitionen als notwendig wahr präsentiert werden. Bealer schreibt:

> In our context, when we speak of an intuition, we mean „rational intuition." This is distinguished from what physicists call „physical intuition." We have a physical intuition that if a house is undermined, it will fall. This does not count as a rational intuition, for it does not present itself as necessary. [...]. When we have a rational intuition, say that if p, then not not p, this presents itself as necessary [...].[14]

Bealer gelangt dann zu der folgenden Analyse rationaler Intuitionen:

> (BI)S hat die rationale Intuition, dass p ↔ S erscheint es als notwendig, dass p.

Es gibt einige Erwägungen, die gegen die Auffassung sprechen, dass philosophische oder rationale Intuitionen tatsächlich einen modalen Gehalt besitzen. Zunächst scheint man jemanden auch dann Intuitionen gegenüber Gehalten, wie sie Bealer im Auge hat, zusprechen zu können, sofern er gar kein Verständnis oder keinen Begriff von Notwendigkeit besitzt. Wenn man davon ausgeht, dass einem Subjekt nur dann eine Überzeugung als notwendig wahr erscheinen kann, sofern es über ein Verständnis oder einen Begriff von Notwendigkeit verfügt, dann wäre diese Person prinzipiell gar nicht in der Lage, rationale Intuitionen zu bilden. Doch es erscheint kontraintuitiv, jemand aufgrund eines derartigen Defizits absprechen zu wollen, dass er überrationale Intuitionen verfügt. Ferner scheint es auch Überzeugungen zu geben, die intuitiv gerechtfertigt sind, die aber lediglich kontingenter Weise wahr sind und deren Gehalt entsprechend gar nicht als notwendig wahr erscheinen kann. Ein paradigmatischer Fall scheinen Kripkes Beispiele für das kontingente A Priori abzugeben, aber auch andere Gehalte, die lediglich kontingterweise wahr sind. Schließlich scheint die Auffassung, dass philosophische Intuitionen zugleich ihren Gehalt als notwendig präsentieren, nach sich zu ziehen, dass derartige Intuitionen zugleich eine Quelle modalen Wissens wären. Die Intuition wäre ein mentaler Zustand, der nicht nur Informationen für den Wahrheitswert einer Proposition liefern würde, sondern zugleich Informationen über deren generellen modalen Status – also ob sie eine notwendige oder bloß kontingente Wahrheit ist.[15] Diese Annahme scheint jedoch relativ stark zu sein, als sie einen Verfechter von Intuitionen darauf verpflichtet, eine spezifische Auffassung über die Quelle modalen Wissens zu machen. Ich bin nicht davon überzeugt, dass man eine derartige Annahme machen sollte und denke, dass man philosophische Intuitionen nicht so charakterisieren sollte, dass sie essentieller Weise einen modalen Gehalt besitzen. Mein Vorschlag wäre die folgende Analyse:

> (IG)S hat die Intuition, dass p ↔ S erscheint es *als wahr*, dass p.

14 Bealer (2000), S. 3.

15 Die Unterscheidung zwischen generellem und spezifischen modalen Status wird in Casullo (2003) eingeführt.

3. Sind manche Intuitionen a priori?

Nachdem ich nun für ein ganz bestimmtes Konzept von Intuitionen plädiert habe, stellt sich noch die Frage, welchen epistemischen Status Intuitionen besitzen. Gibt es Intuitionen, die so etwas wie erfahrungsunabhängige Gründe sind? Manche Autoren haben argumentiert, dass Intuitionen keine apriorischen oder erfahrungsunabhängigen Gründe für philosophische Behauptungen darstellen. Goldman schreibt beispielsweise:

> Indeed, the process of generating classification intuitions has more in common with memory retrieval than with purely intellectual thought or ratiocination, the core of the a priori. The generation of classification intuitions involves the accessing of a cognitive structure that somehow encodes a representation of a category. Of the various sources mentioned above, this most resembles memory, which is the accessing of a cognitive structure that somehow encodes a representation of a past episode. Thus, although I am perfectly willing to allow that application intuitions confer warrant, I don't agree that the type of warrant they confer is a priori warrant.[16]

Goldman ist der Auffassung, dass Intuitionen (oder genauer: die kognitiven Prozesse, die zur Bildung von Intuitionen führen) der Erinnerung ähneln. Wenn man beispielsweise eine Intuition hat, die ein Gedankenexperiment betrifft (etwa dass der Charakter eines Gettier-Beispiels nicht über Wissen verfügt), dann findet laut Goldman ein kognitiver Prozess statt, in dem auf den Gehalt eines Begriffs oder einer mentalen Kategorie zurückgegriffen wird. Laut Goldman ähnelt dies eher der Erinnerung als anderen paradigmatischen Prozessen, die zu a priori gerechtfertigten Überzeugungen führen. Denn in der Erinnerung findet auch ein Zugriff auf kognitiv abgespeicherte Information statt, die etwa vergangene beobachtete Ereignisse betrifft.

Doch ist der Umstand, dass man kognitiv auf einen Begriff zurückgreift, dafür verantwortlich, dass man in einer Überzeugung empirisch gerechtfertigt ist? Selbst wenn Goldman damit Recht hätte, dass die Genese einer Intuition gewisse Ähnlichkeiten zur Erinnerung aufweist, würde dies nicht zeigen, dass Intuitionen empirische Gründe für philosophische Behauptungen darstellen. Denn in der von Goldman skizzierten Funktionsweise von Intuitionen würde dieser erinnerungsähnliche kognitive Prozess lediglich eine *ermöglichende* und nicht eine *konstitutive* rechtfertigende Funktion ausüben.[17] Denn dass man auf eine mentale Kategorie zurückgreift, scheint nicht alleine die Intuition und die Rechtfertigung zu konstituieren, dass etwa ein Charakter eines Gettier-Beispiels nicht über Wissen verfügt. Vielmehr muss die Person, die die Intuition besitzt, den Charakter des Gedankenexperiments im Lichte der im Begriff oder der mentalen Kategorie enthaltenen Information beurteilen. Und letzteres scheint für die Rechtfertigung

16 Goldman (2007), S. 20.

17 Diese Unterscheidung stammt von Burge (1993).

relevant zu sein (und auch erfahrungsunabhängig zu funktionieren) und nicht der bloße kognitive Rückgriff auf die Information, die in einem Begriff enthalten ist. Ähnlich ist das Lesen eines mathematischen Beweises nicht für die Rechtfertigung, die durch den Beweis erzeugt wird, konstitutiv, obwohl der kognitive Prozess, einen Beweis zu führen oder nachzuvollziehen, dadurch gewisse Ähnlichkeiten zur Wahrnehmung aufweist. Die Wahrnehmung würde in diesem Fall lediglich eine die Rechtfertigung ermöglichende Rolle spielen. Obwohl ich Goldmans Überlegungen hier nicht genauer diskutieren kann, denke ich doch, dass Goldmans Überlegungen nicht zeigen, dass Intuitionen nicht a priori sein können.

Doch sind nun manche Intuitionen a priori? Oder alternativ formuliert: Wie müssen Intuitionen geartet sein, damit, falls ein Subjekt seine Überzeugung auf eine Intuition stützt (und eine sog. *no-defeater*-Bedingung erfüllt ist), das Subjekt in der Überzeugung a priori gerechtfertigt ist? Da ich in diesem Aufsatz die Auffassung vertreten habe, dass Intuitionen nicht essentieller Weise modal sind, kann der Umstand, ob eine Überzeugung, die intuitiv gerechtfertigt ist, zugleich a priori gerechtfertigt ist, nicht daran liegen, auf welche Weise mir der Gehalt der Intuition erscheint.[18] Ich denke, dass der Status einer Intuition als erfahrungsunabhängig oder a priori eher davon abhängt, welche Genese oder Ätiologie die fragliche Intuition besitzt. Wird die Intuition *rein verstehensbasiert* gebildet (ohne den Rückgriff auf empirische Information), scheint die Intuition so etwas wie ein apriorischer Grund zu sein. Demgegenüber sind Intuitionen, die nicht rein verstehensbasiert gebildet werden, nicht a priori. Beispiele für Intuitionen der letzteren Art wären Intuitionen wie die folgenden:

(7) Ich habe die Intuition, dass es Bäume gibt.

(8) Ich habe die Intuition, dass, falls ich ein Glas fallen lasse, es zerbrechen wird.

(7) ist deshalb keine apriorische Intuition, da in der Genese dieser Intuition Sinneserfahrungen nicht bloß eine ermöglichende, sondern auch eine konstitutive Rolle spielen. Ebenso im Fall von (8). Auch hier handelt es sich nicht um eine apriorische Intuition, da hier Beobachtungen und empirisch gerechtfertigte Gesetzmäßigkeiten in der Bildung der Intuition im Spiel sind.

Man könnte gegen die Strategie, die Erfahrungsunabhängigkeit einer Intuition an deren Ätiologie festzumachen, einwenden, dass die Ätiologie introspektiv opak ist, d.h. dass wir eigentlich nie erkennen können, wie die Genese der Intuition genau aussieht. Doch dies scheint die hier verfolgte Strategie nicht prinzipiell in Zweifel zu ziehen. Denn ob wir erkennen oder nicht erkennen können, wie die Ätiologie einer Intuition aussieht, ändert nichts daran, wie die Ätiologie

18 Ferner scheint der Umstand, dass eine Intuition den Gehalt als notwendig wahr präsentiert nicht hinreichend dafür zu sein, dass die Intuition a priori ist. Denn es wäre denkbar, dass die Ätiologie der Intuition empirisch ist.

tatsächlich ist. Und letzteres scheint für den Status einer Intuition als a priori relevant zu sein.

4. Konklusion

In diesem Aufsatz habe ich zu argumentieren versucht, dass man Intuitionen nicht auf überzeugungsartige Zustände bzw. Dispositionen zu derartigen Zuständen reduzieren kann. Ferner habe ich zu zeigen versucht, dass das alternative Konzept, wonach Intuitionen eben Einstellungen sui generis sind, besser durch natursprachliche Daten abgedeckt wird. Abschließend bin ich noch auf die Frage eingegangen, welche Phänomenologie Intuitionen besitzen und ob manche Intuitionen so etwas wie apriorische Gründe sind. Meine These war, dass Intuitionen keinen modalen Gehalt besitzen und dass manche Intuitionen durchaus a priori sind.[19]

Literatur

Bealer, George: „The Incoherence of Empiricism". In: *Wagner, S./Warner, R. (Hrsg.):* Naturalism: A Critical Appraisal. University of Notre Dame Press, Notre Dame, 1993. S. 99-138

Bealer, George: „A Theory of the A Priori". Pacific Philosophical Quarterly, 81, 2000. S. 1-29

Burge, Tyler: „Content Preservation". The Philosophical Review, 102, 1993. S. 457-488

Casullo, Albert: A Priori Justification. Oxford University Press, New York, 2003

Earlenbaugh, J./Molyneux, B.: „Intuitions are Inclinations to Believe". Philosophical Studies, 145, 2009. S. 89-109

Goldman, Alvin: „A Priori Warrant and Naturalistic Epistemology". In: *Tomberlin, J. (Hrsg.):* Philosophical Perspectives, 13, 1999. Blackwell, Blackwell, Malden (MA). S. 1-28

Goldman, Alvin: „Philosophical Intuitions: Their Target, Their Source and Their Epistemic Status". Grazer Philosophische Studien, 74, 2007. S. 1-26

19 Ich danke allen Teilnehmern und Diskutanten an meinem Vortrag für ihre kritischen Fragen und Hinweise, insbesondere Thomas Grundmann, Frank Hofmann, Gerhard Ernst und Andreas Kemmerling.

Grice, Hebert P.: „Logic and Conversation". In: *Grice, Hebert P.:* Studies in the Ways of Words. Harvard University Press, Cambridge (MA), 1975, 1989. S. 22-40

Grundmann, Thomas: „The Nature of Rational Intuitions and a Fresh Look at the Explanationist Objection". Grazer Philosophische Studien, 74, 2007. S. 69-87

Lewis, David: Philosophical Papers. Bd. 1. Oxford University Press, New York, 1983

Ludwig, Kirk: „The Epistemology of Thought Experiments: First Person versus Third Person Approaches". Midwest Studies in Philosophy, 31, 2007. S. 128-159

Pust, Joel: Intuitions as Evidence. Garland, New York, 2000

Weatherson, Brian: „What Good are Counterexamples?". Philosophical Studies, 115, 2003. S. 1-31

Weinberg, Jonathan/Nichols, Shaun/Stich, Stephen: „Normativity and Epistemic Intuitions". Philosophical Topics, 29, 2001. S. 429-460

Williamson, Timothy: The Philosophy of Philosophy. Blackwell, Oxford, 2007

Verantwortung und epistemische Rechtfertigung als anfechtbare Begriffe – Strukturelle Übereinstimmungen in Handlungs- und Erkenntnistheorie

Claudia Blöser und Hannes Ole Matthiessen
claudiabloeser@googlemail.com; Matthiessen@em.Uni-Frankfurt.de
Goethe-Universität, Frankfurt /Main

Abstract/Zusammenfassung

We want to point to an interesting and so far neglected similarity between theories of responsibility and epistemic justification. With respect to both concepts, it has been proposed that they display a defeasible structure. First, we present the arguments that support such a view, which can be found in the writings of H.L.A. Hart and Michael Williams. Second, the common conceptual structure will be described.

Following H.L.A. Hart, defeasible concepts in the theory of action are characterized by the fact that they are applicable under certain necessary and *normally* sufficient conditions, their application can be challenge by recourse to certain challenge-conditions. With regard to the concept of responsibility, this structure is plausible: A person is responsible for her action, unless challenge-conditions apply, e.g. physical coercion.

With his Default and Challenge-model, Michael Williams offers an alternative to the traditional model, which ties epistemic justification to the possession of evidence. According to Williams, we are normally (*default*) justified in our beliefs, without having to offer explicit reasons for them. Only the occurrence of challenging reasons can make it necessary to adduce reasons.

Both accounts suggest a default and challenge-model of responsibility and epistemic justification. In both cases, a normative status is normally ascribed without giving reasons for it. Responsibility for actions is normally ascribed to adult persons without argument. Similarly, persons can be justified in their beliefs without having to give reasons for them.

In both cases, the default-status can be challenged, and challenging-reasons are case-specific and exceptional.

These commonalities can be summarized as follows:

A person who fulfills default-conditions has the normative status to be responsible for her action or, respectively, to be justified in her belief, unless there are challenging reasons.

Wir möchten auf eine interessante und bisher unbeachtete Ähnlichkeit in Theorien der Verantwortung und der epistemischen Rechtfertigung hinweisen: Es wurde in Bezug auf beide Begriffe vorgeschlagen, dass ihnen eine anfechtbare Struktur zugrunde liegt. Zunächst werden anhand der Ansätze von H.L.A. Hart und Michael Williams die Argumente herausgearbeitet, die es nahe legen, jeden einzelnen Begriff als anfechtbar zu verstehen. Abschließend wird die gemeinsame Begriffsstruktur dargestellt.

Nach H.L.A. Hart sind anfechtbare Begriffe in der Handlungstheorie dadurch gekennzeichnet, dass sie unter bestimmten notwendigen und *normalerweise* hinreichenden Bedin-

gungen anwendbar sind, aber ihre Verwendung unter Rückgriff auf bestimmte Ausnahmebedingungen angefochten werden kann. In Bezug auf den Begriff der Verantwortung ist diese Struktur plausibel: Eine Person ist verantwortlich für ihre Handlung, es sei denn, es liegen Anfechtungsgründe (z.B. physischer Zwang) vor.

Michael Williams stellt dem traditionellen Modell, das Rechtfertigung an den aktiven Umgang mit Evidenzen bindet, sein Default and Challenge-Modell entgegen; hiernach sind wir im Normalfall (*Default*) in unseren Überzeugungen gerechtfertigt, ohne über explizite oder implizite Gründe (Evidenzen) zu verfügen. Erst das Auftreten von Anfechtungsgründen (*Challenges*) kann das Anführen von Gründen nötig machen.

Nach den beiden Ansätzen lassen sich sowohl Verantwortung als auch Rechtfertigung nach einem *Default and Challenge*-Modell verstehen: In beiden Fällen liegt ein normativer Status im Normalfall (*Default*), der an bestimmte Bedingungen geknüpft ist, ohne Begründung vor. Dass erwachsene Personen normalerweise verantwortlich für ihre Handlungen sind, muss nicht eigens begründet werden. Genauso können Personen in ihren Überzeugungen gerechtfertigt sein, ohne Gründe anführen zu müssen. In beiden Fällen kann der Default-Status angefochten werden (*Challenge*), wobei Anfechtungen sowohl *fallspezifisch* als auch *exzeptionell* sind.

Die genannten Gemeinsamkeiten lassen sich folgendermaßen zusammenfassen:

> *Einer Person, die Default-Bedingungen erfüllt, kommt der normative Status zu, verantwortlich für ihre Handlung bzw. gerechtfertigt in ihrer Überzeugung zu sein, es sei denn es liegen Anfechtungsgründe vor*.

Dieser Vortrag hat das bescheidene Ziel, auf eine interessante und bisher sehr wenig beachtete Ähnlichkeit zwischen Typen von Theorien der Verantwortung und der epistemischen Rechtfertigung hinzuweisen, denen zufolge beide Begriffe als anfechtbare Begriffe aufzufassen sind.

Zunächst soll die grundsätzliche Ähnlichkeit der Ansätze von H.L.A. Hart und Michael Williams aufgezeigt werden. Dabei werden die Argumente zumindest angedeutet, die dafür sprechen, die genannten Begriffe als anfechtbar zu verstehen. Schließlich wird skizzenhaft eine Logik anfechtbarer Begriffe angedeutet werden, die es erlaubt, Verantwortung und epistemische Berechtigung als Spezialfälle eines allgemeinen Typus normativer Begriffe zu betrachten.

Anfechtbarkeit nach H.L.A. Hart

Die These, dass der Verantwortungsbegriff eine anfechtbare Struktur aufweist, geht auf einen Vorschlag H.L.A. Harts zurück, der in seinem Aufsatz „The Ascription of Responsibilty and Rights“ von 1948 den Begriff der Anfechtbarkeit (defeasibility) in die Handlungstheorie einführt. Hart argumentiert dafür, dass rechtliche Begriffe wie „Vertrag“ oder „Mord“ *anfechtbar* sind. Das wichtigste Indiz für das Vorliegen dieser besonderen Begriffsstruktur ist nach Hart die Verwendung des Zusatzes „es sei denn“ („*unless*“): Unter bestimmten notwendigen und *normalerweise* hinreichenden Bedingungen sind die Begriffe an-

wendbar, aber ihre Verwendung kann unter Rückgriff auf bestimmte Ausnahmebedingungen, die unter „es sei denn“ aufgeführt sind, angefochten werden. Hart nennt zwei Möglichkeiten der Anfechtung (*Challenges*): 1) Bestreitung der Fakten, auf deren Basis der Begriff verwendet wird („joinder of issue“); 2) Hinweis auf Umstände, die den Fall zu einer Ausnahme machen. Hart macht die Anfechtbarkeit auch in Bezug auf Begriffe aus, die traditionellerweise mentale Elemente bezeichnen (*mens rea*, Absichtlichkeit, Freiwilligkeit). Da Elemente wie Freiwilligkeit und Absichtlichkeit traditionell als notwendige Bedingungen für Verantwortung gelten, schafft Hart mit seiner Analyse die Grundlage für ein Verständnis von Verantwortung als anfechtbarem Begriff. Für den Begriff der Verantwortung ist insbesondere die Eigenschaft der anfechtbaren Begriffe zutreffend, dass der Begriff so lange legitimer Weise angewendet wird, wie keine Anfechtung bekannt ist: Eine Person ist verantwortlich für ihre Handlung, es sei denn es liegen Anfechtungsgründe vor. Anfechtungsgründe wären beispielsweise, dass die Person geisteskrank ist, dass sie physisch zu der Handlung gezwungen wurde o.ä.

Williams’ Default und Challenge-Modell epistemischer Rechtfertigung

Michael Williams begründet sein *Default und Challenge*-Modell epistemischer Rechtfertigung in erster Linie mit dem Wunsch, dem traditionellen Verständnis von Rechtfertigung als evidentieller Rechtfertigung eine Alternative gegenüberzustellen. Hierzu führt er eine Unterscheidung zwischen zwei Rechtfertigungsbegriffen ein: Eine bestimmte *Überzeugung* ist wohl begründet (adequately grounded), wenn sie nicht bloß zufälligerweise wahr ist. Eine *Person* ist *personal* gerechtfertigt in einer Überzeugung, wenn sie sich bei der Überzeugungsgewinnung epistemisch verantwortungsvoll verhalten hat. Traditionell wurde in der Erkenntnistheorie der subjektive Zugang zur Wohl-Begründetheit als notwendige Bedingung für personale Rechtfertigung verstanden; diese Annahme, so wurde verschiedentlich argumentiert, führt allerdings in den Skeptizismus.

Als Alternative schlägt Williams sein *Default und Challenge*-Modell vor; hiernach kann ein Subjekt personal gerechtfertigt sein, ohne Evidenzen anzuführen (oder auch nur dazu in der Lage zu sein), solange keine relevanten Einwände erhoben werden. Im Normalfall (*Default*) gelten wir in unseren Überzeugungen als (personal) gerechtfertigt, ohne über explizite oder implizite Gründe (Evidenzen) zu verfügen. Erst das Auftreten von *substantiellen* Anfechtungsgründen (*Challenges*) kann das Begründen, die evidentielle Rechtfertigung der Überzeugung, nötig machen.

Während nach Williams personale Rechtfertigung in einer Überzeugung standardmäßig jedem Subjekt zugeschrieben wird, gibt es Ansätze (etwa bei

Brandom 1994; Willaschek 2007), Default-Rechtfertigung an das Vorliegen gewisser (sozial-)externalistischer Kriterien zu binden – es würde demnach nicht jede Überzeugung in jeder Situation prima facie gerechtfertigt sein. Da das Vorliegen dieser Bedingungen jedoch dem epistemischen Subjekt nicht bekannt sein muss und ihm folglich keine diskursiv artikulierbaren Gründe zugänglich sein müssen, stellt dieser Ansatz keinen Rückfall in ein evidentialistisches Rechtfertigungsmodell dar.

Übereinstimmungen zwischen den Genannten

Nach den Ansätzen von Hart und Williams lassen sich sowohl Verantwortung als auch Rechtfertigung nach einem *Default und Challenge*-Modell verstehen: In beiden Fällen handelt es sich um einen normativen Status, der im Normalfall (*Default*) ohne Begründung vorliegt.

Da *Begründungen* bei der Zuschreibung von Verantwortung eine andere Rolle spielen als bei epistemischer Rechtfertigung, werden in den untersuchten Theorien auch unterschiedliche Arten von Begründung als verzichtbar angesehen. Im Fall der Zuschreibung von Verantwortung muss *die zuschreibende Person* nicht nachweisen, dass bestimmte notwendige Bedingungen für Verantwortung seitens der oder des Handelnden erfüllt werden, z.B. Absichtlichkeit oder der Besitz relevanter kognitiver Fähigkeiten. In der Epistemologie hingegen geht eher darum, ob *Personen* in ihren Überzeugungen gerechtfertigt sein können, ohne *selbst* Gründe anzuführen, was Williams bejaht.

Ein weiterer bemerkenswerter Unterschied liegt darin, dass Hart zufolge jede Verantwortungszuschreibung an gewisse Bedingungen geknüpft ist, und sich somit minimale Bedingungen für den Default-Status formulieren lassen. Eine plausible Default-Bedingung wäre zum Beispiel, dass es sich um eine erwachsene Person handeln muss, da kleinen Kindern nicht ohne weiteres Verantwortung zugeschrieben wird. Demgegenüber betont Williams einerseits: „the status of epistemic subjects does not come with mere sentience: it has to be earned through training and education". Andererseits verzichtet er *für den Einzelfall* auf das Vorliegen von positiven Bedingungen: „One is entitled to a belief or assertion [...] in the absence of appropriate ‚defeaters': that is, reasons to think that one is *not* so entitled." (Williams 2001, p. 149) Es erscheint allerdings zweifelhaft, ob dies tatsächlich unsere epistemische Praxis adäquat widerspiegelt. Es scheint eher so zu sein, dass Subjekte in manchen Kontexten in einer Überzeugung per default gerechtfertigt sind, in anderen Kontexten aber nicht. Während ich die Behauptung, dass sich in diesem Zimmer X Personen befinden, nicht begründen muss (oder kann), wäre ich verpflichtet, für die Behauptung, dass im Nachbarraum ebenso viele Zuhörer sitzen, eine Begründung zu liefern. Es

scheint also angemessen, auch für epistemische Rechtfertigung Bedingungen für den Default-Status anzunehmen.

Eine wichtige Übereinstimmung zwischen den genannten Ansätzen ist, dass der ohne Begründung erworbene normative Status angefochten werden kann (*Challenge*), wobei jeweils Anfechtungen sowohl *fallspezifisch* als auch *exzeptionell* sind: Sowohl Zuschreibungen von Verantwortung als auch solche von Rechtfertigung werden nur dann in Frage gestellt, wenn positive Hinweise dafür vorhanden sind, dass *in diesem Fall* eine Bedingung vorliegt, die den normativen Status ausschließt (es genügt nicht der bloße Hinweis auf die logische Möglichkeit). Gleichzeitig haben Anfechtungen einen Ausnahmecharakter: Es kommt in unserer Praxis vergleichsweise selten vor, dass eine Bedingung vorliegt, unter der man jemanden nicht für zurechnungsfähig hält oder die die explizite Begründung eines Wissensanspruchs erfordert. Eine weitere Übereinstimmung scheint darin zu bestehen, dass Anfechtungen unter Umständen zurückgewiesen werden können, dies wiederum unter Hinweis auf spezifische Eigenschaften des Falles.

Die genannten strukturellen Gemeinsamkeiten lassen sich folgendermaßen zusammenfassen:

> *Einer Person, die Default-Bedingungen erfüllt, kommt der normative Status zu, verantwortlich für ihre Handlung bzw. gerechtfertigt in ihrer Überzeugung zu sein, es sei denn es liegen Anfechtungsgründe vor.*

Bevor wir einige zentrale Grundzüge von Default und Challenge-Konzeptionen der Verantwortung und epistemischen Rechtfertigung ausbuchstabieren und einander gegenüberstellen, möchten wir auf ein allgemeines Charakteristikum von Theorien anfechtbarer Begriffe hinweisen. Es liegt nahe, sie mit einer pragmatistischen Bedeutungstheorie zu verbinden, d.h. davon auszugehen, dass die jeweilige Praxis der Begriffsverwendung die Bedeutung anfechtbarer Begriffe konstituiert, so dass für eine im Sinne dieser Praxis korrekte, aber „in Wirklichkeit" falsche Begriffsverwendung kein Raum bleibt. Am analogen Fall der Eigenschaft, ein Schachkönig zu sein, kann diese These illustriert werden: Was es heißt, ein Schachkönig zu *sein*, ist vollständig durch die Regeln des Schachspiels festgelegt, innerhalb dessen einer Figur dieser Status *zugeschrieben* wird. Unabhängig von den Regeln des Schachspiels macht es jedoch keinen Sinn zu fragen, was ein Schachkönig „wirklich" ist. Dafür, dass diese These auch auf die Begriffe der Verantwortung und Rechtfertigung zutrifft, ließe sich folgendermaßen argumentieren: Unsere Praxis der Zuschreibung von Verantwortung und epistemischer Rechtfertigung weist eine Default und Challenge-Struktur auf. Möchte man diese philosophisch ernst nehmen, darf man sich nicht darauf festlegen, dass es von der Begriffsverwendung unabhängige Erfüllungsbedingungen der entsprechenden Begriffe gibt, da damit die Anfechtbarkeit zu einem bloßen Oberflächenphänomen erklärt würde. Dies bedeutet, dass – gemäß dem diskutierten Ansatz – eine Person verantwortlich für ihre Handlung bzw. gerechtfer-

tigt in ihrer Überzeugung *ist*, wenn ihr gemäß den Regeln der Praxis korrekter Weise der jeweilige Status *zugeschrieben* werden kann.

Default- und Challenge-Bedingungen für Verantwortung

Im Folgenden möchten wir mögliche Default- und Challenge-Bedingungen für Verantwortung vorschlagen. Hart wird zu Recht dafür kritisiert, dass er den Verantwortungsbegriff nicht näher erläutert. Wir schlagen vor, den Verantwortungsbegriff eng an den Begriff des Verdienstes von Lob und Tadel zu knüpfen:

> *Wenn Default-Bedingungen erfüllt sind, verdient die Person Lob oder Tadel – ist also verantwortlich – für ihre Handlung, es sei denn es liegen Ausnahme- oder Entschuldigungsgründe vor.*

Da *Zurechenbarkeit* (das StGB spricht heute von *Schuldfähigkeit*) eine notwendige Voraussetzung dafür ist, dass eine Person gelobt oder getadelt werden kann, müssen Default-Bedingungen ausschließen, dass die Person offensichtlich nicht zum Kreis der zurechnungsfähigen Menschen gehört. Ein Element der Default-Bedingungen wäre also: Es handelt sich um eine erwachsene Person, die nicht offensichtlich geisteskrank (oder stark betrunken) ist. Um eine Handlung normativ bewerten zu können, muss in den Default-Bedingungen zusätzlich enthalten sein, dass die Handlung der Person eine Norm verletzt oder in besonderem Maße erfüllt.

Hierbei kann es sich um eine moralische oder rechtliche Norm handeln. Im ersten Fall wird der Person *moralische* Verantwortung für ihre Handlung, im zweiten Fall *rechtliche* Verantwortung zugeschrieben (im negativen Fall entspricht dies dem Sinn von moralischer bzw. rechtlicher Schuld).

Die Zurechenbarkeit und die Verantwortung (im Sinne von Verdienst von Lob und Tadel) können durch bestimmte Anfechtungsgründe in Frage gestellt werden. Anfechtungen der Zurechenbarkeit können entweder temporär sein, z.B. starke Trunkenheit, oder dauerhaft, z.B. schwere geistige Störungen. In Anlehnung an R. J. Wallace' Terminologie nennen wir diese Anfechtungsgründe *Ausnahmegründe* (Wallace, 118). Während Anfechtungsgründe darauf hindeuten, dass die Person relevante kognitive Fähigkeiten nicht besitzt oder nicht ausüben kann – Wallace spricht hier von „Fähigkeiten zur reflexiven Selbstkontrolle" (vgl. z.B. Wallace, 155) – weist die andere Klasse von Anfechtungsgründen darauf hin, dass aufgrund bestimmter Umstände der Handlung die Person nicht mehr oder nur eingeschränkt als tadelnswürdig angesehen wird. Diese *Entschuldigungsgründe*, wie Wallace sie nennt (Wallace, 118), zeigen an, dass die ursprünglich tadelnswürdige Handlung nicht mit *Absicht* ausgeführt wurde. Ein alltägliches Beispiel wäre folgendes: Eine Person, die zu einer Verabredung eine Stunde zu spät kommt, machen wir zunächst dafür verantwortlich – wir halten sie für tadelnswürdig, was sich zum Beispiel in Ärger oder Vorwürfen ausdrückt

– aber wenn wir erfahren, dass sie eine Fahrradpanne hatte, entschuldigen wir den Vorfall.

Default- und Challenge-Bedingungen für epistemische Berechtigung

Um den Gedanken der Anfechtbarkeit in der Epistemologie fruchtbar zur Anwendung zu bringen, sind einige Überlegungen zur geeigneten Terminologie angebracht. Williams gebraucht den traditionellen Begriff der Rechtfertigung (justification). Da dieser fast zwangsläufig Assoziationen mit der Idee des Begründens hervorruft (die ja eben nur in Sonderfällen eine Rolle spielen soll), schlagen wir vor, den Begriff der (epistemischen) Berechtigung (entitlement) an seiner Stelle zu verwenden. In Anlehnung an Michael Williams, der personale Rechtfertigung als einen Status des Subjekts – und nicht der geglaubten Proposition – betrachtet, wollen wir sagen, dass epistemische Berechtigung eines Subjekts zur Überzeugung dass p in dem Recht des Subjekts besteht, selbst dann zu glauben, dass p, wenn es nicht in der Lage ist, seine Überzeugung zu begründen. Neben etlichen Fällen, in denen dieses Recht durch explizites Begründen durch das epistemische Subjekt erworben werden muss (indem es beispielsweise ein neues Theorem beweist), sind die meisten epistemischen Berechtigungen Default-Berechtigungen.

Wir hatten schon auf eine Ambivalenz hingewiesen, die mit den unterschiedlichen Rollen des Zuschreibers in Handlungstheorie und Erkenntnistheorie zusammenhängt. Ist es im ersten Fall die *Zuschreibung* von Verantwortung, die keiner Begründung bedarf, scheint der Vorzug einer Default und Challenge-Theorie epistemischer Berechtigung gerade darin zu bestehen, dass sie *das epistemische Subjekt* von der Begründungslast befreit.

Tatsächlich hängen das begründungsfreie Recht auf Überzeugungen und das begründungsfreie Recht auf Zuschreibungen epistemischer Status jedoch eng zusammen. Dies wird deutlich, wenn man sich vor Augen führt, welches die Default-Bedingungen sind, also die Bedingungen, unter denen wir Subjekten Default-Berechtigungen zubilligen. Unserer Auffassung nach tun wir dies genau dann, wenn es im gegebenen Kontext derart offensichtlich ist, dass man diese Person als einen verlässlichen Zeugen ansehen darf, dass die Frage „Woher weißt du das?“ – an das epistemische Subjekt gerichtet – ebenso wenig Sinn ergibt wie die Frage „Warum sollen wir ihm glauben?“ – an den Zuschreiber gerichtet. Abhängig davon, um welche Art von Überzeugung es sich jeweils handelt, sind ganz unterschiedliche Default-Bedingungen erforderlich. Manchmal gehören bestimmte Wahrnehmungsbedingungen oder besondere Fähigkeiten des Subjekts dazu, in anderen Fällen genügt es, ein menschliches Wesen zu sein (etwa wenn es um den eigenen Namen geht). Meine Default-Berechtigung zu

der Überzeugung, dass X Personen im Auditorium sitzen hängt damit zusammen, dass ich Ihnen bei gutem Licht gegenüber sitze und es keine Versteckmöglichkeiten gibt.

Die ohne Begründungsleistung bestehende epistemische Berechtigung kann unter Verweis auf das Vorliegen mindestens einer Anfechtungsbedingung aufgehoben werden. Es gibt unübersichtlich viele Umstände, die als Basis einer Anfechtung dienen können. Positive Hinweise darauf, dass aktuell Falschgeld im Umlauf ist oder mein Informant ein starkes Interesse daran hätte, mich zu belügen, sind Fälle von unterminierenden Anfechtungen (undercutting/undermining challenges). Hinweise darauf, dass die Proposition, zu der die Berechtigung besteht, falsch ist, wirken als widerlegende Anfechtungen (rebutting/overriding challenges). Beide Typen von Anfechtungen haben gemein, dass in ihrem Lichte sowohl die Frage „Woher weißt du das?“ – an das epistemische Subjekt gerichtet – als auch die Frage „Warum sollen wir ihm glauben?“ – an den Zuschreiber gerichtet – *legitim* wird. Es ist nicht länger offensichtlich, dass das epistemische Subjekt ein verlässlicher Zeuge ist, weswegen die begründungslose Berechtigung erlischt.

Schlussbetrachtung

Wir haben versucht, für die grundsätzliche Plausibilität des folgenden Modells zu werben: Sowohl Verantwortung als auch epistemische Berechtigung dürfen unter Default-Bedingungen zugeschrieben werden. Einer erwachsenen Person können ihre Handlungen normalerweise zugerechnet werden; und wenn eine Handlung eine Norm verletzt, ist Tadel normalerweise angemessen. Unter geeigneten Bedingungen ist man zu seinen Überzeugungen epistemisch berechtigt, ohne in der Lage zu sein, sie zu begründen. In beiden Fällen ist der normative Status anfechtbar. Im Falle der Verantwortung lassen sich die Anfechtungsgründe in *Ausnahme*- und *Entschuldigungsgründe* einteilen. Epistemische Berechtigungen können *unterminiert* oder *widerlegt* werden.

Wechselwirkungen zwischen praktischer und theoretischer Philosophie sind natürlich nichts Neues. Erinnert sei an die in beiden Bereichen vorkommenden Debatten um deontologische, teleologische und Tugendkonzeptionen, sowie die Frage des Grundinternalismus versus -externalismus. Die Untersuchung von Anfechtbarkeit in unterschiedlichen Teildisziplinen der Philosophie bietet einen interessanten neuen Forschungsgegenstand, der sich insbesondere dadurch auszeichnet, dass er mit der kodifizierten Anfechtbarkeit in der Sphäre des Rechts einen Zugang zum Untersuchungsgegenstand bietet, der von strittigen Intuitionen weitgehend unabhängig ist.

Literaturverzeichnis

Brandom, Robert B.: Making It Explicit. Reasoning, Representing & Discursive Commitment. Harvard University Press, Cambridge, 1994

Hart, Herbert L. A.: "The Ascription of Responsibility and Rights", 1948/49. In: *Flew, Antony (Hrsg.)*: Essays on Logic and Language. Oxford University Press, Oxford, 1963. S. 145-166

Wallace, R. Jay: Responsibility and the Moral Sentiments. Harvard University Press, Cambridge, 1994

Willaschek, Marcus: "Contextualism about Knowledge and Justification by Default" . Grazer Philosophische Studien, 74, Rodopi, Amsterdam, 2007. S. 251-272

Williams, Michael : Problems of Knowledge. Oxford University Press, Oxford, 2001

Skeptizismus und epistemische Berechtigung

Sebastian Schmoranzer
schmrnzr@yahoo.co.uk
Universität zu Köln

Abstract/Zusammenfassung

Are we internally justified in believing that sceptical hypotheses are false? If justification is exclusively construed as evidential justification, the prospects for a positive answer look dim.

In order to avoid this unwelcome consequence Crispin Wright suggests to take a more liberal view on internal justification allowing for a kind of non-evidential epistemic warrant: epistemic entitlement. According to Wright we are entitled to trust that we are not the victim of a sceptical delusion even though we lack the respective evidence.

In my article I will explain a) what exactly is meant by epistemic entitlement and b) why Wright thinks it to provide an answer to the sceptic (parts I to III). Finally, I will c) show why Wright's alleged solution fails (parts IV and V).

Sind wir (intern) gerechtfertigt, skeptische Hypothesen für falsch zu halten? Wenn wir unter Rechtfertigung ausschließlich das Vorliegen von wahrheitsindizierenden Anhaltspunkten verstehen, scheint die Antwort negativ ausfallen zu müssen.

Um diese unerfreuliche Konsequenz zu vermeiden, schlägt Crispin Wright vor, den Begriff der Rechtfertigung weiter als bislang üblich zu fassen und um den Begriff der epistemischen Berechtigung zu erweitern. In Wrights Augen sind wir aus bestimmten Gründen epistemisch berechtigt, skeptische Hypothesen für falsch zu halten, obwohl wir keine entsprechenden Anhaltspunkte haben.

In meinem Artikel werde ich a) erläutern, was es mit Wrights Begriff epistemischer Berechtigung auf sich hat und b) wieso Wright glaubt, damit eine Antwort auf den Skeptiker bereitzustellen (Abschnitte I bis III). Abschließend werde ich c) aufzeigen, weshalb Wrights Lösungsvorschlag nicht überzeugt (Abschnitte IV und V).

I

Der Außenweltskeptizismus stellt spätestens seit Descartes ein notorisches und meines Erachtens immer noch ungelöstes Problem dar. Mit einem verblüffend einfachen Argument stellt der Skeptiker unser gesamtes, als selbstverständlich geltendes Wissen über eine von uns unabhängig existierende Außenwelt in Frage: Wissen erfordert Rechtfertigung. Rechtfertigung ihrerseits erfordert den Ausschluss skeptischer Hypothesen. Da man diese aber nicht ausschließen kann, folgt, dass wir kein Wissen haben.

Ich möchte mich hier auf die dritte Prämisse konzentrieren und mich der Frage zuwenden, ob wir im internalistischen Sinn gerechtfertigt sind, skeptische

Hypothesen wie die Gehirn-im-Tank-Hypothese für falsch zu halten. Traditionellerweise hat man zu zeigen versucht, dass sich solche skeptischen Szenarien empirisch oder mit Hilfe von transzendentalen Argumenten ausschließen lassen. Aus Gründen, auf die ich hier nicht eingehen möchte, halte ich solche Versuche jedoch für nicht überzeugend.[1]

In jüngster Zeit ist Crispin Wright zu einer neuen Strategie übergegangen. In seinen Augen besteht ein zentrales Problem in der Skeptizismusdebatte darin, dass wir einen zu engen Rechtfertigungsbegriff haben, bei dem Rechtfertigung immer im Vorliegen von wahrheitsindizierenden Anhaltspunkten bestehen muss. Versteht man unter Rechtfertigung hingegen epistemische Verantwortlichkeit in einem weiteren Sinne, dann, so Wrights Idee, könne man nachweisen, dass wir, wenn auch nicht im klassischen Sinne gerechtfertigt, so doch epistemisch berechtigt sind, skeptische Hypothesen für falsch zu halten.[2]

Dieser Überlegung möchte ich im Folgenden meine Aufmerksamkeit widmen und prüfen, ob wir auf ihrer Grundlage des Skeptikers Herr werden können. Zu diesem Zweck werde ich zunächst ausgehend von einem bestimmten Rechtfertigungsverständnis den Begriff der epistemischen Berechtigung dem der indiziellen Rechtfertigung gegenüberstellen (Abschnitt II). Anschließend werde ich Wrights Position vorstellen (Abschnitt III), um in einem dritten Schritt zwei Einwände vorzubringen (Abschnitte IV und V).

II

Was ist unter epistemischer Berechtigung zu verstehen? Dem Begriff der epistemischen Berechtigung liegt folgendes Verständnis epistemischer Rechtfertigung zugrunde:

> Eine Person S ist genau dann epistemisch gerechtfertigt, eine These p für wahr zu halten, wenn es für S auf der Grundlage der S zugänglichen Informationen und in Hinblick auf die Maxime, möglichst viele wahre und möglichst keine falschen Überzeugungen zu gewinnen, verantwortlich ist, p zu akzeptieren.

Diese Verantwortlichkeit kann auf zwei Arten realisiert sein. Zum einen können S Anhaltspunkte für die Wahrheit von p vorliegen. Wenn ich in der Tageszeitung lese, dass am Abend ein Bundesligaspiel stattfindet, dann habe ich einen

1 Für eine ausführliche Kritik an solchen Versuchen siehe Schmoranzer (im Erscheinen): Kapitel III.1-3.

2 Wright (1985), (1994), (2002), (2003), (2004a), (2004b). Zentral ist Wright (2004b). Zur Auseinandersetzung mit Wrights Ansatz siehe auch Davis (2004), Jenkins (2007), Pedersen (2005), Pritchard (2005). Für eine ausführliche Kritik an Wright siehe Schmoranzer (im Erscheinen): Kapitel III.4. Michael Williams (1996), (2001) vertritt meines Erachtens ebenfalls eine Konzeption epistemischer Berechtigung. Siehe dazu auch Schmoranzer (im Erscheinen): Kapitel III.5.

Anhaltspunkt für die entsprechende Überzeugung. Und folglich ist es – sofern keine gegenteiligen Anhaltspunkte vorliegen – auch epistemisch verantwortlich von mir, diese Überzeugung zu haben. In diesem Fall liegt das vor, was ich eine indizielle Rechtfertigung nenne.

Epistemische Verantwortlichkeit kann der obigen Definition zufolge aber auch noch auf eine andere Weise realisiert sein. Nehmen wir an, Sokrates geht zur Pythia und erbittet von dieser Auskunft über die Tugend. Die Priesterin bietet Sokrates nun folgenden Handel an: Wenn er Goldbachs Vermutung für wahr hält, verrät sie ihm im Gegenzug, was er zu wissen begehrt. Ob Goldbachs Vermutung wahr oder falsch ist, weiß Sokrates natürlich genauso wenig wie die fähigsten Mathematiker. Weder hat er Anhaltspunkte für, noch gegen diese Vermutung. Sokrates weiß aber, dass die Pythia ihr Wort halten wird. In dieser Situation ist er zwar nicht indiziell gerechtfertigt, Goldbachs Vermutung für wahr zu halten. Nichtsdestoweniger ist es in Hinblick auf die Maxime, den Schatz seiner wahren Überzeugungen zu mehren, eine gute Idee, sich durch Autosuggestion dazu zu bringen, die entsprechende These zu akzeptieren. Möglicherweise akzeptiert er damit eine falsche These. Aber der Zugewinn an wichtigen und wahren Überzeugungen dürfte überwiegen. In diesem Fall ist Sokrates epistemisch berechtigt, Goldbachs Vermutung für wahr zu halten.

III

Die nächste Frage ist, wie diese Idee für eine Widerlegung des Skeptikers fruchtbar gemacht werden kann. Unter welchen Bedingungen *genau* liegt eine epistemische Berechtigung vor? Und wieso sind wir epistemisch berechtigt, *skeptische Hypothesen* für falsch zu halten? In „On Epistemic Entitlement – (Warrant for Nothing and Foundations for Free?)“[3] hat Crispin Wright eine Antwort auf diese Fragen zu geben versucht.

Wright unterscheidet vier verschiedene Formen epistemischer Berechtigung, von denen die kognitive Berechtigung („entitlement of cognitive project“) die Grundlage für Wrights Zurückweisung des Außenweltskeptizismus darstellt. Dazu heißt es bei Wright:

> Let me try to harness these ideas to a definite proposal about entitlement. First (to tidy up a bit) a definition: let us say that
>
> P is a presupposition of a cognitive project if to doubt P (in advance) would rationally commit one to doubting the significance or competence of the project. [Voraussetzungsbedingung]
>
> Then the relevant kind of entitlement – an entitlement of cognitive project – may be proposed to be any presupposition of a cognitive project meeting the following two conditions:

3 Wright (2004b).

(i) We have no sufficient reason to believe that P is untrue [Unschuldsbedingung]
and
(ii) The attempt to justify P would involve further presuppositions in turn of no more secure a prior standing ... and so on without limit; so that someone pursuing the relevant enquiry who accepted that there is nevertheless an onus to justify P would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessor. [Regressbedingung]

No doubt that will stand refinement, but the general *motif* is clear enough. If a cognitive project is indispensable, or anyway sufficiently valuable to us – in particular, if its failure would at least be no worse then the cost of not executing it, and its success would be better – [Strategiebedingung] and if the attempt to vindicate (some of) its presuppositions would raise presuppositions of its own of no more secure antecedent status, and so on *ad infinitum*, then we are entitled – may help ourselves to, take for granted – the original presuppositions without specific evidence in its favour. […] [If all these conditions are met] just go ahead and trust that the former [i.e. the presuppositions] are met.[4] (Kursivdruck Wright, meine Unterstreichung)

Diese Ausführungen legen folgende Bestimmung kognitiver Berechtigung nahe:

Ich bin kognitiv berechtigt, darauf zu *vertrauen* („to trust")[5], dass p, wenn gilt:
i) p ist eine Voraussetzung für ein kognitives Projekt P, das heißt: wenn ich bezweifle, ob p der Fall ist, dann muss ich rationalerweise die Bedeutung oder die Angemessenheit des kognitiven Projekts P in Frage stellen, (Voraussetzungsbedingung)
und
ii) das kognitive Projekt P ist epistemisch hinreichend wertvoll für mich, insofern es epistemisch unerlässlich ist oder aber gilt: eine erfolgreiche Durchführung desselben ist aus epistemischer Sicht von Vorteil und im Falle einer erfolglosen Durchführung habe ich keinen größeren Nachteil, als wenn ich es nicht durchführte (Strategiebedingung)
und
iii) es gibt keinen hinreichend guten Grund, p für falsch zu halten[6] (Unschuldsbedingung)
und
iv) eine indizielle Rechtfertigung von p im Rahmen eines anderen kognitiven Projekts hilft mir nicht weiter, da ich Voraussetzungen machen müsste, die zunächst einmal genauso fragwürdig sind wie p und die darum ihrerseits gerechtfertigt werden müssten usw. ad infinitum. (Regressbedingung)

Die für eine epistemische Rechtfertigung geforderte epistemische Verantwortlichkeit ergibt sich gemäß dieser Konzeption der Idee nach aus dreierlei: Zum einen ist es nicht Mangel an epistemischer Sorgfalt, wenn ich p in Abwesenheit von Anhaltspunkten akzeptiere. Eine solche Rechtfertigung ist aufgrund der Regressbedingung nicht möglich. Zweitens urteile ich auch nicht wider besseres Wissen, da ich auch keinen Grund habe, p für falsch zu halten. (Unschuldsbe-

4 Wright (2004b): 191 f.
5 Zu dieser Besonderheit kommen wir später.
6 Wright (2004b): 191 schreibt: „We have no sufficient reason to believe that p is *untrue*". (Mein Kursivdruck) Daraus ergibt sich nur dann ein Unterschied zu meiner Formulierung, wenn man bestreitet, dass ein Satz/eine Aussage entweder wahr oder falsch ist.

dingung) Und drittens scheint es aus epistemischer Sicht eine gute Idee zu sein, p zu akzeptieren. Schließlich ist die Akzeptanz von p Voraussetzung für die positive Bewertung eines kognitiven Projektes, welches wiederum eine dominante Strategie bei meiner Suche nach der Wahrheit darstellt. (Voraussetzungs- und Strategiebedingung)

Ausgehend von dieser Konzeption epistemischer Berechtigung scheinen wir gute Karten in der Auseinandersetzung mit dem Skeptiker zu haben. Es sieht ganz so aus, als seien die genannten Bedingungen hinsichtlich der Annahme erfüllt, dass unsere Sinne uns verlässlich über eine Außenwelt informieren und wir folglich auch keine Gehirne im Tank sind.

Erstens: Mit Hilfe unserer Wahrnehmung wahre Überzeugungen über eine von uns unabhängig existierende Außenwelt zu erlangen, ist ein alltägliches empirisches Projekt. Und man kann es nicht rationalerweise für Erfolg versprechend halten, ohne dabei die Verlässlichkeit der Sinne und somit die Falschheit skeptischer Hypothesen vorauszusetzen.

Zweitens: Es scheint ganz so, als könnten wir nur mit Hilfe des alltäglichen empirischen Projektes überhaupt etwas über die Welt in Erfahrung bringen.

Drittens: Es gibt keinen Grund, unsere Wahrnehmung für unzuverlässig zu halten.[7]

Viertens: Die Verlässlichkeit der Wahrnehmung unter Berufung auf etwaige Anhaltspunkte zu rechtfertigen, setzt bereits die Annahme voraus, dass unsere Wahrnehmung verlässlich ist.

Der Skeptiker scheint somit widerlegt zu sein:

> Entitlement of cognitive project seems to promise well in addressing the challenge of Cartesian scepticism […].[8]

IV

Wie bereits gesagt halte ich dieses Argument aus verschiedenen Gründen allerdings für nicht überzeugend.

---

7 Das gesteht auch der Skeptiker zu. Er behauptet nur, dass es keinen Grund gibt, sie für verlässlich zu halten.

8 Wright (2004b): 105. Siehe auch Wright (2004b): 195. Dort heißt es: „It would follow in particular – provided the very idea of entitlement of project is in good standing – that in all circumstances where there is no specific reason to think otherwise, we are each entitled to take it, without special investigative work, that our basic cognitive faculties are functioning properly in circumstances broadly conducive to their successful operation. If so, that immediately empowers us to dismiss the various scenarios of cognitive dislocation and disablement – dreams, sustained hallucination, envatment and so on – which are the stock-in-the-trade of Cartesian scepticism."

Der erste Einwand hängt mit Wrights Unterscheidung zwischen Vertrauen und Überzeugtsein/Glauben zusammen. Für Wright ist Vertrauen eine schwächere Form des Fürwahrhaltens als Überzeugtsein/Glauben. Aufgrund seiner Bestimmung kognitiver Berechtigung als Berechtigung zu Vertrauen, nicht aber zu einem Überzeugtsein, lädt er den Skeptiker jedoch zu folgender Kritik ein:

(1) Wissen erfordert gerechtfertigte Überzeugungen.
(2) Die gerechtfertigte Akzeptanz einer These über die Außenwelt erfordert die gerechtfertigte Überzeugung, dass wir keine Gehirne im Tank sind.
(3) Wir sind nach wie vor nicht in der Überzeugung gerechtfertigt, dass wir keine Gehirne im Tank sind, selbst wenn wir darauf vertrauen dürfen, keine solchen Geschöpfe zu sein. (Denn kognitive Berechtigung liefert gerechtfertigtes Vertrauen, nicht aber gerechtfertigte Überzeugungen.)
(4) Also haben wir nach wie vor kein Wissen.

Wright weist die zweite Prämisse dieses Arguments zurück. Er geht zwar nach wie vor davon aus, dass unsere Überzeugungen über die Außenwelt nur dann gerechtfertigt sind, wenn wir gerechtfertigt sind zu glauben *oder* darauf zu vertrauen, dass unsere Sinne verlässlich sind. Es ist aber nicht unbedingt erforderlich, dass wir zu Recht in einem starken Sinne davon überzeugt sind, kein Gehirn im Tank zu sein.

Dieser Standpunkt hat jedoch folgende Konsequenz: Obwohl wir z. B. zwar gerechtfertigterweise glauben, eine Hand zu haben, und gerechtfertigterweise glauben, dass Gehirne im Tank keine Hand haben, dürfen wir nur darauf vertrauen, nicht aber glauben, dass wir keine Gehirne im Tank sind. Mit anderen Worten: Der „Glauben-Dürfen"-Operator ist Wright zufolge nicht immer geschlossen unter gerechtfertigter Implikation. Das halte ich – im vorliegenden Fall – für unplausibel. Wenn ich glauben darf, eine Hand zu haben, dann darf ich auch glauben, kein handloses Wesen und ipso facto auch kein handloses Gehirn im Tank zu sein.

Welches Argument führt Wright zugunsten seiner gegenteiligen Auffassung ins Feld? Aus dem wenigen, was er diesbezüglich sagt, lässt sich folgende Überlegung rekonstruieren:

(a) Indizielle Rechtfertigung ist nicht geschlossen unter gerechtfertigter Implikation.
(b) „Glauben-Dürfen" setzt eine indizielle Rechtfertigung voraus.
(c) Also ist „Glauben-Dürfen" nicht geschlossen unter gerechtfertigter Implikation.

Um die erste Prämisse zu begründen, greift Wright auf ein bekanntes Beispiel von Fred Dretske zurück. Ein Vater ist mit seinem Sohn im Zoo und beide sehen ein Zebra im Gehege stehen. Wright bewertet die Situation nun wie folgt: Der Vater ist aufgrund seiner visuellen Wahrnehmung gerechtfertigt, das Tier für ein Zebra zu halten. Schließlich sieht es aus wie ein Zebra. Der Vater ist aufgrund seiner Kenntnisse über Zebras und Maultiere auch in der Annahme gerechtfertigt, dass Zebras keine Maultiere sind. Aber, so Wright, der Vater ist nicht visuell in der Annahme gerechtfertigt, dass es sich bei dem Tier nicht um ein ange-

maltes Maultier handelt. Schließlich sieht das Tier genau so aus, wie es aussähe, wäre es ein angemaltes Maultier. Und folglich ist der Vater zwar hinsichtlich der Zebrahypothese indiziell gerechtfertigt, nicht aber hinsichtlich der Negation der Maultierhypothese.[9]

Gegen diese Ansicht spricht zweierlei. Zum einen gesteht Wright selber zu, dass die Wahrnehmung den Vater nur deshalb in der Zebrahypothese rechtfertigt, weil der Vater unter anderem auch gute Anhaltspunkte für die Auffassung hat, dass Zoos in der Regel keine angemalten Maultiere in ihren Gehegen haben. Wahrnehmung rechtfertigt uns immer nur relativ zu bestimmtem Hintergrundwissen.[10] Das heißt, auch wenn der Vater keine *visuelle* Rechtfertigung für die These hat, dass es sich nicht um ein angemaltes Maultier handelt, hat er doch *andere* Anhaltspunkte, die gegen die Maultierhypothese sprechen. Selbst wenn in der beschriebenen Situation *visuelle* Rechtfertigung nicht geschlossen ist, folgt also noch nicht, dass *indizielle* Rechtfertigung nicht geschlossen ist.

Zweitens teile ich die Ansicht nicht, dass zwar die Zebrahypothese, nicht aber die Negation der Maultierhypothese visuell gerechtfertigt ist. *Wenn* man aufgrund der äußeren Erscheinung das Tier für ein Zebra halten darf, dann darf man auch davon ausgehen, dass es kein anderes Tier und folglich auch kein Maultier ist.

Vor dem Hintergrund dieser Schwierigkeiten stellt sich somit die Frage, warum Wright nicht einfach bereit ist, kognitive Berechtigung für eine vollwertige Form der Rechtfertigung zu halten, so dass wir nicht nur darauf vertrauen, sondern auch in vollwertiger Weise davon überzeugt sein dürfen, keine Gehirne im Tank zu sein.

Der Grund ist folgender: Mit einer kognitiven Berechtigung ist aufgrund fehlender Hinweise auf die Wahrheit der entsprechenden These ein größeres Risiko verbunden als im Falle indizieller Rechtfertigung. Wright schreibt:

> [...] [A]fter all, entitled as we may be, the fact has not gone away that we have no evidence for C [e. g., that we are not brains in a vat].[11] (Mein Zusatz)

Ich teile diesen Vorbehalt. Aber möglicherweise rühren Wrights und meine Zurückhaltung daher, unter dem Jahrhunderte alten Einfluss einer Debatte zu stehen, die von einem unnötig starken Rechtfertigungsbegriff ausgeht. Aus diesem Grund möchte ich mich einer anderen Kritik an Wrights antiskeptischem Argument zuwenden.

9 Wright (2004b): FN 9 und Dretske (1970).
10 Wright (2002), (2003).
11 Wright (2004b): 209.

V

Folgt aus einer kognitiven Berechtigung im Sinne Wrights, dass es für das Erkenntnissubjekt epistemisch verantwortlich ist, skeptische Hypothesen für falsch zu halten? Ich bestreite das. Richten wir erneut unseren Blick auf die Strategiebedingung kognitiver Berechtigung, die besagt:

> ii) das kognitive Projekt P ist epistemisch hinreichend wertvoll für mich, insofern es epistemisch unerlässlich ist oder aber gilt: eine erfolgreiche Durchführung desselben ist aus epistemischer Sicht von Vorteil und im Falle einer erfolglosen Durchführung habe ich keinen größeren Nachteil, als wenn ich es nicht durchführte (Strategiebedingung)

Reicht es aus, dass das Erkenntnissubjekt mit einem kognitiven Projekt *faktisch* eine dominante Strategie verfolgt oder muss es auch Grund zu dieser Annahme haben? Ich optiere für die letzte Antwort. Wenn wir den Blick auf das eine Konzeption der epistemischen Berechtigung stützende Sokratesbeispiel richten, fällt Folgendes auf: In dem beschriebenen Fall ist es nur deshalb verantwortlich von Sokrates, den Handel mit der Pythia einzugehen und Goldbachs Vermutung für wahr zu halten, weil er gute Gründe für die Annahme hat, dass er im Gegenzug wahre Auskünfte über die Tugenden erhält. Hätte Sokrates diese Gründe nicht, dürfte er sich aus epistemischer Sicht nicht auf den Handel einlassen. Und Entsprechendes scheint hier zu gelten. Ohne einen Grund für die Annahme, dass die Durchführung des alltäglichen empirischen Projektes aus epistemischer Sicht eine gute Idee darstellt, darf ich es auch nicht für epistemisch angemessen und somit die Sinne auch nicht für verlässlich halten.

Dieser Punkt wird noch deutlicher, wenn wir uns vergegenwärtigen, dass die Akzeptanz einer These implizit immer auch die Zurückweisung ihrer Negation darstellt. Betrachten wir die folgenden zwei Thesen: 1. Ich bin kein Gehirn im Tank. 2. Ich bin ein Gehirn im Tank. Nehmen wir in Übereinstimmung mit Wright nun einmal an, dass die genannten vier Bedingungen kognitiver Berechtigung für die erste These erfüllt sind. Wie verhält es sich mit der zweiten These? Sie ist in dem von Wright spezifizierten Sinn Voraussetzung für folgendes kognitives Projekt: Interpretiere deine Erfahrung so, dass sie das Ergebnis einer Täuschung ist, so wie sie ein Gehirn im Tank erfährt. Diese Vorgehensweise kann man nur dann positiv bewerten, wenn man zugleich davon ausgeht, ein Gehirn im Tank zu sein. Die Voraussetzungsbedingung kognitiver Berechtigung ist somit auch für die skeptische Hypothese erfüllt. Gleiches gilt für die Unschuldsbedingung. Wir haben keinen Hinweis darauf, dass wir keine Gehirne im Tank sind, dass diese Hypothese also falsch ist.[12] Und schließlich lässt sich die Gehirn-im-Tank-Hypothese auch nur dann empirisch rechtfertigen, wenn wir bereits davon ausgehen, Opfer einer Täuschung zu sein. Die Regressbedingung ist somit ebenfalls erfüllt.

12 Hätten wir solche Hinweise, könnten wir uns Wrights Ausführungen sparen.

Der einzige Unterschied zwischen der skeptischen Hypothese und ihrer Negation besteht per Voraussetzung somit darin, dass die Negation Voraussetzung eines *faktisch* wertvolleren epistemischen Projektes ist. Obwohl *aus Sicht des Erkenntnissubjekts* die beiden Thesen somit gleich gut abschneiden, dürfen wir Wright zufolge die eine These der anderen vorziehen. Das entspricht aber nicht mehr der Konzeption epistemischer Verantwortlichkeit, wie ich sie eingangs definiert habe. Und folglich impliziert kognitive Berechtigung dann auch nicht mehr epistemische Rechtfertigung.

Man muss daher (mindestens) die zweite Bedingung kognitiver Berechtigung wie folgt abändern:

> ii) *ich habe Grund zu der Annahme,* dass das kognitive Projekt P epistemisch hinreichend wertvoll ist, insofern es epistemisch unerlässlich ist oder aber gilt: eine erfolgreiche Durchführung desselben ist aus epistemischer Sicht von Vorteil und im Falle einer erfolglosen Durchführung habe ich keinen größeren Nachteil, als wenn ich es nicht durchführte (Strategiebedingung)

Daraus ergibt sich für Wright jedoch ein Dilemma. Entweder er bestreitet, dass solche Gründe vorliegen. Dann sind wir nicht kognitiv berechtigt, skeptische Hypothesen für falsch zu halten. Oder aber Wright geht davon aus, dass wir solche Gründe haben. Es gilt jedoch: Um zu wissen, dass wir nur mit dem alltäglichen empirischen Projekt zu wahren Überzeugungen über die Außenwelt gelangen oder dass wir auf diese Weise nicht schlechter fahren als mit anderen Projekten, muss man bereits Gründe für die Auffassung haben, in einer normalen Welt zu leben. Denn wären wir Opfer einer perfekten Täuschung, führen wir besser damit, unsere sinnlichen Informationen nicht als Anhaltspunkte für eine von uns unabhängig existierenden Außenwelt zu deuten. Wir führen vielmehr besser damit, unsere sinnlichen Informationen als Anhaltspunkte für eine umfassende Täuschung zu betrachten. Das heißt aber: Wir sind nur dann kognitiv berechtigt, skeptische Hypothesen für falsch zu halten, wenn wir bereits Gründe für diese Auffassung haben. Wrights Konzeption kognitiver Berechtigung wäre somit als Antwort auf das Skeptizismusproblem überflüssig.

VI

Fassen wir zusammen: Selbst wenn man grundsätzlich bereit ist, den Begriff der epistemischen Rechtfertigung im Sinne epistemischer Verantwortlichkeit aufzufassen und neben indizieller Rechtfertigung auch eine epistemische Berechtigung als vollwertige Form der Rechtfertigung zu akzeptieren, hat Wright nicht gezeigt, dass wir epistemisch berechtigt und somit epistemisch gerechtfertigt sind, skeptische Hypothesen für falsch zu halten. Denn zu diesem Zweck muss er seine Bestimmung des Begriffs einer kognitiven Berechtigung in einer Weise

abändern, die seine Antwort auf das Skeptizismusproblem unvollständig oder überflüssig werden lässt.

Und selbst dann, wenn man Wrights Konzeption kognitiver Berechtigung ohne Einschränkung akzeptiert, kann es zwar sein, dass wir *faktisch* kognitiv berechtigt sind, skeptische Hypothesen für falsch zu halten. Aber Wright ist uns damit immer noch den *Nachweis* schuldig geblieben, dass dem so ist. Denn er hat nicht gezeigt, dass es eine gute Idee ist, die Erfahrung realistisch zu interpretieren und sich in einer normalen Welt zu wähnen. Er hat nicht gezeigt, dass es sich dabei um eine dominante Strategie handelt.

Literaturverzeichnis

Davies, Martin: "Epistemic Entitlement, Warrant Transmission and Easy Knowledge". Proceedings of the Aristotelian Society, Supp., 78, 2004. S. 213-245

Dretske, Fred I.: "Epistemic Operators". The Journal of Philosophy", 67, 1970. S.1007-1023

Jenkins, Carrie: "Entitlement and Rationality". Synthese, 157, 2007. S. 25-45

Pedersen, Nikolaj Jang: Entitlement in Mathematics. unveröffentlichte Dissertationsschrift. St. Andrews, 2005

Pritchard, Duncan: "Wittgenstein's On Certainty and Contemporary Anti-Scepticism". In: *Moyal-Sharrock, Danièle/Brenner, William H. (Hrsg.):* Readings of Wittgenstein's On Certainty. New York, 2005. S. 189-224

Schmoranzer, Sebastian: Realismus und Skeptizismus. Paderborn, i. V.

Williams, Michael: Unnatural Doubts – Epistemological Realism and the Basis of Scepticism. Princeton, 1996

Williams, Michael: Problems of Knowledge – A Critical Introduction to Epistemology. New York, 2001

Wright, Crispin: "Facts and Certainty." Proceedings of the British Academy, 71, 1985. S. 429-472

Wright, Crispin: "On Putnam's Proof That We are Not Brains in a Vat". In: *Clarke, Peter/Hale, Bob (Hrsg.):* Reading Putnam, Cambridge (MA), 1994. S. 216-241

Wright, Crispin: "(Anti-)Scepticism Simple and Subtle: G. E. Moore and John McDowell." Philosophy and Phenomenological Research, 65, 2002. S. 330-348.

Wright, Crispin: "Some Reflections on the Acquisition of Warrant By Inference". In: *Nuccetelli, S. (Hrsg.)*: New Essays on Semantic Externalism, Skepticism, and Self-Knowledge. Cambridge (Mass.), 2003. S. 57-77

Wright, Crispin : Wittgensteinian Certainties. In: *McManus, Denis (Hrsg.):* Wittgenstein and Scepticism, London, 2004a. S. 22-55

Wright, Crispin: "On Epistemic Entitlement – Warrant For Nothing (And Foundations for Free)?". In: Proceedings of the Aristotelian Society Supplementary, 78, 2004b. S. 167-211

Assertability or Truth-Values? Prospects for Pragmatic Invariantism

Erik Stei
stei@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität Bonn

Abstract/Zusammenfassung

In current epistemology it is by now hardly disputed that in some sense or other the acceptability of knowledge claims depends on context. Prima facie, what counts as knowledge in one context is denied that status in another. It is, however, very much disputed that, as contextualists or sensitive invariantists claim, the *truth-value* of an utterance depends on contextual factors like salience of error possibilities, interests or stakes.

In this paper, I examine an alternative account that resorts to Gricean pragmatics in order to explain shifting intuitions toward knowledge claims. I argue that so-called *warranted assertability manoeuvres* are implausible, but that there might be other versions of *pragmatic invariantism* that explain the intuitive context-sensitivity of knowledge claims. I suggest that an utterance of a knowledge ascribing sentences might be analysed as triggering conventional implicatures or, in case one is sceptic about this notion, as being multi-dimensional in the sense of Kent Bach's analysis of "but". This would allow for a strict invariantist semantic treatment while also accounting for ordinary language intuitions highlighted by semantic explanations.

1 Introduction

Like so many papers about context-dependency in epistemology, this one starts with a case study. Consider Stewart Cohen's airport case:

> (LO) Mary and John are at the L.A. airport contemplating taking a certain flight to New York. They want to know whether the flight has a layover in Chicago. They overhear someone ask a passenger Smith if he knows whether the flight stops in Chicago. Smith looks at the flight itinerary he got from the travel agent and responds, 'Yes I know– it does stop in Chicago.' (Cohen 1999: 58)

Assume that the flight really stops in Chicago and that Smith gave a sincere answer. Intuitively, Mary and John know that the flight stops in Chicago. But the story continues. Here's Cohen:

(HI) It turns out that Mary and John have a very important business contact they have to make at the Chicago airport. Mary says, 'How reliable is that itinerary? It could contain a misprint. They could have changed the schedule at the last minute.' Mary and John agree that Smith doesn't really know that the plane will stop in Chicago. They decide to check with the airline agent. (Cohen 1999: 58)

Assume that all other features of the situation remain unchanged. Mary and John are in the same epistemic position as before, but – intuitively again – they do not know that the flight stops in Chicago. The intuitions regarding this case are fairly clear. Call them the *initial intuitions*. They can be further divided into the *attribution intuition* concerning the correctness of Mary's knowledge attribution in LO and its incorrectness in HI and the *denial intuition* concerning the correctness of Mary's denial of knowledge in HI and its incorrectness in LO (cf. Brown 2005b). It is less clear, however, how to explain these phenomena.[1]

Contextualists (e.g. Cohen 1999, DeRose 1995, Lewis 1996) notoriously claim that the extension of the verb 'know' depends on the epistemic standards operative in the context of utterance. Thus, because in different contexts the same claim can express different propositions, the seemingly contradictory intuitions can be explained semantically. In the first scenario (LO) Mary's and John's as well as Smith's standards are relatively low. In the formers' ascription as well as the latter's self-ascription of 'knowledge' the content of 'knows' reflects these low standards. By looking at his itinerary Smith meets those standards, while Mary and John meet them by testimony. The propositions expressed by the 'knowledge' claims are true. In the second scenario (HI), Mary's and John's standards rise, causing a different content of 'know' to be expressed by their denial of 'knowledge'. The information delivered by Smith does not meet those standards. The proposition expressed by Mary's and John's denial of 'knowledge' is true. Context makes a difference in truth-values.

Insensitive (or classical) invariantists[2] (e.g. recently, Black 2005, Brown 2005a, Rysiew 2001, Williamson 2000, 2005a) dispute that factors like the salience of error possibilities or the practical interests of a subject or an ascriber have an impact on the truth-value of knowledge claims or ascriptions. Still, the initial intuitions are to be explained. A prominent strategy is to resort to factors that are usually filed under the label *pragmatics*. Given a basic pragmatic apparatus (see § 2), depending on the context in which the assertion is made, the utterance of a sentence may convey additional aspects of meaning that are not subject to truth-conditional evaluation. Accordingly, because of this conveyed meaning, an assertion that is conversationally appropriate in one context may not be so in another. Note that this is independent of whether the assertion itself expresses a truth or a falsehood. The utterance of a falsehood can convey a truth

1 I will disregard the assertion intuitions and the practical reasoning intuitions in this paper.
2 I will not address sensitive invariantism (cf. Hawthorne 2004, Stanley 2005) in this paper.

(consider figurative uses of language, e.g. irony) and *vice versa* (by asserting the truth 'There is a gas station around the corner' I can falsely implicate that it's open and selling gas).

Applying these observations to the airport case, some insensitive invariantists, call them *pragmatic invariantists* (PI), claim that although the proposition expressed by the knowledge ascriptions as well as its truth value is the same in both cases, there is a way to explain the *initial intuition*s, namely by means of the *appropriateness* of the claims. Sceptical invariantists hold that both Mary's knowledge ascription and Smith's self-ascription are false, but that in LO it is conversationally appropriate to assert them. Mary's denial of knowledge in HI is true and conversationally appropriate. Anti-sceptical invariantists hold that both Mary's knowledge ascription and Smith's self-ascription are true, but that in HI it is conversationally inappropriate to assert them. Mary's denial of knowledge in HI is false but conversationally appropriate. Thus, context makes a difference in assertability.

It is the thesis of PI that I want to examine in this paper. In what follows I will sketch the basic ideas concerning non truth-conditional aspects of meaning (§ 2) before focusing on some applications to epistemology (§ 3). I will then draw some conclusions and assess the prospects of such an explanation (§ 4).

2 Pragmatic meaning

Many of the fundamentals now discussed in debates about the semantics/ pragmatics distinction go back to the influential work of Paul Grice (1989b). In contrast to more radical pragmatists, Grice did not dispense with truth-conditional semantics as provided by classical logic, but intended to analyse the broader notion of *speaker meaning* of which truth-conditional meaning is an important part. Here's the conception Grice had in mind:

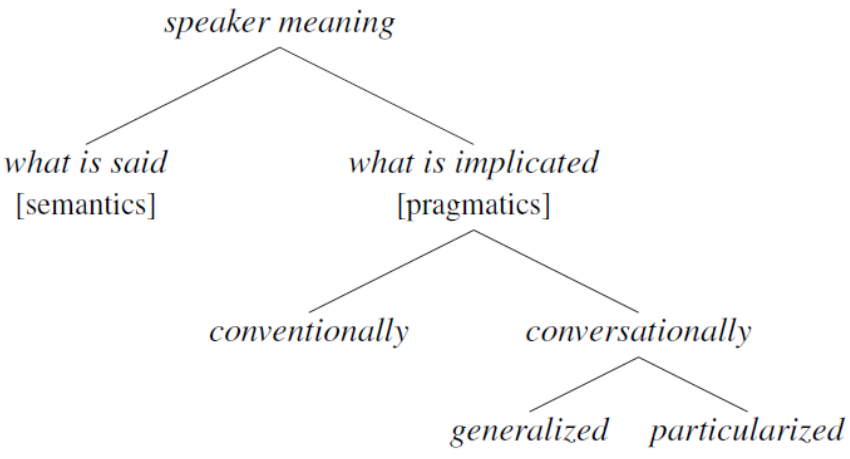

Figure 1: Grice's taxonomy of meaning

While *what is said* represents the level of truth-conditional meaning (i.e. syntax plus compositional semantics with indexicals fixed and rid of ambiguities), the branch of *what is implicated* aims at explaining those kinds of meaning that are not explicitly uttered (in the sense of *what is said*) but are nonetheless communicated by an utterance.

In case of conventional implicature these are attached to certain words. The notorious examples are *but* and *therefore*. Semantically, it is argued, (1) simply expresses a conjunction.

(1) Alice is a philosopher but smart.

The proposition it expresses is true iff Alice is a philosopher and Alice is smart. Pragmatically, however, (1) conveys that usually the properties of *being a philosopher* and *being smart* preclude one another.

Conversational implicatures, on the other hand, are processed by means of conversational maxims on basis of the sentence uttered in a certain context. It is important to keep in mind that *what is said* serves as the input for those derivations. One needs semantics in order to get the pragmatic machinery going in the first place. This machinery is represented by four conversational maxims and the following general cooperative principle, developed in Grice 1989a:

(CP) *Cooperative Principle*
'Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.' (Grice 1989a: 26)

The conversational maxims are *quantity*, *quality*, *relation*, and *manner*. A brief recapitulation might look like this: *quantity* requires the speaker to make his contribution as informative as possible, but not more informative than necessary, *quality* demands the speaker to make a true contribution, based on her evidence, that she does not believe to be false. *Relation* requires the speaker to 'be relevant' – her contribution should match the score of the conversation – and *manner* demands brevity, the avoidance of ambiguity, perspicuity, and the correct order. In contrast to the conventional aspects of meaning the maxims are independent of language. Rather, they represent more basic principles of rational conversational interchange.

We can now see how invariantists can make use of this conception. Semantically, i.e. on the level of *what is said*, the verb 'know' makes a constant contribution – it is invariant, because it always expresses the same relation between a subject *S*, a proposition *p*, and a time *t*. Accordingly, once its variables are fixed a sentence (2) expresses the same proposition, independent of what error-possibilities are salient or what is at stake for the subject/the ascriber. Thus, in compliance with classical semantics (e.g., Kaplan 1989), it is assigned the same truth-value given a fixed world of evaluation.

(2) *S* knows *p* (at *t*).

Pragmatically, however, i.e. on the level of *what is implicated*, the utterance may convey another proposition with a different truth-value. There are several ways to spell out this idea in detail. Some of them will be sketched in what follows. In case one of these strategies is successful, it is easy to see an apparent methodological advantage for an insensitive invariantist: She would have found a way to explain our intuitions without being forced to posit contentious semantic assumptions (e.g. the indexicality of 'know') or the influence of practical interests on epistemic terms (i.e. what Stanley (2005) calls anti-intellectualism).

3 Pragmatic Invariantism

Invariantists are not committed to a specific strategy against the *initial intuitions*. The basic options are to reject (some of) the intuitions as mistaken (e.g., in case of a sceptical invariantist, that Mary appropriately ascribes knowledge in LO) or to resort to a pragmatic explanation. Pragmatic invariantism (PI) chooses the latter. There is a variety of pragmatic answers[3], but given the classical Gricean framework sketched in § 2, PI can either claim that the intuitions are triggered via conventional features of the word 'know' or that they are due to the adherence and/or exploitation of the conversational maxims. It is the latter option that is more prominently discussed in the literature.

3.1 Warranted Assertability Manoeuvres

Patrick Rysiew (2001) develops a sophisticated invariantist explanation of the *initial intuitions* making use of a so called *warranted assertability manoeuvre* (WAM). He relies on a *relevant alternatives* framework, but it involves some modifications, most importantly the distinction between relevant and salient alternatives. He suggests understanding *relevant* alternatives as those which *S* has to be able to rule out in order to meet the invariant epistemic standards for knowledge. *Salient* alternatives, on the other hand, are those that are discussed in a context 'however idiosyncratic, unusual, and peculiar to the conversational setting.' (Rysiew 2001: 489) Quite often, the two do not coincide.

Analogous to this distinction, there are two kinds of meaning of an utterance: what is being $\text{said}_{\text{strict}}$ and what is being $\text{said}_{\text{loose}}$. While, roughly, the former is supposed to correlate with the meaning expressed semantically, i.e. with *what is said,* the latter corresponds to *what is implicated*. Thus, in uttering '*S* knows *p*'

3 For instance, Bach 2005 proposes a so-called belief removal explanation. For criticisms see Brown 2006 and Hawthorne 2004.

the speaker $says_{strict}$ that *S* can rule out all relevant not-p alternatives and $says_{loose}$ that *S* can rule out all salient not-p alternatives (cf. Rysiew 2001: 486).

With this distinction in mind we can now approach the *attribution intuitions* of the airport case. By asserting '*S* knows that the plane has a layover in Chicago' I say_{strict} that *S* can rule out all relevant alternatives in which the plane does not have a layover in Chicago. I say_{loose}, however, that *S* can rule out all conversationally salient alternatives. Thus, although I say_{strict} something true in LO and in HI, I only say_{loose} something true in LO, but not in HI. With respect to the *denial intuition*, on the other hand, I say_{strict} the falsehood that *S* cannot rule out all relevant alternatives to the plane having a layover in Chicago. Yet I say_{loose} something true, namely that *S* cannot rule out all salient alternatives.

Leaving aside comments on the epistemic notions of this approach, let us now see how Rysiew constructs his WAM. The basic idea is to resort to the conversational maxim of relation, i.e. the submaxim *Be relevant.* The meaning conveyed is something along the following lines:

(3) The speaker is able to rule out all not-p alternatives salient in the context in which the knowledge claim is uttered.

If Mary asserts that Smith knows that the plane has a layover in Chicago, implicating that he can rule out all salient not-p alternatives, her assertion is relevant at the stage of the conversation in LO. Only the plane's having a layover or not is at issue, the itinerary meets the standards for deciding on that question, so Mary made a cooperative contribution. The same assertion in HI again implicates that Smith can rule out all salient not-p alternatives. In these circumstances, however, Mary's assertion would be misleading because, by hypothesis, Smith cannot rule out the alternatives salient at this stage of the conversation, e.g. the itinerary containing a misprint. So, although Mary would say something literally true, her utterance would be irrelevant, because a different question is at issue. By saying something true but irrelevant she would give rise to a false implicature as by CP and *relation* the hearer expects her to be relevant. According to the Gricean model, competent speakers have reciprocal knowledge of CP and the maxims, which leads to Mary's denial of knowledge in HI, because she is aware of the false implicature the attribution of knowledge would trigger. As a first approximation, this should give an impression of how PI can deal with the *initial intuitions.*

Brown (2006) proposes a slightly different strategy that incorporates DeRose's conception of epistemic strength in terms of spheres of possible worlds, but the conversational mechanism is parallel to Rysiew's. Black (2005) argues that it is more convincing to resort to the maxim of *quantity*, whereas Leite (2005) takes an explanation involving *quality* to be most promising (although he thinks it fails). In principle, however, the knowledge WAM (K-WAM) appeals to a general conversational mechanism, i.e. one of the Gricean maxims.

Keith DeRose challenged the K-WAM by arguing that it fails to satisfy the following constraints he imposed on WAMs in general (cf. DeRose 2002: 202, fn. 39):

> (C-WAM) *Conditions on successful warranted assertability manoeuvres*
> A successful WAM seeks to (i) explain away an appearance of nontruth in case of a conflict of intuitions. It does so (ii) by means of a general conversational rule. The appearance of non-truth that it seeks to explain away is (iii) balanced by a similar appearance of non-truth that attaches to the opposite statement (Check the negation!).

However, C-WAM is controversial. For instance, Brown rejects condition (i), according to which a successful WAM only explains away intuitions of falsity (cf. Brown 2006: 414-419). And indeed DeRose's requirement seems to be biased. It was already pointed out in § 1 that in figurative uses of language (e.g. irony but also metaphor and hyperbole) a literally false sentence may be appropriate because it conveys a truth. Moreover, as Brown diagnoses, it is 'part of a standard and well-entrenched approach in the philosophy of language' (Brown 2006: 415) that this can happen in non-figurative speech as well. She backs her claim by citing the incorporation of Donnellan's *referential use* of definite descriptions into a Russellian framework by means of pragmatic mechanisms as well as Bach's theory of *implicitures*. Both act on the assumption that uttering a literally false sentence can convey a truth. So there is nothing unusual or *ad hoc* about the rejection of DeRose's condition (i).

DeRose's appeal to a general conversational rule in condition (ii) only makes sense if one takes the implicated meaning go be non-conventional. Thus, C-WAM as well as the invariantists' appeal to Gricean conversational maxims suggest that a WAM is conceptually located at the branch of conversational implicature in Figure 1.

If this is correct, (iii) is equally problematic. It is true that in many cases the negation of the utterance in question triggers the same implicature as the original utterance. If I say that I won all games of chess against Bobby, this implicates that Bobby and I played at least one game. The same holds in case I say that I did not win all games against Bobby. However, there are standard cases of conversational implicature that behave differently: Assume that I ask you whether you will go to the party tonight. If you respond 'I have to work' I can infer by standard Gricean reasoning that you will not go to the party. But assume your answer is 'I don't have to work'. It is not at all clear that this will still convey that you will not go to the party, although this is exactly what DeRose's constraint would require. Rather it is quite natural to assume that you will in fact go to the party.[4]

4 Examples of this kind illustrate that there are indeed cases in which an utterance of "p" conveys that q, while an utterance of "Not-p" conveys that not-q. This parallels the pragmatic invariantist's story, according to which "S knows p" conveys that S can rule out all

So, taking condition (ii) seriously, (i) and (iii) do not seem to state reasonable constraints on WAMs as they exclude clear and uncontentious cases of the phenomenon in question. However, we have the possibility to consult an alternative and less controversial set of constraints in order to avoid inflationary postulation of the phenomenon, namely the ones Grice introduced himself.

3.2 'Know' and Conversational Implicature

The following basic features of conversational implicature are deducible from Grice's seminal 1989a: *calculability, reinforceability, cancellability, context-dependency,* and *nondetachability*. We can now check whether the K-WAM meets those conditions.

Calculability requires that the hearer can in principle infer the *speaker meaning* by means of *what is said*, CP and the conversational maxims. Consider this standard example for a conversational implicature (cf. Grice 1989a: 32) due to adherence of the maxim of relation:

(4) a. [A:] I am out of petrol.
b. [B:] There is a garage round the corner.
+> The garage is open and selling petrol.

Now, in order to derive the implicature, A might reason, roughly, as follows:

(5) a. There is no direct relation between my utterance and the literal meaning of B's response.
b. Usually, however, garages sell petrol.
c. B would be infringing the maxim of relation if I could not get petrol at that garage round the corner.
d. B knows that I can infer this from her utterance and she did not stop me from inferring it.
e. So B wanted to implicate that the garage is open and selling petrol.

Leite (2005: 112-116) provides some reasons to doubt that the K-WAM can provide an analogous inference. Here's the most challenging one: Because in constructing an inference it is required to assume ideal speakers and hearers that are not semantically confused, it is not clear how to bridge the gap from (6d) to the conclusion (6f):

(6) a. By uttering 'S doesn't know p' the speaker said that S cannot rule out all relevant alternatives to p.
b. But S can rule out all relevant alternatives, thus, S knows *p*, and the speaker knows that.
c. So the speaker deliberately uttered a falsehood and she knows that I can infer this from her utterance.

---

salient alternatives while "S does not know p" conveys that S cannot rule out all salient alternatives.

d. So she must have meant something else.
e. …
f. Therefore, the speaker communicated that S cannot rule out the salient alternative q that entails not-p, but which S does not have to be able to exclude in order to know p.

The central problem is, Leite argues, that for justificatory reasons we 'cannot appeal to ancillary knowledge of what speakers usually use "know"-attributions or denials to impart' (2005: 113). Assume that some form of moderate invariantism is correct and that both parties to the conversation know that. It is then hard to see how the hearer can infer (6f) without also assuming that the speaker is confused about 'know' (in thinking that it involves ruling out salient instead of relevant alternatives). This, however, is not an assumption the pragmatic invariantist can make, as calculability demands an idealized derivation against the background of a conversation involving only competent speakers. Advocates of K-WAM will have to provide an answer to this problem and be more explicit about the inferential process. However, let us see how the proposal deals with the other requirements.

What about reinforceability? According to that condition one should be able to make the implicature explicit without redundancy, i.e. it should be appropriate to assert (7a) in LO and (7b) in HI.

(7) a. S knows that the plane has a layover in Chicago. S can rule out the possibility that it is a non-stop flight.
b. S doesn't know that the plane has a layover in Chicago. S cannot rule out the possibility that the itinerary contains a misprint.

The reinforcements seem to be appropriate, which is in accordance with the predictions of the K-WAM. The knowledge claim in (7a) is true as is the implicature it is supposed to trigger. The knowledge claim in (7b) is false, but the implicature it allegedly triggers is true. Thus, although literally false, (7b) is felicitous, *reinforceability* is met.

The cancellability test, however, is less clear. Rysiew 2001 addresses it himself arguing that the implicature conveyed by knowledge claims *can* be cancelled, even though sometimes not quite comfortably. Rysiew finds utterances like (8) unproblematic, but admits that others might not.

(8) I know *p*, but of course I cannot rule out the *bizarre* alternatives to *p*.

However, intuitions seem to be less permissive if one applies the test to more concrete examples. Even if one takes (9a) to be felicitous, (9b) sounds pretty odd:

(9) a. S knows that the plane has a layover in Chicago, but S cannot rule out the possibility that she is a brain in a vat.
b. S knows that the plane has a layover in Chicago, but S cannot rule out the possibility that the itinerary contains a misprint.

Of course, these examples do not provide compelling arguments against an otherwise plausible view. It is the defence against the shaky intuitions regarding (9) which reveals a more important point: Rysiew proposes the following 'general rule': '[T]he more universal the implicature, the less likely it is that it can be cancelled without 'discomfort'' (Rysiew 2001: 496). The rule is certainly not general. Scalar quantity implicatures, one of the most general types of conversational implicature, can be cancelled without oddity as (10b) shows:

(10) a. Some students came to the lecture.
+> Not all students came to the lecture.
b. Some students came to the lecture, maybe even all of them came.

If K-WAM were correct, the implicature in question would indeed be very universal in character as it is supposed to be triggered, in its general form (3), by *every* knowledge claim. However, I can hardly see how this, combined with the cancellability problems, strengthens the case for *conversational* implicature. It rather seems that the meaning conveyed, even granted that it is pragmatic, is far too universal to be a conversational implicature in the first place.

This suspicion is substantiated by the test for context-dependency. It requires that an implicature arises in some contexts, but not in others. The implicature that the speaker can rule out all salient alternatives, however, arises in every context. This is analogous to a different kind of implicature: *but* implicates *conventionally* in every context that the arguments applied to it usually preclude one another. Conversational implicatures normally do not behave like this. Compare knowledge claims with the standard case of conversational implicature presented in (4) and its modification below:

(11) A: I am gathering data for the Yellow Pages. Is there any business in your area?
B: There is a garage round the corner.

Nothing in (11) corresponds to the implicature generated in (4). The alleged conversational implicature 'I can rule out all salient not-*p* alternatives', in contrast, arises in every context. Now, this does not show that 'know' is not context-dependent, but it suggests that there is something conventional going on, which speaks against the kind of implicature postulated by K-WAM.

Similar conclusions are suggested by the nondetachability test, which requires, roughly, that in case an expression e_1 is replaced by another expression e_2 with the same semantic meaning, the implicature should arise nonetheless. It is a question far beyond the scope of this paper whether there is an expression e_k that has the same semantic meaning as *know*, but let us assume that there is. The closest candidate seems to be another non-gradable, factive verb phrase like, e.g.

being aware or, if you like, the notorious *justified true belief*.[5] The test would then be as follows: If (12a) triggers the implicature that *S* can rule out all salient not-*p* alternatives, the same should apply for (12b) and (12c).

(12) a. S knows that the plane has a layover in Chicago.
b. S is aware that the plane has a layover in Chicago.
c. S has the justified true belief that the plane has a layover in Chicago.

Neither of the two seems to involve the requirement that *S* be able to rule out all alternatives salient in a conversational context. By looking at his itinerary Smith might form a justified true belief that the plane has a layover. The intuitive ascription of this condition does not seem to change in case error possibilities are raised, i.e. even if he cannot rule out the possibility that his itinerary contains a misprint this does not change his having a justified true belief. Things are less obvious with (12b), but it seems to me that being aware of something does not intuitively require the ability to rule out given error possibilities. Thus, the implicature does not arise; it is detachable, which also speaks against the K-WAM.

Let me recap the results of the tests: Reinforceability is met. Calculability, however, is problematic. Cancellability fails, and the explanation provided for that failure, together with the negative results of the context-dependency and nondetachability tests, suggests that 'know' does not trigger a conversational implicature. Maybe the K-WAM fails because it focuses on the wrong kind of pragmatic meaning. Yet, this does not mean that PI fails as there are alternatives.

3.3 'Know' and Conventional Implicature

The tests presented in § 3.2 suggest that the *initial intuitions* might be due to the conventional meaning of the word 'know'. It depends on how one draws the semantics/pragmatics distinction whether one takes this to speak in favour of a purely semantic account. According to the Gricean picture assumed here, PI can still appeal to the notion of conventional implicature to make its case. It could be argued that, although sentences containing 'know' are multi-dimensional in the sense that they always express two distinct propositions, only one of these propositions is subject to truth-conditional evaluation. This would then be the proposition expressed by the constant knowledge relation, whereas the propositions liable for the *initial intuitions* are the ones conveyed pragmatically, analogous to the preclusion effect of 'but'.

Generally, it is controversial what to think about *conventional implicature* (CI). Bach (1999) denies their very existence, although he allows for multi-dimensional sentences in the sense sketched above. Potts 2005 aims at rehabilitating the notion, though not for the classical examples like 'but' and 'therefore'.

5 Note, that if there was no such expression this would make the case even harder for the K-WAM.

However, according to the classical picture, the criteria for CI are usually taken to be, roughly, as follows (cf. Grice 1989a: 25 and Potts 2005: 11):

(13) a. CIs are part of the conventional meaning of words and therefore context-independent.
b. CIs are detachable.
c. CIs are speaker oriented.
d. CIs are logically and compositionally independent of what is said.

We have seen in the last paragraph that the meaning conveyed by 'know' as formulated in (3) seems to be context independent and detachable. (13c) is more problematic. At first sight, this criterion is related to the contextualist thesis that the speaker's context is crucial for determining what is expressed by a given knowledge claim. PI is on par with contextualism here. However, Bach (1999: 339) showed that the textbook examples for CIs can occur straightforwardly in indirect quotation. This means that in (14a) 'but' is attributed to the person quoted and *not* to the speaker. The same seems to hold for embedded knowledge claims as in (14b):

(14) a. Marv$_i$ said that Shaq is huge *but*$_i$ that he is agile.
b. Mary$_i$ said that Smith *knows*$_i$ that the plane has a layover in Chicago.

Intuitively, the implicature applies to the alternatives salient in Mary's context and not in the context of the speaker of (14b), but this should not be the case if the CI thesis was correct. According to this test, neither 'but' nor 'know' should be a CI trigger.

Yet, there are several reasons not to take the requirement of speaker orientation too serious: Grice himself did not explicitly consider embedded occurrences of CI triggering expressions, neither did he explicitly mention *speaker orientation* as a condition. Moreover, in the current discussion it has been suggested that even clearer cases of CIs are not necessarily speaker oriented as the following example by Kratzer (cited in Potts 2005: 162), shows:[6]

(15) My father screamed that he would never allow me to marry *that bastard* Webster.

So (13c) does not seem to be enough to reject the CI thesis *per se*. Besides, the problem of speaker orientation threatens the contextualist explanation as well, so this is no criterion to decide between the views.

What does (13d) mean with respect to knowledge claims? Grice uses a negation test to show that 'but' is independent of *what is said*. As pointed out in § 2, the truth of (1) seems to depend only on the truth of its conjuncts and not on whether they in fact preclude each other. Still, (1) is odd in case *being a philosopher* and *being smart* is not preclusive. By parallel reasoning, '*S* knows *p*' is true, but odd in HI. Just as in the case of 'but', competent speakers are likely to

6 The CI meaning is italicized.

mistake the pragmatically conveyed meaning with its truth-conditions. This leads to the *denial intuitions* as speakers mistakenly think that an utterance of '*S* knows *p*' is false in HI. For the same reason they take its denial to be true. This answer might appear to be question begging as this is just the main thesis of the alternative approach for PI I proposed. However, it is a difficult task to prove what exactly the semantic meaning of a given word is and it is not at all clear that competent speakers can always decide whether a certain aspect of meaning is encoded semantically or conveyed pragmatically. Until neither contextualism nor PI has any substantial answer to this problem the decision between the two views has to be made on independent grounds.

A more serious problem for the conventional implicature thesis is that in the case of 'but' speakers are aware of the additionally conveyed meaning, that is, that the conjuncts preclude one another. Apparently, the case is different with 'knows'. The notions of relevant or salient alternatives are first and foremost theoretical. It does not seem very plausible to claim that the everyday use of 'knows' suggests the distinctions in question. But, on the other hand, an alleged indexicality of 'knows' is not very intuitive either as others (e.g. Hawthorne 2004 and Stanley 2005) have pointed out. Again, it seems that PI and contextualism are on par here.

However one decides on these issues, my aim in this paper was to show how PI might accommodate the *initial intuitions*. A theory of knowledge ascriptions that explains all intuitions speakers seem to have about these questions in a completely satisfactory way is still to be developed.

4 Conclusion

I argued that it is possible to defend a strict version of invariantism without being committed to a naïve rejection of all *initial intuitions* as mistaken. Although the construction of a *warranted assertability manoeuvre* was shown to be problematic because the knowledge-WAM flunks central tests for conversational implicature, there appears to be another strategy. Although many points deserve to be spelled out in greater detail, I argued that *prima facie* nothing speaks against resorting to the notion of conventional implicature in order to explain the *initial intuitions*. If this is correct, an important motivation for epistemic contextualism loses much of its bite.

References

Bach, Kent: "The Myth of Conventional Implicature". Linguistics and Philosophy, 22, 1999. S. 327–366

Bach, Kent: “The Emperor’s New ‘Knows’”. In: *Preyer, Gerhard/Peter, Georg (Hrsg)*: Contextualism in Philosophy. Knowledge, Meaning and Truth. Oxford University Press, Oxford, 2005. S. 51–89

Black, Tim: “Classic Invariantism, Relevance and Warranted Assertability Manoeuvres”. The Philosophical Quarterly, 55(219), 2005. S. 328–336

Brown, Jessica: “Adapt or Die: The Death of Invariantism?”. The Philosophical Quarterly, 55(219), 2005a. S. 263–285

Brown, Jessica: “Comparing Contextualism and Invariantism on the Correctness of Contextualist Intuitions”. In: *Blaauw, Martijn (Hrsg)*: Epistemological Contextualism. Rodopi, Amsterdam/ New York, 2005b. S. 71–99

Brown, Jessica: “Contextualism and Warranted Assertability Manoeuvres”. Philosophical Studies, 130, 2006. S. 407–435

Cohen, Stewart: “Contextualism, Skepticism, and the Structure of Reason”. Noûs, 33, 1999. S. 57–89

DeRose, Keith: “Solving the Skeptical Problem”. The Philosophical Review, 104(1), 1995. S. 1–52

DeRose, Keith: “Assertion, Knowledge, and Context”. The Philosophical Review, 111(2), 2002. S. 167–203

Grice, Paul: “Logic and Conversation”. In: Studies in the Way of Words. Harvard University Press, Harvard, 1989a. S. 22–40

Grice, Paul: Studies in the Way of Words. Harvard University Press, Harvard, 1989b

Hawthorne, John: Knowledge and Lotteries. Clarendon, Oxford, 2004

Kaplan, David: “Demonstratives”. In: *Almog, Joseph/Perry, John/Wettstein, Howard (Hrsg)*: Themes From Kaplan. Oxford: Oxford University Press, 1989. S. 481–563

Leite, Adam: “Some Worries for Would-be WAMmers”. In: *Martijn Blaauw (Hrsg)*: Epistemological Contextualism. Rodopi, Amsterdam/ New York, 2005. S. 101–125

Lewis, David: “Elusive Knowledge”. Australasian Journal of Philosophy, 74, 1996. S. 549–567

Potts, Christopher: The Logic of Conventional Implicatures. Oxford University Press, Oxford, 2005

*Rysiew, Patrick: "*The Context-Sensitivity of Knowledge Attributions". Noûs, 35(4), 2001. S. 477–514

Stanley, Jason: Knowledge and Practical Interest. Oxford University Press, Oxford, 2005

Williamson, Timothy: Knowledge and its Limits. Oxford University Press, Oxford, 2000

Williamson, Timothy: "Contextualism, Subject-Sensitive Invariantism, and Knowledge of Knowledge". The Philosophical Quarterly, 55(219), 2005a. S. 213–235

Williamson, Timothy: "Knowledge, Context, and the Agent's Point of View". In: *Preyer, Gerhard/Peter, Georg (Hrsg):* Contextualism in Philosophy. Knowledge, Meaning and Truth, Oxford University Press, Oxford, 2005b. S. 91–114

3 Sprachphilosophie

Kontextualismus: seine Konsequenzen und Widersprüche

Sarah-Jane Conrad
conrad@philo.unibe.ch
Universität Bern, Schweiz

Abstract/Zusammenfassung

Contextualism within the realm of philosophy of language takes the context to be highly important when it comes to determine the semantic, i.e. truth-evaluable content of a linguistic utterance. The question if the context-influence can by handled with systematically, and how it can be theoretically modelled is the bone of contention of an ongoing debate. While representatives of a more radical form of contextualism deny the very possibility of a systematic approach to context and its influence upon uttered sentences, less radical contextualists believe it feasible. Accordingly, the latter plea for a linguistically orientated solution, either anchored in the syntax or the semantic of a language and thereby rejecting a mere pragmatic approach. In any case, all positions operate with a minimal linguistic meaning input. However, if the concept of meaning is totally separated from truth, the very claim that linguistic meaning plays a substantial role for determining the content of an utterance becomes invalid. While trying to save the concept of linguistic meaning, we see that contextualism can't really harm the project of a truth-conditional semantic.

Der Kontextualismus in der Sprachphilosophie betont den starken Einfluss des Kontextes für die Bestimmung des semantischen, d.h. wahrheitswertfähigen Gehaltes von sprachlichen Äußerungen. Ob der Kontexteinfluss systematisch erfassbar ist und wie dieser allenfalls modelliert werden muss, wird kontrovers diskutiert. Während Vertreterinnen eines radikaleren Kontextualismus die Möglichkeit einer systematischen Handhabung rundweg bestreiten, räumen gemäßigtere Kontextualisten diese ein. Entsprechend streben letztere keine pragmatische, sondern eine von sprachtheoretischen Überlegungen geleitete Lösung an, die entweder syntaktisch oder semantisch verankert ist. Trotz aller Unterschiede operieren alle drei Positionen mit einem minimalen sprachlichen Bedeutungsbeitrag. Falls jedoch eine konsequente Ablösung des Wahrheitsbegriffs vom Bedeutungsbegriff vorgeschlagen wird, droht die Annahme eines solchen minimalen Bedeutungsbeitrags unplausibel zu werden. Der Versuch, den Begriff der sprachlichen Bedeutung zu retten, zeigt, dass die kontextualistische Kritik das Projekt einer traditionellen wahrheitskonditionalen Semantik tatsächlich nicht gefährdet.

1. Die Rolle des Kontexts neu bewertet

Dem Kontext wird bei der Bestimmung des semantischen, d.h. des propositionalen bzw. des wahrheitswertfähigen Gehaltes eines geäußerten Satzes seit einigen

Jahren eine immer größere Bedeutung beigemessen.[1] Traditionell gesehen beschränkte sich seine Rolle darauf, bei der Äußerung eines syntaktisch wohlgeformten Satzes mit ambigen Ausdrücken diese zu desambiguieren (siehe Beispiel 1), oder allfällige Bezüge indexikalischer oder demonstrativer Ausdrücke eines Satzes festzulegen (siehe Beispiel 2).

(1) Die Bank ist da vorne.
(2) Das da ist Petra.

Mittlerweile wird dem Kontext jedoch eine immer größere Rolle zugewiesen.[2] Es wird angenommen, dieser sei für die Festlegung des semantischen Gehalts sehr unterschiedlicher Sätze unabdingbar, wie die beiden folgenden illustrieren:[3]

(3) Peter ist groß.
(4) Die Teekanne ist schwarz.

So wird beispielsweise behauptet, der Satz (3) sei bei seiner Äußerung semantisch unvollständig, obwohl er syntaktisch wohlgeformt ist und sein Bezug bekannt sind. Begründet wird diese Behauptung mit dem Hinweis, die Bedeutung des Ausdrucks „groß“ lege die relevante Vergleichsklasse nicht fest, zu der Peters Größe in Beziehung gesetzt werden müsse: Sind es Schulanfängerinnen von sechs Jahren oder aber Tennisspieler der Profiliga? Erst mit diesen zusätzlichen Angaben ist entscheidbar, ob der geäußerte Satz wahr oder falsch ist gemäß einer weit verbreiteten Meinung.

Ebenfalls syntaktisch wohlgeformt ist der Satz (4). Verschiedene Autorinnen[4] sind aber der Meinung, dass nicht nur bekannt sein muss, von welcher Teekanne die Rede ist und ob diese schwarz ist oder nicht, sondern auch, ob sie die Farbe in der richtigen Art und Weise aufweist. Erst wenn bekannt ist, in welcher Hinsicht der Gegenstand schwarz sein muss, könne über die Wahrheit bzw. die Falschheit des geäußerten Satzes befunden werden. Es gilt also zusätzlich abzuklären, ob die fragliche Teekanne in ihrem ursprünglichen Zustand aufgrund ihrer Herstellung schwarz sein soll oder ob sie auch dann als schwarz durchgeht, wenn sie lediglich mit schwarzem Russ überzogen ist; wichtig zu wissen ist zu-

1 Einzelne Autorinnen wie bspw. Kent Bach (1997) oder Robyn Carston (2002) setzen den semantischen Gehalt nicht mit dem propositionalen Gehalt gleich. Ihnen zufolge spielt die Wahrheitswertfähigkeit eine untergeordnete Rolle für die Bestimmung des semantischen Gehalts. Er wird hauptsächlich durch die sprachliche Bedeutung festgelegt und ist damit kontextunabhängig (siehe auch Stojanovic, 2008).

2 Vorläufer der gegenwärtigen Kontextualismusdebatte sind Searle (1978 & 1980) und Travis (1985).

3 Der semantische Gehalt wird stets durch geäußerte Sätze (mit-)bestimmt und der Hinweis auf die Äußerung muss also immer ergänzt werden, falls dieser fehlt.

4 Nebst den bereits erwähnten Autoren Searle und Travis vertritt aktuell Anne Bezuidenhout (2002), aber auch Dan Sperber und Deirdre Wilson (1986) resolut diese Auffassung. Weniger eindeutig ausformuliert findet sich die Position bei Carston (2002).

dem, ob sie innen und außen schwarz sein muss oder ob einzelne schwarze Teile ausreichen, um die Teekanne schwarz nennen zu können.

Die Kenntnis der Bedeutung des Prädikats „schwarz“ reicht offenkundig nicht aus, um angeben zu können, in welcher Hinsicht die Teekanne in einem bestimmten Kontext schwarz sein soll. Schließlich soll die Verwendung des Farbausdrucks unter den richtigen Umständen mit all den genannten Formen von schwarz vereinbar sein. Die relevante Hinsicht, in der die Teekanne schwarz zu sein hat, wird erst durch den Äußerungskontext und den jeweils vorliegenden Kommunikationszweck festgelegt. Erst in Abhängigkeit davon kann die Frage entschieden werden, ob die Äußerung von (4) wahr oder falsch ist.

2. Der Kontexteinfluss und mögliche Folgerungen daraus

Die obigen Beispielsätze (3) und (4) zeigen nach Ansicht verschiedener Autorinnen, dass die syntaktische Wohlgeformtheit eines Satzes nicht länger als Garant für dessen Wahrheitswertfähigkeit angesehen werden kann, trotz vorgängiger Festlegung seiner sprachlichen Bedeutung und erfolgter Bezugsbestimmung. Zwei Fragen drängen sich an diesem Punkt auf: Erstens gilt es zu klären, wie umfassend der Kontexteinfluss für die Bestimmung des wahrheitswertfähigen Gehaltes geäußerter Sätze ist. Sind nur einige wenige Sätze wesentlich kontextabhängig oder sind für die Festlegung des Wahrheitswertes stets zusätzliche Kontextinformationen erforderlich? Zweitens stellt sich die Frage, welche theoretischen Konsequenzen sich aus diesem weitreichenden Kontexteinfluss ergeben.

Weist man die oben beschriebenen Phänomene nicht einfach mit dem Hinweis zurück, dass es sich dabei lediglich um vermeintlich erforderliche Kontexteinflüsse handelt – eine Strategie, die bei den Vertreterinnen des sogenannten Minimalismus in seinen unterschiedlichen Ausprägungen äußerst beliebt ist[5] –, die in Tat und Wahrheit für die Bestimmung des Wahrheitswertes gar nicht erforderlich seien, dann stehen einem grundsätzlich drei Möglichkeiten offen.

Entweder glaubt man, dass aufgrund der Phänomene, die in (3) oder (4) zum Ausdruck kommen, die logische Form von Sätzen neu interpretiert werden muss und demnach eine Revision der geläufigen Syntaxauffassung erforderlich ist, wie das beispielsweise der *Indexikalismus* vorschlägt.[6] Oder man sieht sich veranlasst, eine Neubewertung der Semantik vorzunehmen, welche die Kontexteinflüsse mit Hilfe von zusätzlichen Beurteilungsparametern systematisch zu erfassen versucht, wie das unterschiedliche Ansätze des *Semantischen Relativismus*

5 Vertreten wird der Minimalismus beispielsweise von Emma Borg (2004a) oder Herman Cappelen und Ernest Lepore (2005).

6 Eine Variante des Indexikalismus findet sich bei Jason Stanley (2000).

unternehmen.[7] Vielleicht zieht man aus den beschriebenen Phänomenen aber einfach den Schluss, das Verhältnis von Semantik und Pragmatik müsse insgesamt neu bewertet werden. Die traditionell gesondert betrachteten Bereiche seien nämlich weit enger aufeinander zu beziehen und die Bestimmung des Wahrheitswerts eines geäußerten Satzes sei ebenso Sache der Bedeutung wie des kommunikativen Äußerungskontexts.[8] Entsprechend bestimmen Bedeutung und Kommunikation gemeinsam die Wahrheitsbedingungen.

Freilich spielt die dem dritten Ansatz zu Grunde liegende Vermutung bei den zwei anderen Positionen ebenfalls eine Rolle. Immerhin werden die vorgeschlagenen systematischen Revisionen dadurch begründet, dass es die beschriebenen Kontexteinflüsse (zumindest teilweise) zu berücksichtigen gelte, weshalb die herkömmliche Darstellung des Verhältnisses von Semantik und Pragmatik in einigen Punkten falsch sein müsse. Trotz dieser Einsicht erfolgt aber die Neubestimmung syntaktischer oder semantischer Aspekte von Sätzen in einer Weise, die es erlaubt, diese und ihre Bedeutung weiterhin unabhängig von der Pragmatik zu beschreiben. Genau diese Möglichkeit stellt die dritte Position in Abrede.

Grundsätzlich lassen sich die beiden eingangs dieses Abschnitts formulierten Fragen unabhängig voneinander angehen. Mit anderen Worten: gleichgültig wie umfassend jemand den Kontexteinfluss einschätzt, kann entweder die erste, die zweite oder die dritte theoretische Folgerung vertreten werden. In den meisten Fällen wird aber eine Einschätzung der quantitativ orientierten Frage nicht folgenlos bleiben für die Wahl des theoretischen Ansatzes. Eben diese Verquickung macht es so schwierig, eine Position als eindeutig kontextualistisch auszuweisen. Denn während die einen ein qualitatives Kriterium zu Grunde legen und eine Position aufgrund ihrer systematischen Handhabung des Phänomens als entweder kontextualistisch oder nicht kontextualistisch einstufen, verwenden andere ein quantitatives Kriterium.[9] Diese unterschiedlichen Kategorisierungsvorschläge lassen die Debatte zusätzlich undurchsichtig erscheinen.[10] Wer aber eine Revision der Syntax oder der Semantik für einen gangbaren Weg hält, ist zumeist auch der Ansicht, nicht in sämtlichen Fällen seien zusätzliche Kontextinformationen nötig. Entsprechend werden nur einzelne Satztypen behandelt, beispielsweise solche, mit denen Wissenszuschreibungen vorgenommen, Ge-

7 Den Semantischen Relativismus formulieren beispielsweise Stefano Predelli (2005) oder François Recanati (2007).

8 Dies schlagen die bereits erwähnten Autorinnen Carston, Bezuidenhout und Sperber und Wilson vor.

9 Borg (2007) plädiert für ein qualitatives Kriterium, um eine Position als entweder minimalistisch oder kontextualistisch zu klassifizieren. Cappelens & Lepores (2005) Einordnung dagegen fußt auf einem quantitativen Kriterium.

10 Andere Probleme ergeben sich aus dem sehr uneinheitlich verwendeten Begriff von Sagen oder, wie bereits angesprochen, aus den verschiedenen Interpretationen von Semantik und deren Abgrenzung von der Pragmatik.

schmacksurteile gefällt werden oder die quantifizierende Ausdrücke enthalten. In Bezug auf diese versuchen sie dann die Kontextabhängigkeit in zulässiger Form zu modellieren.

Eine solche Eingrenzung lässt es aussichtsreich erscheinen, das Phänomen des Kontexteinflusses mittels Anpassung der Syntax oder der Semantik in den Griff zu kriegen. Allerdings muss dabei nicht nur gezeigt werden, dass sich die Einflüsse tatsächlich systematisch erfassen lassen; zusätzlich gilt es nachzuweisen, dass sich der Einfluss des Kontexts bei der Bestimmung des semantischen Gehalts nur gerade auf die problematischen Satztypen beschränkt und nicht ein allgemeines Phänomen ist. Erst dann besitzt die Position die erforderliche Stabilität und kann dem Vorwurf begegnen, ein eher moderater Kontextualismus – ein Kontextualismus, der das Phänomen des Kontexteinflusses auf einzelne Satztypen einzugrenzen versucht – drifte zwangsläufig in einen radikaleren Kontextualismus ab, wie das Cappelen und Lepore vermuten.[11]

Wird das Problem des Kontexteinflusses nicht eingegrenzt, dann werden mit einmal zu viele und vor allem auch zu unterschiedliche Formen von Kontexteinfluss unterscheidbar, die eine systematische Handhabung unwahrscheinlich erscheinen lassen. Falls es trotzdem versucht wird, muss gezeigt werden, dass sich die Kontexteinflüsse für sämtliche Sätze mit einigen wenigen Anpassungen der Syntax oder der Semantik erfassen lassen. Andernfalls drohen die Syntax und die Semantik heillos kompliziert zu werden, sodass die Produktivität der Sprache nicht länger plausibel begründet werden kann.

3. Der umfassende Kontexteinfluss und widersprüchliche Folgen

Wie bereits angesprochen, geht die These der universellen Kontextabhängigkeit meist mit der Auffassung einher, die Kontextabhängigkeit stelle ein komplexes Phänomen dar, welches eine systematische Handhabung ausschließe. Darin zeige sich, dass sowohl die sprachliche Bedeutung wie auch Aspekte der Kommunikation ausschlaggebend sind für die Bestimmung des Wahrheitswertes. Damit sind die Wahrheitsbedingungen nur für Äußerungen, nicht aber für Sätze bestimmbar und deshalb wesentlich pragmatisch.

Einige Probleme, welche mit dem scheinbar allgegenwärtigen Kontexteinfluss einhergehen, werden breit diskutiert. Meist ausgeblendet wird dabei die Frage, welche Konsequenzen sich daraus für den Begriff der sprachlichen Bedeutung ergeben und ob die These des universellen Kontexteinflusses tatsächlich vereinbar ist mit der Annahme, die Sprache leiste einen konstanten und systematischen, wenn auch unvollständigen Beitrag zu den Wahrheitsbedingungen

11 Siehe Cappelen & Lepore (2005). Auf diese Autoren geht auch die Gegenüberstellung von moderatem und radikalem Kontextualismus zurück.

von Äußerungen.[12] Die Plausibilität dieser Behauptung soll anhand des Beispielsatzes (4) untersucht werden.

Das Problem

Weiter oben wurde erwähnt, dass mit der Bedeutung des Prädikats „schwarz" noch keinerlei Angaben dazu vorliegen, auf welche Weise die fragliche Eigenschaft gegeben sein muss. Dabei sind nicht einzig die verschiedenen Materialaspekte und die damit zusammenhängenden Farbeigenschaften der Teekanne zu berücksichtigen; ebenso kann die Ausprägung der Farbe von Belang sein, die von tiefschwarz bis zu schwarzgrau etc. reichen kann und die nur einzelne Bestandteile oder aber das Objekt als Ganzes überziehen kann. Erst wenn bekannt ist, in welcher dieser möglichen Hinsichten die Teekanne schwarz sein soll, kann über die Wahrheit der Äußerung von (4) entschieden werden. Und die relevante Hinsicht wird durch den Äußerungskontext bestimmt.

Folgende Annahmen gehen mit diesen Behauptungen einher: Erstens entpuppt sich die sprachliche Bedeutung als weit leistungsschwächer als traditionell angenommen wird. Aspekte, die für die Beurteilung des Wahrheitswertes eines geäußerten Satzes wichtig sind, sind nicht mit der Bedeutung des Farbprädikats gegeben. Die monierte semantische Unterbestimmtheit der fraglichen Sätze ist auf eben diese Schwäche der sprachlichen Bedeutung zurückzuführen. Daraus folgt, zweitens, dass es das Informationsdefizit auszugleichen gilt mit Hilfe des Äußerungskontexts, sodass die Wahrheitsbedingungen vollständig bestimmt werden können. Verbindet man diese beiden Behauptungen mit der Annahme, eine systematische Handhabung des Kontexteinflusses sei ausgeschlossen, weil die verschiedenen möglichen Hinsichten bei den Ausdrücken einer Sprache äußerst vielfältig sind und sie darum unmöglich semantisch erfassbar sind, dann mündet die kontextualistische These in die Behauptung, Wahrheitsbedingungen seien kommunikativ mitbestimmt und in diesem Sinne pragmatisch. Wird diese Behauptung dann noch verallgemeinert, liegt ein Kontextualismus in seiner radikalsten Form vor. Wenn im Folgenden von Kontextualismus die Rede ist, dann in eben diesem Sinne.

Wie bereits angedeutet wurde, sind die beiden zuletzt genannten Schritte keinesfalls zwingend: Die beiden ersten Annahmen können vertreten werden, ohne damit gleich einräumen zu müssen, Wahrheitsbedingungen seien stets wesentlich pragmatisch. Denn die erforderlichen Zusatzinformationen können mittels Anpassung der Syntax oder der Semantik in die Sprache eingebaut und als Bestandteil derselben aufgefasst werden. Genuin pragmatisch sind Wahrheitsbe-

12 Erwähnenswert sind die beiden Ausnahmen MacFarlane (2007) und Imhof (2006), welche für die hier vorgebrachten Überlegungen wegweisend waren.

dingungen erst dann, wenn nicht durch die Sprache selbst indiziert wird, dass zusätzliche Hinsichten erforderlich sind.

Im Folgenden möchte ich zeigen, dass es gute Gründe gibt, die dritte Behauptung über die pragmatischen Wahrheitsbedingungen *nicht* zu vertreten und damit den Kontextualismus abzulehnen. Falls diese Gründe einigermaßen einleuchten, kann weiter gefragt werden, ob das monierte Informationsdefizit der sprachlichen Bedeutung nur vermeintlich vorliegt oder ob damit nicht Ansprüche an die Sprache herangetragen werden, welche diese gar nicht zu leisten braucht.

Das Problem des universellen Kontexteinflusses

Der Kontextualismus – und das ist wichtig – behauptet trotz allem nicht, dass die sprachliche Bedeutung keinen Beitrag zu den Wahrheitsbedingungen eines geäußerten Satzes leiste; der Beitrag der sprachlichen Bedeutung ist lediglich nicht hinreichend. Die Pragmatik mit ihren kommunikativen Zwecken greift also darum ein, weil die Satzbedeutung stets auf den entsprechenden Äußerungskontext relativiert werden muss und die Relativierung dabei äußerst unterschiedliche Formen annehmen kann.

Nun aber stellt sich die Frage, welche Rolle die sprachliche Bedeutung überhaupt noch spielen kann angesichts dieses zwingenden und äußerst vielschichtigen Kontexteinflusses. Im Rahmen des Kontextualismus eine Antwort zu formulieren, ist schwierig, da die eigentliche Bedeutung eines geäußerten Satzes prinzipiell erst auf einen Äußerungskontext relativiert verständlich ist. Gegeben also die Kontextthese, wird es schwierig, die sprachliche Bedeutung eines Ausdrucks unabhängig von seiner Verwendung in bestimmten Äußerungskontexten zu beschreiben.

Diese Schwierigkeit mag der Kontextualismus durchaus einräumen. Gleichzeitig kann er darauf beharren, es handle sich dabei lediglich um ein erkenntnistheoretisches Problem, welches die Annahme der sprachlichen Bedeutung noch lange nicht hinfällig mache. Es könne davon ausgegangen werden, dass es sich bei der dem Ausdruck zugewiesenen Bedeutung quasi um eine Abstraktion von den einzelnen Verwendungen handle.

Tatsächlich steht dieser Ausweg dem Kontextualismus nicht offen. Denn dieser entbehrt nicht nur der Möglichkeit, gehaltvolle Angaben zur sprachlichen Bedeutung eines Ausdrucks zu liefern; seine Behauptung, sprachliche Ausdrücke führten überhaupt wichtige Informationen über ihre zulässige Verwendung mit sich, wie auch die Annahme einer systematisch funktionierenden Sprache wird im Rahmen des kontextualistischen Ansatzes schlicht unhaltbar.

Ein erstes Problem mit dem Problem

In einfacher Weise vorführen lassen sich diese Schwierigkeiten, wenn man berücksichtigt, dass sich mit der Kontextthese im Hintergrund nicht einmal einfachste logische Beziehungen zwischen Sätzen formulieren lassen. So gilt die Folgerung vom Satz in der Prämisse (P) „*x* ist schwarz" zum Satz in der Konklusion (C) „*x* ist schwarz" nicht ausnahmslos, selbst wenn die Einsetzung für *x* jeweils denselben Gegenstand bezeichnet. Schließlich ist es möglich, dass sich die Hinsichten, gemäß welchen der fragliche Gegenstand schwarz zu sein hat, für die Prämisse (P) und die Konklusion (C) unterscheiden und die Prämisse deshalb wahr, die Konklusion hingegen falsch sein kann, wie im Folgenden veranschaulicht:

(P) *x* ist schwarz [*in seinem Originalzustand*]	(wahr)
(C) *x* ist schwarz [*weil mit schwarzem Ruß überzogen*]	(falsch)

So wie die Bedeutung eines Ausdrucks oder Satzes jeweils erst auf den Kontext relativiert verständlich ist, sind laut Kontextualismus auch Folgerungsbeziehungen nur auf einen spezifischen Kontext relativiert entweder gültig oder ungültig; d.h. das Schema „p impliziert p" ist für Sätze nicht allgemeingültig, sondern nur für Sätze relativ zum gleichen Äußerungskontext. Deshalb kann ein Gegenstand sowohl schwarz sein als auch nicht schwarz sein, obwohl stets von demselben Gegenstand die Rede ist – je nach relevanter Hinsicht. Und obwohl ich weiß, dass ein Gegenstand schwarz ist, lässt sich daraus nicht folgern, dass dieser auch tatsächlich schwarz ist. Die Identität der sprachlichen Bedeutung ist nicht hinreichend, es braucht auch die Identität pragmatischer Elemente.

Diese eigentümlichen Auswirkungen der Kontextthese auf die Gültigkeit des Satzes vom Widerspruch sowie auf die Zuschreibung von Wissen und deren metaphysischen Implikationen sind einigermaßen befremdend und es würde sich lohnen, diese eigens zu untersuchen. Bedenklich sind aber auch die Konsequenzen, die sich daraus für die Semantik ergeben. Denn offenkundig fehlt dem Kontextualismus schlicht und einfach die Möglichkeit, gültige Folgerungsbeziehungen zwischen Sätzen zu formulieren. Wie sich die einzelnen Ausdrücke einer Sprache unabhängig von irgendeinem Kontext zueinander verhalten, lässt sich darum nicht länger sagen. Das wiederum macht es schwierig, die Systematizität einer Sprache plausibel zu erklären. Gerade darum werden Wahrheitsbedingungen für die Semantik als unhintergehbar eingestuft.

Ein zweites Problem mit dem Problem

Es ist allerdings vor kontextualistischem Hintergrund nicht nur schwierig, die Systematizität einer Sprache verständlich zu machen. Es stellt sich zudem die Frage, welchen Beitrag die sprachliche Bedeutung überhaupt noch leisten kann,

wenn aus einem Satz wie „*x* ist schwarz“ je nach Kontext sowohl folgen können muss, dass *x* schwarz ist, als auch, dass *x* nicht schwarz ist. Um eine Folgerung auf einen Gehalt wie auch seine Negation zulassen zu können, muss die Information, welche das Prädikat „schwarz“ liefert, sehr minimal sein. Letztlich dürfte sie nicht mehr enthalten, als dass das Prädikat „schwarz“ in etwa bedeutet, dass ein Gegenstand farbig ist. Die sprachliche Bedeutung des Ausdrucks leistet damit gewiss weit weniger, als man von dieser erwarten würde und müsste, wenn sie die Bedingungen der korrekten Anwendung festlegen soll.

Da die Kontextthese gemäß obiger Annahme universell gilt, verschärft sich das Problem zusätzlich, denn in Bezug auf das Prädikat „farbig“ werden dieselben Probleme auftreten, wie sie für „schwarz“ gerade beschrieben wurden: Die eigentliche Bedeutung des generischen Ausdrucks ist nur relativiert auf den Kontext verständlich und seine Bedeutung leistet damit nicht mehr, als was die Bedeutung eines noch allgemeineren Ausdrucks liefert (bspw. „Gegenstand sein“). Und wiederum stellt sich aufgrund der Kontextthese dasselbe Problem für diesen Ausdruck und seine Bedeutung...

Letztlich führt die beschriebene Problematik zur Feststellung, dass die sprachliche Bedeutung vor dem Hintergrund der kontextualistischen Annahmen gänzlich uninformativ ist und ihre Annahme letztlich hinfällig ist. Die seitens Kontextualismus gemachte These, die sprachliche Bedeutung leiste einen wichtigen Beitrag zur Äußerungsbedeutung und zur Festlegung der Wahrheitsbedingungen von Äußerungen, kann demnach nicht länger verteidigt werden. Damit entpuppt sich die These des Kontextualismus, deren intuitive Plausibilität stets gerühmt wird, als gänzlich kontraintuitiv.

4. Das Fazit

Woher rühren aber diese Schwierigkeiten? Ein Vergleich des gerade beschriebenen Falles von Kontextabhängigkeit beim Ausdruck „schwarz“ mit anderen, traditionell als kontextsensitiv eingestuften Ausdrücken hilft womöglich, der Sache auf die Spur zu kommen. Beispielsweise liefert die Bedeutung des indexikalischen Ausdrucks „ich“ klare Angaben darüber, welche Person mit dem Ausdruck bezeichnet werden soll. Ein Irrtum ist ausgeschlossen. Fehlidentifikationen sind freilich immer möglich, weil beispielsweise die Lichtverhältnisse schlecht und die Stimmen zweier Personen kaum unterscheidbar sind. Deshalb kann sich eine Hörerin darüber täuschen, wer nun tatsächlich den Satz mit dem fraglichen Personalpronomen geäußert hat. Solche Fehlleistungen sind aber nicht mit einer Schwäche der sprachlichen Bedeutung des Ausdrucks zu begründen.

Ganz anders zeigt sich nun scheinbar die Angelegenheit beim Farbprädikat und seiner Verwendung in spezifischen Kontexten. Werden nämlich die Wahr-

heitsbedingungen dahingehend aufgefasst, dass diese sowohl durch Aspekte der sprachlichen Bedeutung wie auch Aspekte der Kommunikation wesentlich bestimmt sind, dann ist die Extension eines Ausdrucks erst im Äußerungskontext und nicht bereits mit der Bedeutung gegeben. Das zumindest folgt, wie gezeigt, aus der These der wesentlich pragmatischen Wahrheitsbedingungen. Dann aber stellt sich die Frage, welche Informationen einzelne Ausdrücke oder die Sprache insgesamt noch liefern.

Was lässt sich aus dem drohenden Bedeutungs- und Spracheliminativismus schließen? Offenkundig kann und darf es nicht sein, dass der Beitrag eines Ausdrucks zur Festlegung der Wahrheitsbedingungen wie vom Kontextualismus behauptet uneingeschränkt variiert, soll die Sprache und ihre Semantik nicht ganz in der Pragmatik, also der Kommunikation aufgehen. Gewiss ist es möglich, die Extension eines Ausdrucks zusätzlich einzugrenzen, indem der für die Kommunikationszwecke relevante Bereich in der einen oder anderen Form näher bestimmt wird. Zu behaupten, die Extension müsse zusätzlich eingegrenzt werden, ist aber ganz etwas anderes, als zu behaupten, sie sei mit der Bedeutung eines Ausdrucks gar nicht gegeben. Wenn aber die Extension mit der Bedeutung gegeben ist, dann scheint es nicht länger plausibel, die Gleichsetzung von Bedeutung mit Wahrheitsbedingungen abzulehnen. Vielmehr müssen die zusätzlichen Eingrenzungen anderweitig erklärt werden.[13]

Meines Erachtens sollten Anpassungen der Syntax oder der Semantik nicht voreilig vorgenommen werden. Dass die Bedeutung eines geäußerten Satzes in der Kommunikation zusätzlich bestimmt wird, ist sicherlich ein *locus communis*, der aber nicht ohne Weiteres eine Revision in der Sprache nach sich ziehen sollte.[14] Die oben genannten Schwierigkeiten legen auf jeden Fall nahe, dass das Verhältnis von Bedeutung und Kommunikation nur so eng gefasst werden sollte, dass beide Bereiche noch selbstständig funktionieren.[15]

13 MacFarlane (2007) vertritt aus den oben beschriebenen Gründen die Auffassung, Beurteilungsparameter seien pragmatisch zu interpretieren.

14 Darum hat Herbert Paul Grice (1975) eigens seine Implikaturtheorie formuliert und so eine plausible Erklärungsgrundlage für die beschriebenen und zahlreiche andere Phänomene geliefert.

15 Ein Teil der hier vorgebrachten Überlegungen findet sich auch in meinem Aufsatz „Linguistic Meaning and the Minimalism-Contextualism-Debate“ formuliert, der im Herbst 2010 erscheinen wird. Ganz herzlich danke ich Silvan Imhof, Nathalie Loetscher, Klaus Petrus, sowie den Zuhörerinnen und Zuhörern der GAP 7 für hilfreiche Kommentare zu früheren Versionen dieses Textes, und dem Schweizerischen Nationalfonds (SNF) für die finanzielle Unterstützung (PP001-114821/1).

Literaturverzeichnis

Bach, Kent: "The semantics-pragmatics distinction: what it is and why it matters". Linguistische Berichte, 8, 1997. S. 124-162

Bezuidenhout, Anne: "Truth-Conditional Pragmatics". Noûs 36, 2002. S. 105-134

Borg, Emma: Minimal semantics. Oxford University Press, Oxford, 1. Auflage, 2004a

Borg, Emma: "Minimalism versus contextualism in semantics". In: *Preyer, Gerhard/Peter, Georg (Hrsg.)*: Context-sensitivity and semantic minimalism. Oxford University Press, Oxford, 1. Auflage, 2007. S.339-359

Cappelen, Henri; LePore, Ernest: Insensitive semantics. A defense of semantic minimalism and speech act pluralism. Blackwell, Oxford, 1. Auflage, 2005

Carston, Robyn: Thoughts and utterances. The pragmatics of explicit communication. Blackwell, Oxford, 1. Auflage, 2002

Grice, Herbert Paul: "Logic and conversation". In: Studies in the way of words. Harvard University Press, Cambridge, 1. Auflage, 1975/1989. S. 22-40

Imhof, Silvan: "Literal meaning". http://www.meaning.ch/content/view/101/, 2006

MacFarlane, John: "Semantic minimalism and nonindexical contextualism". In: *Preyer, Gerhard/Peter, Georg (Hrsg.)*: Context-sensitivity and semantic minimalism. Oxford University Press, Oxford, 1. Auflage, 2007. S. 240-250

Predelli, Stefano: "Painted leaves, context, and semantic analysis". Linguistics & Philosophy, 28, 2005. S. 351-374

Recanati, François: Perspectival thought. Oxford University Press, Oxford, 1. Auflage, 2007

Searle, John: "Literal meaning". Erkenntnis, 13, 1978, 207-224

Searle, John: "The background of meaning". In: *Searle, John/Kiefer, Ferenc/Bierwisch, Manfred (Hrsg.):* Speech act theory and pragmatics. Reidel, Dordrecht, 1. Auflage 1980. S.221-232

Sperber, Dan/Wilson, Deirdre: Relevance. Harvard University Press, Cambridge, 1. Auflage, 1986

Stanley, Jason: "Context and logical form". Linguistics and Philosophy, 23, 2000, 391-434

Stojanovic, Isidora: "The semantics/pragmatics distinction". Synthese, 165, 2008. S. 317-319

Travis, Charles: "On what is strictly speaking true". Canadian Journal of Philosophy, 15, 1985. S. 187-229

On the Semantics of Natural Kind Terms: An Examination of Two Kripkean Theses

Luis Fernández Moreno
luis.fernandez@filos.ucm.es
Complutense University of Madrid

Abstract/Zusammenfassung

In the first two lectures of *Naming and Necessity* Saul Kripke is chiefly concerned with proper names, while in the third he deals specially with natural kind terms, being one of his main aims to argue that there are some *similarities* between them and proper names. According to Kripke, two important similarities between natural kind terms and proper names are that both sorts of expressions are rigid designators and appear in identity statements that, if true, are necessary *a posteriori*; Kripke denominates this sort of statements containing natural kind terms "theoretical identities". Now Kripke claims that the latter similarity follows from the former.

My aim in this talk is to examine if it can be maintained that natural kinds are rigid designators and if it is acceptable the justification of the necessity attributed by Kripke to theoretical identities – assuming they are true. I will argue, on the one hand, that the thesis that natural kind terms are, like proper names, rigid designators can be sustained. On the other hand, although I concede that theoretical identities are *a posteriori* statements, I will put into question the justification offered by Kripke of the *necessity* of theoretical identities.

In den ersten zwei Vorlesungen von *Naming and Necessity* beschäftigt sich Saul Kripke hauptsächlich mit Eigennamen, während in der dritten Vorlesung bezieht er sich insbesondere auf Termini für natürliche Arten. Dabei will er die These vertreten, dass es einige Ähnlichkeiten zwischen diesen letzten und den Eigennamen gibt. Zwei wichtige Ähnlichkeiten zwischen Termini für natürliche Arten und Eigennamen sind Kripke zufolge, dass beide Ausdrucksarten starre Bezeichner sind und beide in Identitätsaussagen erscheinen, welche – wenn sie wahr sind − notwendigerweise *a posteriori* sind. Kripke bezeichnet diese Art von Aussagen, welche Termini für natür-liche Arte enthalten, "theoretische Identitäten". Nun Kripke behauptet, dass diese letzte Ähnlichkeit von der erste sich ableitet.

Mein Ziel in diesem Aufsatz ist zu untersuchen, inwiefern Termini für natürliche Arten starre Bezeichner sind und inwiefern die Notwendigkeits-begründung, welche Kripke theoretischen Identitäten – wenn sie wahr sind − zuschreibt, angenommen werden kann. Ich werde einerseits für die These plädieren, dass Termini für natürliche Arten, wie Eigennamen, starre Bezeichner sind. Andererseits möchte ich die Kripke´sche Notwendigkeitsbegründung von theoretischen Identitäten in Frage stellen, wenngleich auch ich zugebe, dass theoretische Identitäten Äußerungen *a posteriori* sind.

1. The contextual setting

In the first two lectures of (1980) Saul Kripke mainly deals with proper names, while in the third, and last one, he pays particular attention to natural kind terms, being one of his aims to point out the existence of certain *similarities* between the latter sorts of terms and proper names. In this paper I will concentrate on a prototypical sort of natural kind terms, terms for chemical substances or, for short, *substance terms*, such as "water" and "gold".

According to Kripke, one of the similarities between natural kind terms and proper names is that both sorts of expressions are involved in identity statements that, if true, are necessary *a posteriori*. This sort of statements containing natural kind terms are called by Kripke "theoretical identifications" and "theoretical identities" – I will opt for the last denomination –, and he exemplifies them through the statements "Water is H_2O" and "Gold is the element with the atomic number 79". Now Kripke claims that this similarity follows from another, consisting in natural kind terms being, like proper names, *rigid designators*. Thus, he asserts:

> Theoretical identities, according to the conception I advocate, are […] identities involving two rigid designators and *therefore* are examples of the necessary *a posteriori*. (1980, p.140; first emphasis added).

In this paper I have twofold aim. Firstly, I will allege that it can be maintained that natural kind terms are, like proper names, rigid designators. Secondly, I will argue that this feature of natural kind terms is not enough to justify the thesis that theoretical identities are statements that, if true, are necessary *a posteriori*. Although I concede that theoretical identities are *a posteriori* statements, I will put into question the justification offered by Kripke in the quoted passage for the *necessity* he attributes to theoretical identities.

It is appropriate to start making some remarks about the context within which Kripke puts forward the thesis that natural kind terms are rigid designators. It is noteworthy that Kripke does not offer a very precise characterization of natural kind terms; he mainly characterizes natural kind terms as *general terms* whose function is to designate *natural kinds*. This characterization can only turn out to be more precise after a clarification of how the reference or designation of natural kind terms is fixed and how natural kinds are conceived of.

In the first lecture of (1980) Kripke introduces the term *designator* as a common denomination for proper names and definite descriptions, the two sorts of *singular terms* he takes into consideration. The definition of a rigid designator is contained in the first and second lectures and therefore it is put forward before Kripke starts considering natural kind terms into account. Since Kripke claims that natural kind terms are, as proper names, rigid designators, but he does not provide an explicit definition of the notion of a rigid designator for natural kind terms, it has to be assumed that the definition of that notion for them will be an extension of the one proposed for singular terms and, in particular, for proper names.

The clearest definition of a rigid designator put forward by Kripke is contained in a letter he sent to David Kaplan. There Kripke affirms that the notion of a rigid designator he pretends is the following: "[A] designator *d* of an object *x* is *rigid*, if it designates *x* with respect to all possible worlds where *x* exists, and *never designates an object other than x with respect to any possible world*" (quoted in Kaplan 1989, p. 569).

This definition leaves two options open. The first one is that a rigid designator designates the same object with respect to all possible worlds; the second, that it designates the same object only with respect to all possible worlds in which the object exists, lacking reference with respect to the rest of possible worlds. Following a usual terminology, initially proposed by Salmon (see 1981, pp. 33 f.), rigid designators satisfying the first characterization are *obstinate* designators, while those fulfilling the second one are *persistent* designators.

Now, though Kripke prefers to leave open that double alternative so as not to get involved in questions arising from the possible non-existence of an object, the definition of a rigid designator applicable to proper names pretended by Kripke is the first one, since in the preface to (1980) he asserts that he considers proper names as *rigid de jure* (1980, p. 21, n. 21). A designator is *rigid de jure* if at fixing its reference it is stipulated that its referent is the same independently of whether we are speaking of the actual world or of a possible world different from it. Since Kripke regards proper names as *rigid de jure*, he claims that a proper name rigidly designates its referent even with respect to possible worlds where its referent would not have existed.

On the opposite, most definite descriptions are *non-rigid designators*, that is, they can designate different objects with respect to different possible worlds, since it may happen that in different possible worlds a different object is the one that possesses the property expressed by the description. Nevertheless, Kripke acknowledges that there are some definite descriptions which are rigid designators, though they are not *rigid de jure* but *rigid de facto*, mentioning as an example the description "the smallest prime", which rigidly designates the number two (*ibid*.). In the case of a rigid *de facto* designator it is not stipulated that there is only one object that is its referent with respect to all possible worlds, but the predicate contained in the description applies to the same object with respect to all possible worlds or at least with respect to all possible worlds where the object exists, depending on whether the definite description in question is an obstinate or a persistent designator. Now, most definite descriptions that are rigid designators are persistent designators; only in the case that the object designated were a necessarily existing object – if there are such sort of objects –, they could be regarded as obstinate designators.

A consequence of the rigidity of proper names is that true identity statements involving two proper names are *necessary*, that is, true with respect to all possible worlds. One of the most famous examples is the identity statement "Hesperus is

Phosphorus". This statement is true, since the names "Hesperus" and "Phosphorus" designate the same object, to wit, the planet Venus. Now, if these proper names are, as Kripke pretends of all proper names, rigid *de jure*, they will designate the planet Venus with respect to all possible worlds, from which it follows that the statement "Hesperus is Phosphorus" is necessary. Nevertheless, it was an empirical discovery that these proper names designate the same object, and this is something we could not know *a priori*. Therefore, the statement "Hesperus is Phosphorus" is, though necessary, true a *posteriori*.

Once we have reached this point, there are two questions to be raised. Firstly, if the definition of a rigid designator for natural kind terms and Kripke's view on natural kinds allow us to maintain that natural kind terms are rigid designators. Secondly, and assuming an affirmative answer to that question, whether the rigidity of natural kind terms provides a justification for the thesis that theoretical identities are, as identity statements involving two proper names, necessary if true – as already pointed out, I will not question their *a posteriori* character.

2. Are natural kind terms rigid designators?

Before we attend to the question concerning whether it can be maintained that natural kind terms are rigid designators, we should mention two other similarities existing, according to Kripke, between natural kind terms and proper names: their *non-descriptionality* and the *historical-causal* character of their reference fixing. Natural kind terms are, as proper names, non-descriptional, that is, they are not synonymous with descriptions usually associated with them by speakers and that would determine their reference. Now if the fixing of the reference of proper names and of natural kind terms is not explained in this way, that is, by means of descriptions with which they were synonymous, an alternative explanation of how their reference is fixed is to be provided. In this regard Kripke sketches a historical-causal theory for both, proper names and natural kind terms. In the case of natural kind terms Kripke claims (see 1980, pp. 135 ff.) that these terms are introduced ostensively in presence of paradigmatic entities of the kind or by means of descriptions that express properties, usually contingent ones, of such entities. In both cases the extension of a natural kind term will comprise all entities of the same kind as those involved in the introduction of the term, and the relation of kind sameness is constituted by underlying or structural properties of such entities – let us say, their internal structure –, whose discovery is subject to empirical research. Once the natural kind term in question has been introduced, it is transmitted to other speakers through causal (historical) chains of communication and the latter speakers will use the term to refer to entities of the kind, although the descriptions they associate with the term do not express identifying properties of the kind.

The non-descriptionality and the historical-causal character of their reference fixing are features that are usually linked to the assertion that proper names and natural kind terms are rigid designators. However, concerning natural kind terms such features are not sufficient to justify the thesis that they are rigid designators, since this justification will also and mainly depend on how the notion of a rigid designator for natural kind terms is defined and how natural kinds, that is, the referents or *designata* of natural kind terms, are conceived of.

Regarding the first question, it is appropriate to take into account the definition of a rigid designator for singular terms pretended by Kripke. Since in the third lecture of (1980) Kripke extends the notion of a rigid designator to natural kind terms, we will have to extend the definition of a rigid designator provided for singular terms to natural kind terms or, in general, to kind terms. In this regard a natural extension, and the only one I will take into consideration, is the following: A designator *d* of a kind *k* is *rigid*, if it designates *k* with respect to all possible worlds where *k* exists, and *never designates a kind other than k with respect to any possible world*.

Although Kripke has not been sufficiently explicit concerning the conditions under which a kind exists in a possible world, I will assume that if a possible world contains entities of a kind, the kind exists in that possible world, so that the claims about the existence of a kind in a possible world are derived from those concerning the existence of entities of the kind in that world.

Regarding the second question, that is, how Kripke conceives of natural kinds, it is pertinent to pay attention to the following passage:

> [I]n general, terms for natural kinds (e.g., animal, vegetable, and chemical kinds) get their reference fixed in this way; the substance is defined as the kind *instantiated* by (almost all of) a given sample. (1980, pp. 135-36; emphasis added).

Concerning this excerpt it is appropriate to make two remarks. Firstly, in that passage the expression "substance" is being understood in a broad sense, to wit, as interchangeable with the expression "natural kind". Secondly, since Kripke characterizes the relationship between a natural kind and the entities of the kind, according to that passage, as a relationship of *instantiation*, he must have conceived of a natural kind as a sort of abstract entity or universal – though he does not introduce any details in this regard.

The view of natural kinds as certain universals or abstract entities, instantiated in concrete entities, is the predominant view at present among the authors that accept the thesis that natural kind terms are rigid designators – see, e.g., Salmon (1981). This view allows us to maintain that natural kind terms are rigid designators according to the definition of a rigid designator for these terms and in general for kind terms proposed above, since the abstract entities designated by natural kind terms will be the same with respect to all possible worlds or at least with respect to all possible worlds where such entities exist. Now, the conception of the referents of natural kind terms as abstract entities instantiated in

concrete entities is applicable to the referents of *all general terms* which apply to concrete entities, what will lead us to accept that not only natural kind terms, but also many other general terms, are rigid designators.

Reached this point it is adequate to introduce certain precisions. As already pointed out, Kripke affirms that proper names, that is, singular terms which are semantically simple, are rigid designators, while most definite descriptions, that is, singular terms which possess a semantically relevant structure, are non-rigid designators. Once conceded that the view of the referents of kind terms as abstract entities makes it possible to maintain that natural kind terms and many other general terms are rigid designators, it can be asserted that general terms that are semantically simple are rigid designators; thus, for example, the terms "water" and "gold" are rigid designators, but so are other sorts of general terms such as "bachelor" or "table". Accordingly rigidity does not enable to *distinguish* natural from non-natural kind terms, since not only non-descriptional general terms, as according to Kripke are natural kind terms, but also paradigmatically descriptional terms, like the term "bachelor", will be rigid designators. A difference between the latter terms and natural kind terms will lay in the other feature already mentioned, that is, the historical-causal character of the reference fixing of natural kind terms.

However, according to that view of the referents of natural kind terms, and in general of general terms, *not* all general terms will be rigid designators. More precisely, most general terms with a relevant semantic structure, that is, most semantically composed general terms will not be rigid designators. Thus, for instance, the terms "liquid employed to wash the hands" or "John's favorite metal" will not be rigid designators – or, at least, they have a non-rigid interpretation –, since they could designate with respect to different possible worlds different liquids or metals, conceived of as abstract entities. Nevertheless, there will be some semantically composed general terms which will be rigid designators, as, for example, the terms "substance (whose samples are) composed of molecules consisting of two hydrogen atoms and one oxygen atom" and "element with the atomic number 79". And once we have reached this point we could *extend* the similarity between the rigidity of singular and general terms claiming that semantically simple general terms are, like proper names, rigid *de jure*, while semantically composed terms, singular or general ones, which are rigid, are only rigid *de facto*.

3. Does the rigidity of natural kind terms justify the necessity of theoretical identities?

Now, if one accepts this last claim, the two examples of theoretical identities mentioned above will contain a rigid *de jure* designator and a rigid *de facto* one, and while all designators of the first sort are obstinate, most of the second sort are persistent. This is a reason why the necessity of theoretical identities could not be es-

tablished following the same procedure as that adduced with regard to identity statements involving two proper names and hence two obstinate designators.

However, from that sort of identity statement it can be obtained a statement that is necessary. This statement has the form of a conditional statement whose antecedent is an existential statement and whose consequent is the identity statement in question. Thus, e.g., the statement "If H_2O exists, water is H_2O" will be, if true, necessary, i.e., true with respect to all possible worlds, even if the term "H_2O" is a persistent designator. For this reason we can concentrate, as so far, on the necessity of theoretical identities. In this regard I will take into account the first example of theoretical identity mentioned above, that is, the statement "Water is H_2O", since similar considerations would apply to the other.

It should be pointed out that this statement can be interpreted as an identity statement, but also as a universally quantified biconditional. Thus interpreted, the statement "Water is H_2O" will be true if and only if for every sample (of the actual world) it holds that a sample instantiates the substance water if and only if that sample instantiates the substance H_2O, and hence if and only if the general terms "water" and "H_2O" are coextensive, i.e., they apply to the same entities. However, from here it follows neither that they are necessarily coextensive nor that such statement is necessary if true.

A similar conclusion is obtained if we regard that statement, as we will do in the following, as an identity statement in which the identity sign (of second order) is flanked by the general terms "water" and "H_2O". This statement will be true if and only if the substances water and H_2O are identical in the actual world, and this will be so only if the instances of these substances are the same in the actual world, i.e., only if the terms "water" and "H_2O" are coextensive.

Put in a general way, theoretical identities, regarded as identity statements involving two natural kind terms, will be true if and only if the kinds designated by those two terms are identical in the actual world, and this will be so only if their instances are the same in the actual world, but from the identity of their instances in the actual world it does not follow that their instances are the same in all possible worlds. The rigidity of the terms "water" and "H_2O" warrants that the samples that are instances in the actual world of the substances designated by them, will be instances of such substances in every world where those instances exist, but the problem arises concerning samples that do *not* exist in the actual world. From the rigidity of the term "water", it follows that a sample of water in a non-actual world will also be an instance of water in every world where that sample exists, but from the rigidity of "H_2O" and the truth of the statement "Water is H_2O" does not follow that that sample will also be an instance of H_2O. Thus the rigidity of the terms "water" and "H_2O" together with the truth of the statement "Water is H_2O" are not sufficient to establish that the instances of the substances designated by those terms are the same in all possible worlds, and hence that this statement is necessary. (For similar considerations see Soames 2002, pp. 257-58).

Kripke pretends to justify the necessity of certain true identity statements, those containing two proper names as well as theoretical identities, resorting to the features that the reference of the terms composing such statements is the same with respect to the actual world and that such terms are rigid designators. This is the procedure used to justify the necessity of the statement "Hesperus is Phosphorus". Kripke's pretended justification for the necessity of theoretical identities proceeds in principle in the same way – see the first passage quoted in section 1. –, but, as already indicated, this procedure by itself is *insufficient* to obtain the desired conclusion. The rigidity of natural kind terms involved in theoretical identities will be only a *necessary condition*, but not a sufficient one, to justify that theoretical identities are necessary if true, since if those terms were non-rigid ones, theoretical identities could not be necessary statements.

We can summarize the foregoing considerations concerning why rigid designation and identity of designation with respect to the actual world, which enable to justify the thesis that identity statements containing two proper names are necessary if true, do not enable, however, the justification of the corresponding thesis regarding theoretical identities, in the following way. The rigidity of proper names entails that if a proper name designates an object with respect to the actual world, it will designate the same object with respect to all possible worlds – proper names are obstinate designators –; thus, the identity statement containing two proper names will be necessary if true. The rigidity of natural kind terms, and in particular of substance terms, entails that if a substance term designates a substance with respect to the actual world, it will designate the same substance with respect to all possible worlds – or at least with respect to all possible worlds where the substance exists. However, the identity of the instances of two substances in the actual world, which is required for the truth of the theoretical identity involving the corresponding substance terms, does not imply that the instances of those substances are the same in all possible worlds, what is required for the theoretical identity to be necessary.

References

Kaplan, David: "Afterthoughts". In: *Almog, Joseph, et al. (Hrsg)*: Themes from Kaplan. Oxford University Press, New York, 1989. S. 565-614

Kripke, Saul: Naming and Necessity. Blackwell, Oxford, 1980

Salmon, Nathan: Reference and Essence. Princeton University Press, Princeton, 1981

Soames, Scott: Beyond Rigidity. The Unfinished Semantic Agenda of Naming and Necessity. Oxford University Press, Oxford, 2002

Freges Adverbialtheorie des Urteilens

Christoph C. Pfisterer
pfisterer@philos.uzh.ch
Philosophisches Seminar der Universität Zürich

Abstract/Zusammenfassung

The present paper examines Frege's notion of judgement, particularly the relation between judgement and truth and the possibility of false judgements. In the first section I will argue that the performance of a judgement consists neither in predicating truth nor in referring to the True; nor do we judge in thinking the Fregean sense of the word "true" in addition to a thought. The second section discusses two problems arising from Frege's standard definition of judgement as acknowledging the truth of a thought. First, I shall argue that judgements do not imply truth; i.e. Frege's use of "acknowledge" is not factive. Second, judgements are not comprised of an act of merely entertaining a thought and an act of acknowledging its truth; i.e. judgements are not cumulative. Rather, a judgement is one single act of acknowledging a thought *as true* or thinking *truly*. For this reason, the last section offers a new interpretation which takes Frege's adverbial definition very serious. I will show that adverbialism with regards to judgements has no factive reading and allows for the normativity of truth. Hence, the adverbial theory of judgement fits logical inferences as well as spontaneous judgements.

Der Beitrag beschäftigt sich mit Freges Urteilsbegriff, insbesondere mit dem Verhältnis zwischen Urteilen und Wahrheit und mit der Möglichkeit falscher Urteile. Im ersten Abschnitt zeige ich, dass ein Urteil weder dem Zusprechen eines Wahrheitsprädikats noch der Bezugnahme auf das Wahre entspricht; auch fällen wir keine Urteile, indem wir den Sinn des Wortes „wahr" zu einem Gedanken hinzudenken. Der zweite Abschnitt ist Problemen gewidmet, die Freges Standardcharakterisierung des Urteilens als ein Anerkennen der Wahrheit von Gedanken anlasten. Erstens ist es nicht richtig, dass das Anerkennen der Wahrheit eines Gedankens impliziert, dass der Gedanke wahr ist; Freges Urteilsbegriff ist nicht faktiv. Zweitens sind Urteilsakte nicht aus einem Akt des bloßen Denkens und einem Akt des Anerkennens zusammengesetzt; Freges Begriff des Urteilens ist nicht kumulativ. Vielmehr ist ein Urteil ein einziger Akt des Anerkennens eines Gedankens *als wahr* oder des *wahrerweise* Denkens. Aus diesem Grund unterbreite ich im letzten Abschnitt einen Vorschlag, der von Freges adverbialer Bestimmung des Urteilens ausgeht. Für die Adverbialtheorie sind Urteile nicht faktiv, sondern normativ; sie eignet sich daher sowohl für Urteile im Kontext von logischen Schlüssen als auch für spontane Urteile.

Was heißt es zu urteilen?

Es gehört zu den Verdiensten Freges, die Logik von den traditionellen Urteilsformen befreit zu haben: Logische Verhältnisse sind auf Negation, Konditional und Quantifikation rückführbar und bestimmen den Inhalt von Urteilen. Diese

Verlagerung macht die Unterscheidung zwischen verschiedenen Urteils*formen* überflüssig, jedoch nicht den Urteils*begriff*.

Angesichts der Tatsache, dass Urteilen eine „logische Urtätigkeit" (NS 16) ist, gibt Frege erstaunlich wenig Auskunft darüber, was es heißt zu urteilen. Im Wesentlichen erfahren wir, dass Urteilen vom Denken ganz verschieden ist: „Ein Urteil ist mir nicht das bloße Fassen eines Gedankens, sondern die Anerkennung seiner Wahrheit" (SB 34; vgl. G 62). Gelegentlich scheint Frege sogar in Erwägung zu ziehen, dass einem Urteilsakt ein Akt des bloßen Denkens zeitlich vorausgeht: „Man kann einen Gedanken nicht als wahr anerkennen, bevor man ihn gefasst hat" (NS 271; vgl. NS 8). Und manchmal charakterisiert er das Urteilen als „Fortschreiten von einem Gedanken zu seinem Wahrheitswert" (SB 35; vgl. WB 96). Solche Formulierungen suggerieren, dass Urteilen gewissermaßen *mehr* ist als bloßes Gedankenfassen: Wer urteilt, denkt nicht nur, sondern anerkennt zusätzlich die Wahrheit des Gedachten. Es entsteht der trügerische Eindruck, dass Urteile Zweiakter sind, bestehend aus einem bloßen Akt des Denkens und einem Akt des Anerkennens. In der Frege-Literatur findet der *kumulative* Urteilsbegriff, Wittgensteins Mahnen ungeachtet (PU §22), weite Verbreitung.

Eine weitere Hürde für das richtige Verständnis von Freges Urteilsbegriff stellt dessen breiter Anwendungsbereich dar. Unter *Urteil* fallen für Frege sowohl logische Schlüsse und Axiome als auch Wahrnehmungsurteile und spontane Urteile, wie sie etwa beim Behaupten gefällt werden. Da die Bedingungen dafür, etwas ein Urteil zu nennen, in der Logik, der Erkenntnistheorie und in der Sprachphilosophie sehr verschieden sind, scheint es um eine universale Konzeption des Urteilens schlecht bestellt zu sein. Mit der Adverbialtheorie des Urteilens will ich einen Vorschlag unterbreiten, der im dargelegten Sinn universal ist und die Probleme des kumulativen Urteilsbegriffs vermeidet.

Urteile stehen in einem besonderen Verhältnis zur Wahrheit, soviel steht fest. Der Zusammenhang zwischen Urteilen und Wahrheit wird oft bildlich als „internal link" (Sluga 2001, 80) oder „intimate relation" (Heck/May 2007, 18) beschrieben; die Begriffe Urteil und Wahrheit seien „intertwined" (Ricketts 1996, 130), da Urteile auf Wahrheit „abzielen" würden (vgl. Kremer 2000, 579, Burge 2005, 16 und Reck 2007, 158). Solche Metaphern sind allerdings nur bedingt erhellend und lassen Neugierige über die rätselhafte Verbindung zwischen Urteilen und Wahrheit im Ungewissen. Drei etwas handfestere Antworten scheiden ebenfalls aus: Urteilen besteht weder im *Prädizieren* von Wahrheit, noch im *Referieren* auf Wahrheit, und auch nicht im *Hinzudenken* des Sinns von „wahr". Auf alle drei Vorschläge werde ich im Folgenden kurz eingehen.

Die Einführung der Unterscheidung zwischen Sinn und Bedeutung erlaubt es Frege, seinen Formalismus zu präzisieren. Die Begriffsschrift-Revision soll sich auch auf den Urteilsbegriff auswirken, denn fortan versteht Frege unter einem Urteil nicht mehr das „Bejahen" eines beurteilbaren Inhalts, sondern das Aner-

kennen der Wahrheit dessen, was er den *Sinn* eines Aussagesatzes nennt. Die *Bedeutung* eines solchen Satzes ist ein Wahrheitswert, und Wahrheitswerte sind Gegenstände – das Wahre und das Falsche. Da Gegenstände nicht prädiziert werden können, wird Frege nicht gemeint haben, dass wir beim Urteilen die Wahrheit prädizieren. Der gegenstandstheoretische Wahrheitsbegriff schließt freilich nicht aus, dass Frege einem Wahrheitsprädikat vollends abgeneigt ist. Tatsächlich drückt er sich in seinen späten Schriften gelegentlich so aus, als ob Wahrheit auch ein Prädikat sein könnte, doch vermutlich ist ihm vorrangig daran gelegen, dass das Wort „wahr" sprachlich als Eigenschaftswort „erscheint" (G 59). Jedenfalls wird beim Urteilen kein Wahrheitsprädikat zugesprochen, wie Frege mit der folgenden Überlegung klarstellt:

> Man gelangt durch die Zusammenfügung von Subjekt und Prädikat immer nur zu einem Gedanken, nie von einem Sinne zu dessen Bedeutung, nie von einem Gedanken zu dessen Wahrheitswerte. (SB 35)

Das Resultat einer Wahrheitsprädikation ist kein Urteil, sondern ein neuer, komplexerer Gedanke, von dem selbst wiederum Wahrheit prädiziert werden müsste. Da dies in einen *infiniten Regress* führen würde, scheidet der prädikative Urteilsbegriff aus.[1]

Die soeben grob skizzierte Semantik verleitet zu der Annahme, dass Frege unter einem Urteil nicht das Zusprechen eines Wahrheitsprädikats, sondern die *Referenz* auf das Wahre versteht: „judging that *p* is attempting to refer, by thinking that *p*, to the True" (Heck/May 2007, 19-20). Der Vorschlag eines referentiellen Urteilsbegriffs hat vor allem einen Haken – er ist nicht allgemein genug. Erstens lässt er offen, was Frege *vor* der Einführung der Unterscheidung zwischen Sinn und Bedeutung unter einem Urteil verstanden hat. Zweitens ist Referenz eine semantische Beziehung zwischen Zeichen und Bezeichnetem – über diese Tatsache kann auch der Einschub „by thinking that *p*" nicht hinwegtäuschen. Von Freges Begriffschriftsätzen lässt sich problemlos sagen, dass sie Versuche sind, auf das Wahre zu referieren. Aber nicht alle Urteile beinhalten Zeichen; inwiefern sollen wir etwa beim Wahrnehmen auf das Wahre referieren?

Ein Urteil wird auch nicht dadurch gefällt, dass man zu einem Gedanken den Sinn von „wahr" hinzudenkt. Das Wort „wahr" hat zwar einen Sinn, denn sonst „hätte auch ein Satz, in dem ‚wahr' als Prädikat vorkäme, keinen Sinn" (NS 272), aber dieser ist inhaltsleer und trägt zum Sinn eines Satzes nichts bei: Das Wort „wahr" liefert durch seinen Sinn „keinen wesentlichen Beitrag zum Gedanken" (NS 271). Daher ist es einerlei, ob wir behaupten „Es ist wahr, dass Meerwasser salzig ist" oder einfach nur „Meerwasser ist salzig" – beide Sätze drücken denselben Gedanken aus. Obendrein argumentiert Frege mehrfach für die These, dass Wahrheit immer mitbehauptet oder mitgedacht wird, wenn etwas

1 Auf diesen infiniten Regress hat bereits Heck (2002, 86) hingewiesen.

behauptet oder gedacht wird (vgl. NS 140, G 61). Diese Beobachtung wird unter dem Etikett *Omnipräsenzthese* verhandelt (vgl. Burge 2005, 127; Künne 2003, 34). Wenn der Sinn von „wahr" gedanklich allgegenwärtig ist und das Anerkennen der Wahrheit eines Gedankens das Urteilen vom bloßen Denken unterscheidet, dann kann Urteilen nicht im Hinzudenken dieses Sinns bestehen, denn sonst wäre jeder gefasste Gedanke gleichsam ein Urteil.

Die Möglichkeit falscher Urteile

Urteile würden dann eine möglichst enge Bindung mit der Wahrheit eingehen, wenn sie wie Wissen Wahrheit *implizieren* würden. Unter der Voraussetzung eines *faktiven* Urteilsbegriffs wäre es schließlich unmöglich, falsche Urteile zu fällen; aus meinem Urteil, dass *p*, folgt, dass *p* wahr ist. Dennoch haben mehrere Autoren dafür argumentiert, dass Freges Urteilsbegriff faktiv ist. Carl schreibt Urteilen denselben epistemischen Status wie dem Wissen zu: „to make a judgement is not just to make a claim to knowledge, such a judgement is really knowledge that a particular thought is true" (Carl 1994, 144). Nach Ricketts (1986, 78) ist die Beziehung zwischen Urteilen und Wahrheit nicht „casual", Frege habe daher für die Charakterisierung des Urteilens bewusst das faktive Verb „anerkennen" gewählt. Er stützt seine Interpretation mit Wörterbüchern aus dem 19. Jahrhundert, wo „anerkennen" als „stärkeres erkennen" – also eindeutig faktiv – bestimmt wird und kommt zu dem Schluss: „To make a judgment is to acquire a piece of knowledge" (Ricketts 1996, 131).

Der faktive Urteilsbegriff wird oft vorschnell als unplausibel abgetan. Erstens ist das Verb „anerkennen" mit dem faktiven „erkennen" etymologisch verwandt (vgl. Stepanians 1998, 83) und wird in der englischsprachigen Fachliteratur oft mit faktiven Verben wie „recognize" oder „acknowledge" wiedergegeben. Zweitens, und viel wichtiger, ist ein faktiver Urteilsbegriff vor dem Hintergrund von Freges Logizismus und dem damit einhergehenden Schlussbegriff nicht abwegig:

> Ein Schluss [...] ist eine Urteilsfällung, die auf Grund schon früher gefällter Urteile nach logischen Gesetzen vollzogen wird. Jede der Prämissen ist ein bestimmter als wahr anerkannter Gedanke, und im Schlussurteil wird gleichfalls ein bestimmter Gedanke als wahr anerkannt. (KS 303-4)

Mit dem zweiten Satz deutet der Vater der modernen Logik an, dass für ihn beim logischen Schließen mehr auf dem Spiel steht als dies ein Logiker heute zugestehen würde. Noch deutlicher wird Frege in einem Brief an Philip Jourdain: „Was als Prämisse eines Schlusses dienen soll, muss wahr sein" (WB 127). Diese Forderung ist nachvollziehbar, wenn man bedenkt, dass Frege in erster Linie darum bemüht war, den arithmetischen Wahrheiten zu einem sicheren Fundament zu verhelfen. Logische Schlüsse sind für Frege stets Urteile in-

nerhalb von Schlussketten, deren erste Glieder jeweils Axiome sind – das sind Gedanken, die unmittelbar selbst einleuchten. Kurz: Schließen ist Beweisen und „beweisen“ ist ein faktives Verb, denn falsche Beweise gibt es genau so wenig wie falsches Wissen.[2]

Der Logizismus erklärt bestenfalls Freges Präferenz für wahre Urteile, aber er rechtfertigt sie nicht. Auf die Feststellung, dass die Gründe, welche die Anerkennung einer Wahrheit rechtfertigen, „oft in anderen schon anerkannten Wahrheiten“ (NS 3) liegen, ist man zu erwidern geneigt: Oft, aber eben nicht immer! Wie bereits gesagt, spricht Frege nicht ausschließlich im Kontext von Schlüssen über das Urteilen, doch für Wahrnehmungsurteile und Behauptungen taugt der faktive Urteilsbegriff nicht. Über die Möglichkeit falscher Urteile sagt Frege hingegen herzlich wenig: „Wir können irren“ (NS 2) und „Beim Wahren ist ein Irrtum möglich“ (NS 143). Soll Freges Urteilstheorie etwa just jener Prüfung nicht standhalten, welcher sich gemäß Russell alle philosophischen Theorien zu unterziehen haben?

> A good many philosophers [...] have constructed theories according to which all our thinking ought to have been true, and have then had the greatest difficulty in finding a place for falsehood. (Russell, 1912, 70)

Dem Problem von falschen Urteilen oder Fehlurteilen schenkt Stepanians (1998) als einer der ersten Interpreten die erforderliche Aufmerksamkeit. Er bestreitet die Faktizität von Freges Urteilsbegriff und verweist auf den *juristischen* Sinn des Wortes „anerkennen“ in Freges Standardcharakterisierung. Wer ein Urteil fällt, anerkennt die Wahrheit eines Gedankens im Sinne einer Legitimation oder Billigung; so wie etwa die Macht eines Königs anerkannt werden kann, so wird beim Urteilen der Anspruch eines Gedankens auf Wahrheit anerkannt (vgl. Stepanians 1998, 83ff.). Belegt wird die These, dass Gedanken einen Wahrheitsanspruch erheben, mit der gedanklichen Omnipräsenz des Sinns von „wahr“: Für den Denkenden ist es einerlei, ob er denkt, dass *p*, oder ob er denkt, dass es wahr ist, dass *p* – der Gedanke, unabhängig ob wahr oder falsch, präsentiert sich ihm als wahr. Frege spricht sogar davon, dass gefasste Gedanken dazu „drängen“, die Frage nach ihrem Wahrsein zu beantworten (vgl. NS 183). Die Omnipräsenz des Sinns von „wahr“ deutet Stepanians als Wahrheitsanspruch von Gedanken: „Wer urteilt, dass *p*, der bestätigt gewissermaßen nur noch, was der Gedanke ausdrückt: dass es wahr ist, dass *p*“ (Stepanians 1998, 89).

Diese Interpretation weist viele Vorzüge auf, allen voran schließt sie die Möglichkeit falscher Urteile nicht aus (Ansprüche werden manchmal fälschlicherweise oder zu Unrecht eingeräumt), sie hat aber auch ihre Schwächen. Erstens ist es kontraintuitiv, dass *alle* Gedanken, also auch falsche, einen Anspruch auf Wahrheit erheben; selbst wenn alle Gedanken zum Urteilen auffordern wür-

2 Bereits Anscombe (1959, 115) hat darauf hingewiesen, dass Freges Doktrin, dass die Prämissen von Schlüssen wahr sein müssen, im Kontext von Beweisen annehmbar ist.

den, dann müssten sie über ein performatives Element verfügen und nicht nur das allgegenwärtige konstative „es ist wahr, dass...“ mit sich führen. Zweitens ist es falsch, dass *jeder gefasste* Gedanke einen Wahrheitsanspruch erhebt. Wenn Gedanken in einem Konditional eingebettet sind, erhebt nur das gesamte Gedankengefüge einen Anspruch auf Wahrheit, nicht aber die einzelnen Gedanken, gleichwohl werden die Teilgedanken gefasst.[3] Aus diesem Grund ist es auch nicht richtig, die Omnipräsenz des Sinns von „wahr“ mit dem Wahrheitsanspruch von Gedanken gleichzusetzen. Drittens sind es nicht Gedanken, die einen Wahrheitsanspruch erheben, sondern Menschen, die Gedanken fassen; der Vorschlag steht unter dem Verdacht, Gedanken zu anthropomorphisieren. Viertens liegt dem Vorschlag, trotz Stepanians' überzeugender Kritik am „Zwei-Stufen-Modell“, ein kumulativer Urteilsbegriff zu Grunde; das Hervorbringen eines Anspruchs und das Stattgeben dieses Anspruchs sind ebenso zwei verschiedene Tätigkeiten, wie das Ausarbeiten und das Unterschreiben eines Vertrags.[4]

Adverbialismus

In diesem letzten Teil schlage ich eine Alternative vor: die Adverbialtheorie des Urteilens. Will man über Freges Urteilsbegriff etwas in Erfahrung bringen, kommt man um seine Standardcharakterisierung nicht herum und diese hat genau genommen zwei Ausprägungen:

(N) die Wahrheit eines Gedankens anerkennen
(A) einen Gedanken als wahr anerkennen

Mehrere Gründe sprechen dafür, die adverbiale Konstruktion (A) der nominalen Konstruktion (N) vorzuziehen. Erstens fällt auf, dass Frege (A) weitaus häufiger verwendet als (N).[5] Zweitens erweckt (N) den Anschein, Wahrheit zu prädizieren, so wie in „den Bart eines Kunden rasieren“ die Eigenschaft bärtig zugesprochen wird.[6] Die Konstruktion in (A) hingegen lässt die Idee erst gar nicht aufkommen, dass von einem Gedanken Wahrheit prädiziert wird, da der adverbiale Ausdruck „als wahr“ die *Art und Weise* bestimmt, wie der Gedanke anerkannt wird. Auf den dritten und größten Vorteil von (A) gehe ich im folgenden

3 Siehe hierzu Freges Kommentar zum Beispielsatz „wenn jetzt die Sonne schon aufgegangen ist, ist der Himmel stark bewölkt“ (SB 46).

4 Mit diesem juristischen Vergleich erläutert Hare (1989, 25) die Funktion von Freges Urteilsstrich.

5 In meiner Zählung verwendet Frege in 70% der Fälle (A); für eine vollständige und kommentierte Liste sämtlicher Vorkommnisse von Freges Konstruktionen mit „anerkennen“ siehe Künne (2010, 432ff.).

6 Damit will ich nicht behaupten, dass *jedes* Nomen ein Prädikat seines Genitivattributs ist.

Abschnitt ausführlicher ein; er besteht darin, dass (A) keinen faktiven Charakter aufweist.[7]

Die beiden Linguisten Kiparski und Kiparski (1970) haben festgestellt, dass Sätze mit faktiven Verben wie „wissen“, „lernen“, „erinnern“ oder „bedauern“ ihre Präsuppositionen bewahren, wenn sie negiert oder zu Fragesätzen umgeformt werden: „Charlotte bedauert (nicht), dass es regnet“ und „Bedauert Charlotte, dass es regnet?“ setzen beide voraus, dass es regnet. Werden (N) und (A) auf dieses Kriterium hin überprüft, gelangt man zum Ergebnis, dass (N) einen faktiven Charakter hat, (A) hingegen nicht.[8] Da „anerkennen“ sowohl in (A) als auch in (N) vorkommt, kann der faktive Charakter von (N) nicht vom Verb herrühren, sondern muss auf die in (N) enthaltene bestimmte Kennzeichnung zurückgeführt werden. Bereits Strawson (1950) hat darauf aufmerksam gemacht, dass mit der Verwendung von bestimmten Kennzeichnungen die Existenz des Bezeichneten präsupponiert wird. Hierin unterscheidet sich die Kennzeichnung „der König von Frankreich“ nicht von der Kennzeichnung „die Wahrheit des Gedankens“. Die bestimmte Kennzeichnung in (N) hat denselben Effekt wie die Kennzeichnung „die Tatsache, dass *p*“ – sie sorgt dafür, dass selbst nichtfaktive Verben wie „glauben“ in faktiven Konstruktionen vorkommen können („Glaubst du an die Tatsache, dass es schneit?“). Diese drei Gründe sprechen für die Präferenz von (A).

Was heißt es nun, einen Gedanken *als wahr anzuerkennen*? Ein Vergleich mit der Adverbialtheorie der Wahrnehmung ist hierfür hilfreich. Die Adverbialtheorie der Wahrnehmung bricht mit der traditionellen Unterscheidung zwischen dem Akt des Wahrnehmens und dem wahrgenommenen Objekt und begreift den Inhalt der Wahrnehmung als Modus des Wahrnehmungsaktes. Ducasse schildert diese Position wie folgt:

> To sense blue is then to sense bluely, just as to dance the waltz is to dance “waltzily” (i.e., in the manner called “to waltz”) to jump a leap is to jump “leapily” (i.e., in the manner called “to leap”) etc. Sensing blue, that is to say, is I hold a species of sensing – a specific variety of the sort of activity generically called “sensing”. (Ducasse 1942, 232-3)

Die Behauptung, das Wahrnehmen einer Farbe sei eine Tätigkeit wie Tanzen oder Springen, ist kontrovers, doch der Vergleich macht eines deutlich: Walzertanzen ist eine *Art und Weise* des Tanzens und diese Art ist *konstitutiv* für das Walzertanzen. Wer nicht nach einer bestimmten Schrittfolge tanzt und sich nicht in einer bestimmten Art und Weise im Rhythmus dreht und dabei eine bestimmte Körperhaltung einnimmt, der tanzt vielleicht, aber er tanzt keinen Walzer.

---

7 Ich ziehe es vor, von einem faktiven „Charakter“ zu sprechen, da weder (A) noch (N) faktiv sind, wie ich im nächsten Abschnitt zu zeigen versuche.

8 Vgl. hierzu die Satzpaare „Sie anerkennt die Wahrheit (nicht), dass *p*“/„Anerkennst du die Wahrheit, dass *p*?“ und „Sie anerkennt *p* (nicht) als wahr“/„Anerkennst du *p* als wahr?“

Entsprechendes ließe sich auch vom Urteilen und Denken sagen: So wie das Walzertanzen eine Art des Tanzens ist, so ist das Urteilen eine Art des Denkens. Wer nicht in einer bestimmten Art und Weise denkt, urteilt nicht, denn für ein Urteil reicht es nicht, bloß zu denken, dass *p*, sondern man muss *p als wahr* denken oder *wahrerweise* denken, dass *p*.

Man wird diesem Vorschlag vermutlich nur dann etwas abgewinnen können, wenn es gelingt, das wahrerweise Denken eines Gedankens schärfer zu umreißen. Ansonsten sind wir genau so ratlos, wie ein Einsteiger nach der Anweisung des Tanzlehrers „dance *waltzily*!". Ich erlaube mir, Ducasses Vergleich für einen weiteren Schritt in meiner Argumentation zu strapazieren, und gebe zu bedenken, dass es gute Tanzlehrer natürlich nicht bei dieser unnützen Anweisung belassen, sondern in der Lage sind, die Art und Weise des Tanzens vorzuführen; sie zeigen wie man tanzen muss, wenn man walzern will. Hinter der adverbialen Bestimmung „wahrerweise" steckt im Grunde nichts anderes – Urteilen und Schließen sind Teil einer regelgeleiteten Praxis, und wer sich an dieser beteiligen will, muss wahrerweise denken. Wahrheit ist ein normatives Ziel von Urteilsakten, ein Ziel, das man beim Vollzug dieser Akte im Auge haben sollte, und nur wer wahrerweise denkt, verfolgt dieses Ziel. Frege bestimmt die Wahrheit oft als Ziel – man denke etwa an den Anfang der *Logischen Untersuchungen* – und fragt: „Wie muss ich denken, um das Ziel, die Wahrheit, zu erreichen?" (NS 139). Frege fragt nicht, *was* muss ich denken, um das Ziel zu erreichen, sondern *wie* muss ich denken. Das Urteilen ist eine Art und Weise des Denkens, die darum bemüht ist, das Ziel, die Wahrheit, zu erreichen. Weil die Logik Mittel bereitstellt, die helfen, dieses Ziel zu erreichen, erinnert Frege daran, „[d]ass die logischen Gesetze Richtschnuren für das Denken sein sollen zur Erreichung der Wahrheit" (GGA, XV).

Die Adverbialtheorie des Urteilens weist zusammenfassend eine ganze Reihe von Vorzügen auf. Erstens trägt sie dem Umstand Rechnung, dass die adverbiale Konstruktion (A) Freges präferierte Standardcharakterisierung für das Urteilen ist. Ferner steht sie im Einklang mit Freges frühen Theorie des Urteilens als Bejahen. Es ist nämlich nicht plausibel anzunehmen, dass Frege im Zuge der Revision seiner Begriffsschrift den intuitiv nachvollziehbaren Begriff des Bejahens durch einen rätselhaften Begriff des Anerkennens ersetzt. Zweitens haben wir gesehen, dass die Adverbialtheorie auch falsche Urteile zulässt, indem sie nicht einen faktiven, sondern einen normativen Urteilsbegriff unterstützt; d.h. Urteile sind nicht immer wahr, aber sie sollten es sein. Drittens taugt die Adverbialtheorie für alle Arten von Urteilen (Schlüsse, Wahrnehmung etc.). Sie hängt auch nicht von Zeichen ab, mittels derer wir beim Urteilen auf das Wahre referieren, und ist daher im eingeforderten Sinne universal. Viertens muss die Adverbialtheorie kein Wahrheitsprädikat bemühen, da die adverbiale Konstruktion (A) in keiner Weise suggeriert, dass Wahrheit eine Eigenschaft von Gedanken ist; die Adverbialtheorie ist nicht prädikativ. Schließlich vermeidet die Adverbialtheorie

den kumulativen Urteilsbegriff, indem sie das Urteilen nicht als kategorial verschieden vom Denken, sondern als eine bestimmte Art und Weise des Denkens begreift. Wer ein Urteil fällt, tut nicht zwei Dinge, sondern verrichtet nur eine Sache – aber auf eine bestimmte Art und Weise.

Literaturverzeichnis

Anscombe, Elizabeth: An Introduction to Wittgenstein's Tractatus. Hutchinson, London, 1959

Burge, Tyler: Truth, Thought, Reason. Essays on Frege. Oxford University Press, Oxford, 2005

Carl, Wolfgang: Freges Theory of Sense and Reference. Cambridge University Press, New York, 1994

Ducasse, Curt J.: „Moore's The Refutation of Idealism". In: *Schilpp, P. A. (Hrsg.):* The Philosophy of G. E. Moore. Northwestern University Press, Evanston, 1942

Frege, Gottlob [SB]: „Über Sinn und Bedeutung". Zeitschrift für Philosophie und philosophische Kritik, 100, 1892. S. 25-50; hier zit. nach *Frege, Gottlob* [KS]: Kleine Schriften. hrsg. von Angelelli, I.. Georg Olms, Hildesheim, 1990

Frege, Gottlob [GGA]: Grundgesetze der Arithmetik. Jena, 1893; zit. nach: Grundgesetze der Arithmetik I/II. Georg Olms, Hildesheim, 1998

Frege, Gottlob [G]: „Der Gedanke". Beiträge zur Philosophie des Deutschen Idealismus I. 1918/19, S. 58-77; hier zit. nach *Frege, Gottlob* [KS]: Kleine Schriften. hrsg. von Angelelli, I.. Georg Olms, Hildesheim, 1990

Frege, Gottlob [WB]: Gottlob Frege: Wissenschaftlicher Briefwechsel. hrsg. von Gabriel, G./Hermes, H./Kambartel, F./ Thiel, C./Veraart, A.. Felix Meiner, Hamburg, 1976

Frege, Gottlob [NS]: Gottlob Frege: Nachgelassene Schriften. hrsg. von Hermes, H./Kambartel, F./Kaulbach, F.. Felix Meiner, Hamburg, 1983

Frege, Gottlob [KS]: Kleine Schriften. hrsg. von Angelelli, I.. Georg Olms, Hildesheim, 1990

Hare, Richard: „Some sub-atomic Particles of Logic". Mind, 98, 1989

Heck, Richard: „Meaning and Truth-Conditions: A Reply to Kemp". The Philosophical Quarterly, 52, 2002

Heck, Richard/May, Robert: „Frege's Contribution to Philosophy of Language". In: *Lepore, E., Smith, B.C. (Hrsg.)*: The Oxford Handbook of Philosophy of Language. Oxford University Press, Oxford, 2007

Kiparski, Paul/Kiparski, Carol: „Fact". In: *Bierwisch, M./Heidolph, K.E. (eds.):* Progress in Linguistics. Mouton, The Hague, 1970

Kremer, Michael: „Judgement and Truth in Frege". Journal of the History of Philosophy, 38, 2000

Künne, Wolfgang: Conceptions of Truth. Oxford University Press, New York, 2003

Künne, Wolfgang: Die Philosophische Logik Gottlob Freges. Klostermann, Frankfurt/M., 2010

Reck, Erich: „Frege on Truth, Judgment, and Objectivity". Grazer Philosophische Studien, 75, 2007

Ricketts, Thomas: „Objectivity and Objecthood: Frege's Metaphysics of Judgement". In: *Haaraparanta, L./Hintikka, J. (Hrsg.):* Frege Synthesized. Reidel, Dordrecht, 1986

Ricketts, Thomas: „Logic and Truth in Frege". The Aristotelian Society, Suppl. Vol. 70, 1996

Russell, Bertrand: The Problems of Philosophy. Oxford University Press, New York, 1912/2001

Sluga, Hans: „Frege on the Indefinability of Truth". In: *Reck, E. (Hrsg.):* From Frege to Wittgenstein. Oxford, 2002. S.75–95

Stepanians, Markus S.: Frege und Husserl über Urteilen und Denken. Mentis, Paderborn, 1998

Strawson, Peter F.: „On Referring". Mind, 59, 1950

Wittgenstein, Ludwig [PU]: Philosophische Untersuchungen. Suhrkamp, Frankfurt/M., 1984

Hybride Propositionen und aposteriorische Notwendigkeit

Silvère Schutkowski
silvere.schutkowski@uni-erfurt.de
Universität Erfurt

Abstract/Zusammenfassung

The paper examines arguments for the thesis that necessity does not imply apriority, i.e. that there are a posteriori necessities. The result of the discussion impinges on the question whether necessity might be explainable on the basis of analyticity (section 2). To evaluate these theses and arguments I shall specify the a priori/a posteriori distinction in greater detail and introduce the notion of hybrid propositions (section 3). As an immediate result, the problem of alethically unknowable propositions arises for the thesis that necessity implies apriority (section 4). Ignoring those propositions, I then turn to the standard argument in favour of the existence of a posteriori necessities. It will be argued that the standard argument faces a dilemma (section 5). Subsequently, I discuss an alternative argument by Kitcher against the thesis that necessity implies apriority (section 6). Finally, I deal with the question whether Kitcher's argument faces a similar dilemma (section 7).

In dieser Arbeit gehe ich den Argumenten für die These nach, dass Notwendigkeit nicht Apriorität impliziert bzw. dass es aposteriorische Notwendigkeiten gibt. Das Ergebnis dieser Diskussion hat Auswirkung auf die Frage, ob Notwendigkeit vermittels Analytizität erklärt werden kann (Abschnitt 2). Für die Bewertung der Thesen und Argumente muss zunächst die A priori-a posteriori-Unterscheidung genauer bestimmt und der Begriff der hybriden Proposition eingeführt werden (Abschnitt 3). Für die These, dass Notwendigkeit Apriorität impliziert, ergibt sich daraus umgehend das Problem alethisch unerkennbarer Propositionen (Abschnitt 4). Nach Ausklammerung der alethisch unerkennbaren Propositionen wende ich mich dem Standardargument für die Existenz aposteriorischer Notwendigkeiten zu und konfrontiere es mit einem Dilemma (Abschnitt 5). Anschließend wird ein alternatives Argument Kitchers gegen die These vorgestellt, dass Notwendigkeit Apriorität nach sich zieht (Abschnitt 6). Abschließend wende ich mich der Frage zu, ob auch Kitchers Argument einem analogen Dilemma ausgesetzt ist (Abschnitt 7).

1. Analytizität, Notwendigkeit, Apriorität

Analytizität, Notwendigkeit und Apriorität sind u.a. Eigenschaften von Propositionen.[1] Zwei Beziehungen zwischen diesen Eigenschaften sind recht unumstritten und werden auch hier vorausgesetzt. Erstens impliziert die Analytizität einer

1 Propositionen werden im Folgenden als Bedeutungen von Aussagesätzen, als Objekte propositionaler Einstellungen sowie als Wahrheitswertträger verstanden.

Proposition ihre Apriorität. Zweitens impliziert die Analytizität einer Proposition ihre Notwendigkeit[2]. Weitere Beziehungen zwischen diesen Eigenschaften sind umstritten, bspw. ob aus der Notwendigkeit einer Proposition ihre Apriorität folgt. Bekanntlich behauptet Saul Kripke (1980), die Implikationsthese:

(Impl.) Notwendigkeit impliziert Apriorität

sei falsch. Hingegen hält er die Existenzthese:

(Exist.) Es gibt aposteriorische Notwendigkeiten

für wahr.

2. Eine theoretische Konsequenz aposteriorischer Notwendigkeit

Es könnte behauptet werden, unser Verständnis von Analytizität sei besser als unser Verständnis von Notwendigkeit. Daher wäre es erfreulich, wenn die Notwendigkeit einer Proposition vermittels ihrer Analytizität erklärt werden könnte.[3] Dass Notwendigkeit Analytizität impliziert, ist eine notwendige Bedingung für diese Erklärung.

Der Nachweis aposteriorischer Notwendigkeiten würde u.a. zeigen, dass diese Erklärung unerreichbar ist. Weil Analytizität wie angenommen Apriorität impliziert, gilt nämlich: Würde Notwendigkeit Analytizität implizieren, dann würde Notwendigkeit Apriorität implizieren. Gibt es jedoch aposteriorische Notwendigkeiten, wie Kripke behauptet, dann kann Notwendigkeit nicht Apriorität implizieren. Folglich kann Notwendigkeit nicht Analytizität implizieren.[4]

3. Die negative[5] A priori-a posteriori-Unterscheidung

Wenn es gelingt, die Existenz aposteriorischer Notwendigkeiten nachzuweisen, dann ist die Erklärung von Notwendigkeit vermittels Analytizität zum Scheitern

2 Die Proposition *p* ist genau dann analytisch, wenn sie analytisch wahr oder analytisch falsch ist. Analytisch zu sein ist ein allgemeiner, analytisch wahr zu sein ein spezifischer semantischer Status. Analog ist notwendig zu sein ein allgemeiner, notwendig wahr zu sein ein spezifischer modaler Status. Vgl. Casullo (2003), S. 91.

3 Es ist umstritten, ob und in welchem Sinne Notwendigkeit überhaupt erklärt werden kann. Vgl. u.a. Blackburn (1986) sowie Hale (2002).

4 Vgl. Williamson (2007), S. 51.

5 Im Folgenden werden die Ausdrücke „a priori“ und „a posteriori“ negativ bestimmt, um größtmögliche Neutralität zu gewährleisten. Ich werde also nur von der Abhängigkeit oder Unabhängigkeit der Erkenntnis von Erfahrung reden, jedoch offenlassen, was genau hier unter „Erfahrung“ zu verstehen ist.

verurteilt. Für eine Bewertung der umstrittenen Implikations- (Impl.) und Existenzthese (Exist.) sowie der entsprechenden Argumente muss deren Gehalt genauer bestimmt werden. Und dies setzt u.a. eine genauere Bestimmung apriorischer Erkenntnis voraus.

Einer ersten Charakterisierung zufolge ist eine Proposition genau dann a priori, wenn wir den Wahrheitswert dieser Proposition unabhängig von Erfahrung erkennen können.[6] Und eine Proposition ist genau dann a posteriori, wenn wir ihren Wahrheitswert vermittels Erfahrung erkennen können.

Ich werde eine Proposition genau dann als *hybrid* bezeichnen, wenn wir ihren Wahrheitswert sowohl unabhängig von als auch vermittels Erfahrung erkennen können. Hybride Propositionen, sofern es welche gibt, sind gemäß der ersten Charakterisierung sowohl a priori als auch a posteriori. Die Unterscheidung wäre somit nicht ausschließend. Paul Boghossian und Christopher Peacocke tragen hybriden Propositionen Rechnung, indem sie die erste Charakterisierung entsprechend anpassen:

> An a priori proposition is one such that *there is a way* of coming to know it under which the thinker's entitlement to accept the proposition does not involve the character of the thinker's experience. An a posteriori proposition is one such that *any way* of coming to know it will involve an entitlement which does concern the character of the thinker's experience. (Boghossian, Peacocke (2000), S. 2, meine Hervorhebung)

Hybride Propositionen werden als a priori klassifiziert:

(Def. 1) Eine Proposition ist genau dann a priori, wenn wir ihren Wahrheitswert erkennen können und mindestens eine der Erkenntnisweisen unabhängig von Erfahrung ist.

(Def. 2) Eine Proposition ist genau dann a posteriori, wenn wir ihren Wahrheitswert erkennen können und jede Erkenntnisweise Erfahrung einschließt, d.h. wenn wir ihren Wahrheitswert *nur* vermittels Erfahrung erkennen können.

Die umstrittenen Thesen besagen somit:

(Impl.) Wenn eine Proposition notwendig ist, dann können wir ihren Wahrheitswert unabhängig von Erfahrung erkennen.

(Exist.) Es gibt notwendige Propositionen, deren Wahrheitswerte wir *nur* vermittels Erfahrung erkennen können.

Zwei Konsequenzen dieser Definitionen sind hervorzuheben. Erstens sind Propositionen, deren Wahrheitswerte für uns unerkennbar sind, weder a priori noch a posteriori. Diese Eigenschaften kommen per definitionem nur Propositionen zu, deren Wahrheitswerte wir erkennen können. Innerhalb der Klasse aller Propositionen ist diese Unterscheidung also nicht erschöpfend.

---

6 Den Wahrheitswert einer Proposition erkennen, heißt wissen, dass *p*, falls die Proposition *p* wahr ist, bzw. wissen, dass nicht-*p*, falls die Proposition *p* falsch ist.

Zweitens ist innerhalb der Klasse der alethisch erkennbaren Propositionen jedes Element entweder a priori oder a posteriori. Hinsichtlich der alethisch erkennbaren Propositionen ist die Unterscheidung somit ausschließend und erschöpfend.[7]

4. Alethisch unerkennbare Propositionen

Die erste Konsequenz der Charakterisierung legt bereits eine Schwachstelle der Implikationsthese offen. (Impl.) besagt, dass wir den Wahrheitswert jeder notwendigen Proposition unabhängig von Erfahrung erkennen können. Von einem realistischen Standpunkt aus betrachtet, sollten jedoch notwendige Propositionen zugelassen werden, deren Wahrheitswerte für uns gar nicht erkennbar sind. Unter diesen Umständen gibt es notwendige Propositionen, die *nicht a priori* sind, weil sie gar nicht erkannt zu werden vermögen. (Impl.) ist falsch.

Den Freunden aposteriorischer Notwendigkeit ist damit jedoch nicht gedient. Denn die nicht-erkennbaren, notwendigen Propositionen sind den Definitionen entsprechend auch *nicht a posteriori*. Der Schluss von der Nicht-Apriorität einer notwendigen Proposition auf ihre Aposteriorität (Exist.) ist blockiert.[8] Wenn wir einen realistischen Standpunkt einnehmen, folgt somit die Falschheit der Implikationsthese (Impl.), jedoch nicht ohne Weiteres die Wahrheit der Existenzthese (Exist.). Die These, Notwendigkeit impliziere nicht Apriorität, erweist sich als *nicht äquivalent* mit der These, es gäbe aposteriorische Notwendigkeiten.[9]

Während Realisten bezüglich des Nichterkennbaren die Existenz notwendiger, aber alethisch unerkennbarer Propositionen akzeptieren dürften, behaupten Antirealisten bezüglich des Nichterkennbaren, dass es zwar notwendige Propositionen gibt, deren Wahrheitswerte de facto nicht erkannt wurde, jedoch keine,

---

7 Es gibt ein Verständnis von „a priori“ und „a posteriori“, in dem diese Ausdrücke auch auf Propositionen zutreffen, deren Wahrheitswerte wir nicht erkannt haben, und zwar unabhängig davon, ob wir sie erkennen können. Die Goldbachsche Vermutung ist bspw. eine mathematische Proposition, deren Wahrheitswert uns unbekannt ist. Die Wahrheitswerte mathematischer Propositionen, deren Wahrheitswerte wir erkannt haben, konnten wir unabhängig von der Erfahrung erkennen. Es scheint daher: Wir können den Wahrheitswert der Goldbachschen Vermutung unabhängig von Erfahrung erkennen, wenn wir ihn überhaupt erkennen können. Also ist es vernünftig, nach einer erfahrungsunabhängigen Rechtfertigung (sprich: einem Beweis) dieser Proposition zu suchen. Selbst wenn die Goldbachsche Vermutung eine Proposition sein sollte, deren Wahrheitswert für uns unerkennbar ist, kann sie in diesem Sinne dennoch als a priori klassifiziert werden. Ich werde dieses Verständnis im Folgenden unberücksichtigt lassen.

8 Nicht-Apriorität impliziert nur zusammen mit Erkennbarkeit Aposteriorität.

9 Dies gilt nicht für das weite Verständnis von „a priori“ und „a posteriori“ (siehe Fußnote 7). Laut des weiten Verständnisses folgt aus der Nicht-Apriorität einer Proposition ihre Aposteriorität.

deren Wahrheitswerte prinzipiell nicht erkannt werden können. Von einem antirealistischen Standpunkt aus betrachtet ist die These, Notwendigkeit impliziere nicht Apriorität, äquivalent mit der These, es gäbe aposteriorische Notwendigkeiten.

Obgleich die Einnahme eines realistischen Standpunktes nicht ohne Weiteres die Existenzthese begründet, wird dadurch die Erklärung von Notwendigkeit vermittels Analytizität dennoch widerlegt. Betrachten wir noch einmal die Widerlegung: Nehmen wir zum Zweck der Reduktion an, Notwendigkeit impliziere Analytizität. Weil Analytizität Apriorität impliziert, würde Notwendigkeit Apriorität implizieren. Von einem realistischen Standpunkt aus betrachtet, impliziert Notwendigkeit jedoch nicht Apriorität. Also impliziert Notwendigkeit nicht Analytizität. In dieser Widerlegung spielt die Existenz aposteriorischer Notwendigkeiten keine Rolle. Die Widerlegung kann mithin von einem realistischen Standpunkt aus auch dann begründet vollzogen werden, wenn der Nachweis aposteriorischer Notwendigkeit noch aussteht oder sogar scheitert.

In der folgenden Diskussion der Existenzthese werde ich das Problem der alethisch unerkennbaren Propositionen ausklammern und mich auf die alethisch erkennbaren Propositionen beschränken.

5. Das Standardargument

Nachdem die Gehalte der umstrittenen Implikations- (Impl.) und Existenzthese (Exist.) präzisiert sind, kann nun das Standardargument für die Existenz aposteriorischer Notwendigkeiten bewertet werden. Es wird sich zeigen, das Standardargument begründet nicht schlüssig die Existenz aposteriorischer Notwendigkeiten. Dieses Argument unterbindet die Erklärung von Notwendigkeit vermittels Analytizität mithin nicht prinzipiell.
Das Standardargument lautet wie folgt:

(P1) Die Proposition
(1) Hesperus ist Phosphorus
ist notwendig.

(P2) Wir haben den Wahrheitswert der Proposition (1) tatsächlich vermittels Erfahrung erkannt.[10]

(K) Also gibt es zumindest eine notwendige Proposition a posteriori.

So dargestellt ist das Standardargument jedoch ungültig, falls es hybride Propositionen gibt. Denn es ist möglich, den Wahrheitswert einer hybriden Propositi-

10 Wohlgemerkt, es geht nur um die Erkenntnis der Wahrheitswerte *von* notwendigen Propositionen, d.h. um die Erkenntnis, dass *p*, falls die Proposition *p* wahr ist, bzw. dass nicht-*p*, falls die Proposition *p* falsch ist. Es geht nicht um die Erkenntnis, *dass* es eine notwendige Proposition ist bzw. dass der Proposition ihr Wahrheitswert notwendig zukommt.

on tatsächlich vermittels Erfahrung zu erkennen. Hybride Propositionen sind den Definitionen zufolge jedoch a priori. Dass der Wahrheitswert einer Proposition tatsächlich vermittels Erfahrung erkannt wurde, ist also nicht hinreichend für die Aposteriorität der Proposition, sofern es hybride Propositionen gibt.

Die Gültigkeit des Standardargumentes kann durch eine weitere Prämisse gesichert werden:

(P3) Wir können den Wahrheitswert der Proposition (1) *nur* vermittels Erfahrung erkennen.

Wenn die Prämisse (P3) wahr ist, dann ist (1) keine hybride Proposition. Folglich gibt es zumindest eine aposteriorische Notwendigkeit.

Die Prämisse (P3) sichert zwar die Gültigkeit des Standardargumentes, doch lässt sich nun die Schlüssigkeit des Arguments bestreiten. Die Prämisse (P3), so wird nämlich behauptet, sei falsch. Es wird eingewendet, wir wären im Stande den Wahrheitswert der Proposition (1) unabhängig von Erfahrung zu erkennen. Denn die Proposition (1) sei doch identisch mit der Proposition

(2) Hesperus ist Hesperus.

Und dass Hesperus Hesperus ist, vermögen wir unabhängig von Erfahrung zu erkennen.

Beim Versuch, diesen Einwand zu entkräften, sehen sich die Vertreter des Standardarguments einem Dilemma gegenüber: Einerseits können sie nicht die Identität der Proposition (1) mit der Proposition (2) akzeptieren; die Aposteriorität der Erkenntnis wäre verloren. Aber was kann diese Propositionen unterscheiden? Wenn die singulären Terme „Hesperus“ und „Phosphorus“ direkt referieren, dann besteht ihre semantischer Beitrag zur Proposition einzig im bezeichneten Objekt. Und wenn die Proposition (1) wahr ist, dann sind die bezeichneten Objekte identisch und die Propositionen ununterscheidbar.[11] Andererseits könnte der semantische Beitrag der singulären Terme über die bezeichneten Objekte hinausgehen, indem er bspw. auch Arten des Gegebenseins enthält. Doch droht nun die Notwendigkeit der Proposition (1) verloren zu gehen:

> One could, of course, avoid this conclusion by adopting the assumption (foreign to Kripke) that – in addition to predicating identity of Venus and itself – the proposition [(1)] also predicates the properties of being visible in the evening and being visible in the morning of Venus. But then, the proposition will be *contingent*. (Soames (2006), S. 294)

Das Dilemma besagt also: Entweder können wir den Wahrheitswert der Proposition (1) auch unabhängig von Erfahrung erkennen, dann wäre (1) nicht a posteriori, oder wir können den Wahrheitswert der Proposition (1) *nur* vermittels Er-

11 Zum selben Ergebnis kommt man auch, wenn man Propositionen mit Mengen möglicher Welten identifiziert. Vgl. Williamson (2007), S. 67.

fahrung erkennen, dann wäre (1) aber kontingent. Der Nachweis aposteriorischer Notwendigkeiten misslingt in beiden Fällen.

Die Vertreter des Standardarguments sind jedoch nicht geschlagen. Sie könnten zu Recht einwenden, die Beweislast liege bei ihren Gegnern. Die schwer zu erbringende Rechtfertigung der Prämisse (P3) sei nämlich nur dann nötig, wenn die Existenz hybrider Propositionen angenommen wird. Gibt es keine hybriden Propositionen, dann ist der Umstand, dass der Wahrheitswert einer Proposition tatsächlich vermittels Erfahrung erkannt wurde, hinreichend für die Aposteriorität der Proposition. Die Gegner des Standardarguments müssen daher zunächst Gründe für die Existenz hybrider Propositionen vorlegen, bevor die Rechtfertigung der Prämisse (P3) überhaupt relevant wird.

Als erste Replik kann der Gegner des Standardarguments darauf hinweisen, dass die Existenz hybrider Propositionen (zumindest implizit) weitestgehend angenommen wird. Erstens ist die Existenz hybrider Propositionen in die Definitionen apriorischer und aposteriorischer Propositionen eingeflossen. Boghossian und Peacocke definieren die Aposteriorität einer Proposition als ihre Erkennbarkeit *nur* vermittels Erfahrung.[12] Die Anwesenheit des Ausdrucks „nur“ in der Definition wäre ohne die Existenz hybrider Propositionen überflüssig.

Zweitens wird die Prämisse (P3) auch von den Freunden aposteriorischer Notwendigkeiten als relevant und ihre Begründung als nötig erachtet.[13] Dieser Umstand deutet auf ein (zumindest implizites) Annerkennen hybrider Propositionen hin.

Als eine zweite Gegenmaßnahme kann der Gegner des Standardarguments die folgende Überlegung ins Feld führen: Einerseits können wir den Wahrheitswert der Proposition, dass 2 + 2 = 4, durch Ableitung aus bestimmten Axiomen und Definitionen erkennen, d.h. unabhängig von Erfahrung. Andererseits können wir den Wahrheitswert dieser Proposition durch Zählexperimente, d.h. vermittels Erfahrung erkennen. So schreiben wir Mathematikern der Antike durchaus berechtigt das Wissen zu, dass 2 + 2 = 4, obgleich kein Mathematiker der Antike über eine Ableitung aus den Peano- und Identitätsaxiomen verfügte. Unbestritten führen beide Erkenntnisweisen zu verschiedenen Graden an Glaubwürdigkeit. Dennoch scheint es sowohl erfahrungsunabhängige als auch erfahrungsabhängige Weisen zu geben, das Wissen zu erlangen, dass 2 + 2 = 4. Diese Proposition ist mithin hybrid.

Nicht nur wird die Existenz hybrider Propositionen also von beiden Seiten (zumindest implizit) anerkannt, die Existenz hybrider Propositionen ist überdies recht plausibel. Die schwer zu erbringende Rechtfertigung der Prämisse (P3) ist daher wesentlich für den Erfolg des Standardarguments und steht aus. Die Erklä-

12 Vgl. u.a. Boghossian, Peacocke (2000), S. 2.

13 Vgl. u.a. Kitcher (1980), S. 98; Soames (2006), S. 294.

rung von Notwendigkeit vermittels Analytizität ist durch das Standardargument also nicht widerlegt.

6. Kitchers Argument

Die Erklärung von Notwendigkeit vermittels Analytizität wird jedoch durch ein weiteres Argument bedroht. Dieses Argument von Philip Kitcher[14] beruht auf der Verstarrung von Sätzen vermittels des Aktualitätsoperators und scheint unabhängig von Beispielen die Implikationsthese (Impl.) direkt zu widerlegen.[15]

Kitchers erste Prämisse besagt:

(P4) Wenn der *F G* ist, dann ist es notwendig wahr, dass der aktuale *F* aktual *G* ist.

Betrachten wir Sätze der Form „Der *F* ist *G*“ und ihre erstarrten Varianten „Der aktuale *F* ist aktual *G*“. „Der aktuale *F*“ referiert, so wie wir den Ausdruck verwenden, in jeder Welt auf den Referenten des Ausdrucks „der *F*“ in der aktualen Welt, sagen wir: *b*. Die Extension des Ausdrucks „ist aktual *G*“, so wie wir den Ausdruck verwenden, ist in jeder Welt dieselbe wie die Extension des Ausdrucks „ist *G*“ in der aktualen Welt, sagen wir: die Menge *β*. Sätze der Form „Der aktuale *F* ist aktual *G*“ sind in einer Welt w_i nur dann wahr, wenn *b* ein Element von *β* ist. Wenn „Der *F* ist *G*“ in der aktualen Welt wahr ist, dann ist *b* ein Element von *β*. Also: wenn „Der *F* ist *G*“ wahr ist, dann ist „Der aktuale *F* ist aktual *G*“ notwendig wahr.

Kitcher nimmt an:

(P5) Der *F* ist *G*.

Aus (P4) und (P5) folgt:

(K1) Es ist notwendig wahr, dass der aktuale *F* aktual *G* ist.

Als Reduktionsannahme fügt Kitcher eine Instanz der Implikationsthese (Impl.) hinzu:

(Impl.) Wenn es notwendig wahr ist, dass *p*, dann können wir unabhängig von Erfahrung wissen, dass *p*.

Aus (K1) und (Impl.) folgt:

(K2) Wir können unabhängig von der Erfahrung wissen, dass der aktuale *F* aktual *G* ist.

14 Vgl. Kitcher (1980), S. 99.
15 Siehe jedoch Fußnote 18.

Soweit zunächst Kitchers Argument. Inwiefern bedroht seine Konklusion (K2) die Implikationsthese? Es gibt sicherlich Instanzen von (K2), die wahr sind, bspw.:

(K2*) Wir können unabhängig von Erfahrung wissen, dass die aktuale kleinste Primzahl aktual gerade ist.[16]

Wir verfügen über diese apriorische Erkenntnis, weil wir über die beiden folgenden apriorischen Erkenntnisse verfügen:

(P4) Wenn der *F G* ist, dann ist es notwendig wahr, dass der aktuale *F* aktual *G* ist

und

(P5*) Die kleinste Primzahl ist gerade.

Wir verfügen also über die apriorische Erkenntnis, dass die aktuale kleinste Primzahl aktual gerade ist, weil wir dies aus den beiden apriorischen Erkenntnissen (P4) und (P5*) schließen können.

Wenn (Impl.) wahr ist, dann muss (K2) jedoch für alle Instanzen wahr sein, für die gilt, dass der *F G* ist, nicht nur für diejenigen Instanzen, von denen wir unabhängig von Erfahrung wissen, dass der *F G* ist. Bspw. muss auch gelten:

(K2**) Wir können unabhängig von Erfahrung wissen, dass der aktuale 44. Präsident der USA aktual Demokrat[17] ist.

Wäre (Impl.) wahr, dann müssten wir den Wahrheitswert der Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, unabhängig von Erfahrung erkennen können. Dies mutet absurd an, weil es, analog zur vorangehenden Überlegung, zu erfordern scheint, dass wir unabhängig von Erfahrung das Folgende wissen können:

(P5**) Der 44. Präsident der USA ist Demokrat.

Weil nicht ersichtlich ist, wie wir unabhängig von Erfahrung (P5**) wissen können, haben wir laut Kitcher allen Grund (Impl.) zu verwerfen:

(K3) (Impl.) ist falsch.

16 Dass die kleinste Primzahl gerade ist, ist bereits notwendig. Die Verstarrung ist in diesem Fall also redundant.

17 „Demokrat zu sein“ wird als Abkürzung für „Mitglied der Demokratischen Partei der USA zu sein“ verwendet.

7. Ein analoges Dilemma?

Kitchers Argument scheint robuster als das Standardargument. Jedoch setzt Kitchers Argument voraus, dass es *nur* einen Weg gibt, wie (K2**) wahr sein kann, nämlich *nur* dann, wenn wir unabhängig von Erfahrung (P5**) wissen können.

Wie im Falle des Standardargumentes könnte jedoch eingewendet werden, dass es eine Proposition gibt, die erstens identisch ist mit der Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, deren Wahrheitswert zweitens aber unabhängig von Erfahrung erkannt werden kann. Wenn es eine solche Proposition gibt, dann gibt es einen Weg, wie (K2**) wahr sein kann, ohne dass wir unabhängig von der Erfahrung (P5**) müssten wissen können. Es ist mithin kein absurdes Ergebnis, so der Einwand, wenn mithilfe von (Impl.) abgeleitet zu werden vermag, wir könnten unabhängig von der Erfahrung wissen, dass der aktuale *F* aktual *G* ist. Kitchers Argument liefere, so der Einwand, daher keinen Grund, (Impl.) zu verwerfen.[18]

Die folgende Überlegung könnte diesen Einwand stützen: Betrachten wir den Satz „Der aktuale 44. Präsident der USA ist aktual Demokrat". Wenn der semantische Beitrag zur Proposition des rigiden Designators „der aktuale 44. Präsident der USA" einzig im bezeichneten Objekt, Barack Obama, besteht und wenn der semantische Beitrag zur Proposition des Ausdrucks „ist aktual Demokrat" einzig in der Menge aller Demokraten der aktualen Welt, also {*x* | *x* ist Demokrat in der aktualen Welt}, besteht, dann scheint Kitcher recht zu haben. Denn wie sollten wir unabhängig von Erfahrung wissen können, dass Barack Obama ein Element von {*x* | *x* ist Demokrat in der aktualen Welt} ist?

Es besteht jedoch kein Grund, die Extension von „ist *aktual* Demokrat" durch „{*x* | *x* ist Demokrat in der aktualen Welt}" anzugeben. Denn diese Extension ist in der aktualen Welt eindeutig bestimmt, endlich und *variiert nicht* zwischen möglichen Welten. Die Mengenkennzeichnung „{*x* | *x* ist Demokrat in der aktualen Welt}" muss daher auch nicht von Welt zu Welt neu ausgewertet werden. Wir müssen die betreffende Menge nicht durch eine Eigenschaft festlegen, sondern können schlicht ihre Elemente aufzählen. Daher kann die Extension von „ist aktual Demokrat" auch durch „{Tim Kaine, Barack Obama, Harry Reid}" angeben werden.[19]

Wenn der semantische Beitrag zur Proposition des Ausdrucks „ist aktual Demokrat" also einzig in der Menge {Tim Kaine, Barack Obama, Harry Reid}

18 Hier wird ersichtlich, dass Kitchers Argument letztlich doch nicht unabhängig von Beispielen aposteriorischer Notwendigkeiten (Impl.) angreift, sondern lediglich eine andere *Art* (vermeintlicher) aposteriorischer Notwendigkeiten ins Feld führt, nämlich dass der aktuale *F* aktual *G* ist. Dies ist aber, wie sich noch zeigen wird, unter Umständen ein Vorteil von Kitchers Argument.

19 Der Einfachheit halber tue ich hier so, als hätte die Demokratische Partei der USA lediglich drei Mitglieder.

besteht, dann resultiert daraus die Proposition, dass Barack Obama ein Element von {Tim Kaine, Barack Obama, Harry Reid} ist.

Die Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, ist, wenn sie wahr ist, diesem Einwand zufolge also identisch mit der Proposition, dass Barack Obama ein Element von {Tim Kaine, Barack Obama, Harry Reid} ist. Und dass Barack Obama ein Element von {Tim Kaine, Barack Obama, Harry Reid} ist, können wir offensichtlich unabhängig von Erfahrung wissen.

Laut diesem Einwand schien es nur so, als könnten wir nicht unabhängig von Erfahrung wissen, dass der aktuale 44. Präsident der USA aktual Demokrat ist, weil angenommen wurde, diese apriorische Erkenntnis müsse das Ergebnis eines Schlusses aus apriorischen Erkenntnissen sein, weil also vorausgesetzt wurde, dies sei der einzige Erkenntnisweg. Dem Einwand zufolge ist diese Erkenntnis jedoch nicht das Ergebnis eines Schlusses, sondern schlicht die Erkenntnis, dass Barack Obama ein Element von {Tim Kaine, Barack Obama, Harry Reid} ist.[20]

Soweit die schlechte Nachricht für Kitcher. Nun zur (vermeintlich) guten. Die Vertreter des Standardargumentes wurden mit einem Dilemma konfrontiert: Entweder können wir den Wahrheitswert der Proposition (1) auch unabhängig von Erfahrung erkennen, dann wäre (1) nicht a posteriori, oder wir können den Wahrheitswert der Proposition (1) *nur* vermittels Erfahrung erkennen, dann wäre (1) aber kontingent. Dem Nachweis aposteriorischer Notwendigkeiten war in beiden Fällen nicht gedient.

Sind die Vertreter von Kitchers Argument mit einem vergleichbaren Dilemma konfrontiert? Das erste Horn des ursprünglichen Dilemmas lässt sich, wie gerade gezeigt worden ist, durchaus übertragen. Wie steht es um das zweite Horn des Dilemmas? Wenn man die Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, von der Proposition, dass Barack Obama ein Element von {Tim Kaine, Barack Obama, Harry Reid} ist, unterscheidet, dann müsste sich die Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, für ein analoges Dilemma als kontingent erweisen.

Es ist nicht ganz einsichtig, ob sich das zweite Horn des Dilemmas übertragen lässt. Um über den modalen Status von „Der *F* ist *G*“ zu entscheiden, muss man in jeder möglichen Welt w_i überprüfen, ob der Referent von „der *F*“ in w_i ein Element von $\{x \mid Gx\}$ in w_i ist. Um über den modalen Status von „Der aktuale *F* ist aktual *G*“ zu entscheiden, muss man *nur* in der aktualen Welt überprüfen, ob *b* ein Element von $\{x \mid Gx\}$ ist. Der Aktualitätsoperator verstarrt die

20 Die apriorische Erkenntnis, dass die aktuale kleinste Primzahl aktual gerade ist, scheint deshalb das Ergebnis eines Schlusses zu sein, weil die Extension von „ist aktual gerade“ nicht durch „{2, 4, 6, ...}“ angegeben werden kann. Denn es gibt unendlich viele gerade Zahlen. Daher muss die Extension von „ist aktual gerade“ durch eine Eigenschaft festgelegt werden: $\{x \mid x$ ist in der aktualen Welt gerade$\}$.

Kennzeichnung und das Prädikat. Weil einer Menge ihre Elemente notwendig zukommen, gilt: Wenn Barack Obama Demokrat ist, dann ist die Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, notwendig wahr. Wenn er kein Demokrat ist, dann ist sie notwendig falsch. Dass der aktuale 44. Präsident der USA aktual Demokrat ist, scheint mithin nicht kontingent sein zu können!

Der Vertreter von Kitchers Argument scheint daher in einer ungleich besseren Situation als der Vertreter des Standardarguments. Er muss lediglich eine Konzeption von Propositionen vertreten, welche die Identität der Proposition, dass der aktuale 44. Präsident der USA aktual Demokrat ist, mit der Proposition, dass Barack Obama ein Element von {Tim Kaine, Barack Obama, Harry Reid} ist, ausschließt. Mit einer solchen Konzeption von Propositionen scheint er sich nicht dem Vorwurf auszusetzen, dass diese Erkenntnis zwar *nur* vermittels Erfahrung erlangt werden könne, jedoch eine *kontingente* Proposition betreffe.

Auf der anderen Seite muss es dem Vertreter von Kitchers Argument Unbehagen bereiten, dass wir es hier mit Propositionen zu tun haben, deren modaler Status nicht von der Menge aller möglichen Welten abhängt, sondern lediglich von der aktualen Welt. Diese Propositionen scheinen in einem gewissen Sinne also doch kontingent zu sein. Nun hängt der modale Status der Propositionen, dass Barack Obama ein Mensch ist in einem gewissen Sinne zwar auch nicht von der Menge aller möglichen Welten ab, sondern nur von der Menge der möglichen Welten, in denen Barack Obama existiert. Doch selbst wenn er nur in der aktualen Welt existieren würde, müsste man, um den modalen Status zu eruieren, dennoch in jeder möglichen Welt w_i überprüfen, ob Barack Obama in w_i ein Element von $\{x \mid x$ ist ein Mensch$\}$ in w_i ist. Im Lichte des eigenartigen modalen Status von Propositionen der Art, dass der aktuale F ist aktual G, ist nicht ganz einsichtig, ob sich das Dilemma nicht doch auf Kitchers Argument übertragen lässt.[21]

Literaturverzeichnis

Blackburn, S.: „Morals and Modals“. In: *Macdonald, G./Wright, C. (eds.):* Fact, Science and Morality. Blackwell, Oxford, 1986. S. 119-141

Boghossian, P./Peacocke, C.: „Introduction“. In: *Boghossian, P./Peacocke, C. (eds.):* New Essays on the A Priori. Clarendon Press, Oxford, 2000. S. 1-10

Casullo, A.: A Priori Justification. Oxford University Press, Oxford, 2003

21 Ich danke den Teilnehmern des *Philosophischen Kolloquiums* (Erfurt), des *Seminaro del Venerdi* (Parma) und der *GAP 7* (Bremen) für hilfreiche Hinweise zu früheren Fassungen. Besonderer Dank gilt Alex Burri, Marco Santambrogio, Andrea Bianchi und Wolfgang Huemer.

Hale, B.: „The Source of Necessity“. Philosophical Perspectives, 16, 2002. S. 299-319

Kitcher, P.: „Apriority and Necessity“. Australasian Journal of Philosophy 58, 1980. S. 89-101

Kripke, S.: Naming and Necessity. Blackwell, Oxford, 1980

Soames, S.: „The Philosophical Significance of Kripkean Necessary Aposteriori“. Philosophical Issues 16, 2006. S. 288-309

Williamson, T.: The Philosophy of Philosophy. Blackwell, Oxford, 2007

The Theory of Appearing and the Problem of Hallucination

Alexander Staudacher
Alexander.Staudacher@ovgu.de
Otto-von-Guericke-Universität, Magdeburg

Abstract/Zusammenfassung

The majority of philosophers of perception actually prefer a version of direct realism to a kind of sense datum theory. One way to be a direct realist is to opt for the Theory of Appearing. Adherents of this theory claim that it is superior to the Adverbial Theory or to Intentionalism when it comes to do justice to the core intuition behind direct realism, namely the claim that perception puts us in direct or immediate contact with physical objects (cf. Alston 1999, 2005). The proponents furthermore claim that this theory gives the most adequate account of perceptual experience as a kind of mental state endowed with a phenomenal character (Alston 1999, Langsam 1997). According to this theory perception includes a basic relation of appearing between the subject and the perceived object which is not accessible to any further analysis. This creates a problem of how to deal with hallucinations where no such object is present. In order to solve this problem adherents of the Theory of Appearing either have tried to establish the claim that even in the case of hallucination there is a kind of object which appears to us (Alston 1999) or they have defended a particular version of the disjunctivist conception of experience according to which hallucinations are mental states of a completely different kind in that they relate us only to portions of empty space (Langsam 1997). It will be argued that both options are not viable: The first option leads to highly implausible results and the second one has insurmountable problems when it comes to spell out the relation of appearing in more detail. Finally, a further option is considered as a way out. But this option is shown to be incompatible with the way the basic and unanalysable relation of appearing has to be conceived of if the Theory of Appearing is true.

In der Philosophen der Wahrnehmung genießen gegenwärtig Versionen des direkten Realismus gegenüber Sinnesdatentheorien den Vorzug. Die Theorie des Erscheinens stellt eine Form des direkten Realismus dar. Die Anhänger dieser Theorie behaupten, dass sie der Adverbialtheorie oder dem Intentionalismus überlegen ist, wenn es darum geht, der Kernintuition hinter dem direkten Realismus Gerechtigkeit widerfahren zu lassen, nämlich der These, dass uns Wahrnehmung einen direkten oder unmittelbaren Kontakt mit physischen Objekten ermöglicht (vgl. Alston 1999, 2005). Darüber hinaus wird behauptet, dass diese Theorie erlaubt, den phänomenalen Charakter von Wahrnehmungserfahrungen angemessen zu berücksichtigen (Alston 1999, Langsam 1997). Gemäß der Theorie des Erscheinens beinhaltet Wahrnehmung eine grundlegende Beziehung des Erscheinens zwischen dem Subjekt und dem wahrgenommenen Objekt, die keiner weiteren Analyse mehr zugänglich sein soll. Es stellt sich in diesem Zusammenhang die Frage, wie dann Halluzinationen zu verstehen sein sollen, wo kein solches Objekt vorhanden ist. Zur Lösung dieses Problems haben die Anhänger der Theorie des Erscheinens entweder behauptet, dass es auch im Fall der Halluzination eine Art

von Objekt gibt, welches uns erscheint (Alston 1999), oder sie haben eine bestimmte Version von Disjunktivismus verteidigt, wonach Halluzinationen mentale Zustände ganz anderer Art sind als veridische Wahrnehmungen, indem hier die Beziehung des Erscheinens zu Teilen des leeren Raums bestehen soll (Langsam 1997). Es wird argumentiert, dass beide Optionen nicht gangbar sind: Die erste Option führt zu höchst unplausiblen Ergebnissen und die zweite Option sieht sich unüberwindlichen Problemen gegenüber, wenn es darum geht, den Charakter der Beziehung des Erscheinens genauer zu präzisieren. Abschließend wird noch eine weitere Option als Ausweg erwogen. Doch hier zeigt sich, dass diese Option nicht mit dem Verständnis verträglich ist, auf welches die Theorie des Erscheinens bezüglich der Relation des Erscheinens als einer unanalysierbaren Relation festgelegt ist.

The Theory of Appearing

The Theory of Appearing (ToA) is a theory of perceptual consciousness. Theories of perceptual consciousness typically have two aims: *First of all*, they want to give an account of how perceptual consciousness is best to be understood. Usually, this includes an answer to the question whether perceptual consciousness differs in any notable respects from other kinds of consciousness. Given the answer to this question is positive, the obvious task then is to spell out these differences in further detail. *Secondly*, they try to find a solution for the so-called "Problem of Perception"[1], that is, they try to answer the question what kind of objects we "immediately" or "directly"[2] perceive. Concerning the first point the ToA holds that perceptual consciousness consists (perhaps among other things) in a genuine relation, the relation of appearing, between the mind of the perceiving subject and the perceived object, which is taken to be basic in the sense that it doesn't allow any further analysis; we can't do better than to admit that there is such a relation in play when we are perceptually conscious. Concerning the second point the ToA holds that the immediate or direct objects of perception are the physical objects of our surroundings. To enjoy perceptual consciousness for the ToA therefore implies that the mind of the perceiving subject is related to the perceived physical object by the basic relation of appearing.

In the first half of the 20th century when sense-data-theories dominated the scene the ToA was considered to be one of the most important alternatives to the

1 Cf. Crane 2006, Quinton 1956/65, and A.D.Smith 2002.

2 There has been considerable dispute in the literature on the question how we have to understand the terms "immediate" and "direct" here (cf. Jackson 1977, Armstrong, Martin 2005). Some have even disputed that there is a clear cut or at least context independent answer to this question (cf. Austin 1961, Martin 2005, Clarke 1965). For our purposes we need only a very superficial notion of direct or immediate perception: An object of is perceived directly iff it isn't perceived in virtue of some other object. If you watch Obama in the TV you don't perceive him directly, because you perceive him in virtue of something else, namely the ordered dots on the TV-screen. (For further discussion on direct perception see also see also Staudacher 2007)

sense-datum-thesis, according to which we never are immediately aware of physical objects in perception.[3] More recently, the ToA has been defended by W. Alston and H. Langsam.[4] In the following, the discussion will concentrate on these two versions of the ToA.

The ToA wants to do justice to *two* well-entrenched intuitions: (i) *the phenomenal character* of perceptual states: perceiving is different from thinking and believing. To see a red patch is different from thinking of a red patch or believing that a red patch is present. That is, perception has a kind of phenomenal character which is lacking in the case of thinking and believing. Colors, extension etc. appear in perception phenomenally in a way which is absent in the case of thinking and believing. (ii) The *"direct-presence-intuition"*: in perception physical objects of our environment (or at least parts of them as their surfaces) are somehow "directly present" to the perceiver. The awareness of the physical object is not mediated by the awareness of another object; e.g. a sense datum, a sensation or whatever.

As has already been noted, the ToA takes the fact that an object appears in a certain way to a subject as a *basic* relation between the mind of the perceiving subject and the perceived object which *cannot be analyzed any further*. For example, if a tomato looks red to *s* then there is a relation of appearing red between the mind of *s* and the tomato. Furthermore, it is this relation of appearing red which is responsible for the phenomenal character of the experience in question. This claim surely deserves further scrutiny and discussion that, however, cannot be delivered here due to limitations of space.

Note that the relation of appearing is a "real" or "true" relation, that is, a relation which can only be instantiated when the relata of the relation exist. That *o* appears *R* to *s* is, therefore, not to be confounded with the fact that *s* represents *o* to be *R* or takes *o* to be *R*. The latter is possible even when *o* is not *R* or when *o* does not exist. On the other hand, the ToA does not have to deny that a perceptual state has representational or intentional content; it will deny only that perceptual *consciousness* can be analyzed with the help of the notion of representational content. This leaves open the possibility that perceptional states are complex states involving an appearing-relation in the sense of the ToA *and* representational content.

3 Cf. Prichard 1909, ch.4 and in a sense Hicks 1913/14 and 1938, ch.1-3. See also the relatively respectful critical discussion of the ToA by sense-datum-defenders as Moore 1918-19/65 and Broad 1925, 187ff.

4 Cf. Alston 1999, Langsam 1997; with some justification Alston notes that Dretske's view at least in his early work on perception (Dretske 1969) shows considerable similarity to the ToA (Alston 1999, Fn.3)

The ToA and other accounts of perceptual consciousness

As has been indicated above, the ToA is not the only account of perceptual consciousness. Other well known accounts are the sense-datum theories, the Adverbial Theory and Representationalism or Intentionalism. Why should one prefer the ToA to these other accounts of perceptual consciousness? According to defenders of the ToA like Alston only the ToA can adequately deal with the *direct-presence-intuition"*:

> What does it take to see a certain physical object, over and above being in a certain state of sensory consciousness?
> When we think of the problem in this last form, a striking difference between TA [that is ToA] and its two rivals [the adverbial theory and representationalism] comes to light. For those cases in which it is an external physical object that is appearing in a certain way, TA already specifies a perceptual relation to a physical object in its account of sensory consciousness itself. So in those cases TA's answer to the question: "What has to be added to sensory consciousness to get a perception of an external object?", is "Nothing".5

And after Alston has pointed out that the rivals[6] of the ToA (including also the sense-datum theory of an indirect-realist variant) will specify the perceptual relation between the subject and the perceived object mainly in causal terms, he continues with the following question:

> Would having an experience […] causally related in that way to x *constitute* seeing x? NO. Now matter how x *causally* contributes to the production of an experience, I do not *see* […] x in having that experience unless x *presents itself to my experience* as an object. […] no causal relation of x to the experience could make it true that I *see* x or, indeed that I am *aware* of x in any way at all. Causality is no substitute for awareness.[7]

According to the ToA we are directly related to the appearing physical object by the relation of appearing. The physical object is present to us in the mode of perceptual consciousness because *it* appears to us in *that specific way*.

Other accounts of perceptual consciousness don't deny that in perception a physical object appears to us. But they don't consider the fact that it appears to us as a basic relation which can't be further analyzed. To the contrary, they analyze this relation in different ways. The general structure of these proposals, however, is always the same: The subject enjoys a certain mental state with a phenomenal character that can be present in the case of veridical perception as well as that of hallucinations, where no object is present. The relation to the physical object is due (i) to the right kind of causal relationship between this ob-

5 Cf. Alston 1999,192f.

6 Alston doesn't mention Intentionalism in this context, but it is pretty clear that his argument applies as well to the currently popular versions of this account (cf. e.g. Tye 1995, Dretske 1995). For a criticism of these accounts see Alston 2005.

7 Cf. Alston 1999, 193f.

ject and the perceiver and/or (ii) a belief or perhaps a non-conceptual representation of the object (which in turn is often explained with help of a certain kind of causal dependence).[8]

According to sense-datum theories the direct-presence intuition can't be true with respect to physical objects, it is true only for sense data, typically thought of as mind-dependent objects. Adverbialists and representationalists, on the other hand, say that we experience physical objects directly. When it comes to explain this further they would point to the causal relation between perceived object and the subject and/or the belief or the non-conceptual representation of the object. A consequence of claims of this kind is that it is not perceptual consciousness *as such* (that is, the kind of conscious state the subject is enjoying while perceiving or even hallucinating) that makes us aware of physical objects but only the causal and/or representational relation between object and perceiver. In contrast, the ToA identifies perceptual consciousness with the basic relation of appearing and concludes that it can do justice to the direct presence intuition in a way the other theories can't do.[9] Now, even if we assume for the sake of argument that this is true, there is an obvious difficulty for the ToA: How to account for hallucinations, where no relation of appearing in the sense of the ToA can take place because one of the relata, the object, is lacking?

An obvious Problem for the Theory of Appearing: Hallucinations

The notion of hallucination relevant in this context is the notion of a state where no object is present but the subject nevertheless takes herself to be in a state with the same kind of phenomenal character as a veridical perception. How can the ToA accommodate this kind of case? There are the following various possibilities:

Option 1: The subject enjoys the same kind of phenomenal state in the case of veridical perception and the case of hallucination. We have only to *revise* our notion of hallucination; even in cases of hallucination there is an object we are aware of.

Option 2: The fact that the subject credits the hallucination with the same kind of phenomenal character does not imply that hallucinations are the same kind of mental state as veridical perceptions; therefore, the ToA-analysis for veridical perceptions might be correct even if it is not applicable to hallucinations.

There are at least *two* ways to flesh out this latter thought further: (i) In the case of a hallucination the subject only *erroneously believes* to enjoy a mental state

8 See e.g. Tye 1995 and Dretske 1995. For a severe criticism of accounts along these lines see Alston 2005; For his criticism that perception implies belief see Alston 1999, 184ff.

9 Cf. Alston 1999, 182f.

with a certain phenomenal character;[10] This is a move no defender of ToA actually makes and, therefore, it won't be pursued here any further.[11] (ii) In the case of hallucination the subject enjoys a state with a certain phenomenal character, though a state of a different kind as in the case of veridical perception; if the subject takes them to be alike it has a false belief *in this respect*.[12]

Option 1: revised notions of hallucinations

If we have to revise our notion of hallucination, so that even in the case of hallucination there is an object which appears to the subject, what kind of object can this be? Alston mentions two candidates[13]: (i) the air at the location, where the hallucinated object is hallucinated to be; (ii) a mental image. Both these, however, have unacceptable consequences. The first candidate, the air, arbitrarily excludes hallucinations in the vacuum. The answer invoking the second candidate founders on the so-called Argument from Hallucination (AfH).

The argument from hallucination

The upshot of the AfH can be rendered as follows: If one admits that in the case of hallucination one is aware of a mental image one has to admit also that one is aware of a mental image in the case of veridical perception. The AfH, presented by H. Robinson[14], has *two* premises, the first of which is a fairly plausible speculation based on empirical findings, the second one is an application of a widely held principle, the principle that the *s*ame proximal *c*auses will have the *s*ame *e*ffects (scse-principle):

(1) It is theoretically possible by activating some brain process which is involved in a particular type of perception to cause an hallucination which exactly resembles that perception in its subjective character.

(2) It is necessary to give the same account of both hallucinating and perceptual experience when they have the same neural cause. Thus, it is not, for example, plausible to say that the hallucinatory experience involves a mental image or sense-datum, but that the perception does not, if the two have the same proximate - that is, neural - cause.

(C) These two propositions together entail that perceptual processes in the brain produce some object of awareness which cannot be identified with any feature of the external world - that is, they produce a mental image or a sense-datum.

---

10 A.D. Smith ascribes a view of this kind to Evans and McDowell. See chapter 8 of his book. This ascription might be contested, however. In any case, Evans and McDowell aren't adherents of the ToA.

11 For a criticals discussion of this thought see Staudacher 2007.

12 Cf. Langsam 1997, 38ff.

13 Cf. Alston 1999, 191.

14 Cf. Robinson 1994, 151.

The AfH discredits the proposal that in the case of hallucinations we are aware of mental images because there seems to be no way how one can effectively escape the conclusion that the same takes place in the case of veridical perception.[15]

But The AfH seems *also* to discredit the variant (ii) of Option 2 according to which the subject in the case of hallucination enjoys a state with a certain phenomenal character, though *a state of a different kind* as in the case of a veridical perception. Langsam has argued, however, that this variant can be given a form that is immune to the AfH.

Langsam‘s solution

Put in a nutshell, Langsam‘s argument runs as follows: The weak point of the AfH is the second premise, because the ToA can circumvent the principle behind, the scse-principle. The scse-principle is not universally valid because it holds only for „intrinsic changes“. Where “Intrinsic changes include changes in intrinsic properties of objects, and changes in relations obtaining between objects whose intrinsic properties have changed”[16], those different changes, however, that follow the stimulation of our brain in the veridical and the hallucinatory case, are not intrinsic according to the ToA. Therefore, the AfH can’t do any harm to the claim that we enjoy different kinds of phenomenal states in the case of veridical perception and hallucination.

The scse-principle

Langsam doesn’t want to deny that billiard balls obey the scse-principle in the following way: playing a ball with the white ball in the same way in different situations will always lead to the same change with respect to that ball.

What he takes the scse-principle to exclude here in particular is the case that the ball suffers in one case a certain kind of change of its *intrinsic properties* (from rest to movement) while not in the other. But there are other cases, cases with different effects from the same causes without a violation of the scse-principle. Compare the following two cases:

> *Case 1*: Let a worm being situated at the center of a ten foot long table crawl to the left edge of the table.
> *Case 2*: Put a soda can on the right edge of this table and let the same worm crawl from the same position in the same direction as before, so that it will be finally at a distance of 10 feet from the soda can.[17]

15 For further discussion see Robinson 1990, ch. 6.
16 Cf. Langsam 1997,43, Fn. omitted.
17 Cf. Langsam 1997, 44.

In a sense we have two different effects resulting from the identical crawling of the worm here: in case 2 the worm will be 10 feets away from the soda can in case 1 not, because there is no can.

Why the scse-principle isn‘t violated here?

According to Langsam there will be simply no violation of the scse-principle because the change doesn‘t involve *a change of the intrinsic properties* of the soda can. A problem with Langsam‘s explanation is that there seem to be cases, where the change affects the involved object in Langsam‘s sense, but where there is still no violation of the scse-principle. Consider again two cases:

Case 3: John points his rifle in direction *R* and pulls the trigger with the effect that the bullet flies in direction *R* and falls to the ground after some time.

Case 4: same situation with the exception that at some distance in direction *R* a man is standing.

The effect in case 4 two is obviously different from the one in case 3: the man will be hit and die. But his dying is obviously a change in of one of his intrinsic properties.

A better explanation why we have no violation of the scse-principle here seems to be that the application of the scse-principle obviously presupposes a ceteris paribus clause: we will demand that the effects have to be the same only if the same conditions hold. But the conditions in the two cases are obviously so different that this clause is not fulfilled here. And the same consideration can be also applied to Langsam’s soda-can-worm-cases, so that the recourse to ceteris paribus clauses delivers an explanation which is more encompassing than the one given by Langsam.

Be this as it may, we are now in a position to see how the general line of defense of the ToA towards the AfH will run: if one can show that the ceteris paribus clause isn’t fulfilled with respect to veridical perceptions and hallucinations, the argument will lose its bite. To see how this might work in detail let us once more take a look at the cases of veridical perception and hallucination: in both cases the visual cortex is stimulated in a certain way and in both cases the effect consists in conscious states the subject takes to be of the same kind. In the case of veridical perception the subject *s* becomes related to a physical object *o* which is appearing to it, whereas in the case of hallucination the subject *s* doesn‘t become related to a physical object. Thus, here we have different “effects” resulting from the same proximate causes. On the other hand, there is clearly no violation of the scse-principle because the situation can be seen as analogous to the worm-soda-can-case or the rifle-case. Therefore, the mere fact that the same kind of brain stimulation in one case leads to a relation of appearing of a physical object to *s* but not in the other doesn‘t show that the ToA produces a violation of the scse-principle in the face of hallucinations.

But there remains the question as to what is happening in the case of hallucinations. It is easy to see why the ToA still owes us a a positive account of these cases even if they differ in that in the one case a physical object is present which is lacking in the other one: if in the case of hallucinations we are aware of a mental image, it is difficult to see how this could be otherwise in the case of veridical perceptions even if in these latter cases a physical object is present. The stimulation of the brain process will be the same in both cases and it is not easy to see how one can be conscious of something else than a mental image in the veridical case even if in this case a physical object will be present in one's environment. So the question whether the scse-principle might be violated in the relevant way is still threatening.

Therefore, Langsam offers a positive conception of hallucinations to dispel any worries of this kind.[18] According to him there is no violation of the scse-principle if we take a hallucination to be a relation of appearing between a subject and *a portion of empty space*. The different effects that will result here from the same kind of stimulation of our visual cortex in the veridical and the hallucinatory case won't count as different effects in the sense of the scse-principle.

In the veridical case we have a relation of appearing between the subject and a physical object and in the hallucinatory case a relation of appearing between the subject and a portion of empty space. In both cases certain phenomenally conscious states will be the result; these states are of a different type, however, because they involve different relations of appearing, although the subject won't be able to tell them apart. Again the situation is analogous to the worm-soda-can-case and the rifle-case. Thus, again we have no violation of the scse-principle.

A critical examination of Langsam's account of hallucination

An obvious and important question which comes to the mind immediately is whether we can really make sense of the claim that a portion of empty space is a suitable candidate for a relatum which is entering the relation of appearing. Perhaps one might have some reservations here towards such a claim because empty portions of space are so to speak not "substantial" enough for us to consider them as relata in the first place. But this judgment seems rash: After all, it makes perfect sense to say that a portion of space appears empty or replete.

But there is still another difficulty which makes the claim highly questionable that we are related to portions of empty space in cases of hallucinations: The fact whether *R* is a *relation* between *x* and *y* and not only a *property* of *y* which it does instantiate in the presence of *x* is generally supposed to be dependent on the

18 Cf. Langsam 1997, 47f.

following question: Will *R* be changed if we change the other properties (including other relations) of *x*? For example, the question whether a wall appears red to *s* or not will depend on the fact whether the wall has been painted with the right kind of pigment or not. If we change the pigment, we expect that the wall appears different. If such facts concerning the properties and relations of *x* have no bearing on the fact whether *R* holds between *x* and *y* we would normally conclude that we haven't encountered a relation where *x* is one of the relata.[19] But in the case of empty portions of space there is no such influence; if you stimulate the brain of the subject in the right way it will experience a red tomato irrespective of the fact whether something blue, green or nothing at all has been put there in the first place.

Related to this difficulty we find yet another one. In order to see this more clearly we should ask the following question: if we have two different relations of appearing in the two cases of veridical and hallucinatory experience, what is responsible here for the difference between these two? There seems to be only one answer at hand: The respective relations are to be distinguished by their respective types of objects: in the one case a physical object enters into the relation of appearing, in the other it is an empty portion of space. And furthermore, it seems to be the mere presence of these different "objects" in these situations which is sufficient for the fact that we have two different kinds of relations here.

But this leads to the following unwelcome consequence: It is now difficult to see how one can distinguish between veridical perceptions and "*veridical hallucinations"*, that is, cases where a hallucination of an object *o* at position *p* is brought about by artificial brain stimulation while there is in fact an object of the same kind at position *p*. Intuitively, in the case of veridical hallucinations we wouldn't neither say that the object *o* appears to us nor that we perceive it. This intuition is backed by the observation that changes in *o* won't lead to a different appearance of *o*. After all, the way the world appears to the subject is wholly determined by the artificial stimulation of its brain. The ToA, on the other hand, seems to drive us to the highly counterintuitive conclusion that in cases of veridical hallucinations the physical object in question is appearing to us or to put it differently that there is no difference between veridical perceptions and veridical hallucinations.

Appearing as an external relation - A way out?

But perhaps we can evade the problem that changes of one of the relata have no influence on the character of the relation of appearing in the case of hallucina-

19 Perhaps there are cases where the influence of a change of *x* can be neutralized by a change of *y*. So let's assume that we hold the properties of *y* and the other relations it is entering fixed.

tions if we hold that the relation of appearing is an *external* relation. Consider for example David Armstrong's definition of external relations:

> Two or more particulars are externally related iff there are no properties of the particular which logically necessitate that the relation, or any relation which is part of the relation, holds.[20]

If external relations hold independently from the fact what other properties the relata possess, it seems possible that these properties change while the relation in question will persist. It is controversial which relations if any we should consider as external; spatial relations probably are among the least controversial ones. But the claim that the appearing relation is basic and not further analyzable seems to fit well with the assumption that it is an external relation. Therefore, it might seem tempting to counter our objection with the claim that appearing is an external relation.

It is easy to see, however, that this maneuver is of no help here. The required kind of independence of the relation of appearing from changes in the relata is a kind of independence according to which *any change whatsoever* happening to the portion of space must not change the relation of appearing in any way. And this is surely a much more demanding requirement than the requirement that the instantiation of the relation is not necessitated by the other properties and relations of the relata. The required kind of independence demands furthermore the following: *given*, the relation obtains in the one form or another, its *determinate* form will be independent of the other properties and relations of the relata. That is, given a portion of empty space appears in *some way* to the subject, a change with respect to this portion of space won't lead to *another way* this portion of space appears to the subject. This has to be so, because the way things appear to the subject in the case of hallucinations is fixed by the stimulation of its brain.

Now, it is easy to see that this stronger requirement is not fulfilled even in the case of external relations. Take for example spatial relations which are the least controversial cases for external relations. Suppose, *X* and *Y* stand in a certain distance to each other, and let *W* be an exact duplicate of *X* (with respect to color, extension, movement etc.) and *Z* an exact duplicate of *Y* (with respect to color, extension, movement etc.). If distance is an external relation, *W* and *Z* need not stand at the same distance to each other as *X* and *Y* do. One might even argue that facts about the color, the extension and the movement of *X* and *Y* as such will never necessitate that they are spatially related at all. But, given, *X* and *Y* *are* spatially related then facts of this kind will surely determine *how* they are spatially related: Consider for example the case where *X* is *to the left* of *Y* at time *t1* and *X* begins to grow continually at this time. In such a case *X* will eventually *enclose* Y at *t2*. Thus, the determinate kind of spatial relation between *X* and *Y* is

20 Armstrong 1978, 85.

in a way dependent on their further properties even if spatial relations are external relations.

But if in the case of hallucinations portions of space appear to the subject no change whatsoever concerning these portions of space will lead to a new determinate appearing relation (e.g. a change from appearing blue to appearing red). Therefore, the claim that the appearing relation is external won't be of any help for the ToA here.

A different kind of relation?

The assumption that appearing is a relation of a different kind which is not even subjected to the mentioned requirement for external relations seems to be entirely ad hoc. We could then consider *any* property of an object we like as a relation between this object and another object. Take for example my property of being happy. If relations can obtain between objects in the sense that no other property of one the relata has any influence whatsoever on the fact what specific form this relation will take, there seems to be no obstacle to see this property of mine as a relation which holds between me and the surrounding air. After all, the fact that the air can change in any way without thereby modifying my state of happiness, wouldn't count against such a strange proposal.

So, we have good reason to conclude that it doesn't make sense in the case of hallucinations to assume a relation of appearing between subjects and portions of space. Therefore, we have no reason to accept Langsam's defense of the ToA in the face of the AfH and the scse-principle.

No relation at all in the case of hallucination?

But perhaps the ToA can avoid these problems if it doesn't follow Langsam's proposal according to which we have two different kinds of relations in the case of veridical perception and hallucination. Perhaps we have a relation of appearing *only in the case of veridical perception* (and illusion for that matter). Whereas on the other hand in the case of hallucination we are simply in a certain mental state with a certain phenomenal character the subject is unable to tell from the veridical case.

This new proposal, however, leads to the question of how the relation of appearing is constituted in the case of veridical (or illusory) perceptions. Now part of it must lie in the stimulation of the brain which is the same in both cases. After all, without any stimulation of the brain there will be no conscious state at all. The relation then must be due to the fact that additionally in these cases the required object is present. And a certain phenomenal state becomes a relational state just in those cases in which the object in question is present at the right place. In effect, this comes down to giving up the central claim of the ToA that

the relation of appearing is basic and not accessible to further analysis. That this is so can be also seen, if one pays due attention to the fact that on such an account of the constitution of the relation of appearing the ToA seems to lose much of its greater intuitive appeal which it is supposed have when compared with other accounts of perceptual consciousness: What is then the difference, if there is any at all to adverbial or representational accounts of perceptual consciousness? According to these positions we enjoy a certain form of perceptual consciousness because we enjoy a certain mental state, which could be given even if the perceived object isn't present. Now if the relation of appearing consists in nothing more than the fact that such a state occurs and the further fact that the object in question is present at the right place, the only difference to these accounts seems to consist in the fact that the adherent of the ToA here chooses to speak of the obtaining of a relation of appearing with a certain emphasis. Proponents of the other accounts on the other hand try to specify what is responsible for the fact that the object appears to the subject (that is, among other things why the subject is enjoying the state in question and why this state counts as a perception of the present object). Typically, accounts of this kind will do this by pointing to the right kind of causal relationship between the object and the perceptual state of the subject and/or by trying to specify how the intentional content of this state can be seen as a content that has especially this object as its intentional object. Adherents of the ToA typically think that maneuvers of this kind prevent us from gaining the proper and most natural understanding of what it means that an object is present to us in perception. But it is difficult to see how a version of the ToA could do any better here which takes the relation of appearing as being composed of the phenomenal state of the subject and the mere fact that the object in question is present in the right place.

To give up Langsam's suggestion that we have to assume two different relations in the case of veridical perception and hallucination respectively, seems to lead to an abandonment of crucial elements of the ToA: the claim, that the relation of appearing isn't accessible to further analysis and the claim that it can deal better with our intuition that in perception objects are immediately present to the observer.

Conclusion

The ToA plausibly admits that hallucinations have a certain phenomenal character at least similar to veridical perceptions. But it is unable to give a plausible account of hallucinations in the face of the AfH and the scse-principle.

References

Alston, William: "Back to the Theory of Appearing". Philosophical Perspectives, 13, 1999. S. 181-203

Alston, William: "Perception and Representation". Philosophy and Phenomenological Research, 70, 2005. S. 253-289

Armstrong, David: A Theory of Universals Vol II. Universals and Scientific Realism. Cambridge University Press, Cambridge, 1978

Armstrong, David: "Immediate Perception". In: *Armstrong, David*: The Nature of Mind and other Essays. Cornell University Press, Ithaca/NY, 1976/80

Austin, John L.: Sense and Sensibilia. Oxford University Press, Oxford, 1961

Broad, Charles D.: The Mind and Its Place in Nature, Kegan Paul, London, 1925

Clarke, Thompson: "Seeing Surfaces and Physical Objects". In: *Black, M. (Hrsg.):* Philosophy in America. George Allen, London 1965

Crane, Tim: "The Problem of Perception". In: *Edward N. Zalta (Hrsg.):* The Stanford Encyclopedia of Philosophy (Winter 2006 Edition). http://plato.stanford.edu/archives/win2006/entries/perception-problem. 2006

Dretske, Fred: Seeing and Knowing. Routledge and Kegan Paul, London, 1969

Dretske, Fred: Naturalizing the Mind. MIT Press, Cambridge/MA, 1995

Hicks, Dawes: "Appearance and Real Existence". Proceedings of the Aristotelian Society N.S. 14, 1913/14. S. 1-48

Hicks, Dawes: Critical Realism. Studies in the Philosophy of Mind and Nature. Macmillan and Co, London, 1938

Jackson, Frank: Perception. A representative theory. Cambridge University Press, Cambridge, 1977

Langsam, Harold: "The Theory of Appearing Defended". Philosophical Studies, 87, 1997. S. 33-59

Martin, M.G.F.: "Perception". In: *Jackson, F. and Smith (Hrsg.):* The Oxford handbook of Contemporary Philosophy. Oxford University press, Oxford, 2005

Moore, George E.: "Some Judgments of Perception". In: *Swartz, R.J. (Hrsg.):* Perceiving, Sensing, Knowing. Anchor Books, Garden City/NY, 1918-19/65. S. 1-28

Prichard, Harold A.: Kant's Theory of Knowledge. Clarendon Press, Oxford, 1909

Quinton, A.: "The Problem of Perception". In: *Swartz, R.J. (Hrsg.):* Perceiving, Sensing, Knowing. Anchor Books, Garden City/NY, 1956/65. S. 497-526

Robinson, Howard: Perception. Routledge, London, 1994

Smith, A. D.: The Problem of Perception. Harvard University Press, Cambridge (MA), 2002

*Staudacher, Alexande*r: „Spielarten und Probleme des Disjunktivismus". In: *Bohse, H., Walter, S. (Hrsg.):* Ausgewählte Sektionsbeiträge der GAP.6. Mentis, Paderborn, 2007

Tye, Michael: Ten Problems of Consciousness. MIT Press, Cambridge (MA), 1995

Semantic Values, Beliefs, and Belief Reports

Clas Weber
Clas.Weber@anu.edu.au
Australian National University, Canberra

Abstract/Zusammenfassung

It is commonly assumed that propositions play a number of different theoretical roles. At one and the same time, propositions are supposed to be the semantic values of sentences, the objects of propositional attitudes, the contents of that-clauses, the contents of illocutionary acts, as well as the bearers of truth-values and modal properties. In this paper, I will focus on the first three of these roles. I will inquire into the question of whether propositions can play all three roles simultaneously and what kind of objects they have to be to do so. I am going to argue that they cannot be standard propositions, i.e. entities whose truth-value varies only relative to a possible world, to play any one of the roles. As it turns out, we will need rather non-standard entities, if we want something to do all the jobs at once.

The semantic value of a sentence in a context is a function from indices to truth-values. In contrast, we should conceive of the content of propositional attitudes as functions from contexts to truth-values. I attempt to show that attitudes ascriptions reflect this fine-grained content of propositional attitudes and that their content is therefore at least as fine-grained as that of the ascribed attitudes themselves. This observation presents a problem for our semantic framework. According to Kaplan's prohibition against monsters, all semantic composition takes place at most at the level of sets of indices (i.e. Kaplanian *contents*). However, we cannot capture the required content in terms of sets of indices. I present an alternative to a monstrous answer to this challenge. The central idea is to augment indices with a context component. Doxastic and epistemic operators then shift this contextual index parameter. This enables us to treat the "believe"-predicate as a context shifter, thereby reflecting the fine-grained content of attitudes, while retaining Kaplan's prohibition against monsters. If we thus conceive of propositions as functions from these extended indices to truth-values, we have found something that can play all of the three above-mentioned theoretical roles. Propositions, thus understood, are the semantic values of sentences in contexts, the content of that-clauses, and they also encompass the content of the corresponding propositional attitudes.

Es ist eine verbreitete Annahme, dass Propositionen eine ganze Reihe verschiedener theoretischer Rollen spielen. So wird davon ausgegangen, dass Propositionen zugleich die semantischen Werte von Sätzen, die Objekte propositionaler Einstellungen, die Gehalte von dass-Sätzen, die Inhalte illokutionärer Akte, sowie die Träger von Wahrheitswerten und modaler Eigenschaften sind. Ich werde mich in diesem Beitrag auf die ersten drei dieser Rollen konzentrieren. Dabei gehe ich der Frage nach, ob Propositionen alle drei Rollen gleichermaßen spielen können und wie sie dafür beschaffen sein müssen. Ich werde argumentieren, dass keine der drei Rollen von Standard-Propositionen gespielt werden kann, i.e. Entitäten deren Wahrheitswert lediglich relativ zu einer möglichen Welt variiert. Um alle drei Funktionen auf einmal zu erfüllen, benötigen wir ziemlich ungewöhnliche Entitäten.

Der semantische Wert eines Satzes an einem Kontext ist eine Funktion von Indizes in Wahrheitswerte. Den Gehalt propositionaler Einstellungen sollten wir hingegen als Funktionen von Kontexten in Wahrheitswerte auffassen. Ich versuche zu zeigen, dass Zuschreibungen von propositionalen Einstellungen diesen feinkörnigigen Gehalt von Einstellungen reflektieren und ihr Gehalt somit mindestens ebenso feinkörnig ist, wie jener der zugeschriebenen Einstellungen selbst. Diese Erkenntnis stellt unser semantisches System vor ein Problem. Gemäß Kaplans Monsterverbot findet jegliche semantische Komposition höchstens auf der Ebene von Mengen von Indizes statt (i.e. Kaplansche *contents*). Der benötigte Gehalt lässt sich jedoch nicht mithilfe von Mengen von Indizes modellieren. Ich schlage eine Alternative zu einer monströsen Antwort auf diese Herausforderung vor. Die zentrale Idee besteht darin, Indizes um eine Kontextkomponente zu erweitern. Doxastische und epistemische Operatoren verschieben dann diesen kontextuellen Indexparameter. Dies erlaubt uns das „glauben"-Prädikat als einen Kontextverschieber zu behandeln und damit den feinkörnigen Gehalt von Einstellungen wiederzugeben, jedoch gleichzeitig an Kaplans Monsterverbot festzuhalten. Wenn wir also Propositionen als Funktionen dieser erweiterten Indizes in Wahrheitswerte auffassen, haben wir etwas gefunden, was in der Lage ist, alle drei oben genannten theoretischen Rollen zugleich zu spielen. Propositionen, so verstanden, sind die semantischen Werte von Sätzen in Kontexten, der Gehalt von dass-Sätzen und sie umfassen darüber hinaus den Inhalt propositionaler Einstellungen.

1. Introduction

According to conventional philosophical wisdom, there are certain abstract objects – propositions – that are able to play a plethora of distinct theoretical roles. Propositions are supposed to figure as: the semantic values of sentences, the objects of attitudes, the contents of that-clauses in attitude reports, the contents of illocutionary acts, the bearers of truth and modal properties, and maybe more.[1] It is commonly further assumed that the objects that fulfil all these different roles are entities whose truth-value varies only relative to a possible world; they are conceived of either as functions from possible worlds to truth-values, or as structured entities that determine such functions. Let us call these entities *standard propositions*.[2]

In this paper, I will concentrate on the interconnection between the first three mentioned theoretical roles: the semantic values of sentences, the objects of attitudes, and the content of the sentences with which we reports these attitudes. It is intuitively plausible that there is a close connection amongst the three roles. We use *sentences* to ascribe attitudes to others and compositionality suggests that the contents of the involved that-clauses have to correspond to the semantic values of the embedded sentences. Moreover, the central function of language is

1 See: (Salmon and Soames, 1988, Introduction); (Stalnaker, 1999, p. 36); (King, 2007, pp. 1-3); (Cappelen and Hawthorne, 2009, p.1).

2 They are also often referred to as *eternal propositions*.

to enable us to share our beliefs, wishes and hopes with each other. A good semantic theory should be able to account for these points.

In Section 2, we will consider the question of what the semantic values of sentences are. Propositions in their role as compositional semantic values of sentences in contexts should be identified with functions from indices to truth-values. In Section 3, we will take a rather condensed look at what the objects of beliefs are. I suggest that we should conceive of them as functions from contexts to truth-values. Focussing on *de se* reports, I argue in Section 4 that that-clauses in belief reports are sensitive to this kind of content, and hence at least as fine-grained as the contents of beliefs. As I point out in Section 5, this observation turns out to be problematic for our semantic framework. According to Kaplan's prohibition against monsters, all semantic composition takes place at most at the level of sets of indices. However, we cannot capture the content of *de se* beliefs in terms of sets of indices. One possible reaction is to abandon Kaplan's prohibition against monsters. In Section 6, I will present my alternative to a monstrous account of belief reports. *The crucial idea is to enrich indices with a context component.* This allows us to treat the "believe"-predicate as a context shifter, while at the same time obeying Kaplan's prohibition against monsters. Finally, I compare my account of *de se* reports to a related account offered by David Chalmers and present a short consideration as to why the account proposed here is preferable to the latter. The arrived at semantic values of sentences correspond to the contents of that-clauses and they also encode the fine-grained content of attitudes. Our semantic framework is thus able to explain the connection between all three theoretical roles.

2. Semantic Values

2. 1. Contexts and Indices

How shall we decide what kind of entities the semantic values of sentences are?[3] I think we are well advised to follow the methodological strategy of David Lewis:

> In order to say what a meaning *is*, we may first ask what a meaning *does* and then find something that does that. (Lewis, 1970, p. 20)

There are two main tasks for meanings, i.e. semantic values, to perform:

1. They determine the truth-values of sentences depending on the way the world is.
2. They determine, in conjunction with syntactic structure, the semantic values of larger expressions in which they are embedded.

3 This section largely follows (Lewis, 1980). I will omit repeated reference to that work.

I will consider the first role first. It is not hard to find entities that determine truth-values relative to various possible ways the world might be: functions from possible worlds to truth-values do just that. However, the truth-values of many sentences vary within the same world. For instance, the sentence *It is raining* is true in rainy situations, and false when it is not raining. We therefore need more fine-grained entities as inputs to the functions. In general, we can say that the truth-values of sentences, which are variable in the described way, depend on the context in which these sentences are uttered. One can tentatively think of contexts as space-time points within a world, i.e. as triples of the form: <time, place, world>.[4] These locations model potential situations of utterance and they metaphysically determine the multifarious features that are relevant to settle the truth-values of context-sensitive sentences. Given their first role, we can thus identify semantic values with *functions from contexts to truth-values*.

What about semantic values in their second, compositional role? Sentences combine with other expressions to form larger sentences. A reasonable explanation of how we are able to communicate under time constraints using these sentences had better assume that they do so in a systematic, i.e. compositional, manner.[5] The semantic values of larger sentences are determined by the semantic values of their parts and their syntactic make-up.

Certain linguistic constructions conjoin with sentences in such a way that the truth-value of the resulting sentence systematically depends on what the truth-value of the embedded sentence is in other situations. We can say that those expressions "shift" us to different situations. For instance, the expression *Yesterday, it was the case that* shifts us one day back; i.e. *Yesterday, it was the case that S* is true iff *S* was true yesterday. The same mechanism is operative in certain modal and locational constructions. Hence, to do their second job, semantic values have to incorporate information about the truth-values of sentences at other times, places, and worlds. Apparently, that is just the same information semantic values had to provide to fulfil their first task: to model context-dependence, they needed to tell us about the truth-values of sentences at differ-

4 There are certain technical difficulties here: do we have to include a speaker parameter in the tuple to account for situations in which e.g. one hungry and one satiated ghost utter the sentence *I am hungry* at the same space-time location within the same world? On the other hand, one might not want to include a speaker parameter, if one thinks that sentences like *There is a speaker* should not come out as true at every context. This issue is especially pressing for those who want to use *truth at every context* to model some notion of the *a priori*; see the discussion in: (Chalmers, 2006, Section 2). Even without a speaker parameter, there is a problem for this project within the present framework with sentences like: *Space–time exists*.

5 For this way of arguing for compositionality, see: (Pagin and Westerståhl, forthcoming, Section 4.6).

ent times, places, and worlds.[6] It thus seems that functions from contexts to truth-values fit the bill for the second, compositional role of semantic values just as well. Unfortunately, things are not that simple.

Intensional operators shift single features of the context. But as Lewis notes: "No two contexts differ by only one feature. Shift one feature only and the result is not a context at all." (Lewis, 1980, p. 29). If that is true, we end up at points that do not correspond to contexts anymore. Hence, we need something over and above contexts at which to evaluate sentences. However, *prima facie* it seems that there are many contexts that differ by one feature only. According to our present framework, and Lewis's alike, contexts are time-place-world triples. Now, if one solely shifts the time feature of e.g. the triple <today, Canberra, $world_{@}$> by one day forward, one ends up with the triple <tomorrow, Canberra, $world_{@}$>. That looks very much like another context. Is Lewis confused?

No. Lewis is making a valid point. Contexts determine *sets of contextual features*. The triple <now, Canberra, $world_{@}$> determines in addition to the time, place and world, e.g. me as speaker, Tobias as addressee, a certain object as demonstratum, the history of the universe up to the time of the context, etc. If we now shift a single element of this set of contextual features - we move the place to Berlin, say - we get another set of contextual features that differs from the first set only in its place feature. However - and this is the point Lewis is getting at - there will be no corresponding context triple that determines this specific set of contextual features. Since e.g. Tobias and I are in Canberra at the present moment in the actual world, there will not be a triple <now, Berlin, $world_{@}$> that determines this place, time, and world, plus me as speaker, Tobias as addressee, etc. Since no context triple will determine this specific set of contextual features, it will not correspond to a context.

Even though this reasoning will generalize, Lewis' claim that no two contexts differ by a single feature only is still too strong. There might be e.g. symmetrical universes, or worlds with eternal recurrence where a shift of a single contextual parameter results in a list of contextual features for which there is in fact a corresponding context triple. These minor difficulties notwithstanding, we now have a principled reason for expanding our semantic values - we cannot get by just with contexts. We need to evaluate sentences at points that do not correspond to contexts; i.e. we need so called *indices* as inputs to our semantic functions. The need for indices arises, because of the existence of *shifting* expressions in our language. We may therefore conceive of indices as *sets of shiftable contextual features*. For now, we have considered time, place, and world shifters. We can accordingly regard indices as arbitrary triples of the form <time,

6 Provided we follow Lewis in considering contingency as just another form of context-dependence.

place, world>. Index triples will, in contrast to context triples, not necessarily correspond to real (actual or possible) situations.

2. 2. Two objections against the argument for relativized semantic values

In this section, I will shortly consider two recent objections against the just given conception of semantic values. According to the above argument, the semantic value of a sentence type in a context will be a function from world, time, place triples to truth-values.

In (King, 2003) and (King, 2007), Jeffrey King objects to this account of semantic values. According to King, the semantic values of sentences are required to also function as objects of attitudes. However, King claims, the objects of attitudes do not vary in truth-value across time and location, but only across worlds - they are standard (i.e. eternal) propositions. If the above argument for time- and location-relative truth conditions is sound, we have to conclude that semantic values do in fact vary in truth-value across times and location. However, since time- and location-relative semantic values cannot, in King's eyes, play the role of objects of attitudes, we would then need a *further* class of objects serving as objects of attitudes: standard propositions. But this requires abandoning the view that there is a single class of objects that plays both roles.[7]

King's main objection against the above argument for temporally and locationally relativized semantic values is based on syntactic considerations. He points out that the argument for relativized semantic values relies on a *modal* interpretation of the relevant constructions; i.e. we interpreted the relevant expressions as sentential operators that shift certain contextual parameters. King thinks that this interpretation is inferior to an *extensional* conception, according to which temporal expressions function as *quantifiers* that bind tacit time and place variables (or alternatively as referential expressions). According to that conception, English sentences make implicit reference to times (and locations).

It is important to notice that King's own extensional alternative does not give him what he is after. His main motivation in arguing against the modal account was to preserve a unified picture of propositions, according to which standard proposition are both objects of attitudes and compositional semantic values. However, on King's own account, temporal operators likewise do *not* take standard propositions as arguments. Rather, they operate on the semantic value of *open sentences*. The semantic value of an open sentence is commonly taken

7 I will argue in the next section that King's assumption about the objects of attitudes is wrong, and that the contents of beliefs do indeed change their truth-values across times and locations.

to be a set of assignments that make the open sentence true.[8] Hence, whatever the advantages of the extensional framework may be, it does not in general allow us to identify the objects of attitudes with compositional semantic values, as King intended. Standard propositions will play the role of objects of attitudes, while sets of assignments will play the role of compositional semantic values for the constructions under consideration. On the other hand, if it turns out that, contrary to what King supposes, the contents of beliefs are in fact time- and location-relative, then there is no need to give up the unified picture of propositions. This is what I will argue for in section 3. Thus, ironically, opting for a modal framework and the corresponding relativized semantic values does preserve the chance of having a unified picture, while it is the extensional account that has to give up on this prospect.

Another objection against the above argument for relativized semantic values can be found in a recent monograph by Herman Cappelen and John Hawthorne.[9] Our argument relied on the premise that the relevant time and location shifters are to be treated in the same way as ordinary modal operators. In particular, we parsed them as *sentential* operators. Cappelen and Hawthorne object on syntactic grounds to this premise of the argument. They discuss the *Sententiality* premise in relation to the temporal expression "On Tuesday" and claim that it is implausible to interpret that expression as accepting sentential arguments. This may be true. However, it should be obvious that establishing that much does not demonstrate that there are no sentential temporal operators in English. A far more important case for the temporal dimension are tenses. And it is standard practice in linguistic theory to treat past and future tenses syntactically as sentential operators. In standard x-bar theory, past or future tenses are interpreted as heads of the inflection phrase, which c-commands the verb phrase. Tenses thus syntactically dominate a sentence-like entity; i.e. they behave as sentential operators.

Moreover, there are complex constructions that uncontroversially take sentences as arguments, such as: *It has been the case that*. We find the same complex constructions for the locational and modal dimension as well, such as: *In Australia it is the case that*, *It is necessary that*. These constructions compose with sentences and they have just the truth conditional effect that our theory predicts. Thus, I take it that Cappelen and Hawthorne's objection against the discussed syntactic premise fails.[10]

---

8 In King's framework, things look slightly different. Simplifying somewhat, we can say that the semantic value of an open sentence is a structured entity with a hole in it (corresponding to the open argument place); see: (King, 2007, Appendix).

9 See: (Cappelen and Hawthorne, 2009, chapter 3).

10 Cappelen and Hawthorne reconstruct our argument for relativized semantic values as relying on four premises, of which I here could only discuss one, because of limitations of space. Their basic argumentative strategy is to present counterexamples to all four premis-

2. 3. Double Indexing

I have attempted to show that we have to move beyond contexts as inputs to the interpretation function. Since all contexts are indices, but not *vice versa*, we might relativize the interpretation function simply to indices and consider semantic values of sentences to be functions from indices to truth-values. However, we cannot just rest here, either. The need for a further modification of our framework derives from the interaction of intensional operators and certain "freezing" expressions that resist intensional shift. Some expressions, indexicals being the paradigm example, are such that even under intensional embeddings, their contribution to the truth conditions depends on aspects of the actual utterance context and not on features of the points to which the operator shifts us. Compare the following sentences:

(1) It will always be the case that it is raining.
(2) It will always be the case that it is raining now.

In (2), the indexical *now* has frozen the time feature and protected it from being shifted by the time shifter *It will always be the case that*. To evaluate (2), we have to go back to the time of utterance and assess the weather there. Given that it rains at that time, (2) will be true regardless of what the weather will be in the future. In contrast, (1) is true only if it will go on raining until the end of time.

To account for freezing expressions within intensional embeddings, semantic values must store information about the actual utterance context, such that we are able to recover it at the shifted points. Semantic functions therefore have to take two elements as inputs: the original contexts and the indices to which we are shifted.[11] Consequently, semantic values are functions from two triples of the form <time, place, world> to truth-values, where one triple corresponds to a context and the other to an index. Conclusively, we have good reason to conceive of semantic values of sentences as *functions from contexts and indices to truth-values*. Propositions in their role as semantic values of sentences *in contexts* can then be understood as *functions from indices to truth-values*.

---

es, i.e. sample expressions for which one of the premises fails. However, all that is needed to run the argument is *one* convincing example for each intensional dimension, and they have not succeeded in demonstrated that such examples do not exist. The reader can convince herself that all premises hold for the following expressions: *It has been the case that* (time); *In Australia it is the case that* (location); *It is necessary that* (world). Hence, we should relativize semantic values to times, locations, and worlds.

11 At least this need arises under slightly more complicated constructions (for the simplest cases, we might get by with having a designated index which serves as the context of utterance). As (Cresswell, 1990) has shewn, if we complicate the constructions even further, semantic values will have to take even more elements as inputs. I will ignore this complication here.

3. Beliefs

Let us next examine the role of propositions as objects of attitudes; I will here focus on beliefs. A number of philosophers have convincingly argued that the objects of beliefs cannot in general be given by standard propositions, i.e. entities whose truth-value is relative only to a possible world.[12] One prominent case for that conclusion is provided by (Lewis, 1979).[13] Lewis presents an example of two gods who know exactly which possible world they inhabit - they are omniscient with respect to standard propositions. Still, they are ignorant about who, where and when they are. This implies that the information they acquire when they learn who, where, and when they are cannot be captured in standard propositional terms.[14] Hence, not all beliefs can be modelled by standard propositions.

A different case against capturing the content of attitudes with the help of standard propositions is exhibited by subjects who agree in all their beliefs and desires, when those are cashed out in non-indexical terms - they bear the same relation to all standard propositions. Such individuals may nevertheless differ significantly in their behavioural dispositions. If you want to have the last piece of cake and I also want you to have the last piece of cake, I will offer it to you, while you will eat it.[15] Given that our practice of ascribing attitudes to others is rooted in their role in explaining and rationalizing behaviour, those attitudes will be, at least partly, individuated by their causal-functional roles. Two persons that differ in their respective behavioural dispositions will thereby also differ in what attitudes they have. It follows that the standard picture does not provide an adequate theory of all attitudes.

Lewis has offered a very elegant way of modifying the standard picture. Instead of using sets of possible worlds to model the content of belief, he suggested that we should rather characterize it in terms of *centered* possible worlds. Centered possible worlds correspond to contexts from the previous section: they are time-place-world triples. As Lewis has demonstrated, the centered content account subsumes the standard one: each standard proposition corresponds to a centered proposition, but not *vice versa*.

According to Andy Egan, the best case for construing belief content in terms of centered worlds is provided by so-called "arguments from similarity". Only centered content, Egan claims, allows us to account for certain commonalities among different thinkers:

12 See: (Perry, 1979); (Lewis, 1979); (Chalmers, 2009a).

13 See: (Lewis, 1979).

14 See: (Stalnaker, 2007, Chapter 3) for a strategy to resist Lewis's conclusion. This strategy, however, is committed to the rather controversial thesis of *haecceitism*.

15 See: (Perry, 2006, p. 215).

> Take all of the well–informed inhabitants of Cleveland. They've all got something doxastically in common. They all believe that they are in Cleveland. [...] This doxastic similarity, though, is one that we can't capture on the Simple Picture [i.e. the standard conception, C. W.]. (Egan, ms, p. 7)

Intuitively, there is a belief which everybody who believes to be an inhabitant of Cleveland shares. We can only capture this commonality, so Egan argues, in terms of centered content, not in standard terms. However, I do not think that this is the best, or even a very good case for the centered content conception. We have a trade-off: we gain some similarities by moving to centered content, but we lose others at the same time. If Tobias thinks to himself *I am sick* and I think *Tobias is sick*, there is an intuition that we share a common belief. Importantly, the centered content picture does not secure this similarity, since there is no centered proposition that we share just in virtue of having the above beliefs. Admittedly, when Tobias has the first-personal belief, he will typically also believe *Tobias is sick*, so that in normal circumstances there is indeed a common object of belief even on the centered conception. Importantly, however, he *need not* believe this, e.g. when he is confused about who he is. Some of our intuitions about shared beliefs seem to go one way - other intuitions go the other way. Therefore, I do not think that intuitions about shared beliefs provide an adequate means to adjudicate the issue before us.

However, we have already encountered two independently compelling cases for centered belief content. I thus conclude that we should follow Lewis and take the content of beliefs to be sets of centered worlds, or equivalently sets of contexts. Thus, propositions in their function as objects of beliefs should be understood as *functions from contexts to truth-values*.[16]

4. Belief Reports

It would seem that once we establish what the objects of beliefs are, the question about the contents of that-clauses within belief reports would be easily answered, as well. Belief reports report beliefs, thus their respective contents have to be the same. Regrettably, things are once again not that straightforward. It is not self-evident that our English "believe"-predicate is sensitive to each and every difference among our beliefs. Indeed, it seems that the self-locating nature of essentially indexical beliefs is lost in English attitude reports. We both report the first-personal belief Tobias has about himself *My pants are on fire*, as well as

16 In interpreting the above considerations as straightforward arguments for a specific account of the *content* of beliefs, I am assuming that Perry's strategy of distinguishing *belief contents* from *belief states* is eventually unsatisfactory. Our practice of ascribing certain attitudes to each other is, I take it, silent on the representational vehicles of the ascribed contents.

the third-personal belief he has unknowingly also about himself *His pants are on fire*, with the following belief report:

(3) Tobias believes that his pants are on fire.

Thus, the English "believe"-predicate apparently does not track the corresponding differences in the underlying attitudes.[17] This observation has belief reports comply with a law of semantic theory decreed by David Kaplan, the so-called "prohibition against monsters". According to that prohibition there are no natural language constructions that are sensitive to the *character* of an expression. Or to put it in our terminology: semantic composition takes place at most at the level of functions from indices to truth values, i.e. after the functions are saturated with the respective context argument. No linguistic construction shifts the context argument.[18]

Kaplan assumes that in the described situation, both of Tobias' beliefs will determine the same possible world proposition: the set of worlds in which Tobias' pants are on fire.[19] Our ascription practice, reflected in statement (3), suggests that only this coarse-grained propositional object gets reported. And this coarse-grained proposition corresponds to the Kaplanian *content* of the complement sentence of (3), in accord with Kaplan's prohibition against monsters.

However, appearances are deceptive. Firstly, there is a form of attitude reports, which exclusively serve to report essentially self-locating attitudes. Secondly, one can show that (3) is ambiguous between a *de re* and a *de se* reading. To illustrate the first point, consider the sentence:

(4) Tobias wants to set his pants on fire.

Reports like (4), with a subject control PRO element in the underlying Logical Form (LF), can only be used to report *de se* attitudes.[20] For the second point, imagine the following scenario: out of a group of four men whose pants are on fire, three are watching themselves in the mirror. Without recognizing themselves in the mirror, they are each thinking about themselves: *His pants are on fire*. Tobi-

17 Again, it is evident that there are two different attitudes involved, since the two beliefs are associated with rather different dispositions to act.

18 According to von Stechow and Zimmermann, Kaplan's prohibition against monsters is "the most important restriction for the interpretation of natural languages [...]." (von Stechow and Zimmermann, 2005, p. 212). The truth conditions Kaplan himself gives for indirect speech reports do involve characters. However, they are only quantified over. On his account, belief reports are not sensitive to the specific character of an attitude, they only state that there is some character or other entertained by the subject that yields the right content. See: (Kaplan, 1989, Section XX.).

19 This assumption is problematic, since even Tobias himself will be ignorant about some of his essential properties. For this problem, see: (Lews, 1981).

20 See: (Chierchia, 1989).

as is the only one that entertains a belief of the form: *My pants are on fire*. It seems that we can correctly report this situation with the sentence:

(5) Only Tobias thinks that his pants are on fire.

Given that (5) is a correct characterization of the above scenario, (5) and (3) must have a dedicated *de se* interpretation.[21]

These cases demonstrate that at least some attitude reports are sensitive to and therefore at least as fine-grained as the centered content of the corresponding attitude, against Kaplan's suspicion. We can then tentatively assume that the contents of that-clauses within belief reports are (at least in some cases) equivalent to the contents of the beliefs they report. Thus, propositions, as the contents of that-clauses, have to be at least as fine-grained as functions from contexts to truth-values.

Let me recapitulate where we stand so far. I have tried to establish the following claims:

1. Propositions in their role as semantic values can be identified with functions from indices to truth-values.
2. In their role as objects of beliefs, propositions emerge as functions from contexts to truth-values.
3. Since belief reports are sensitive to the centered content of beliefs, propositions understood as contents of that-clauses will correspond to something at least as fine-grained as functions from contexts to truth-values.

We now have to integrate these results into our semantic framework. This task turns out to be rather difficult, since the right content for belief reports cannot in general be derived from sets of indices. A suggested solution is to break Kaplan's law and opt for a monstrous account of attitude reports. I will suggest that this proposal is not a promising one for English attitude reports. Therefore, I am going to present a non-monstrous alternative in section 6.

5. Semantic Values, Belief Reports, and Monsters

How does the claim that that-clauses capture the centered content of attitudes mesh with our semantic framework? Consider again sentence (3) on its *de se* interpretation:

(3) Tobias believes that his pants are on fire.

21 This case is based on the one found in (Percus and Sauerland, 2003). Their claim that the ambiguity of (3) is explained by two distinct underlying LFs is contested by (Anand, 2006, Section 1.3). Importantly, both parties agree that (3) does indeed have a *de se* reading over and above the *de re* interpretation.

We saw that this report reflects Tobias' first-personal belief *My pants are on fire*. The content of this belief is the set of contexts in which the agent of the context has burning pants. Now, take the corresponding sentence as used by Tobias:

(6) My pants are on fire.

How does the semantic value of that sentence relate to the content of Tobias' belief? The content of a sentence in a context is a set of indices. In contrast, the content of his belief is a set of contexts. Can the "believe"-predicate still yield the right content, given that contexts are identical to indices? Maybe our "believe"-predicate picks out just those indices that correspond to the right contexts?

Regrettably, we cannot capture the right content in terms of indices. Assume that I could not possibly (nor anytime or anywhere) be identical to David Hume. That implies that the set of indices associated with my use of the sentence *I am David Hume* will be the empty set. However, I might believe without contradiction that I am David Hume. The set of contexts associated with my belief can therefore not be empty and the "believe"-predicate can thus not pick out a special subset of the associated set of indices.

There is another relation between the content of Tobias' belief and the semantic value of the corresponding sentence (6). The belief content corresponds to the so-called *diagonal* of the sentence. We can obtain the diagonal of a sentence with the help of the diagonal operator δ. This operator replaces the context parameter with the index-parameter:

$$[\![\delta\alpha]\!](\text{context}_i)(\text{index}_i) = [\![\alpha]\!](\text{index}_i)(\text{index}_i)$$

As we can see, δ is a monster - it shifts the context parameter. Does the English "believe"-predicate contain a hidden monstrous operator?

Apart from a violation of Kaplan's monster prohibition, there are two difficulties with this suggestion. The first problem concerns the behaviour of indexicals in belief contexts. In a Kaplanian framework, indexicals receive their interpretation from the context argument. Since δ shifts the context element, we would expect to find shifted indexicals within English belief reports. However, the evidence for shifted indexicals in English belief reports is dim to non-existent.[22]

22 There are clear examples of shifted indexicals within belief reports of other languages; see: (Schlenker, 2003); (Anand and Nevins, 2004). Even if there were examples of shifted indexicals in English belief reports, they might have non-monstrous explanations, e.g. anaphoric or quantificational binding. These explanations are already needed for other examples such as: *Every football fan went to a **local** bar to watch the Superbowl*. For cases like these, see: (Partee, 1989). Such a case of binding is arguably what accounts for the only example of a shifted indexical in English indirect speech that Schlenker presents:

The second difficulty is this: even though (6) would be an appropriate sentence for Tobias to articulate his first-personal belief, it is not the sentence that is used in the complement clause of (3): *Tobias believes that **his** pants are on fire*. Hence, even a monstrous account would not straightforwardly help us in deriving the right content for the belief report (3). If the "believe"-predicate was a real monster, we would instead have to use a belief report involving (6) to characterize Tobias' first-person belief:

*(7) Tobias believes that **my** pants are on fire.*

This is how belief reports work in languages that purportedly have a monstrous "believe"-predicate, but it is obviously not how English belief reports function. The second difficulty thus boils down to two points:

1. The monstrous account has to give the right content for (3).
2. It also has to explain why (7) cannot be used to report Tobias's belief.[23]

6. Context, Index and Context

Given these difficulties, I think it is worthwhile to explore an alternative to the monstrous approach. We want the "believe"-predicate to yield the right set of contexts from the semantic value of the complement sentence. As we have seen, this is a problem for the present account, since the relevant set of contexts cannot in general be derived from the set of indices associated with the complement clause. *I therefore suggest enriching indices with a context element*.

So far, our indices are time-place-world triples. Why do indices contain just these features? The reason is our motivation for and our understanding of indices. We needed indices to give a compositional semantics for shifting constructions; indices were therefore functionally characterized as *packages of shiftable features*. Since we have so far found only time, place, and world shifters, we have construed indices as the corresponding triples.[24]

"John has told me repeatedly over the years that he was sick *two days ago*". See: (Schlenker, 2003, p. 64, my emphasis).

23 Of course, if we followed the assumption in (Schlenker, 2003) that pure indexicals in English have a special feature that protects them from being shifted by the allegedly monstrous belief operator, we would have an explanation for why we cannot use (7) to report Tobias' belief. (We might also have an answer to the first mentioned difficulty.) However, this strategy is somewhat artificial, because all the clear evidence for monsters in English is guaranteed not to surface in virtue of the postulated special features of English indexicals.

24 (Lewis, 1980) also mentions standard of precision shifting operators, such as "Strictly speaking". I am ignoring these operators here.

Following an idea from Jaakko Hintikka, we can treat the "believe"-predicate as a shifting operator, too.[25] According to Hintikka, the belief operator quantifies over a subject's belief worlds. A belief report *S believes that p* is true iff *p* is true in all of S's belief worlds. However, we cannot in general treat the belief operator as a *world* shifter. Neither can we interpret it as a complete *index* shifter - there are sentences that correspond to the empty set of indices, while still serving to ascribe non-trivial beliefs.

Rather, the belief operator shifts *contexts*. To account for this, I propose to add a context element to the index, such that they will now be quadruples of the form: <time, place, world, **context**>. The belief operator then quantifies over those contexts that correspond to the subject's doxastic alternatives. The original context arguments are still needed as anchors for indexicals expressions, since, even in belief reports, indexicals are not shifted. Here is the basic idea (where 'c' denotes the original context and '$\mathbf{c}_*$' the newly introduced context element within the index argument):

> [[x believes that φ]] (c) (<t, p, w, $\mathbf{c}_*$>) = True, iff [[φ]] (c) (<t, p, w, $\mathbf{c}_*$**'**>) = True at all $\mathbf{c}_*$**'** that are x's doxastic alternatives in w_c at t_c.

We now have three different roles for contexts:

1. They are arguments for the semantic values of sentence types.
2. They model the objects of beliefs.
3. They are the element of the index that the belief operator shifts.

This account gives us an explanation for why we cannot use: *(7) Tobias believes that my pants are on fire*, to report Tobias' first-personal belief. Since indexicals are still anchored in the original context argument, the pronoun *my* will contribute the actual speaker to the truth conditions of (7).

What about the report: *(3) Tobias believes that his pants are on fire*, on its *de se* reading? We do not want the pronoun *his* to refer to Tobias in all of Tobias' doxastic alternatives. That would make (3) report the third-personal belief *Tobias' pants are on fire*. To get the right result, we need for the pronoun *his* to denote the respective agents of the belief contexts the belief operator shifts us to. These agents will not always be identical to Tobias.

There are different ways to achieve this. One option is to plead for ambiguity. The idea is that in *de se* reports, the complement contains special *logophoric* pronouns which in English superficially look like ordinary personal pronouns. We might even have constructions in English that make this explicit: I vs. I myself, she vs. she herself; his vs. his own, etc.[26] We can then say that these logo-

25 See: (Hintikka, 1962).

26 Compare Castañeda's 'he*' in (Castañeda, 1966). (Anand, 2006, Section 2.4.1) objects in a nearby context to related "pronoun-centered" approaches. However, his objections rely on peculiarities of the languages he considers, and do not straightforwardly apply to Eng-

phoric pronouns work differently than ordinary indexicals. More specifically, they denote the subject of the relevant belief contexts and equally so in the case of "I myself", "you yourself", "she herself", etc.

A more elegant and general strategy can be found in (von Stechow, 2002).[27] To illustrate the basic idea, consider the first person equivalent of (3):

(8) I believe that my pants are on fire.

Von Stechow's proposal is that the pronoun *my* in the complement is only morphologically, but not semantically present. The reason is that the pronoun's semantic features - e.g. that it refers to the speaker of the actual utterance context - get deleted under semantic binding. He formulates the responsible principle thus: "Feature deletion under semantic binding: Delete the features to [sic] all variables that are semantically bound." (von Stechow, 2002, p. 380). In our case, *my* is bound by the context shifting operator introduced by the "believe"-predicate. Since the semantic features of *my* are deleted, what remains at the level of LF is an individual variable. This variable is then bound by the context shifter and assigned to the respective subjects of the different belief contexts, as desired. The same mechanism is operative for other personal pronouns, as well. Hence, the semantic features of the third-person pronoun *his* in (3) are likewise deleted and the remaining variable gets assigned to the subjects of the respective belief contexts, too. We thus have an account that also yields the right content for (3). The complement clauses of (3) and (8) express the same semantic value after the pronouns' features are deleted. This reflects the fact that both clauses report the first personal belief: *My pants are on fire*.

However, in English the feature deletion principle is only operative under morphological agreement. This explains why the semantic features of *my* in the complement clause of (7) do not get deleted. Again, there is our explanation as to why we cannot use (7) to report Tobias' first personal belief.

We now have an account of *de se* reports for all personal pronouns that does not postulate a genuine ambiguity and thereby complies with Grice's razor. Moreover, the account generalizes to mood and tense features in attitude reports. It thus accounts e.g. for so-called "sequence of tense" phenomena. Finally, feature deletion is not restricted to attitude reports, i.e. "verbal quantification", but extends to many examples given by (Partee, 1989) and (Heim, 1991), such as the "sloppy reading" of:

(9) Only I did my homework.

Hence, the account is independently motivated.

---

lish. The ambiguity approach would assimilate the current framework to some extent to the account of (Schlenker, 2003).

27 Similar accounts were given by: (Kratzer, 1998); (Heim, 1991) and (Heim, 2009).

However, the feature deletion approach gives rise to a question about my suggestion to add a context element to the index. The motivation for this move derived from the observation that the original indices could not give us the right content for e.g.:

(10) I believe that I am David Hume.

The set of indices associated with the sentence *I am David Hume* turned out to be empty, whereas the content of the corresponding belief was not. However, according to the present suggestion, this sentence is not semantically a part of that belief report. On the level of LF, the complement rather looks like this: *x is David Hume*. Does not this observation undermine the motivation for a context parameter within the index? I do not think it does. The belief operator introduces a context quantifier. The semantic binding explanation for the personal pronoun case requires there to be an element that quantifies over individuals. I suggest that we interpret the context quantifier as a lambda abstractor that abstracts over the corresponding triples: $\lambda<x_e, t_i, w_s>$.[28] The operator quantifies over individual, time, world triples. We thus have to modify our conception of contexts, and construe them as triples of the form: <speaker, time, world>, to account for the feature deletion processes of person, tense, and mood. However, sticking to our functional characterization of indices, they are still inapt as elements for the belief operator to quantify over. There is no dedicated shifting operator for individuals in English and hence no independent motivation to have a subject element within the index that the individual element of the lambda abstractor $\lambda<\mathbf{x_e}, t_i, w_s>$ could bind. Hence, the motivation to have a context element within the index is not undermined.

7. Comparison with Chalmers' Framework

The presented account opens up the possibility for an approach, which has a close resemblance to that of (Chalmers, 2009b). The original sets of indices are similar to Chalmers' *secondary propositions*, while the added sets of contexts resemble Chalmers' *primary propositions*.[29] The extended sets of indices thus correspond roughly to Chalmers' *enriched propositions*. One advantage of both

28 I prefer this representation to one where we have three different λ-abstractors - one over individuals, one over times, one over worlds. The belief operator does not quantify over arbitrary individual, time, world triples, but only over triples that correspond to contexts/centered worlds. This feature is meant to be represented by letting the belief operator quantify over the corresponding triples.

29 There are important differences, though. Chalmers secondary and primary propositions are structured entities, and his primary propositions are defined over a space of *epistemic scenarios*.

accounts is that they have the belief operator operating on the same object as the other intensional functors. What is more, the parallel to Chalmers' account suggests that the account can be generalized from the *de se* case to other forms of belief reports. The rough idea is that expressions do not only have a *modal* (temporal and locational) intension as standardly assumed in the Lewisian and Kaplanian framework. Moreover, they also have a *contextual* (or in Chalmers' terms: *epistemic*) intension, which is their semantic contribution to epistemic and doxastic operators. We might then explain typical Frege puzzles by arguing that "Hesperus" and "Phosphorus" are associated with different contextual intensions. Then there will be belief contexts where "Hesperus" and "Phosphorus" having different extensions, yielding a non-trivial content for the corresponding complement.

This suggestion also makes room for a different story about *de se* reports. One could claim that "I" likewise has both a modal and a contextual intension. The contextual intension of "I" will be a function from belief contexts to the respective speakers of the contexts, and the "believe"-predicate in *de se* reports will be sensitive to this contextual intension. This would give us the right result for the report: (8) *I believe that **my** pants are on fire*. However, the explanation will not carry over to report: (3) *Tobias believes that **his** pants are on fire*. "I" (or "my") and "he" (or "his") will have different contextual intensions, while the reported beliefs in (3) and (8) are the same. Chalmers (2009a) accepts this result. He accounts for the data by introducing a context-sensitive coordination relation between the content of the belief and the content of the corresponding that-clause.[30] The content of the complement clause does not have to be identical to the content of the corresponding belief; rather, the two contents have to stand in a certain relation to each other. How close the correspondence between what is expressed by the complement and the content of the reported attitude has to be depends on the context.

However, for most *de se* reports there is a principled mismatch. In most cases, the content of the complement clause and the content of the corresponding belief can never be in harmony. *De se* reports express first personal beliefs. Using some contestable terminology, we can say that the believer employs an EGO-concept. The only pronouns that express EGO-concepts are first person singular pronouns, like "I" and "my".

Hence, on the current conception, only in cases of self-ascriptions of *de se* beliefs can the content of the report truly match the content of the corresponding attitude.

> *(+match)*
> *I believe that my pants are on fire.*
> *I believe that I am Hume.*

30 This idea is inspired by the account in (Richard, 1990).

For all other *de se* reports, there cannot be a complete match, since no other pronoun expresses an EGO-concept.

> *(−match)*
> *You believe that your pants are on fire.*
> *They believe that their pants are on fire.*
> *He believes that he is Hume.*

A possible account of *de se* reports in Chalmers' framework might be to interpret them as a special subclass of *de re* reports. We have seen that *de se* reports require a reading that is stronger than the standard *de re* interpretation. We might explain the difference between ordinary *de re* reports and the special class that correspond to *de se* ascriptions by a difference in what is required of the contextually relevant acquaintance relation: *de se* reports require the subject to be acquainted with the relevant *res* in an especially intimate way. As in standard *de re* reports, the pronoun in the complement clause will merely specify who the relevant belief is about, i.e. the *res*, but there will be an extra restriction requiring the subject to be acquainted with the corresponding *res* by the relation of *identity*. Even if that is a viable option, I believe that the account proposed here is superior, since it enables *de se* reports to faithfully reflect the contents of the reported beliefs for all personal pronouns.[31]

Before concluding, we have to deal with a final wrinkle. As (Chalmers, 2009b) has pointed out, some belief reports are sensitive to more than the contextual/epistemic intension of the complement sentence. The truth-values of belief reports can thus not in general depend merely on sets of contexts alone. The reason is that some belief reports characterize the reported belief in relation to the ascriber's environment. If there is an expression in the complement that depends for its meaning on the situation the ascriber is in, the belief report will likewise depend for its truth on the actual extension of the relevant expression. It is arguable, however, that extensions plus sets of contexts suffice to account for the semantics of belief reports. And the extension of an expression is contained within the set of extended indices - it is the expression's denotation when evaluated at the index that corresponds to the context of utterance. Hence, even if sets of contexts are not enough, sets of extended indices seem to be sufficient to give an adequate account of belief reports.

If the conception I have proposed here does indeed give us a reasonable account of *de se* reports, we can eventually conclude that propositions in their role

31 There might appear to be something problematic with the proposal suggested on behalf of Chalmers of reducing *de se* reports to *de re* ones. The relevant acquaintance relations are plausibly egocentric relations. So the situation seems to be the opposite: the *de re* is actually a species of the *de se*; for such a verdict, see also: (Cresswell and von Stechow, 1982). Bearing in mind, however, the distinction between beliefs and belief reports, that is not an immediate problem for the proposed account, since we do not attempt to give a semantic analysis of the quantified over egocentric acquaintance relations.

as contents of that-clauses can be understood as functions from extended indices to truth-values.

8. Conclusion

In this paper, I have considered three different roles for propositions: the semantic values of sentences, the objects of beliefs, and the contents of that-clauses in attitude reports. The fact that sentences embed under time, location, and world shifting operators suggests that their semantic values in contexts are functions from time, location, and world tuples to truth-values. I then shortly discussed the question of what the objects of beliefs are, arguing that we should follow Lewis and conceive of them as functions from contexts to truth-values. Evidence from *de se* reports demonstrated that some of our belief reports track the fine-grained, centered content of attitudes. This observation posed a problem for our semantic framework: we could not model the right kind of content in terms of sets of indices. I proposed an alternative to a monstrous solution to this problem. The basic idea was to interpret the "believe"-predicate as a further shifting operator. However, it was not adequate to treat this operator as shifting just the world parameter of the index; neither could we interpret it as shifting whole indices. Rather, it has to shift contexts. Given our functional characterization of indices and Kaplan's prohibition against monsters, my proposal was to add a context parameter to the index, while also retaining the original context argument as an anchor for indexical expressions. The resulting system turned out to have a close resemblance to the account of attitude ascriptions recently developed by David Chalmers. However, I presented a short argument as to why an account of *de se* reports relying on von Stechow's feature deletion principle was superior to that of Chalmers. The picture we end up with is this: the semantic values of sentences in contexts are functions from extended indices to truth-values. These semantic values correspond to the contents of that-clauses and they also encode the fine-grained, *de se* content of attitudes. Our system is thus able to give a unified account of the three discussed theoretical roles.[32]

References

Anand, Pranav: De de se. Ph. D. thesis. MIT, 2006

32 I would like to thank David Chalmers, Leon Leontyev, Dan López de Sa, Dan Marshall, Brian Rabern, Jonathan Schaffer, Wolfgang Schwarz and Tobias Wilsch for discussion.

Anand, Pranav/Andrew Nevins: "Shifty operators in changing contexts". In: *R. Young (Hrsg.)*: Proceedings of SALT XIV, CLC Publications, Cornell, NY, 2004

Cappelen, Hermann/Hawthorne, John: Relativism and Monadic Truth. Oxford University Press, Oxford, 2009

Castañeda, Hector: "'He': A study in the logic of self-consciousness". Ratio 8, 1966. S. 130-57

Chalmers, David: "Two-dimensional semantics". In: *Lepore, E./ Smith, B. C. (Hrsg.)*: Oxford Handbook of Philosophy of Language. Oxford University Press, New York, 2006. S. 574-606.

Chalmers, David: Frege's puzzle and the objects of credence. Unpublished manuscript, 2009a

Chalmers, David: "Propositional attitudes: A Fregean account". Noûs, 2009b, i.V.

Chierchia, Gennaro: "Anaphora and attitudes de se". In: *Bartsch, R./ van Benthem, J./van Emde Boas, P. (Hrsg.)*: Semantics and Contextual Expression, Volume 11 of Groningen-Amsterdam Studies in Semantics, Foris, 1989

Cresswell, M./ von Stechow, Arnim: "De Re belief generalized". Linguistics and Philosophy, 5 (4), 1982. S. 503-35

Cresswell, Maxwell: Entities and Indices. Kluwer, Dordrecht, 1990

Egan, Andy: Three grades of self-involvement. Unpublished manuscript

Heim, Irene: Notes on the first person. Class notes, 1991

Heim, Irene: "Features on bound pronouns". In: *Harbour, D./Adger, D./Bejar, S. (Hrsg.)*: Phi-Theory. Oxford University Press, Oxford, 2009. S. 35-56,

Hintikka, Jakkoo: Knowledge and Belief. Cornell University Press, Ithaca (NY) 1962

Kaplan, David: "Demonstratives". In: *Almog, J./Perry, J./Wettstein, H. (Hrsg.):* Themes from Kaplan, Oxford University Press, New York, 1989. S. 481-614

King, Jeffrey: "Tense, modality, and semantic values". Philosophical Perspectives 17, 2003. S. 195-245

King, Jeffrey: The Nature and Structure of Content. Oxford University Press, New York, 2007

Kratzer, Angelika: "More structural analogies between pronouns and tenses". In: *Strolovich, D./Lawson, A. (Hrsg.):* Proceedings of SALT VIII, CLC Publications, Ithaca, NY, 1998. S. 36-54

Lewis, David: “General semantics”. Synthese 22, 1970. S. 18-67
Lewis, David: “Attitudes de dicto and de se”. The Philosophical Review 88, 1979. S. 513-43
Lewis, David: “Index, context, and content”. In: *Kanger, S./Ohman, S. (Hrsg.):* Philosophy and Grammar, D. Reidel, Dordrecht, 1980. S. 79-100
Lewis, David: “What puzzling Pierre does not believe”. Australasian Journal of Philosophy, 59, 1981. S. 283-89

Pagin, Peter/Westerståhl, Dag:. “Compositionality”. In: *von Heusinger, K./ Maienborn, C./Portner, P. (Hrsg.):* Semantics. An International Handbook of Natural Language Meaning. Mouton de Gruyter, Berlin, i.V.
Partee, Barbara: “Binding implicit variables in quantified contexts”. In: *C. W. et al. (Hrsg.)*: CLS, Volume 25, Chicago, 1989. S. 342-65
Percus, Orin/ Sauerland, Uli: “On the LFs of attitude reports”. In: *Weisgerber, M. (Hrsg.):* Proceedings of SUB 7, Konstanz, 2003
Perry, John: “The problem of the essential indexical”. Noûs 13, 1979. S. 26–49
Perry, John: “Stalnaker on indexical belief”. In: *Thomson, J. J./ Byrne, A. (Hrsg.):* Content and Modality: Themes from the Philosophy of Robert Stalnaker. Oxford University Press, New York, 2006. S. 204-221

Richard, Mark: Propositional Attitudes. Cambridge University Press, New York, 1990

Salmon, Nathan/Soames, Scott (Hrsg.): Propositions and Attitudes. Oxford University Press, Oxford, 1988
Schlenker, Philippe: “A Plea for Monsters”. Linguistics and Philosophy 26, 2003. S. 29-120
Stalnaker, Robert: “Pragmatics”. In: *Stalnaker, R.:* Context and Content, Oxford University Press, New York, 1999. S. 31-46
Stalnaker, Robert: Our Knowledge of the External World. Oxford University Press, New York, 2007
Stechow, Arnim von: “Binding by verbs: Tense, person and mood under attitudes”. In: *Kadowaki, M. / Kawahara, S. (Hrsg.)*: Proceedings of NELS 33, MA: GLSA, Amherst, 2002. S. 379-403
Stechow, Arnim von/Zimmermann, Ede: “A problem for a compositional account of de re attitudes”. In: *Carlson, G./Pelletier, F. (Hrsg.):* Reference and Quantification: The Partee Effect. Stanford University Press, Stanford, 2005. S. 207-228

4 Philosophie des Geistes

First-Person Authority: The Case for the Constitutive Approach

Benedikt Kahmen
benedikt.kahmen@uni-bielefeld.de
Universität Bielefeld, Bielefeld

Abstract/Zusammenfassung:

Present-tense self-ascriptions of mental states are commonly treated remarkably unlike ascriptions of mental states to other persons. The aim of this essay is to outline the anomalous features of such self-ascriptions and to assess different explanations of them, especially of the supposed authority of the first person with regard to her own mental states. The different approaches considered are behaviour-based inference accounts as well as expressivist, dualist and functionalist theories.

I argue that accounts of first-person authority in terms of inferences to the best explanation on the basis of observation and interpretation of one's own behaviour and expressivist positions are unattractive because they cannot capture their core idea in their own terms due to lack of a privileged position for the first person to have self-referential beliefs. Dualist and functionalist theories of first-person authority are implausible, I try to show, because they are incompatible with regarding the relation between first-order mental states and self-ascriptive beliefs as reason-transferring, that is, as such that a reason for changing one's beliefs constitutes a reason for believing something else.

I contend that reason-transferability can only be maintained if questions about mental states are transparent to first-order, outward-directed questions. This transparency is achieved in the constitutive approach to first-person authority as proposed by Jane Heal and Richard Moran, which views second-order beliefs as being formed only *via* making up one's mind on first-order queries. This approach can also avoid the earlier criticisms of behaviour-based inference accounts and expressivist positions.

The constitutive approach has beliefs about one's cognitive attitudes as its paradigm case. In order to accommodate conative and phenomenal attitudes as well, I propose slight alterations to the constitutive approach that are modelled on impersonal redescriptions of first-person self-ascriptive statements suggested by Elizabeth Anscombe. With these modifications, I conclude, the constitutive approach is the most promising candidate for explaining the anomalous features of present-tense self-ascriptions of mental states.

Selbstzuschreibungen gegenwärtiger mentaler Zustände werde für gewöhnlich bemerkenswert anders behandelt als Zuschreibungen mentaler Zustände zu anderen Personen. Ziel dieses Aufsatzes ist es, die anomalen Merkmale solcher Selbstzuschreibungen zu skizzieren und verschiedene Erklärungsansätze einzuschätzen, insbesondere die der anscheinenden Autorität der ersten Person in Bezug auf ihre eigenen mentalen Zustände. Es werden behavioristische, expressivistische, dualistische und funktionalistische Positionen diskutiert.

Es wird nahe gelegt, dass der behavioristische Ansatz, dass erstpersonale Autorität durch Schlüsse auf die beste Erklärung auf der Grundlage der Beobachtung und Interpretation eige-

nen Verhaltens erklärt werden kann, sowie die expressivistische Sichtweise ihre eigene zentrale Idee nicht erfassen können. Denn hier fehlt die Grundlage für die Fähigkeit, selbstreferentielle Überzeugungen auf eine privilegierte erstpersonale Weise zu haben. Es wird versucht zu zeigen, dass dualistische und funktionalistische Theorien erstpersonaler Autorität nicht überzeugend sind, da sie nicht mit der Tatsache vereinbar sind, dass die Beziehung zwischen mentalen Zuständen erster Ordnung und Überzeugungen, in denen wir uns diese Zustände selbst zuschreiben, grund-transferierend ist. "Grund-transferierend" bedeutet, dass ein Grund, seine Überzeugungen zu ändern, auch einen Grund darstellt, von etwas Anderem überzeugt zu sein.

Es wird dafür argumentiert, dass Grund-Transferabilität nur dadurch gewährleistet werden kann, dass Fragen über mentale Zustände transparent zu außengerichteten Fragen erster Ordnung sind. Diese Transparenz wird im konstitutiven Ansatz zur Erklärung erstpersonaler Autorität, wie ihn Jane Heal und Richard Moran vorschlagen, erreicht. Nach dem konstitutiven Ansatz formen wir Überzeugungen zweiter Ordnung, indem wir uns über unsere Antworten auf Fragen erster Ordnung klar werden. Dieser Ansatz ist darüber hinaus immun gegen die Kritikpunkte, die zu Beginn an behavioristischen und expressivistischen Positionen angebracht wurden.

Der paradigmatische Fall für den konstitutiven Ansatz sind Überzeugungen über die eigenen kognitiven Einstellungen. Um auch konative und phänomenale Einstellungen zu erfassen, wird eine leichte Modifikation des Ansatzes vorgeschlagen, die Elizabeth Anscombes Modell der impersonalen Reformulierung erstpersonaler, mentale Zustände zuschreibender Aussagen folgt. Es soll gezeigt werden, dass mit dieser Modifikation der konstitutive Ansatz der überzeugendste Kandidat für eine Erklärung der anomalen Merkmale von Selbstzuschreibungen gegenwärtiger mentaler Zustände ist.

1 The Anomaly of Self-Ascriptions

First-personal, present-tense statement with which a speaker ascribes some mental state to himself are treated unlike most third-personal statements in several respects: Usually, a speaker is supposed to be authoritative about her mental states. Challenges to such statements center on her sincerity or her understanding of language. But she is usually not supposed to be capable of plain, honest error like it is possible in third-personal judgements, for example due to misleading evidence.[1]

The reason for this is that they are usually not thought of as being based on evidence and inference, neither observation and interpretation of one's own behaviour[2], nor psychological inference from others of one's mental states.[3] The alleged self-knowledge manifest in such statements is hence supposed to be immediate. Furthermore, a speaker is usually treated as omniscient about her mental states: If such a state occurs, she cannot fail to notice it.[4] As these features do

1 See (Heal 2001-2), p.1.
2 See (Moran 2001), pp.10 ff.
3 See (Fricker 1998), p.157.
4 See (Alston 1971), p.229 and (Wright 1998), p.15.

not occur in third-personal statements, they are internally related to first-personality.

Supposing it to be the usual case that someone is wrong about her mental states, serious questions arise as to the integrity of the alleged person making statements about her mental states.[5] The radical abrogation of first-person authority dissociates the current of one's mental states from one's explicit thinking, issuing in the rejection of unified consciousness or personhood.[6] Thus, the anomalies to present-tense self-ascriptions of mental states seem essential to the concept of a person. Nevertheless, it seems that our linguistic practice concerning first-person present-tense statements about mental states allows for failures in authority due to self-deception, confusion, wishful thinking and the like. If our philosophical account of first-person authority is to do justice to our linguistic practice, a weak version of the privileges of the first person in self-ascriptions of mental states is most convincing: Usually, but not in all cases, one is authoritative about one's mental states, knows immediately about them and cannot fail to notice them.

2 Attempts at Explanation

2.1 Self-ascriptions Based on One's Own Behaviour

One attempt to account for the anomaly of first-person present-tense statements suggests that self-ascription of mental states is based on inference to the best explanation of one's own behaviour and oneself has most and best evidence for making these inferences, for one has and more complete and detailed knowledge of one's own behaviour than anyone else.[7] Hence, one is authoritative on the question which mental states one is in because one is in the best position to make the inference to the best explanation of our behaviour in terms of mental states. Such an account would have to claim that we are misled in assuming that one's knowledge of one's mental states is immediate.

But being in the best position to make the inference by being necessarily present when one acts is not enough to be authoritative in ascriptions of mental states based on behavioural evidence: We need at least to pay attention to our own behaviour and to interpret it in order to arrive at plausible ascriptions of mental states. Yet a third person, for example a psychoanalyst, may be more attentive to our behaviour than we are and may more skilled at interpreting it. In such a cases, our self-ascriptions are not authoritative, even though they are

5 See (Heal 2001-2), p.2.
6 See (Moran 2001), p.123.
7 See (Wright 1998), pp.13 f.

based on an inference to the best explanation of our behaviour and we are in a principally privileged position to make these inferences.
Consequently, the correctness of self-ascriptive statements with regard to mental states is conditional upon the fact that no behavioural evidence is found that renders another explanation more plausible and that there are no mistakes in interpretation. At every point of the process of self-observation and self-interpretation, a person's actual behaviour can prove her wrong about her own mental states.

This leads to a potential dissociation of the person from her own thinking, as Moran (2001), p.123, points out: "On this view, what [a person] announces confidently as the conclusion of his thinking is one thing […] but his actual belief, as an empirical psychological matter, is another." Hence, even a person's statements with which she ascribes mental states to herself or another person can only be taken as *evidence* of the mental states she is really in. But then, the proposed account falls under its own scope: Any inference to the best explanation of someone's behaviour we make is itself primarily evidence of our mental states. Any interpretation or further inference we might give, however, can again only be taken as evidence. As every interpretation stands itself in the need of interpretation, we cannot succeed in ascribing mental states. If the proposed account cannot even intelligibly ascribe mental states, it cannot claim that behaviour indicates mental states, because there is nothing on this account that behaviour is indicative *of*.

2.2 Expressivism

Another attempt at explanation is expressivism: Self-ascriptive statements are only expressions of the first-order mental states they are about. The sincere statement "I am in pain" is a linguistically more sophisticated form of "Ouch!", as Wittgenstein (1953) argues in § 244, and expresses just the mental state of being in pain. As Fricker observes, expressivism regards self-ascriptions and first-person authority related to them as nothing more than "artefacts of grammar"[8]. We should not succumb to the philosophical need for explanations here.[9] Rather, our anomalous treatment of self-ascriptions of mental states is everyday linguistic practice and not indicative of an underlying relation between first-order mental states and second-order beliefs.[10]

Expressivism thus leads to a redescription of first-person present-tense statement about mental states: Utterances of the form "I believe that p" express, and from the first-person perspective, amount to nothing more than "p". Hence, psychological terms change their meaning and role with first- and third-person us-

8 See (Fricker 1998), p.160.
9 See (Wright 1998), pp.43 ff.
10 See (Fricker 1998), p.186.

age.[11] This implies what Moran calls the "perverse idea"[12] that everyone but oneself can talk about one's mental states. Therefore, the only route to genuinely self-directed knowledge is via third-person statements. These statements can only be based on inference on the basis of behaviour. So the prospect of self-knowledge in expressivism vanishes for the reasons counting against the behaviour-based-inference explanation of first-person authority above.

Calling the consequence of the expressivist suggestion that everyone but oneself can ascribe mental states to oneself a "perverse idea" is, however, circular as an objection: It is the whole point of expressivism that we are wedded so closely to our own thoughts that we cannot speak about them in the present tense and ascribe them to ourselves in the same way that others do. While others are speaking *about* our mental states, we are speaking from *within* them, speaking of the states we are at the same time in. This is the expressivist core idea: Our self-ascriptions of mental states are expressions of our being in the mental states we are ascribing to ourselves. Thus, they indicate that we are in these mental states. So if we are sincerely speaking about our own mental states, we cannot normally fail to be in these states. The normality condition allows for weak authority argued for in section 1.

Still, self-referential knowledge can only be obtained, by the expressivist's lights, from a third-person point of view. This point of view bases its ascriptions of mental states on behaviour. As argued in section 2.1, this method is unworkable if regarded as the only way how ascriptions of mental states are arrived at. Hence, it undermines the expressivist core notion of an *expression*: If the expressivist cannot intelligibly ascribe mental states, what could she regard an expression to be an expression *of*?

The problems of expressivism do not end here. Heal (2001) points out that the expressivist's core idea does not ensure that someone uttering what looks like a self-ascriptive statement is in the corresponding first-order mental state. For nothing in the expressivist's position precludes the possibility of false positives in someone's training to express her mental states, i.e. that she utters a self-ascriptive statement spontaneously and in good faith without being in the corresponding mental state.[13] Furthermore, expressivism as it is presented here cannot readily account for the inferential relations which run smoothly between first-person present-tense statements and non-first-person or non-present-tense ones, like between "I believe that p" and "Someone believes that p" or "I believed that p".[14] Although a more sophisticated version[15] might be able to cope with this

11 See (Heal 2001-2), p.3.
12 (Moran 2001), p.106.
13 See (Heal 2001-2), p.9.
14 See (Heal 2001-2), p.8.
15 See, for example, (Jacobsen 1996), pp.25 ff.

problem, expressivism still fails to make its own core notion intelligible and exclude the possibility of false positive under normal circumstances.

2.3 Dualism and Functionalism

It might be thought that expressivism gets into the above trouble because it cannot offer any substantial epistemic route to one's own mental states as the basis for self-ascription. Avoiding this problem, dualism might initially seem attractive: There is a special realm of entities exclusively accessible to oneself via introspection.[16] We should, of course, be aware of the potential problems of dualist metaphysics: How can private, non-physical entities stand in causal relations to physical ones?[17] Where and when does this interaction take place if essentially private entities are, as non-physical entities, not located in space and time? Furthermore, a dualist would have to cope with Wittgenstein's private language argument for the contention that essentially private entities are irrelevant for the correctness of the use of mental state language.[18]

In contrast to this, functionalist accounts maintain that first-order mental states reliably cause self-ascriptive beliefs about them. This causal relation of course only holds for the person who is in the first-order mental states concerned. Thus, her self-ascriptive (second-order) beliefs are authoritative. Functionalism hence tries to explain first-person authority without the appeal to a potentially objectionable metaphysics of essentially private entities.[19]

However, both introspective and functionalist accounts of first-person authority cannot account for certain rational relations between first-order mental states and second-order beliefs about them, as pointed out by Burge (1996) and Moran (2001). Since authority seems to be bound to these rational relations, introspective and functionalist approaches are no promising candidates for explaining this phenomenon.

The common basis of introspective and functionalist accounts is the assumption that authoritative beliefs about one's mental states reliably causally track the states they are about: This is explicit in functionalist approaches as described by Fricker: "When I am in a given first-level mental state S, it is contingently (and happily!) the case that this tends, given suitable cueing, to cause in me the belief that I am in S"[20]. It is implicit in the dualist regarding introspection as a kind of quasi-perception of essentially private entities, because in perception reliable causal relations between perceptual beliefs and their physical subject matter ensure the veracity of the former.

16 See (Heal 2001-2), p.3.
17 See (Heal 2001-2), p.4.
18 See (Wittgenstein 1953), § 270. See also (Ayer 1963), p.41.
19 See (Fricker 1998), pp.178-179.
20 (Fricker 1998), p.176.

Basing authoritative knowledge about one's own mental states only on causal tracking, however, puts a person in the role of a purely theoretical observer with regard to her first-order mental states: She is authoritative about her mental states because the latter have a certain effect on her that issues in the discovery of a fact about her mental life of which she was ignorant. But this purely theoretical attitude towards oneself threatens to undermine first-person authority: Moran (2001), p.64, notes that a person's beliefs about her mental states need to be in line with what she thinks about "outward-directed" issues, i.e. what she maintains as true, desirable, frightening, etc. for authority to hold. Otherwise it would be possible for this person to discover that she has, for example, the belief that not p, while she still thinks that p is true. This possibility is due to the fact that in trying to discover a pre-existing truth about her mental life, she conceives the answer to a psychological question about herself to be separate from the impersonal, factual question "Is p true?". She might then truthfully claim "p, but I don't believe it", which undermines her authority since her very utterance betrays a conviction contrary to her belief. We then may and usually will challenge any self-ascription of the belief that not p on which this paradoxical claim is based.

In order to keep her beliefs about her mental states and what she thinks true, desirable, frightening, etc. in line, what a person considers in order to determine whether p is true, desirable, frightening, etc. has also to be the ground of her answering psychological questions about herself. Since whatever she uses to make up her mind about p we might call her *reasons* – broadly construed – for doing so, this comes down to the requirement that the relation between first-order mental states and second-order beliefs about them has to be *reason-transferring*[21]. This does not mean that there are no causal relations between first- and second-order mental states, but it does suggest that first-person authority has to be spelled out in terms of reasons-responsiveness rather than of cause and effect.

Reason-transferingness not only works from first-order to second-order level, but also in the opposite direction, as Burge (1996), p.109, emphasizes: It is "constitutive of critical reasoning that if reasons or assumptions being reviewed are justifiably found wanting by the reviewer, it *rationally follows immediately* that there is prima facie reason for changing or supplementing them". If the relation between a person's first-order mental states and the second-order beliefs about these mental states were not reason-transferring in this way, the outcome of her assessment and reasoning about her mental states could not have any rational impact on what she believes, desires, etc. Such a person could not be regarded as assessing, revising or discarding her beliefs or desires and would be dissociated from her mental life on reflection.

21 The term is due to (Burge 1996).

This is just what would happen if the relation between second-order beliefs and first-order mental states was solely one of causal tracking: Causal tracking might go wrong under unfavourable circumstances, even if the believer meets all reasonable standards of epistemic entitlement, valid reasoning and well-functioning faculties.[22] Hence, ensuring on critical reflection that beliefs about one's mental states meet all relevant standards cannot ensure that they are true. But it is the point of first-person authority to exclude such plain, honest error: If a person meet all epistemic and rational requirements that can reasonably be imposed on her, i.e. if she is not subject to self-deception, confusion, wishful thinking and the like, her beliefs about her mental states are veridical. Reason-transferingness is thus supposed to ensure that a person is authoritative on questions concerning her mental life.

3 The Constitutive Approach

The above criticism of dualist and functional theories is based on the theoretical a person is supposed to take towards her own mental life that entails the independence of first-order mental states from second-order beliefs. It therefore seems more attractive to adopt an approach to first-person authority that takes a person to understand questions about her own psychology in a more deliberative spirit, doing justice to the reason-transferability between the first- and the second-order level.

Reason-transferability can only be ensured if, from the first-person perspective, the second-order level question "Do you believe that p?" is answered *via* making up one's mind on the question "p?". For only then reasons for coming to a certain verdict on the first, psychological issue immediately translate into reasons for coming to a corresponding verdict on the second, factual query, which is what reason-transferability demands. Thus, reason-transferability can only be maintained if questions about mental states are *transparent*[23] to first-order, "outward-directed" questions. The core claim of the constitutive approach is that self-ascriptions of mental states constitute representations of the world from the perspective of the person self-ascribing mental states. This approach thus contains a methodological element – the idea of a method how we answer psychological questions – that ensures that "the existence of a second-level belief about a first-level psychological state is itself what makes it true that the first-level state exists"[24]. For a person's belief that she believes that p normally is sufficient to make it true that she believes that p because her belief that she believes that p is formed by her determining whether or not p. Thus, her belief that p is what

22 See (Burge 1996), p.107.
23 The term is due to Moran (2001).
24 (Heal 2001-2), p.4.

normally grounds her belief that she believes that p. Thus, the constitutive approach exhibits an expressivist mode of thought in that it regards self-directed psychological statements to transport statements on a factual matter: "So it [a first-person present-tense statement] is a judgement which represents the world as being a certain way in representing the self as being a certain way."[25]

In this way, the constitutive approach can explain what is wrong about statements like Moore's paradox "The cat is on the mat but I don't believe that the cat is on the mat". The problem is that such statements sound somehow wrong, but they are not a formal contradiction, for the different parts of the sentence have a different subject matter. Transparency may provide the clue to explaining what is strange about these statements: If, from the first-person perspective, psychological statements are transparent to factual ones and therefore are also factual commitments, the latter part of the above sentence is internally related to a commitment contradictory to the former part. The constitutive relation, on the picture drawn by this approach, also has to work in the opposite direction: One's factual commitments, one's "outlook on the world", are embedded in a psychological perspective which allows the transition of transparency in the first place. Because of this embedment, one's factual claims also represent oneself a being a certain way. So *each* part of the above sentence is, from a first-person perspective, internally related to a commitment that is contradictory to the other part. However, since this contradiction only arises between the one part of the paradoxical assertion and the commitment arrived at from the other via transparency, the contradiction is fundamentally elusive and the two parts have a different subject matter – a factual and a psychological.

Despite its affinity to expressivism, the constitutive approach should not be taken to hold first-person present-tense statements to be non-assertive. Transparency does not amount to reduction, elimination or equation; it is conceived as a method of answering to questions about one's mental states by responding to reasons for answering a question about what is true or desirable. Thus, the constitutive approach can make room for genuinely self-referential second-order beliefs. After the methodological claim above, this belief claim is the second ingredient of the constitutive approach. The belief claim renders the constitutive approach immune to the above criticism of expressivism, for a person is on this account in a privileged position to acquire self-referential beliefs.

In order to provide a plausible account of first-person present-tense statements, authority in the constitutive approach has to be weak: Epistemic and/or rational failure has to be possible. Thus, self-ascription of first-order mental states as constitutive of being in the corresponding first-order states has to be surrounded by conditions pertaining to the possibilities of self-deception, confusion and the like: "[…] the occurrence of the second-level belief contributes a

25 (Heal 2001-2), p.17.

necessary element to a set of conditions which are jointly sufficient for the first-level state."[26] This avoids the further problem which might have been thought to arise in the constitutive approach that one can only be in first-order mental states if one self-ascribes these to oneself. But self-ascription, under favourable conditions, is only sufficient, and not also necessary for being in first-order mental states. Yet, because the constitutive relation works in both directions, one *can* – but need not – shift from one's perspective on the world to the psychological perspective in which the factual one is embedded.

3.1 Problems for the Constitutive Approach

One of the strengths of the constitutive approach to first-person authority is clearly in explaining this phenomenon with regard to self-ascriptions of belief. For a belief is commonly supposed to be concerned with what someone regards more or less as fact. Self-ascriptions of belief are the paradigm case for the application of the methodological element of the constitutive approach: We normally answer psychological questions by making up our minds about factual ones.

How this method is supposed to work, however, in the case of mental states not concerned with the way the world is, is not so clear. Desires do not have factual issues as their propositional content. Whether or not one desires that p cannot be answered by coming to a verdict on whether or not p. The problem is aggravated if we try to account for non-propositional attitudes like pain or pleasure in the constitutive approach, for these attitudes do not readily indicate how shift from the psychological level to any factual or phenomenological one. In these cases, it seems rather hard to apply the method of the constitutive approach. The terminology of "outward-directedness" seems to loose its sense here.

Furthermore, it might be objected that the constitutive approach renders first-order states and self-ascriptive beliefs too close: One is authoritative about one's own mental states by definition. Therefore, no genuine cognitive achievement is involved in making first-person present-tense statements, an objection discussed by Fricker (1998), pp.173-175. Finally, we might worry whether the constitutive approach avoids the earlier criticism of dualism, that it is committed to an unattractive metaphysics, by relying too heavily on unclear notions of a person's perspective and (cognitive) abilities.

The strategy for integrating conative and phenomenal attitudes is to recast them, from the first-person perspective, into impersonal statements. Following Anscombe (1981), psychological statements like "I am in pain" are translated into the first-person perspective as "It pains". Likewise, questions such as "Do you desire that p?" are answered by making up one's mind on "Is it desirable that

26 (Heal 2001-2), p.5.

p?".[27] This ensures reason-transferability between the subject matter level and the psychological level. Hence, a desirer has not only and perhaps not necessarily to think of herself as desiring, but she has to be engaged in desiring, she has to be minimally committed to what she desires, if she is rational. The metaphor with which the methodological element of the constitutive approach can be circumscribed – that a person answers questions about her mental states by "looking to the world" – should therefore not be taken too literally. Rather, the point of the methodological claim is that a person does not answer psychological questions about herself by trying to find out something about herself, but by coming to a verdict on an issue her psychological states are no part of.

The objection that there is no cognitive achievement in self-ascribing mental states can be answered by looking again at transparency and the belief claim of the constitutive approach: Self-ascriptions of mental states is based on making up one's mind on a subject matter, outward-directed issue. If coming to a verdict whether or not the world is a certain way is not a cognitive achievement, nothing is. What is more, self-ascription involves the exercise of the capacity to think of oneself as a thinker, to conceptualize and embed one's verdict on a subject matter issue in a second-order belief about one's mental states. Hence, there is a cognitive achievement in self-ascriptions both in the methodological and in the belief claim of the constitutive approach.

Furthermore, this approach is not committed to an unattractive metaphysics, relying on the mysterious capabilities of a person. It is committed to person's having a unified perspective with can be modified as suggested above and the capabilities to come to a verdict on a subject matter issue and to think of oneself as a thinker.[28] Though it is of course not entirely unproblematic, this is far from the unattractiveness of dualist metaphysics.

Most importantly, the constitutive approach avoids the above criticisms of inference-from-evidence and expressivist accounts of first-person authority: It places a person in a privileged position to form genuine self-referential beliefs about her mental states. Because of this, the constitutive approach does not face the problem of the inference-from-evidence account that every utterance and every explicit thought could only be taken as further evidence of the mental state one is in, in effect preventing one from making any inference to mental states at all. Rather, an utterance or explicit thought self-ascribing mental states is normally enough to make it true that one is in these mental states because such an utterance or thought is transparent to the mental states it is about.

Thus, it seems the most convincing explanation of the anomaly of first-person authority is that self-ascriptions of mental states are transparent to the corresponding first-order, outward-directed questions, which are one's own

27 Compare (Stampe 1987), pp.361-362.

28 See (Heal 2001-2), p.18.

business to make up one's mind about. Evidence, theory and inference are therefore under normal circumstances only needed to decide on the subject matter issue, not on the psychological one. For the psychological question is already decided from the viewpoint of the first person in deciding on the subject matter question. Likewise, the relevant conditions being fulfilled, one cannot fail to know how one has made up one's mind if one indeed has done so.

References

Alston, W.P.: "Varieties of Privileged Access". American Philosophical Quarterly, 8, 1971. S. 223-241

Anscombe, G.E.M.: The First Person. In: The Collected Philosophical Papers of G.E.M. Anscombe, Vol. 2. Blackwell, Oxford, 1981. S. 21-36

Ayer, A.J.: Can There Be A Private Language?. In: The Concept of a Person. Macmillan, London, 1963. S. 36-51

Burge, T: "Our Entitlement to Self-Knowledge". Proceedings of the Aristotelian Society, 96, 1996. S. 91-116

Fricker, E.: "Self-Knowledge: Special Access versus Artefact of Grammer – A Dichotomy Rejected". In: *Wright, C./Smith, B.C./Macdonald, C. (Eds.)*: Knowing Our Own Minds. Oxford University Press, Oxford, 1998. S. 155-205

Heal, J.: "On First-Person Authority". Proceedings of the Aristotelian Society, 102, 2001-2002. S. 1-19

Jacobsen, R.: "Wittgenstein on Self-Knowledge and Self-Expression". The Philosophical Quarterly, 46 (182), 1996. S. 12-30

Moran, R.: Authority and Estrangement. Princeton University Press, Princeton (NJ), 2001

Stampe, D.W.: "The Authority of Desire". The Philosophical Review, 96(3), 1987. S. 335-381

Wittgenstein, L.: Philosophische Untersuchungen. In: Werkausgabe, Band 1. Suhrkamp, Frankfurt/M., 1984, 1953

Wright, C.: "Self-Knowledge: The Wittgensteinian Legacy". In: *Wright, C./Smith, B.C./Macdonald, C. (Eds.)*: Knowing Our Own Minds. Oxford University Press, Oxford, 1998. S. 1-45

Mirror Neuron Systems and Understanding Mental States: an Expanded Simulationist Framework

John Michael
joal@dpu.dk
Gnosis Research Center, Aarhus University, Copenhagen

Abstract/Zusammenfassung

In this paper, I investigate the claim that the discovery of mirror neuron systems (MNSs) provides empirical support for simulation theory (ST). This idea involves two claims: (1) that MNSs are involved in understanding others' intentions or emotions; and (2) that the way in which they do so supports a simulationist viewpoint. I will be giving *qualified* support to both claims. After giving theoretical and empirical reasons to doubt the claim that the kind of simulation process instantiated by mirror neurons could be *sufficient* for understanding intentions or emotions, I will present theoretical and empirical points in support of the view that they nevertheless play a substantial role and are perhaps *necessary* although not sufficient for understanding at least some intentions or emotions. Turning to claim (2), I will sketch an expanded simulationist framework in which this role can be understood on simulationist terms. I will argue that the work on MNSs best supports a fairly weak version of ST, according to which social cognition involves simulation simply because conceptual thought in general has a simulationist component. In elucidating this idea, I appeal to Lawrence Barsalou's theory of concepts (1999, 2005). Note that the term "simulation" here refers not to simulations of a target agent's experience, nor even specifically to one's own experience in a similar counterfactual situation, but to simulations of experience in general - activating sensory, motor, proprioceptive, affective, and introspective representations that match representations one would have when perceiving, carrying out actions, experiencing emotions, etc. The appeal to empirical work on MNSs in support of ST is therefore a two-edged sword; making this appeal persuasive requires us to modify our understanding of simulation to make it line up with the empirical work.

Mein Ausgangspunkt ist die Behauptung, dass die Entdeckung von Spiegelneuronensystemen (MNSs) seinen Beleg fuer die Simulationstheorie (ST) darstellt. Die Behauptung, dass Studien zu MNSs die ST unterstuetzen, laesst sich in 2 Teilbehauptungen zergliedern: (1) dass das Spiegeln tatsaechlich etwas mit dem Verstehen der inneren Zustaende anderer Menschen zu tun hat; und (2) dass die Art und Weise, wie es damit zusammenhaengt, im Sinne der ST ist. Ich bestreite, dass MNSs fuer das Verstehen von Handlungsabsichten oder Emotionen ausreichen koennten. Aber ich ich versuche die These stark zu machen, dass sie notwendig fuer das Verstehen mancher Handlungsabsichten und Emotionen sein koennten. Da das Verstehen von Handlungsabsichten andere Ressourcen als MNSs involvieren muss, die moeglicherweise keinen simulativen Charakter haben, stellt sich die Frage, ob eine adaequate Theorie immer noch mit der ST in Einklag gebracht werden kann. Ich vertrete die These, dass man die anderen notwendigen Ressourcen sowie das Zusammenspiel der diversen Ressourcen mit einem erweiterten Simulationsbegriff integrieren kann. Diesen erweiterten Simulationsbegriff

entlehne ich aus der Begriffstheorie Lawrence Barsalous. Die wesentliche Erweiterung besteht darin, dass Simulation nicht nur als sozial-kognitiven Mechanismus angesehen wird, sondern als einen grundlegenden Mechanismus des begrifflichen Denkens ueberhaupt. Im speziellen Falle der sozialen Kognition, um den es mir hier geht, wird nicht nur die aktuelle Erfahrung eines anderen Menschen simuliert, sondern kontrafaktische Erfahrungen ueberhaupt, d.h. auch die eigenen vergangenen oder kuenftigen Erfahrungen, wodurch unter anderem motorische, sensorische, affektive, propriozeptive und introspective Repraesentationen aktiviert werden koennen. Entsprechend koennen MNSs auch dadurch einen Beitrag zum Verstehen der Handlungabsichten anderer Menschen leisten, indem sie Teile von eigenen kontrafaktischen, beispielsweise vergangenen oder kuenftigen, Handlungen, simulieren.

1. Introduction

My starting point is the idea that the discovery of mirror neurons systems (MNSs) provides empirical support for simulation theory (ST), the justification being that ST would predict that processes occurring when we are deciding upon, planning and executing actions or having emotions would also underlie our understanding of others' actions or emotions. This idea involves two claims:

(1) MNSs are involved in understanding others' intentions or emotions.
(2) The way in which they do so supports a simulationist viewpoint.

I will be giving *qualified* support to both claims. I will start with claim (1). After giving theoretical and empirical reasons to doubt the claim that the kind of simulation process instantiated by MNSs could be *sufficient* for understanding intentions or emotions, I will present theoretical and empirical points in support of the view that they nevertheless play a substantial role and are perhaps *necessary* although not sufficient for understanding at least some intentions or emotions. Turning to claim (2), I will sketch an expanded simulationist framework in which this role can be understood on simulationist terms. I will argue that the work on MNSs best supports a fairly weak version of ST, according to which social cognition involves simulation simply because conceptual thought in general has a simulationist component. In elucidating this idea, I appeal to Lawrence Barsalou's theory of concepts (1999, 2005)[1]. Note that the term "simulation" here refers not to simulations of a target agent's experience, nor even specifically to one's own experience in a similar counterfactual situation, but to simulations of experience in general - activating sensory, motor, proprioceptive, affective, and introspective representations that match representations one would have when perceiving, carrying out actions, experiencing emotions, etc. The appeal to empirical work on MNSs in support of ST is therefore a two-edged sword; making this appeal persuasive requires us to modify our understanding of simulation to make it line up with the empirical work.

1 For a philosophical cousin of Barsalou's psychological theory, see Jesse Prinz (2002).

2. Mirror neuron systems not sufficient

There are plenty of theoretical and empirical reasons to doubt that a matching relation in the sense of mirroring could suffice for understanding intentions or emotions. Let me start with two theoretical considerations:

1) The first has to do with action understanding and thus with the motor mirroring. Action understanding appears to require a more abstract kind of representation (i.e. conceptual processing) than motor representation (since one action can be carried out with different movements and different actions can be carried out with one and the same movement in different contexts). (Jacob and Jeannerod 2005). Note that it is not clear that this objection applies to understanding emotions by mirroring.
2) Understanding an intention or an emotion involves ascription of a representation of that intention or that emotion – simply mirroring (being in the same state as) as someone else does not count as understanding that that state refers to them rather than to oneself (representation is asymmetric) (Goldman 2005).

There is also plenty of empirical work that casts doubt on robust interpretations of MNs. I will just mention one widely cited study here to give you the flavor. Brass et al. (2007) used fMRI to measure brain activity in human subjects watching videos of an actor performing unusual actions, such as using her knee to turn on a light. The videos differed with respect to the ease with which the action could be interpreted. In one set, the actor had her hands full, so it was obvious why she was using her knee. In the other, her hands were free, so the use of her knees was opaque. The authors argue that the latter condition should activate any system involved in action understanding more than the former condition, since it is a more challenging case for action understanding. And it turned out not to be areas associated with MNSs but areas associated with context-sensitive inferential processes of rationalization or mentalizing that are based on the visual processing of the stimuli – STS, TPJ, aFMC (anterior fronto-median cortex) and pCC (posterior cingulate cortex). They conclude that MNSs are not substantially involved in understanding global or prior intentions.

3. Is mirroring necessary for understanding intentions and/or emotions?

As compelling as these critical considerations are, they are still compatible with the view that MNSs are crucially involved in understanding intentions and emotions, e.g. that they are necessary although not sufficient for understanding at least some intentions and/or emotions. What would motivate such a view?

Focussing on actions for the moment, one may not want to make a strict theoretical distinction between representations of prior intentions and representations of motor plans for realizing intentions. It seems more parsimonious to suppose

that the representation of a prior intention includes at least a representation of a motor plan rather than being a distinct, wholly abstract representation. So, granted that matching activation in the motor system is not sufficient for identifying prior intentions, it may still be involved in identification, i.e. in combination with other processes.

I will mention a few empirical studies that indeed suggest that this is the case. In an experiment conducted by Hamilton and Grafton (2006), MNSs were found to habituate to repeated actions but not to repeated movements. In the experiment, there were two objects and an actor that grasped the objects. In test trials, the actor reached for either the same object (repeated action) as in the previous trial, or for the other object. Sometimes the same object was in the same location and sometimes the locations were reversed, so that the reaching for the same object in the other location would count as a repeated action constituted by a novel movement, reaching for the other object but in the same location as before would count as same movement, different action, etc. The result was that areas having neural groups with mirroring properties, in particular the intraparietal sulcus (IPS), habituate to repeated actions even if those actions are performed with slightly different movements, but not to the repeated movements constituting different actions, which suggests a higher level of representation than mere representation of bodily kinematics. So it seems that the activation of MNSs does reflect differences in actions (i.e. not just movements).Why should this be the case is if they are not making any contribution to action understanding?

This is not enough, though.A skeptic might consider that MNS activation is correlated with understanding intentions or emotions because it is caused by understanding rather than causally contributing to understanding. The claim that MNSs make a causal contribution would be supported either by finding that individuals who are poor at social cognition have compromised MNSs or that people whose motor systems or emotional experience is compromised are impaired at understanding the corresponding actions or emotions of others. There is some support for both of these predictions.

There are some studies linking impaired motor skills (Parkinson's) to some impaired conceptual abilities (Boulenger et al., 2008). But there have also been a number of studies on apraxia that have established that patients who have specific motor impairments are unimpaired at recognizing pantomimes of the actions they cannot perform (Mahon and Caramazza, 2005). In short, the jury is still out on the extent to which the motor systems might be necessary (or something close to necessary) for understanding some kinds of intentions, but it seems unlikely to be sufficient.

As for emotions, the evidence is clearer. The basic emotion that has been researched most extensively is disgust, which is relatively easy to trigger in the laboratory. Wicker et al. (2003) found an overlap in activation between scenarios where subjects experienced foul odors and when they saw others sniffing the

same foul odors. That this overlap is essential for understanding that the target person is experiencing disgust is implied by the fact that two subjects who cannot experience disgust because of a lesion in IFO are also impaired in their ability to recognize disgust in others (Adolphs et al. 2003).

Pain has also been studied relatively extensively. Numerous studies have shown that the same areas are activated in third-person scenarios. For example, Singer et al. (2004) found this mirroring phenomenen when subjects were informed via a symbol on a screen that their romantic partner was receiving a painful stimulus (note that they observed neither the pain-inducing event nor the facial expression of pain.

4. MN systems and ST

4.1 Simulation Theory

The common denominator of the various versions of ST is: understanding others' actions and/or emotions involves undergoing (simulating) the same procedures that we would undergo if we ourselves were deciding upon, planning or executing an action in the same circumstances or experiencing the same emotion (Matching relation between first- and third-person scenarios). Note that there is no version of ST that merely asserts such a matching relation and leaves it at that, it is just that they differ (importantly) with respect to what they add to this. Leaving these fine points aside for a moment, proponents of ST should agree in predicting that some instantiations of the matching relation implied by the concept of simulation should be found by neuroscientists. Here is a prediction that any proponent of ST is entitled to make:

(p1): Predicting and/or understanding someone else's actions (or emotions) involves undergoing the some of the same first-order states and processes as one would undergo if one were carrying out the same action (having the same emotion, etc.).

Note that although (p1) predicts a kind of overlap of the states and/or processes involved in two different cases, namely acting or experiencing emotions, on the one hand, and interpreting others' actions or emotions on the other, it does not specifically predict an overlap of the interpreter's states and processes with those of the target. The prediction of such an interpersonal overlap - call it (p2) - is compatible with (p1) but is not entailed by it. Supplementing (p1) with the additional premise that the interpreter is relevantly similar to the target would make (p2) a plausible prediction, but adding this additional premise requires additional argumentation, and this additional argumentation would differ depending upon which version of ST one adopts.

There is also a third possibility, namely (p3): that so-called mirror neuron systems[2] could enable the interpreter to simulate the experience of performing some other action which is related to the observed action or some other emotional state which is related to the observed emotional state, for example an appropriate response. In this case, one would still be simulating an experience, and this simulation may still be used in understanding the other person's intention or emotion, but clearly the kind of simulation at issue in this sort of case is further removed from the idea that we simulate the other person's experience as they act or have an emotion.

These three alternatives are compatible. If an instance of mirroring fulfills (p2), it will also fulfill (p1) and (p3). If it fulfills (p1), it will fulfill (p3) and may fulfill (p2) but need not. If it fulfills (p3), it may fulfill (p2) but need not, and may also fulfill (p1) but need not. Moreover, different groups of neurons or neural circuits could be instantiating mirroring in different senses, either independently of each other or in a complementary fashion.

4.2 ST and MNSs

Do mirror systems instantiate simulation in the sense of (p2)?

Csibra (2008) points out that only a subset of MNs is strictly congruent. Strictly congruent MNs fire when observing or performing one and the same action (same type of grasp and same object). Many other MNs are responsive to multiple actions. They may be active during the execution of only one action but active during the observation of several actions, or active during the execution of several actions but to the observation of only one action. Beyond this, many MNs fire when one action is executed or when a functionally related action is observed. Taken together, they constitute the class of "broadly congruent" MNs. Altogether, broadly congruent MNs make up something like 60% or 70% of all MNs. The upshot of Csibra's criticism here is that only the strictly congruent MNs would actually successfully match an observed action with the activity patterns that are present when the same action is executed. If understanding an action involves (or, more robustly, just means) being in state that one is in when performing the action, then MNSs are a highly unreliable means of understanding.

Do mirror systems instantiate simulation in the sense of (p1)?

Many people, such as Pierre Jacob and Gergely Csibra (Jacob 2008, Csibra 2008), conclude that this matching business probably does not play a role in identifying or ascribing intentions (they do not say much about emotion mirror-

2 I say "so-called" mirror neuron systems because the use of the highly suggestive term "mirroring" for such cases is probably entirely unjustified, although the neural circuits in question may nevertheless be simulating.

ing) or emotions, but perhaps in predicting the ongoing motor realization of prior intentions, which are ascribed by other means. In proposing this Jacob preserves a key simulationist idea that the observer's decision-making and action-planning resources are employed when she seeks to predict the agent's behavior, and in fact MNSs are thereby granted a role in social cognition that accords with ST. But they are not playing a role in ascription of a prior intention.

Goldman (2006) suggests that retrodiction (that is ascribing an intention or an emotion on the basis of observed behavior and/or observed emotional expressions, such as facial expressions) could proceed via a "generate-and-test" method in which mirror systems play a role: A hypothesis is generated about what intention or emotion the other person may have, and then a simulation begins, whereby an action or an emotional expression is chosen that one would oneself produce if one really had the intention or emotion in question. If this action or emotional expression matches what the other person is doing, the hypothesis is confirmed and action understanding is achieved. If not, a new hypothesis is formed, etc. But where do the hypotheses come from and how are they limited? Goldman himself considers this issue unresolved, and suggests that theory-style resources are probably involved, which is not a problem for his hybrid theory. So here again, MNSs have a role in social cognition that can be understood in simulationist terms. But they are in fact involved in deriving (pretend) behavior and/or expressions on the basis of (pretend) intentions and/or emotions, and are not constitutive of intentions and/or emotions, nor or of understanding others' intentions and/or emotions.

Do mirroring systems instantiate simulation in the sense of (p3)?

Some empirical findings that suggest this. For example, Newman-Norlund et al. (2007) found that the "human mirror neuron system" (specifically: right inferior frontal gyrus and bilateral inferior parietal lobes) is more active when observers are simultaneously preparing a complementary action than when they are preparing an imitative action. They take this finding to suggest that the function of these neurons lies in "dynamically coupling action observation to action execution". Note that, if this is the case, they would not be simulating in the sense of (p1) or (p2) but they may be simulating in the sense of (p3).

If it does make sense to think of the contribution of MNSs as instantiating simulation in the sense of (p3), what do they have to do with ST in the usual sense?

5. An expanded simulationist framework

5.1 Simulationist theories of concepts

Let me start by saying a bit about simulationist theories of concepts. The basic idea is that conceptual thought, rather than taking place in an amodal symbolic code such as a "language of thought", involves the same modality-specific neural activity as perception. This can include, alongside perceptual representations, also motor representations, proprioception, and Barsalou also mentions "introspection," by which he means representation of one's own emotions and other internal states, as well as representation of one's own cognitive processes. Let me give a simple example to illustrate Barsalou's theory:

When one sees a car, neural feature detectors are active in the visual system. Conjunctive neurons in a nearby area conjoin the active features and store them in memory. These sets of conjunctive neurons also account for the trans-modal nature of concepts, namely by integrating the feature detection activity that occurred during visual perception of the car with feature detection activity that was active in other modality-specific systems, such as the auditory system. Later on, when one reasons about the car or about cars in general, the conjunctive neurons activate the neurons in the visual system and/or in other modality-specific systems that were active when the car was perceived, thereby simulating the sensory perception of the car.

5.2 What about abstract concepts?

That is all well and good for concrete concepts, but there is an obvious objection having to do with abstract concepts. Perception-based theories of concepts have always had difficulties with abstract concepts, and since I have already emphasized that the use of abstract concepts is a part of social cognition that mirror systems seem especially ill-equipped to account for, it may seem pointless to turn to perception-based theories of concepts to address skeptical concerns about the role of mirroring in social cognition. To some extent, this objection indeed marks a limit to how far one can push a robust interpretation of mirror systems. But perception-based theories like Barsalou's are not totally hopeless when it comes to abstract concepts, and it is worth checking to see whether some of the resources they have in store for abstract concepts might be appropriate in the case of social cognition, i.e. in particular for mental concepts. Just to give an idea of how Barsalou would go about accounting for abstract concepts: According to Barsalou, abstract concepts also involve concrete representations, but combinations, or sequences, of such concrete representations become more important at increasing levels of abstraction. Although I am not confident that Barsalou can explain highly abstract concepts

satisfactorily, at least some of the elements of his theory may be applied to the concepts that feature centrally in social cognition. For example, Barsalou ascribes a prominent role to introspection in abstract concepts. This is especially interesting in light of Goldman's pursuit of a theory of mental concepts in which introspection would play a key role. Let's have a look at how Barsalou thinks introspective contents may feature in conceptual processing.

5.3 Introspection

In discussing introspection, Barsalou speaks of proprioception, representation of one's emotion states, and representation of what he calls "cognitive operations", including "rehearsal, elaboration, search, retrieval, comparison, and transformation" (1999, 585). I would like to say a bit about each of these.

To start with proprioception, Barsalou and Wiemer-Hastings (2005) found that, when asked to list typical features of concepts, subjects cited more introspective contents for abstract concepts than for concrete concepts. So, for example, people would be more likely to mention hunger when listing features for FOOD than for specific kinds of food, such as CHEESE. This is intuitively plausible: where concrete perceptible features are not readily available, people tend to attend to introspectable internal states.

Turning to emotions, Barsalou gives the example of ANGER. The concept of ANGER involves the recollected introspection (simulation) of one's own past affective state(s) of anger. This recollected introspection content could be combined with other components in simulating anger. A simulation, then, may also activate perceptions of angry-behavior, and perhaps also typical cognitive operations underlying judgments typically associated with ANGER, such as the judgment that a given action has caused harm or was unfair.

As for third sort of introspection that Barsalou mentions, namely representation of one's cognitive processes, I would like to focus on what isc called metacognition. The term "metacognition" refers to cognitive *monitoring and control* of first-order cognitive processes. In other words, it refers to cognitive processes that target other cognitive processes as opposed to events or properties in the world. The targeted cognitive processes include judging the adequacy of a particular response, correcting that response, evaluating one's ability to carry out a particular task, evaluating the ease or difficulty of learning some new information or of recalling some previously learned information (Proust 2006, 18-19). Ease of learning has probably been the most extensively researched area: subjective assessments of ease of learning prior to tasks, judgment of learning during and after learning tasks, and feeling of knowing – i.e. judgment about whether a currently non-recallable item will be remembered in a subsequent test. Metacognition could thus give us access to internal cues that enable us to distinguish between degrees of certainty (i.e. knowing first-hand, knowing via some-

one else's testimony making an educated guess, just guessing, etc.) This may plausibly be considered a component of mental concepts, since it appears to be present in some species that lack ToM abilities, whereas there are no species that have ToM abilities but lack metacognitive abilities (Proust 2006).

In sum, although perceptual theories have trouble with abstract concepts, the sort of combination of diverse representations that Barsalou proposes for abstract concepts may work for some of the abstract concepts that are important for social cognition, such as intentions, emotions and various propositional attitudes.

5.4 Simulation and social cognition

How does this framework help respond to the two theoretical challenges posed in section 2.1 (context-sensitivity and ascription)? Mirroring alone does not represent actions or emotions in a way that is context-sensitive and abstract. But mirroring may still contribute substantially to understanding intentional actions and emotions. Focussing on intentional actions for a moment, activation in perceptual areas plus activation in motor areas give a representation of a movement and of a context, and this constellation taken together suffices to yield a representation of the target's intention.

If one is strongly inclined toward an embodied approach, one might say that the representation of the action/intention is constituted by this constellation of various kinds of representation. Alternatively, one could think of the tokening of a concept of an action/intention as being caused by these various kinds of representation. This latter option can be formulated as a modification of the "generate-and-test" model envisioned by Goldman, since it enables the motor representations provided by motor mirroring to actually contribute to the generation of hypotheses rather than merely testing them. If you observe someone reaching, motor neurons are active that constitute the basic concept of reaching, and you also have perceptual representations attained from the situation, as well as background knowledge about the target person. Either way, these various kinds of representation could be linked up by Hebbian learning.

An account of the role of mirror systems in social cognition must, like any simulationist approach, include a response to the second theoretical challenge directed at interpretations of mirroring systems, namely how shared representations are *ascribed* to a target person. Minimally, an account should pick out functional differences between the roles played by shared representations in first- versus third-person cases. These functional differences will of course be reflected in differences in the neural activation that occurs in tandem with the activation of the shared representations.

One possibility is to appeal to central monitoring theory (Jeannerod and Pacherie 2004) Briefly, in action performance, an efference copy of the motor com-

mand (a dynamic representation or simulation of the action) is retained in a comparator. The state of the motor system and also perceptions arising during the course of the action are compared to this model for control purposes, i.e. so that modifications can be made on the fly to the effect that the actual action matches the dynamic model. What this means with respect to third-person ascription is that you have two ingredients to action understanding: a shared representation of the action, which is the same in first-person and third-person cases, but in the first-person case you also have the feedback from an internal action model, or simulation. This latter difference could help distinguish between action performance and action understanding.

6. Conclusion

I have defended the thesis that the discovery of MN systems constitutes empirical support for ST by corroborating a prediction made by ST. This has involved affirming two claims: (1) that MN systems are involved in social cognition and (2) that they do so in a way that instantiates simulation. With respect to (1), I have argued that MN systems are likely to be substantially involved in (perhaps necessary for) understanding many intentions and emotions, although they are not likely to be sufficient.

So it is true that we need to use concepts for identifying and ascribing prior intentions and likely also emotions. But MNSs may play a role in understanding intentions insofar as they partially constitute the relevant concepts. Specifically, the incorporation of additional features outside the motor system can be accomplished by expanding the concept of simulation to include simulations of one's own past or imagined (perceptual, motor, proprioceptive, emotional, metacognitive) experiences.

As for (2), in what sense does mirroring instantiate simulation and therefore constitute empirical corroboration of a prediction made by ST? I have tried to show that the most obvious sense in which this could be the case, namely that we mirror a target person's intention or emotion by getting into the same state as they are in, is not well-founded, especially with respect to intentions (i.e. understanding actions). ST is not committed to this anyway; ST is committed to a weaker claim that we get into the same state that we would be in if we were having that intention/emotion.

Much of the work on MNSs in fact fits best with a third sense of simulation, which is broader and therefore weaker than the other two. This is the sense of simulation employed in "simulationist" theories of concepts, such as that espoused by Barsalou (1999, 2005) and, in philosophy, by Jesse Prinz (2002). According to such theories, conceptual thought in general has a simulationist component, but the term simulation here refers not to simulations of a target's expe-

rience, nor even specifically to one's own experience in a similar counterfactual situation, but to simulations of one's one past experiences in general - activating sensory, motor, proprioceptive, affective, and introspective representations that match representations one would have when perceiving, carrying out actions, experiencing emotions, etc.[3] My suggestion is the following: instances of mirroring that instantiate simulation in the sense of ST are a special case of a broader class of phenomena that instantiate simulation in the broader sense of Barsalou.

References

Adolphs R.: "Cognitive neuroscience of human social behaviour". National Review of Neuroscience, 4 (3), 2003. S. 165-178

Barsalou L. W.: "Perceptual symbol systems". Behavioral and Brain Sciences, 22, 1999. S. 577-609

Barsalou, L. W./Simmons, W. Kyle/Barbey, Aron K./Wilson, Christine D.: "Grounding conceptual knowledge in modality-specific systems". Trends in Cognitive Sciences, **7** (2), 2005. S. 84-91

Barsalou, L. W./Wiemer-Hastings, K.: "Situating abstract concepts". In: *Pecher, D./Zwaan R. (Hrsg)*: Grounding Cognition: the role of perception and action in memory, language and thought. Cambridge University Press**,** New York, 2005. S. 129-163

Boulenger, V.; Mechtouff, L.; Thobois, S.; Broussolle, E.; Jeannerod, M.; Nazir, T.A.: "Word processing in Parkinson's Disease is impaired for action verbs but not for concrete nouns". Neuropsychologia 46, 2008. S. 743–756

Csibra, G.: "Action Mirroring and action understanding: an alternative account". In: *Haggard P./Rossetti, Y./ Kawato M. (Hrsg)*: Sensorimotor Foundation of Higher Cognition: Attention and Performance, XXII. Oxford, OUP, 2008

Goldman, A.: Simulating Minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press, Oxford, 2006

Goldman, A.: "Interpretation Psychologized". In: *Stone, T./Davies, M. (Hrsg)*: Folk Psychology: The Theory of Mind Debate. Blackwell, Oxford, 1995

3 Just a side note: obviously, as simulation theorists such as Goldman and Gordon have said all along, there may also be simulation processes (i.e. high-level simulation) that do not involve mirror systems at all; I will leave that to the side.

Goldman, A.: "The Psychology of Folk Psychology". In: *Goldman, A. (Hrsg)*: Readings in Philosophy and Cognitive Science. MIT Press, Cambridge (MA), 1993. S. 347-380

Gordon, R.: "Simulation without introspection or inference from me to you". In: *Stone, T./Davies, M. (Hrsg):* Mental Simulation: Evaluations and Applications. Blackwell, Oxford, 1995

*Hamilton, A. F. d./ Grafton, S.: "*Goal Representation in Human Anterior Intraparietal Sulcus". The Journal of Neuroscience 26(4), 2006. S. 1133-1137

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

Jacob, P./Jeannerod, M.: "The motor theory of social cognition: a critique". http://jeannicod.ccsd.cnrs.fr/ijn_00000573, 2005

Jeannerod, M./Pacherie, E.: "Agency, Simulation and Self-Identification". Mind and Language, 19 (2), 2004. S. 113-146

Newman-Norlund, Roger D./van Shie, Hein T./van Zuijlen, Alexander M.J./Beckering, Harold: "The mirror system is more active during complementary compared with imitative action". Nature Neuroscience, 10(7), 2007. S. 817-818

Oberman, L./Hubbard, E./McCleery, J./Altschuler, E./Ramachandran, V./Pineda, J.: "EEG evidence for mirrow neuron dysfunction in autism spectrum disorders". Cognitive Brain Research, 24, 2005. S. 190-98

Pineda J.: "Sensorimotor Cortex as a critical component of an 'extended' mirror neuron system: Does it solve the development, correspondence, and control problems in mirroring?". Behavioral and Brain functions, 4, 47, 2008

Pineda. J./Brang, D./Hecht, E./Edwards, L./Carey, S./Bacon, M./Futagaki, C./ Suk, D./Tom, J.; Birnbaum, C./Rork, A.: "Positive behavioral feedback and electrophysiological changes following neurofeedback training in children with autism". Research in Autism Spectrum disorders, doi:10.1016/j.rasd., 2007.12.003, 2008

Prinz, J.: Furnishing the Mind: Concepts and their Perceptual Basis. MIT Press, Cambridge (MA), 2002

Proust, J.: "Rationality and metacognition in non-human animals". In: *Hurley, S./Nudds, M. (Hrsg)*: Rational Animals. Oxford University Press, Oxford 2006

Singer, T./Seymour, B./O'Doherty, J./Kaube, H./Dolan, R. J./Frith, C. D.: "Empathy for pain involves the affective but not sensory components of pain". Science, 303, doi:10.1126/science.1093535, 2004. S. 1157–1162.

Southgate, V./Hamilton, Antonia F. de C.: "Unbroken mirrors: Challenging a theory of Autism". Trends in Cognitive Sciences, 12(6), 2009. S. 225-229

*Wicker, B./Keysers, C./Plailly, J./Royet, J. P./Gallese, V./Rizzolatti, G.: "*Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust". Neuron, 40, doi:10.1016/S0896-6273(03)00679-2, 2003. S. 655–664.

Explaining the gap intuition

Bruno Mölder
bruno.moelder@ut.ee
University of Tartu, Estonia

Abstract/Zusammenfassung

An explanatory gap ensues when the truths constituting the explanans do not entail the explanandum. Attempts to give a physicalist account of consciousness seem to generate an explanatory gap, which is special in the following psychological sense. In other cases, it is possible to bridge or close the gap by regimenting or eliminating the respective concepts. In the case of consciousness, however, there is a pervasive intuition that the gap remains even when one works out a notion of consciousness that allows the entailment of truths about consciousness from physical premises. The intuition is expressed by the sense that something essential about consciousness is left out from any such conceptual regimentation.

This paper defends the view that this intuition is generated by the way our experiences are given to us. Every experience has a component that is not captured by any specification. The gap intuition is due to this ineffable component in experiences. Since conscious experiences are partly ineffable, it seems that no account of consciousness captures conscious experiences entirely and so would develop the intuition that something has been left out.

An alternative explanation to the gap intuition is provided by the "phenomenal concepts' strategy". According to this strategy, it is the peculiar nature of phenomenal concepts that feeds the intuition. The main problem with such an approach is that phenomenal concepts are more coarse-grained than the experiences themselves and thus their application still does not entail what it is like to have a particular phenomenal experience.

Eine Erklärungslücke entsteht, wenn diejenigen Wahrheiten, die das Explanans einer Erklärung konstituieren, das Explanandum nicht implizieren. Physikalische Theorien des Bewusstseins scheinen eine solche Erklärungslücke zu generieren, die im folgenden psychologischen Sinne besonders ist: während es in anderen Fällen oft möglich ist, eine Erklärungslücke dadurch zu schließen, dass relevante Begriffe präzisiert bzw. eliminiert werden, haben wir im Fall des Bewusstseins die stabile Intuition, dass die Erklärungslücke auch dann bestehen bliebe, wenn ein Bewusstseinsbegriff entwickelt wäre, der eine Deduktion der Wahrheiten über Bewusstsein aus physikalischen Prämissen erlaubte. Diese Intuition drückt sich darin aus, dass wir das Empfinden haben, eine solche Begriffsklärung müsste etwas Essentielles über Bewusstsein zwangsläufig auslassen.
In diesem Aufsatz wird die Auffassung verteidigt, dass diese Intuition durch die Weise, in der uns unsere Erfahrungen gegeben sind, generiert wird. Jede Erfahrung besitzt eine Komponente, die durch keine Spezifikation vollständig erfasst werden kann. Die Intuition vom Vorhan-

The work on this paper has been supported by the grant ETF7163 from the Estonian Science Foundation and the targeted financing scheme SF0180110s08. I am grateful to Daniel Cohnitz for helpful comments.

densein einer Erklärungslücke ist auf diese unbeschreibbare Komponente zurückzuführen. Weil bewusste Erfahrungen zum Teil unbeschreibbar sind, scheint es, dass keine Analyse phänomenalen Bewusstseins jemals bewusste Erfahrungen ganz erfassen kann, was die Intuition produziert, etwas würde durch die Analyse ausgelassen.
Eine alternative Erklärung der Erklärungslückenintuition stellt die Phänomenale-Begriffe-Strategie dar. Dieser Auffassung zufolge ist es die besondere Natur unserer phänomenalen Begriffe, die für die Intuition verantwortlich ist. Das Hauptproblem eines solchen Ansatzes besteht darin, dass selbst phänomenale Begriffe noch grobkörniger sind als die Erfahrungen selbst, weshalb auch phänomenale Begriffe nicht implizieren, wie es ist eine bestimmte Erfahrung zu haben.

1. Gaps: ontological, epistemic and explanatory

Some philosophers think that physical accounts of phenomenal consciousness are gappy. Our attention is brought to various kinds of gaps between the phenomenal and the physical, ranging from the epistemic to the ontological gap. In this paper, I approach the gap issue from a somewhat different angle. Instead of trying to find ways of closing the gap, I discuss why do we think (and cannot help thinking) that there is a gap. The main claim defended here is that our disposition to assume the existence of mental-physical gaps has roots in our psychological constitution.

Phenomenal consciousness is the kind of consciousness that as a matter of conceptual necessity involves experience that has qualitative or phenomenal aspects. There is nothing mysterious about this necessity, it is just the way the notion of phenomenal consciousness is defined. That is, when one experiences something, then usually this experience is about something, but besides this intentional aspect, it also feels in some way to have that experience. This feeling is what makes one's consciousness of something a phenomenal consciousness. Being phenomenally conscious is not just acquiring some information about the object one is conscious of, it is undergoing an experience.[1]

One way to introduce the notion of a gap between phenomenal consciousness and the physical (which includes the entities and relations postulated by all natural sciences) is to point out that for each physical basis of phenomenal consciousness proposed, one can conceive its being present without consciousness being present. This can be interpreted in the weaker and in the stronger way. In the weaker sense, this shows that even if phenomenal consciousness is physical, we do not understand how it can be physical. In the stronger sense, an ontological moral is drawn. Namely, it is concluded that physical facts do not entail facts about phenomenal consciousness. The arguments for such conclusions are very well known in the philosophy of mind (see e.g., Chalmers 2003). For example,

1 See Block (1995) for a notion of phenomenal consciousness, which he distinguishes from the notion of access consciousness.

there is the much discussed knowledge argument that Frank Jackson (1982) formulated in terms of the knowledge possessed by the neurophysiologist Mary.[2] Jackson's argument begins by introducing a neuroscientist Mary, who knows all the physical facts about colour vision. But she does not know what it is like to experience something red. Presumably then there are facts about colour vision that she does not know. When she knows all the physical facts, but still does not know all the facts about colour vision, then it is concluded that there are facts about colour vision beyond the physical facts.[3] This can be taken to show the existence of a gap between the physical facts and facts about what it is like to experience colours.

David Chalmers (2003: 108) has called the arguments like the knowledge argument and the arguments that rely on the conceivability considerations "epistemic arguments against materialism". Epistemic arguments involve postulating two kinds of gaps – the epistemic and the ontological – and the existence of the ontological gap is inferred from the presence of the epistemic gap. In Chalmers' terms, the *epistemic gap* is present when physical truths do not entail truths about phenomenal consciousness. For instance, in the case of Mary, the knowledge of all physical facts was not enough for deducing the knowledge concerning one particular fact about phenomenal consciousness. Once the epistemic gap is established, these arguments proceed to the conclusion that there is an *ontological gap*. That is, it is concluded that our inability to deduce certain kind of truths from truth of other kinds is based on the distinctness of the respective kinds of worldly facts. (Chalmers puts this in terms of the failure of necessitation: some truths about phenomenal consciousness are not necessitated by all physical truths). Recall the conclusion of (the ontological reading of) the knowledge argument: there are facts about colour vision that are not physical.

I do not think that it is right to infer the ontological distinctness from the epistemic premises. However, this is not what the present paper sets out to discuss. The aim is rather to explain why we think that there is an epistemic gap between the phenomenal and the physical.

Since the classic paper by Joseph Levine (1983), it is quite common to tackle the epistemic matters under the heading of "explanatory gap". Levine concedes that the epistemic arguments do not establish the existence of an ontological gap, but they nevertheless show that there is a kind of epistemic gap, namely an *explanatory gap*. This is to say that we lack a satisfying account of the physical basis of the phenomenal consciousness. We still do not completely understand how can the phenomenal arise from the physical or how it can be identical with

2 For predecessors and earlier variants of the knowledge argument, see Chalmers (2003: 135, n.7).

3 This is the ontological version of the knowledge argument (cf. Chalmers 2003: 107). Instead of facts, Jackson's own version was put in terms of information, which can be read either epistemically or ontologically (cf. Horgan 1984).

it. The explanatory gap is more specific than the epistemic gap. It relates to the explanatory failures, more specifically, to the failure of the reductive explanation of the phenomenal consciousness. Chalmers (2007: 169) relates the explanatory gap to the epistemic gap that he understands in terms of the entailment in the following manner: if there is a truth about phenomenal consciousness such that all the physical truths do not entail it, then we cannot reductively explain such a phenomenal truth in the physical terms. In this way, the explanatory gap results from the epistemic gap, from our failure to deduce phenomenal truths from all the physical truths. For the purposes of the following discussion, let us thus assume that there is an explanatory gap in case the truths constituting the explanans do not entail truths that constitute the explanandum and that in case of the current worry, the explanans consists of all the physical truths and, depending on the explanatory problem, the explanandum contains truths or some particular truth about the phenomenal consciousness.[4]

2. The gap intuition and ineffable experience

The explanatory gap in the case of consciousness is special in the psychological sense. Elsewhere, when explanatory gaps occur, they can be closed or bridged by adding bridge-laws or by regimenting or eliminating the respective concepts. In the case of phenomenal consciousness, however, there is a pervasive intuition that the gap remains open even when we have found such an account of consciousness that allows the entailment of truths about consciousness from physical premises. Although, given the entailment, we are not faced with a gap anymore, the intuition is expressed by the sense that something essential about phenomenal consciousness is still left out from any such account. There is an intuition that any such attempt to close the gap is unsuccessful, since the explanandum has not been specified in a satisfactory manner.

The present paper is devoted to this intuition, the *gap intuition*.[5] What needs explaining is not the emergence of a gap, but rather the intuition that there is a gap. This intuition is persistent: intuitively, the gap remains even in our best accounts of phenomenal consciousness and will remain also in any future account.

4 Explanandum is what requires explaining and the explanans is what gives the required explanation. See Hempel and Oppenheim (1948) for the classic account. However, besides the terminology, no commitments to their specific model of explanation are assumed here, unless noted otherwise.

5 For this term, see e.g., Kiverstein (2007: 336). In his version, the gap intuition is the intuition that the knowledge of one's experiences from the first-person point of view cannot be deduced from the knowledge about the brain that consists in the neuroscientific descriptions from the third-person point of view. Since the first-person / third-person distinction is not always entirely clear, I would not link it to the notion of gap intuition.

This intuition is usually conveyed, when people say that we do not understand how could consciousness be physical or think that they can conceive phenomenal consciousness as being distinct from its physical basis (according to our best theory of it).

Here is the explanation to the gap intuition that I favour. The intuition has psychological roots. By this I mean that it is generated by the way our experiences are presented to us. Namely, every experience has a component that is not uniquely captured by any conceptual specification. It may well be the case that this ineffable component is not introspectible separately, that is, in isolation of the conceptual elements in experience, but it can be singled out as a theoretical construction. Experiences appear to us as more fine-grained than the concepts that can be used to describe these experiences.

The gap intuition is due to this ineffable component in experience. As conscious experiences are partly ineffable, then while one considers the accounts of phenomenal consciousness, it seems to one that they do not capture all the aspects of conscious experience. There is something that such accounts do not convey, although one cannot put into words what exactly has been left out, for it is ineffable. But this is enough to give rise to the intuition that something has been left out.

The ineffability of experiences could be explained in different ways. I list some options below. All of them are more or less simplified versions of possible positions, also known to the unsympathetic reader as "straw men".

1) *The externalist account*: the ineffability of experience consists in the very rich informational content of the external property that is phenomenally represented. An account along these lines is proposed by Dennett (1988), who conceives ineffable experiences as our idiosyncratic dispositions to detect informationally rich properties in the environment.
2) *The nonconceptualist account*: our experiences are not encoded fully in propositional or conceptual form. Some contents of experience are non-conceptual in the strong sense that they cannot be conveyed by concepts.[6] Such contents are ineffable, since they do not have a propositional form. They also cannot be fully encoded in memory, since the scope of our memory is limited. Hence, they tend to be elusive as well.[7]
3) *The syntactical account*: the information that is represented has not been processed "semantically". It has been given purely "syntactical" processing, for example, by intervals or by other physical quantities of the stimulus. The results of such processing

6 There is another sense of nonconceptual content in which content is nonconceptual to someone if one does not possess the concepts that characterise the content (for an overview, see e.g., Crane 1992).

7 For such a position concerning memory, see Raffman (1995). See also Raffman (1993), for a computational account of the ineffability of musical knowledge.

are not connected with conceptual representations, the identification of a stimulus is accomplished by the blind matching of subpersonal representations.[8]

4) *The missing concepts' account*: the experience is ineffable, since it is an experience of something such that the description of it requires conceptual resources that one has not acquired. One's inability to describe what one experiences consequently gives rise to the feeling of ineffability, but if one could master the relevant concepts, the experience would not seem to be ineffable anymore.[9]

These possible explanations are obviously sketchy, some of them involve rather dubious assumptions and I would not subscribe to any of them at present. The point of presenting them here is to indicate that the ineffability of experiences is not itself a *sui generis* property. It is possible to give a naturalist explanation as to why experiences have such a property. Of course, such an explanation has to integrate various explanatory levels, and not remain at the personal level (as in (1), (2), (4)) or at the functional level (as in (3)). Also the implementation of the ineffability in the brain requires an explanation. In this sense, the outlined approach to the gap intuition does not itself generate a gap between the physical properties and the ineffable experiences. Physicalism is not compromised, when we assume that to have experiences that are partly ineffable is just to have certain subpersonal processes going on in one's brain. Although experiences appear ineffable at the personal level, the relevant features of the processing will be decomposed at the subpersonal level. It may well be that eventually we can explain the ineffable appearance of experience as originating from the features of the format of a mode of presentation of subpersonal processes. In such an account, having a phenomenally conscious experience would be a matter of undergoing some brain processes and the ineffability of the experience would be explained by the special format of such subpersonal processes.

I do not see reasons to think that such a subpersonal explanation is in principle impossible. Given the assumption that consciousness is basically a physical phenomenon, no additional threat of ontological gap is brought in by the ineffability of consciousness. Once we have a subpersonal account of the ineffability of experience, we can couple it with the here provided account of the psychological roots of the gap intuition. When we add all this to our best theory of phenomenal consciousness *T*, we can explain away the intuition of the epistemic gap as well. This can be done by showing why, even given *T*, the emergence of a gap intuition can be expected.

---

8 A different syntactic explanation to ineffability is provided by Jakab (2000). In his account, experiences are ineffable, since their representations are atomic, lacking a syntactic structure.

9 This account differs from (2) and (3) as it views ineffability a property of experience that it could lose. Note that it is also compatible with (1) in this sense. In (1), there is no principled reason why one's detection-dispositions could not become more and more fine-grained to the point that the ineffability disappears.

3. Phenomenal concepts

In the previous section, I attempted to provide a psychological explanation to the gap intuition that is based upon the features of phenomenal experience. A psychological approach is also adopted by those philosophers who try to explain the gap intuition in terms of the special properties of phenomenal concepts.

To appreciate the specificity of concepts by which we refer to phenomenal states, let us recall an account of reductive explanation that is available when the phenomena can be functionally analysed. Assume that we need to explain why does some object x bear a mental property *M*. Let us also assume that *M* has a functional analysis in terms of a functional or a causal role. If that is the case, then the reductive explanation, which deduces *x*'s having the property *M* at a certain time *t* from the physicalist and functionalist explanans, would take the following form (Kim 2005: 111):

(a) x has a physical property *P* at *t*
(b) *P* plays the causal role C
(c) To have *M* is defined as to have the property that plays the causal role *C*
(d) Hence, x has *M* at *t*

Now, the very problem with the phenomenal states, according to several philosophers is that they lack the functional definition as in (c). For example, Levine (1983) claims that phenomenally conscious events like pains cannot be given a reductive explanation, since our concepts of such phenomena are not functional concepts.

> "…there is more to our concept of pain than its causal role, there is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firings is *why pain should feel the way it does*!" (Levine 1983: 357)

While Levine points out in this quote that a single concept could have different aspects (the causal and the qualitative), the later discussions presume that we are dealing with different kinds of concepts. According to this proposal, we refer to phenomenal properties like the feeling of pain not by functional concepts, but by using concepts of a special kind, phenomenal concepts (see e.g., Loar 1990, Papienau 2002). Physicalists who invoke phenomenal concepts, presume that phenomenal properties are physical properties. Consequently, phenomenal concepts refer to physical properties, but those properties are given to one under the phenomenal mode of presentation.

In what follows, I will very briefly review the main theories of phenomenal concepts.

Brian Loar (1990, 1997) and Peter Carruthers (2004) conceive phenomenal concepts as *recognitional concepts*. When the phenomenal concept is applied, its referent is recognised as a token of a certain type. The concept refers directly to the property that triggers its application and the property serves as its own mode

of presentation, but the concept does not reveal the constitution of a property that triggers their application. Namely, it is not revealed that the phenomenal property is in fact a physical property. This yields one explanation to the explanatory gap. Loar (1997: 609) notes that we assume the existence of a gap, since we are in the grip of the "illusion of expected transparency: a direct grasp of a property ought to reveal how is it internally constituted, and if it is not revealed as physically constituted, then it is not so".

Another approach is taken by William Lycan (1996), who treats phenomenal concepts as *indexical concepts*. The idea is that some neural states are given indexically to the one that has them and such indexical reference cannot be deduced from the non-indexical facts. According to Lycan, one is aware of one's mental states by monitoring one's first-order states. Thereby one forms higher-order representations of those states and "no one else can use a syntactically similar representation to represent the very first-order state token (of S's own) that is the object of S's own representations" (Lycan 1996: 60). He uses this account to explain why other people cannot tell how one's feelings feel to one: namely, they cannot token the very same representations. Those representations have no semantic relations to expressions in a public language, thus no such expression can replace those indexical representations.

Phenomenal concepts have also been analysed as *quotational concepts* (Papineau 2002). According to this theory, the phenomenal state is represented as "The experience: ___ ", where the blank contains the experience itself or its recreation in imagination. The concept thus involves quoting the very state of experience. More specifically, phenomenal concepts are "formed by entering some state of perceptual classification or re-creation into the frame provided by a general experience operator 'the experience: ___ '" (Papineau 2002: 116).

According to most theories of phenomenal concepts, they are conceptually isolated from all other kinds of concepts. This is to say that they have no a priori conceptual links to other concepts. Some (e.g., Carruthers 2004: 163) use this feature of phenomenal concepts to explain why it *seems* to us that there is an explanatory gap, if the latter is understood as a failure of entailment. Physical (and other relevant) truths do not entail phenomenal truths (truths that involve phenomenal concepts) simply because phenomenal concepts resist linking them to physical and other concepts and provide no information about the physical circumstances in which the respective phenomenal property would occur. Hence, their occurrence cannot be deduced from the physical premises.[10]

Papineau (2008) does not agree with this explanation, although he too advances a psychological explanation in the sense that he traces the origins of the

10 It should be noted that Carruthers (2004) does not assume that this generates a real explanatory gap, since according to him, all facts individuated "thickly" can be deduced and "thinly" individuated facts (when there is a fact for every mode of presentation) do not need to be reductively explained.

intuition to the features of our psychological make-up. According to Papineau (2008: 59), the gap intuition is not a result of the lack of conceptual connections. He points out that this is clear from those cases where we are also unable to establish a priori entailments, but have no feeling that there must be an explanatory gap (such are the cases involving indexical reference or identity claims that involve proper names).

Papineau's own explanation to the gap intuition is that this is due to our dualistic thinking that we succumb to unintentionally. According to Papienau (2002), we have an intuition – an "intuition of distinctness" as he calls it – that the mental is something distinct from the physical, that mind and brain cannot be identical. Papienau has attributed the ground of this intuition to our different ways of thinking about conscious mental states, or more exactly, to the special feature of such thinking.[11] On the one hand, we may think about our conscious states and refer to them by physical concepts in the neural vocabulary and on the other hand, we may think about those states by applying phenomenal concepts. Since in Papineau's account, the phenomenal concepts quote the very experience (or its replica), their application certainly feels very different from the application of the physical concepts.[12] Hence, we come to think that concepts from these two different kinds cannot refer to the same object. To use the term of Papineau (1995), we commit an "antipathetic fallacy", that is, fail to identify phenomenal experiences with states of the brain.

4. The qualitative gap

Although I have some suspicions concerning the general viability of the phenomenal concepts strategy, I do not deny that we can indeed pick out our experiences by phenomenal concepts or demonstratives such as "That feel". In this section I raise one problem for attempts to explain the gap intuition by invoking phenomenal concepts. The aim is to strengthen the alternative explanation to the gap intuition that does not rely on phenomenal concepts as well as to clarify the account further.

The problem that I have in mind is the following. Even phenomenal concepts fail to capture the experience fully as they are more coarse-grained than the experiences themselves. Thus when phenomenally conscious experiences are characterised by using phenomenal concepts, they are still prone to give rise to the

11 More recently, Papineau has acknowledged that the intuition of distinctness could be also due to several factors of which the antipathetic fallacy is just one among the others and that taken individually, the explanations provided by each factor can be criticised (see Papineau 2009).

12 In the terminology he seems to favour, the phenomenal concepts also *use* the feelings that they *mention*, whereas other concepts only mention them.

intuition that something essential about the experience has been left out. Accordingly, we encounter a gap. To distinguish this from the gaps already discussed, let us call it the "qualitative gap". As with other gaps, we may initially formulate also this gap in terms of the entailment. We may say that the premises that include physical truths and truths expressed in part by phenomenal concepts still do not close the qualitative gap: they do not entail what it is like to have a particular phenomenal experience. If the phenomenal concepts' approach would be on the right track, then the gap intuition would be grounded on the way we think about the experience and this would be fully accounted for in this approach. Thus, there should be no further qualitative gap. But if there is, then phenomenal concepts do not provide the whole story.

The present approach to the gap intuition provides an easy explanation to the qualitative gap. Since what it is like to have an experience is partially ineffable, it has an ineffable component, which cannot be entailed by any set of truths. However, the qualitative gap is not problematic in the same way as other gaps and it does not lead to the epistemic or the explanatory gap, as its existence follows from an account of the phenomenal experience, which – as presumed – could be given a subpersonal explanation compatible with physicalism.

Why is not the qualitative gap problematic? We should not worry about closing the qualitative gap at it cannot be closed and does not have to be closed to give a proper account of conscious experience. Whereas the epistemic gap, which involves the failure of the entailment of one set of truths from another set of truths, can in principle be closed, the qualitative gap cannot be closed as it is a gap between concepts and experience. To my mind, this shows that it is not even right to formulate the qualitative gap in terms of the entailment of truths. An experience itself cannot be placed to the right-hand side of the entailment, for it cannot be fully expressed by any truth, since truths have conceptual form. Recall that already in the classic account of explanation by Hempel and Oppenheim (1948), the explanandum was always the sentence that describes the event, not the phenomenon or event itself (see Hempel and Oppenheim 1948: 137). This shows that there is no reason to expect from a theory of phenomenal consciousness to entail the experiences themselves.

But this does not mean that there is no such chasm between concepts and partially ineffable experience. The chasm is still there. No set of concepts, even if they involve phenomenal concepts, can yield or substitute the phenomenal experience. It is just that when we formulate the issue solely in terms of the failure of entailment, it may mislead us to think that there is a real issue to be solved, namely, that by fiddling with the premises the conclusion could somehow be derived. The qualitative gap does not need to be closed, we only need to explain why it is inevitable. This shows that phenomenal concepts are not the right remedy to the gap problem. They do not seem to yield a comparable explanation to the qualitative gap.

As noted, one main crucial difference between concepts and experiences is the ineffability of the latter. Although there are other differences as well, for example, while experiences are something that we undergo or live through, concepts are the means by which we think and refer, the ineffability is important in the present context as it explains why concepts are too coarse-grained to capture the experiences fully. And this explains why even applying phenomenal concepts does not remove the intuition that we have not succeeded to explain everything about phenomenal consciousness. It leaves a qualitative gap, which in turn does not let the gap intuition fade away.

In sum, my suggestion is that the ensuing qualitative gap is not a kind of gap that needs to be closed as it is based on a chasm between concepts and the experience. However, one's account should be able to predict it and the approaches that invoke phenomenal concepts seem to have no resources for this. The explanation provided earlier to the gap intuition can be used to predict the qualitative gap – it is the ineffability of the experience that does not allow experiences to be captured by concepts.

Somewhat similar views have been also developed by José Musacchio (2002, 2005a, 2005b). For instance, he acknowledges the ineffability of experiences and claims that experiences cannot be fully communicated through descriptions. However, in his account, the ineffability of experiences results from the difference between the experiences and their descriptions: "I propose that the phenomenal is intrinsically ineffable because experiences are physical processes that cannot be realized in other brains by the propositional description of the processes themselves." (Musacchio 2002: 342). In my approach, the explanation would go the other way: the ineffability of experiences is a part of the explanation why experiences cannot be captured by conceptual descriptions. The ineffability itself would be explained by certain specific features of the subpersonal processing of the experiences. If the matter would be solely that the processes cannot be realized by their descriptions, then the ineffability should be much more widespread. Take the case of rain for example. It is true that the theory of rain does not "realize" rain itself, but this does not mean that raining as a physical process is ineffable, in the sense that something would be left out from the description of the rain in physical terms. If Musacchio were right about the origins of ineffability, the latter should be the case as well. At this point, one might object that rain is indeed ineffable to us, for we cannot fully describe what it is like to experience the rain, but this is just an instance of the ineffability of the experiences, an account of which need not rely solely on the process-description difference.[13]

13 In addition, Musacchio subscribes to various views that I am not prepared to accept, for example that the meaning of words is grounded in phenomenal experiences or that we perceive only the internal brain nerve signals. Musacchio (2002: 361) also argues that the explanatory gap argument involves a "fallacy of equivocation that results from ignoring

I pointed out that the qualitative gap is not predicted by accounts that invoke phenomenal concepts. I should have said "by most accounts" as there is a one theory of phenomenal concepts, where arguably no such gap occurs. It is Papineau's quotational account, presented in Papineau (2002) and further modified and elaborated in Papineau (2007). I conclude this paper with a short discussion of his account, and concentrate here on the later version. Papineau (2007) takes phenomenal concepts to be perceptual concepts that use stored sensory templates to think about experiences. This is to say that phenomenal concepts use an instance of the experience itself to think about it. However, in his account, the phenomenal feature of experience do not determine the reference of phenomenal concepts. That is determined teleologically by the entity that the concept is supposed to give information about.

I have two critical remarks about this account as well as one more general consideration. Let us take the critical points first. In Papineau's theory, phenomenal concepts refer to types, i.e., to repeatable types of experience (Papienau 2007: 123). The reason for this is that token experiences are not repeatable or not persistent enough to allow applying the same concept to the same token of experience. However, referring to types presumes a rather sophisticated ability to classify token sensory templates into types and has the problematic consequence that it makes the overall account too intellectualistic. After all, phenomenal concepts are supposed to be among the most primitive kind of concepts, which application does not require elaborate mental abilities. We are supposed to share such concepts also with animals, but typing token experiences is a complex ability that presumes that one is able to keep track of one's experiences with the help of memory and is able to recognise them later. However, I do not exclude the possibility that animals could have such abilities too, so this objection is perhaps not too damaging.

But there is another worry, which is more relevant to the present concerns. To refer to token experiences, on Papineau's account, one has to use descriptions such as "the particular experience I am having now" (Papineau 2007: 123). (He also acknowledges that this is a more complex ability than the ability to refer to types.) Note that the actual token experience itself is not part of that description, the description only refers to it. This gives rise to two problems. First, the description need not pick out phenomenal properties uniquely. There is no guarantee that a single phenomenal property of the experience is picked out and not any other that it may also have (cf. Tye 2009: 47). Second, this manoeuvre leaves room for a qualitative gap, for here the application of a description still leaves out the experience. While we think about experience, we refer to it. In case of the experience-types, some version of the experience is involved via the

---

the epistemological as well as the neurobiological differences between phenomenal and propositional knowledge". This sets the issue at the level of knowledge. I criticise such a move in the main text.

application of a phenomenal concept, but in case of experience-tokens, no experience is involved.

I conclude with bringing the attention to one general feature of the accounts that attempt to explain the gap intuition. The accounts that rely on phenomenal concepts try to explain our intuitions that occur when we *think about* experiences or have *knowledge* of our own phenomenal states. The approach that I recommend, on the other hand, relies on the features of our experiences themselves, not on the features of our thoughts about those experiences. Thinking about experiences is by its nature a second-order state with respect to the experiences. Why is it better to remain at the first-order? In that way we can deal with the source of our intuition more directly. When the explanation is given at the level of our knowledge or awareness of our experiences, it presumes that the gap intuition is only due to some faults in our thinking about our own experiences. This prevents us from searching for the roots of the gap intuition among the properties of experiences. But this is where the real roots of the gap intuition reside. The intuition comes from the ineffability of the experience, which is a feature of the experience, not a feature of our thinking about the experience.

5. Conclusion

The explanation to the pervasive gap intuition should be a part of a complete account of consciousness. The latter should be coupled with a story about what generates such intuitions. I attempted to account for the gap intuition by pointing to the ineffability of our experiences. This feature gives rise to the feeling that no account of phenomenal consciousness can be complete. Alternative accounts to the intuition that invoke special features of phenomenal concepts are also on the right track, for they seek to explain the gap intuition in psychological terms. However, these accounts do not pay sufficient attention to the fact that concepts cannot substitute experiences and they are formulated at the wrong level: instead of tackling the experiences directly, they target the features of our thinking about the experiences.

References

Carruthers, Peter: "Reductive explanation and the 'explanatory gap'". Canadian Journal of Philosophy, 34, 2004. S. 153-174

Chalmers, D. J.: "Consciousness and its place in nature". In: *Stich & Warfield (Hrsg.)*: The Blackwell Guide to Philosophy of Mind. Blackwell, Oxford, 2003. S. 102-142

Chalmers, D. J.: “Phenomenal concepts and the explanatory gap”. In: *Alter, T./ Walter S. (eds.)*: Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford University Press, Oxford, 2007. S. 167-194

Crane, T.: “The nonconceptual content of experience”. In: Crane, T. (Hrsg.): The Contents of Experience. Essays on Perception. Cambridge University Press, Cambridge, 1992. S. 136-157

Dennett, D.: “Quining qualia.” In: *Marcel, A./Bisiach, E. (Hrsg.):* Consciousness in Contemporary Science. Oxford University Press, Oxford, 1988. S. 42-77

Hempel, C./Oppenheim, P.: “Studies in the logic of explanation”. Philosophy of Science, 15, 1948. S. 135-175

Horgan, T.: “Jackson on physical information and qualia”. Philosophical Quarterly, 32, 1984. S. 127-136

Jackson, F.: “Epiphenomenal qualia”. Philosophical Quarterly, 32, 1982. S. 127–136

Jakab, Z.: “Ineffability of qualia”. Consciousness and Cognition, 9 (3), 2000. S. 329-351

Kim, J.: Physicalism, or Something Near Enough. Princeton University Press, Princeton, 2005

Kiverstein, J.: “Consciousness, the minimal self and brain”. Synthesis Philosophica, 44, 2007. S. 335-360

Levine, J.: “Materialism and qualia: the explanatory gap”. Pacific Philosophical Quarterly, 64, 1983. S. 354–361

Loar, B.: “Phenomenal states.” In: *Tomberlin, J. (Hrsg.)*: Philosophical Perspectives IV: Action Theory and Philosophy of Mind. Ridgeview, Atascadero, 1990. S. 81-108

Loar, B.: “Phenomenal states (Second version)”. In: *Block, N./ Flanagan O./Güzeldere G. (Hrsg.):* The Nature of Consciousness: Philosophical Debates. Oxford University Press, Oxford, 1997. S. 597-616

Lycan , W.: Consciousness and Experience. M.I.T. Press, Cambridge, MA, 1996

Musacchio, J. M.: “Dissolving the explanatory gap”. Brain and Mind, 3, 2002. S. 331-365

Musacchio, J. M.: “The ineffability of qualia and the word-anchoring problem” . Language Sciences, 27, 2005a. S. 403-435

Musacchio, J. M.: "Why do qualia and the mind seem nonphysical?" Synthese, 147, 2005b. S. 425-460

Papineau, D.: "The antipathetic fallacy and the boundaries of consciousness". In: *Metzinger, T. (Hrsg.):* Conscious Experience. Ferdinand Schöningh, Paderborn/Imprint Academic, Exeter, 1995. S. 259-270

Papineau, D.: Thinking About Consciousness. Clarendon Press, Oxford, 2002

Papineau, D.: "Phenomenal and perceptual concepts." In: *Alter, T./Walter, S.* (Hrsg.): Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford University Press, Oxford, 2007. S. 111-144

Papienau, D.: "Explanatory gaps and dualist intuitions." In: *Weiskrantz, L./Davies M. (Hrsg.):* Frontiers of Consciousness. Oxford University Press, Oxford, 2008. S. 55-68

Papineau, D.: "What is the explanatory gap?". A keynote lecture at ASSC 13, Berlin, 08.06.2009: http://www.kcl.ac.uk/content/1/c6/06/13/82/GapBerlin.ppt

Raffman, D.: Language, Music, and Mind. MIT Press, Cambridge (MA), 1993

Raffman, D.: "On the persistence of phenomenology" In: *Metzinger, T. (Hrsg.):* Conscious Experience. Ferdinand Schöningh, Paderborn/Imprint Academic, Exeter, 1995. S. 293-308

Tye, M.: Consciousness Revisited. Materialism without Phenomenal Concepts. MIT Press, Cambridge (MA), 2009

Warum wir in einem grundlegenden Sinn biologische Lebewesen und nicht Personen sind

Gerson Reuter
g.reuter@em.uni-frankfurt.de
Institut für Philosophie, J.W. Goethe-Universität, Frankfurt/Main

Abstract/Zusammenfassung

Who or what is the subject of mental episodes? For example, who or what is it that feels the warming sun on a mild winter morning dwelling on the thought that spring is in the air? It will be argued that every time we think or feel something it is the biological (human) organism which thinks and feels. For each of us is *numerically identical* with an biological (human) organism. Moreover, each of us is *essentially* such an organism. These claims, characterizing a position known as *Animalism*, have rather view adherents. However, these claims can be accounted for on the basis of two assumptions which are far less contentious than *Animalism* itself. The first assumption says that in the course of interpreting oneself and other people we ascribe mental and physical (non-mental) properties to *one and the same particular*. The second assumption brings the *concept of person* into play, roughly in the tradition of Peter Strawson. It states that it is persons to which we ascribe mental as well as non-mental properties. But this second assumption is at best an intermediate step in answering the question of which fundamental or 'essential' kind we are. For the assumption leaves the problem of our diachronic identity untouched. Therefore, the two assumptions will be placed in the context of the so-called *problem of personal identity*. For this purpose, I will discuss an (fictional) exemplary case of someone who at a particular time is a person but subsequently loses the status of a person. The central thrust of the paper consists in showing that *if* one is bent on sticking to the two assumptions one will be almost inevitably pushed in the direction of *Animalism*. To be sure, we, at least many of us, are persons. But, strictly speaking, the subjects of thoughts and experiences are, at least in our case, biological organisms. For each of us is identical with an biological organism and also is essentially such an organism.

Wer oder was ist das Subjekt von mentalen Episoden, wenn beispielsweise jemand die wärmenden Strahlen der Mittagssonne spürt und dabei dem Gedanken nachhängt, dass sich nun langsam der Frühling ankündigt? Plädiert werden soll für die Behauptung, dass dann, wenn wir denken oder etwas empfinden, es der jeweilige biologische Organismus ist, der denkt oder Empfindungen hat. Denn wir sind jeweils *numerisch identisch* mit einem biologischen (menschlichen) Organismus. Und wir sind auch jeweils *wesentlich* ein solcher Organismus. Diese Kernthesen des sogenannten *Animalismus* stoßen bei vielen auf Ablehnung. Es kann jedoch gezeigt werden, dass sie sich vor dem Hintergrund zweier Annahmen motivieren lassen, die weit weniger kontrovers sind als der *Animalismus* selbst. Die erste der beiden Hintergrundannahmen besagt, dass wir in Selbst- und Fremdinterpretationen jeweils *ein und demselben Gegenstand* sowohl mentale als auch körperliche (nicht-mentale) Eigenschaften zuschreiben. Die zweite Annahme bringt – in der Tradition Peter Strawsons

– den *Personenbegriff* ins Spiel und behauptet, dass es Personen sind, denen wir sowohl mentale als auch körperliche Eigenschaften zuschreiben. Diese zweite Annahme ist jedoch allenfalls ein Zwischenschritt in der Beantwortung der Frage, zu welcher grundlegenden Art von Einzeldingen wir gehören. Denn diese Annahme lässt offen, welche transtemporalen Identitätsbedingungen wir haben. Aus diesem Grund werden in einem nächsten Schritt die beiden Hintergrundannahmen im Kontext eines Beispielsfalls diskutiert, in dem jemand zu einem früheren Zeitpunkt eine Person ist, diesen Personenstatus jedoch später verliert. Das zentrale Ziel dieser Überlegungen besteht in dem Nachweis, dass dann, *wenn* man an den beiden Hintergrundannahmen festhalten möchte, sich die ontologische Auskunft des Animalismus aufdrängt: Zwar sind wir – oder zumindest viele von uns – auch Personen, aber das Subjekt von Gedanken und Empfindungen ist streng genommen der biologische Organismus, der wir jeweils wesentlich sind.

Wer oder was denkt und empfindet, wenn ich einem tröstlichen Gedanken nachhänge, mich Tagträumen hingebe oder auch einen bevorstehenden Zahnarztbesuch zu verdrängen versuche? Genereller – und etwas technischer – ausgedrückt: Wer oder was ist das Subjekt von mentalen Episoden? Natürlich sind jeweils *wir* es, die etwas denken oder empfinden. Aber über welche *grundlegende Art* von Einzeldingen reden wir, wenn wir uns und andere interpretieren? Etwas anders gefragt: Was sind wir *wesentlich*?

Auf diese Frage bieten sich zahlreiche Antwortkandidaten an. Wir könnten beispielsweise grundlegend immaterielle Einzeldinge sein, Ansammlungen von physikalischen Kleinstpartikeln oder auch Gehirne. All diese Kandidaten haben auch Fürsprecher. Und wollte man der Frage, was wir grundlegend sind, angemessen nachgehen, müssten all diese Vorschläge natürlich berücksichtigt und ausgiebig diskutiert werden. Konzentrieren werde ich mich im Folgenden jedoch auf lediglich zwei Kandidaten: darauf, dass wir wesentlich *Personen* oder wesentlich *biologische Lebewesen (Organismen) der Spezies Homo sapiens* sein könnten.[1] Damit wird der Diskussionsrahmen ziemlich eng abgesteckt. Aber immerhin dürften es zwei der aussichtsreichsten Kandidaten dafür sein, was wir grundlegend sind.

Diese Einschätzung basiert zugegebenermaßen auf Vorannahmen, die ihrerseits begründungsbedürftig sind. Wäre man primär daran interessiert, unsere Welt möglichst durchweg *physikalistisch* zu beschreiben und zu erklären, würde man vermutlich zuallererst ausloten, wie weit man mit der Annahme kommt, wir seien Ansammlungen von physikalischen Kleinstpartikeln. Eine Auseinandersetzung mit dieser Annahme kann sich nicht damit bescheiden zu vermerken, man teile dieses Interesse an einem physikalistischen Weltbild nicht. Denn welche Antwort auf die Frage, was wir wesentlich sind, richtig ist, hängt gewiss nicht von unseren Interessen ab! Ich kann hier jedoch nicht mehr tun, als die

1 Hier und im Folgenden verwende „biologisches Lebewesen“ und „lebendiger biologischer Organismus“ als gleichbedeutende Ausdrücke.

Überzeugung zu artikulieren, dass unsere *alltägliche Urteils- und Interpretationspraxis* der Startpunkt und die Korrekturfolie für theoretische Vorschläge sein sollte.[2] Und wählt man diesen Startpunkt, stößt man sofort auf die beiden Antwortkandidaten, dass wir in einem grundlegenden Sinn Personen oder auch biologische (menschliche) Organismen sein könnten. Denn zu den integralen Bestandteilen unseres *Selbstverständnisses* gehört zweifellos, dass wir uns als Personen und auch als biologische Lebewesen auffassen. Und beide Selbstbeschreibungen nehmen wir mit einer Selbstverständlichkeit vor, dass womöglich die Annahme verwundert, es handele sich überhaupt um *konkurrierende* Antwortkandidaten.

Der vorliegende Text ist ein Plädoyer für die Behauptung, dass wir im grundlegenden Sinn lebendige biologische (menschliche) Organismen sind. Diese Behauptung mag auf den ersten Blick kaum mehr als eine Trivialität artikulieren. Denn, wie gerade gesagt, begreifen wir uns ja (auch) als biologische Lebewesen! Gemeint ist aber letztlich, und der Anschein von Trivialität dürfte sich dann auch verflüchtigen, dass jeder und jede von uns jeweils *numerisch identisch* mit einem biologischen (menschlichen) Organismus ist. Wir ‚haben' also nicht lediglich einen biologischen Organismus, wir *sind* – im Sinne strikter numerischer Identität – ein solcher Organismus. Und wir sind auch insofern grundlegend biologische (menschliche) Organismen, als wir jeweils *wesentlich* ein Organismus der Art *Homo sapiens* sind – und *nicht* etwa wesentlich eine Person.[3]

Diese Thesen gehören zum Kern einer Position, die mittlerweile als *Animalismus* bekannt geworden ist.[4] Vertraut ist diese Position hauptsächlich aus den Debatten rund um das Problem der personalen Identität. In diesen Debatten hat der Animalismus nicht gerade viele Anhänger. Diese Unbeliebtheit hat verschiedene Gründe. Ein wichtiger Grund liegt darin, dass der Animalismus sich mit intuitiven Einschätzungen unserer *transtemporalen Identitätsbedingungen* beißt – also, sehr grob gesagt, Einschätzungen dazu, welche Veränderungen wir überleben können und welche nicht.[5] Verdeutlichen lässt sich das an einem der beliebtesten Gedankenexperimente in der Diskussion über personale Identität –

2 Das schließt natürlich nicht aus, dass eine ausgearbeitete philosophische Theorie auch einige ‚revisionistische Elemente' enthält.

3 Für ein Verständnis dieser These ist klarerweise entscheidend, was mit dem Ausdruck „wesentlich" gemeint ist. Darauf werde ich in Kürze näher eingehen.

4 Siehe insbesondere Olson (1997), Olson (2007), Snowdon (1990), Snowdon (2003) und van Inwagen (1990).

5 Ich mache hier keinen Unterschied zwischen der Rede von *transtemporalen Identitätsbedingungen* und der von *Überlebensbedingungen*. Unterstellt ist also, dass ein beliebiger Mensch m genau dann bestimmte Veränderungen überlebt und zu einem Zeitpunkt t existiert, wenn es ein Einzelding x zu t gibt und wahr ist, dass m = x. (Den bekanntesten Versuch, einen hinreichend substantiellen Begriff des Überlebens vom Begriff der transtemporalen (numerischen) Identität zu entkoppeln, hat Derek Parfit vorgelegt (siehe Parfit (1971).)

dem der *Gehirntransplantation*. Aus der Perspektive der ersten Person erzählt, geht dieses Gedankenexperiment folgendermaßen: Während mein Großhirn noch gesund ist, sind weite Teile meines restlichen Körpers unheilbar krank. Meine Ärzte schlagen mir vor, mein Großhirn – den Teil, der für mein mentales Leben entscheidend ist – von meinem Körper zu trennen und in einen anderen Körper zu verpflanzen. Nach der Operation soll mein alter, kranker Körper zerstört werden. Die aus der Operation resultierende Person wird also mein Großhirn haben. Und weil mein Großhirn nomologisch hinreichend für mein ‚mentales Leben' ist, wird diese Person die mentalen Eigenschaften aufweisen, die ich vor der Operation habe. (Es hat demnach den Anschein, als werde diese Person *meine* Erinnerungen, Überzeugungen, Wünsche etc. haben.)

Es scheint so zu sein, dass die meisten Leute, ob nun bereits philosophisch vorgebildet oder nicht, auf Nachfrage ‚spontan' zu dem Urteil neigen, sie würden ein derartiges Prozedere überleben.[6] Gemäß dem Animalismus würden wir aber gerade nicht überleben – zumindest dann, wenn der Organismus, dem das Gehirn entnommen wurde, diese Operation tatsächlich nicht überlebt. (Und die vorgestellte Version des Gedankenexperiments sieht ja vor, dass ‚der Körper', dem man das Gehirn entnimmt, zerstört wird.) Denn naheliegenderweise würde durch die skizzierte Operation kein biologischer Organismus – kein biologisches Lebewesen – verpflanzt, sondern nur ein *Organ*, wenn auch zugegebenermaßen ein nicht gerade unwichtiges. Und somit würden *wir* nicht durch eine solche Operation verpflanzt.

Aber immerhin: Auch gemäß dem Animalismus gibt es eine Chance, eine derartige Operation zu überleben. Dazu müsste man aber eine Variante des Gedankenexperiments erzählen, in der der ‚alte Körper' gerade nicht zerstört würde, vielmehr alle lebensnotwendigen biologischen Funktionen des Organismus intakt blieben.[7] In einem solchen Fall würde der biologische Organismus überleben und somit auch wir. (Allerdings stünden wir dann traurigerweise ohne Großhirn da, was natürlich keine wirklich tröstliche Aussicht ist.)

Genau diese Kommentare zum Gedankenexperiment der Gehirntransplantation halten viele für extrem kontraintuitiv. Was folgt daraus? Es ist sicherlich strittig, welches theoretische Gewicht ein Abfragen von Intuitionen in Reaktion auf Gedankenexperimente haben kann. Ganz und gar nicht klar ist also, was man aus theoretischer Sicht mit solchen ‚Befunden' anstellen sollte. Allerdings: Müsste man nicht gerade dann, wenn unsere alltägliche Urteils- und Interpretationspraxis Ausgangspunkt für theoretische Vorschläge sein soll, mit derartigen intuitiven Einschätzungen zu theoretisieren beginnen?

Ich denke, solche intuitive Einschätzungen artikulieren nur in Ausnahmefällen *tief verankerte Überzeugungen*, die eine bestimmte philosophische Deutung

6 Das entspricht auch der Meinungsverteilung in den einschlägigen philosophischen Texten.

7 Gewährleisten ließe sich das vermutlich durch einen weiterhin funktionierenden Hirnstamm. (Aber das ist natürlich eine empirisch zu entscheidende Frage.)

erzwängen. ‚Hinter' diesen Intuitionen stehen wohl selten fest umrissene Überzeugungen darüber, was wir sind, geschweige denn Überzeugungen, die dann auch noch in ein umfassenderes Bild der Welt – von dem, ‚was es gibt' – kohärent integriert wären. Auch jemand, für den es eine Selbstverständlichkeit darstellt, uns Menschen als eine Art von Tieren zu betrachten und uns Seite an Seite mit anderen Tieren in der Geschichte der Evolution zu verorten, könnte glauben, dass er sich immer dort aufhält, wo sich sein Großhirn befindet. Vielleicht hat er noch nie darüber nachgedacht, wie seltsam die Annahme ist, Tiere könnten mit einem bestimmten Organ ‚mitwandern'! Intuitive Einschätzungen von Gedankenexperimenten scheinen demnach nicht arg belastbar zu sein. Isoliert betrachtet, können sie ganz sicher keine *Begründungsfunktion* übernehmen.[8]

Und es gibt auch Gründe dafür, intuitiven Widerständen gegen den Animalismus zu misstrauen. Im Folgenden möchte ich einen Überlegungsstrang vorstellen, der vor allem zeigen soll, dass sich die Kernthese des Animalismus mit Hilfe zweier ziemlich unkontroverser Annahmen, die in unserer alltäglichen Interpretations- und Urteilspraxis verankert sind, zumindest gut motivieren lässt. Im Zuge dieses Motivierungsversuchs sollen keine sonderlich ‚theorielastigen' Argumente bemüht werden. Zwar wird der Überlegungsgang nicht wirklich ohne die eine oder andere theoretische Voraussetzung auskommen. Aber gezeigt werden soll, wie *nahtlos* im Grunde die Kernthese des Animalismus an zentrale und recht allgemeine Merkmale unserer Urteils- und Interpretationspraxis anschließt.[9]

Der Überlegungsgang verfolgt noch ein weiteres, wenn auch eher untergeordnetes Ziel: Wie noch deutlich werden wird, artikulieren die Annahmen, die das Argument in Gang bringen sollen, nicht nur zentrale Aspekte unserer alltäglichen Urteilspraxis, sondern auch den Kern des Personenbegriffs Peter Strawsons.[10] Vielleicht ist es übertrieben zu behaupten, dieser Personenbegriff sei der

---

8 Trotzdem sollte natürlich eine ausgearbeitete Theorie personaler Identität, sofern sie derartige Intuitionen nicht ‚bestätigen' kann, zumindest eine Diagnose liefern können, die aufzeigt, warum diese Intuitionen derart weit verbreitet sind.

9 Einige Voraussetzungen werde ich im weiteren Verlauf meiner Überlegungen explizit machen und kurz zu plausibilisieren versuchen. Auf andere Voraussetzungen werde ich hingegen nicht weiter eingehen. Drei davon sollten hier aber zumindest genannt werden: Erstens unterstelle ich die Angemessenheit des sogenannten *Drei-Dimensionalismus*: Wir (und auch viele andere Einzeldinge) haben also keine zeitlichen Teile, sondern sind persistierende Einzeldinge. (Zu einer Verteidigung dieser Annahme siehe beispielsweise Olson (2006).) Zwei weitere Voraussetzungen betreffen das Verständnis des Identitätsbegriffs: Ich gehe davon aus, dass dann, wenn zwei vorgeblich verschiedene Einzeldinge tatsächlich numerisch identisch sind, sie *notwendigerweise* identisch sind. (Vgl. dazu z.B. die Diskussion in Kripke (1971).) Ferner ist hier unterstellt, dass der Begriff der *relativen Identität* unplausibel ist. (Eine überzeugende Kritik dieser von Geach (in Geach (1967)) eingeführten Idee findet sich in McGinn (2000).)

10 Siehe Strawson (1959).

Standardbegriff in philosophischen Debatten. Aber er ist sehr einflussreich. Und vor allem arbeiten mit diesem Begriff viele Autoren, die nicht gerade in Verdacht stehen, Anhänger des Animalismus zu sein.[11] Es wäre also ein interessanter Nebeneffekt der Überlegungen, ließe sich zeigen, dass man gerade auf der Basis zentraler Aspekte dieses Personenbegriffs für die Kernthesen des *Animalismus* argumentieren kann.

Welche zwei Annahmen sollen den Motivierungsversuch nun in Gang bringen? Die erste – und letztlich entscheidende – der beiden Annahmen ist in Schlussfolgerungen verankert, die wir alltäglich zuhauf vollziehen und für zweifelsfrei gültig halten. Wenn ich beispielsweise mitbekomme, dass meine Tochter Olivia fröhlich ist, ich ferner sehe, dass Olivia gerötete Wangen hat, dann darf ich trivialerweise schließen, dass es *Olivia* ist, die *sowohl* fröhlich ist *als auch* gerötete Wangen hat. Zuschreibungen von *mentalen* und *körperlichen (nicht-mentalen)* Eigenschaften sind häufig auf diese oder ähnliche Weise inferentiell vernetzt. Dabei unterstellen wir, dass die verschiedenen Vorkommnisse des betreffenden Eigennamens immer auf *ein und dasselbe* Einzelding Bezug nehmen. (In meinem Beispiel sind es die Verwendungen des Eigennamens *Olivia*.) Gleiches gilt für *Selbstzuschreibungen* mit Hilfe des Ausdrucks „ich“. Wir schreiben in Selbst- und Fremdzuschreibungen jeweils ein und demselben Einzelding – eben jeweils uns – sowohl mentale als auch körperliche Eigenschaften zu. Und wir glauben auch, dass wir damit richtig liegen. Die erste Annahme lautet demnach:

(1) Wenn eine Selbst- oder Fremdzuschreibung von mentalen und körperlichen (nicht-mentalen) Eigenschaften wahr ist, dann kommen jeweils *ein und demselben* Einzelding *sowohl* die zugeschriebenen mentalen *als auch* die zugeschriebenen körperlichen (nicht-mentalen) Eigenschaften zu.

Natürlich kann man theoretische Gründe gegen diese Annahme – oder zumindest gegen eine *buchstäbliche Lesart* dieser Annahme – mobilisieren. So gibt es Autoren, die meinen, uns kämen mentale und nicht-mentale Eigenschaften auf *unterschiedliche Weise* zu.[12] Die Idee lautet, grob gesagt, dass uns, die wir Personen sind, mentale Eigenschaften in einer *direkten* Weise zukommen, nicht-mentale Eigenschaften hingegen nur indirekt. Und zwar kommen sie uns indirekt dadurch zu, dass wir einen Körper *haben*, dem diese Eigenschaften in einer direkten Weise zukommen.[13] Ich unterstelle einmal, dass wir in unserer alltäglichen Urteilspraxis einen solchen Unterschied *nicht* machen. Mit einer buchstäb-

11 Siehe z.B. die Verwendung des Personenbegriffs in Tugendhat (1979), Spitzley (2000) und Hacker (2002).

12 Siehe z.B. Baker (2000).

13 Lynne Baker formuliert diesen Unterschied in einer anderen Terminologie; aber das ist hier unerheblich. Auf die Idee, dass es zwei Weisen des ‚Habens von Eigenschaften' gibt, werde ich am Ende des Textes noch kurz eingehen.

lichen Lesart der Annahme (1) ist dementsprechend gemeint, dass uns mentale und nicht-mentale Eigenschaften *in gleicher Weise zukommen* – gleich ‚direkt', wenn man so will.

Annahme (1) lässt noch offen, *von welcher Art* das Einzelding ist, auf das wir mit unterschiedlichen sprachlichen Mitteln (z.B. mit den Ausdrücken „ich" oder „du") Bezug nehmen und dem wir Eigenschaften zuschreiben. Natürlich: Wenn ich mich beispielsweise (wahrheitsgemäß) daran erinnere, gestern traurig gewesen zu sein, dann bin *ich* es, der traurig war. Damit ist aber noch nichts darüber gesagt, welche Art von Einzelding ich bin. (‚Iche' bilden gewiss keine Art von Einzeldingen!) Und an genau diesem Punkt kommt die zweite Annahme ins Spiel. Auch sie knüpft an unsere alltäglichen Interpretations- und Urteilspraktiken an. Und darüber hinaus hat sie einflussreiche philosophische Theorien im Rücken – insbesondere in der Tradition von Peter Strawson. Sie lautet:

(2) Wir sind *Personen*, und es zeichnet Personen gerade aus, dass ihnen *sowohl* mentale *als auch* körperliche (nicht-mentale) Eigenschaften zukommen.[14]

Was soll hier mit dem Personenbegriff genauer gemeint sein – abgesehen davon, dass Personen solche Einzeldinge sind, die körperliche und mentale Eigenschaften haben?[15] Es kursieren etliche, in ihren Details voneinander abweichende Erläuterungen dieses Begriffs. Klärungsbedürftig ist vor allem die Frage, *welche* mentalen Eigenschaften (oder Fähigkeiten) für den Personenstatus entscheidend sind. Einsetzen könnte man beispielsweise den wohl prominentesten Vorschlag, dass sich Personen durch den Besitz von *Selbstbewusstsein* auszeichnen (der Fähigkeit, selbstbewusste Gedanken zu denken).[16] Und ein derarti-

14 Strawson schreibt: „What I mean by the concept of a person is the concept of a type of entity such that both predicates ascribing states of consciousness and predicates ascribing corporeal characteristics, a physical situation &c. are equally applicable to a single individual of that single type." (Strawson (1959), S.101f.)

15 Strawson hat sicherlich nicht geglaubt, mit seinen Erläuterungen eine befriedigende *Definition* des Personenbegriffs vorgelegt zu haben.

16 Vertreter eines solchen Personenverständnisses berufen sich meist auf John Locke. Die einschlägige Textstelle bei Locke lautet: „Meiner Meinung nach bezeichnet dieses Wort [*Person*] ein denkendes, verständiges Wesen, das Vernunft und Überlegung besitzt und sich selbst als sich selbst betrachten kann. Das heißt, es erfaßt sich als dasselbe Ding, das zu verschiedenen Zeiten und an verschiedenen Orten denkt. Das geschieht lediglich durch das Bewußtsein, das vom Denken untrennbar ist und, wie mir scheint, zu dessen Wesen gehört. (Locke (1690/1988) Bd.1, S.419f.) Einen solchen Personenbegriff kann man auch den Überlegungen Strawsons entnehmen. Denn damit einem Wesen x sogenannte P-Prädikate (also letztlich mentale Eigenschaften) wahrheitsgemäß zugeschrieben werden können, muss x laut Strawson in der Lage sein, sie sich selbst (und anderen) zuzuschreiben. Und diese Fähigkeit setzt sicherlich *Selbstbewusstsein* voraus (oder ist sogar mehr oder weniger die gleiche Fähigkeit wie die Fähigkeit, selbstbewusste Gedanken zu denken). (Siehe Strawson (1959), Kap. 3.)

ges (oder verwandtes) Verständnis des Personenbegriffs soll an dieser Stelle auch vorausgesetzt werden.[17]

Ist nun mit Annahme (2) eine Beantwortung der Ausgangsfrage vorentschieden? Mitnichten. Auch diese zweite Annahme ist allenfalls ein *Zwischenschritt* in einer Analyse, die zeigt, was wir grundlegend sind. Denn (2) sagt nichts darüber aus, ob wir, die wir Personen sind, *wesentlich* Personen sind.

Der Ausdruck „wesentlich" ist nun bereits mehrfach gefallen. Es ist also an der Zeit, etwas über seine Verwendung zu sagen und damit auch einige Voraussetzungen explizit zu machen. Letztlich handelt man sich mit diesem Ausdruck (und verwandten Ausdrücken) gewiss jede Menge metaphysischen Ballast – und damit natürlich theoretische Probleme – ein. Einführen lässt sich die Rede davon, wesentlich ein Einzelding einer bestimmten Art zu sein, aber auf ziemlich harmlose und intuitiv recht plausible Weise. Starten kann man mit der unkontroversen Annahme, dass wir – oder auch andere Einzeldinge – manche Veränderungen überleben können, andere hingegen nicht. *Warum* aber können wir manche Veränderungen überleben und andere nicht? Das hat offenbar etwas damit zu tun, *was* für Einzeldinge wir sind – eben von welcher *Art* wir sind.

Tische oder Stühle beispielsweise können andere Veränderungen überstehen (oder nicht überstehen) als Menschen oder auch Katzen. Und das liegt sicherlich daran, dass wir es einmal mit Tischen (als eine Art von Artefakten), das andere Mal mit Menschen oder auch Katzen zu tun haben – eben mit unterschiedlichen Arten von Einzeldingen. Wären wir beispielsweise ‚Cartesische Egos', die nur kontingenterweise mit einem Körper ausgestattet sind, hätten wir andere Überlebensbedingungen als beispielsweise Katzen. Wir könnten den ‚biologischen Tod' wohl überleben, weil der nur unseren Körper beträfe. Eine besondere Art von Geistwesen zu sein, *erklärte* also, warum etwas, ich beispielsweise, eine solche Veränderung überleben könnte. Und Katzen könnten eine solche Veränderung nicht überleben, *weil* sie Katzen – eine bestimmte Art von biologischen Lebewesen – sind.

Jedes Einzelding kann man natürlich verschiedenen ‚Arten von Einzeldingen' (oder neutraler: Klassen) zuordnen. Und klarerweise geben nicht alle derartigen Zuordnungen Aufschluss über die Überlebensbedingungen eines Einzeldings. Dass viele Menschen Kinder und Westeuropäer sind, tangiert deren transtemporale Identitätsbedingungen nicht im Geringsten. Niemand hört beispielsweise dadurch auf zu existieren, dass er kein Kind mehr ist. Darüber nachzudenken, ob jemand dadurch aufhören könnte zu existieren, dass er kein Mensch mehr wäre (oder auch eine Person), ist hingegen ganz und gar nicht abwegig. Und zu behaupten, wir seien *wesentlich* Menschen oder auch Personen, soll besagen (zu-

17 Wie man den Personenbegriff im Detail weiter erläutern sollte oder könnte, ist für die folgenden Überlegungen unerheblich. Auch auf die Frage, ob der Personenbegriff eine *normative* Dimension hat, werde ich nicht weiter eingehen. (Locke zufolge hat er diese Dimension.)

mindest partiell), dass wir nicht (weiter) existieren könnten, ohne ein Mensch bzw. eine Person zu sein.[18] Unterstellt ist hier, dass wir zumindest *irgendetwas* – von irgendeiner Art – wesentlich sind.

Von diesen Behauptungen bis zu einer ausgearbeiteten ontologischen Theorie – die vermutlich auf eine Variante eines sogenannten *sortalen Essentialismus* hinausliefe – ist es noch recht weit.[19] Hier möchte ich nur einen ersten Schritt in diese Richtung gehen. (Und mehr ist an dieser Stelle auch gar nicht nötig.) Dieser Schritt besteht aus einem Bündel von drei Thesen, die den bereits angedeuteten Zusammenhang zwischen der Artzugehörigkeit und den transtemporalen Identitätsbedingungen von Einzeldingen präzisieren sollen:

(i) Für jedes ‚genuine' Einzelding x gilt: Es gibt mindestens eine Art A (von Einzeldingen), so dass x von dieser Art A ist und x *wesentlich* ein A-Ding ist. Und wenn x wesentlich ein A-Ding ist, folgt, dass x notwendigerweise so lange ein A-Ding ist, so lange x existiert.[20]

(ii) Alle Einzeldinge, die wesentlich A-Dinge sind, haben dieselben transtemporalen Identitätsbedingungen. Und derartige A-Dinge haben diese transtemporalen Identitätsbedingungen (partiell) aufgrund der Tatsache, dass sie A-Dinge sind.[21]

---

18 Offen bleiben soll, ob der Begriff der wesentlichen Eigenschaft oder der wesentlichen Artzugehörigkeit restlos durch Modalbegriffe analysiert werden kann. Unterstellt ist hier nur: Wenn einem Gegenstand eine Eigenschaft (oder Artzugehörigkeit) wesentlich zukommt, dann kommt sie ihm notwendigerweise zu (in allen möglichen Welten, in denen er existiert). Dass – in umgekehrter Richtung – Eigenschaften, die einem Gegenstand notwendigerweise zukommen, zwangsläufig auch wesentliche Eigenschaften dieses Gegenstands sind, kann man hingegen mit guten Gründen bezweifeln. (Solche guten Gründe hat beispielsweise Kit Fine in Fine (1994) herausgearbeitet.)

19 Eine solche Position hat beispielsweise David Wiggins entwickelt (siehe Wiggins (2001)).

20 Die vage Einschränkung ‚genuin' soll andeuten, dass nicht behauptet wird, (i) treffe auf alle Gegenstände zu, die wir so landläufig als Einzeldinge betrachten. Entscheidend ist hier jedoch im Grunde nur, dass das Gesagte auf *uns* zutrifft – ob wir nun wesentlich Personen, Organismen oder was auch immer sind.

21 Eine Komplikation sollte an dieser Stelle erwähnt werden: Angenommen, wir wären wesentlich Organismen der biologischen Art *Homo sapiens*. Das wäre dann, so eine recht übliche Redeweise, unsere *grundlegende Art*. (Die Auskunft, wir seien Exemplare der Art *Homo sapiens*, wäre die Antwort auf die Frage, *was wir grundlegend sind*.) Daraus folgte jedoch nicht, dass wir *andere* transtemporale Identitätsbedingungen hätten als beispielsweise Katzen oder Klammeräffchen. Und in der Tat spricht eher vieles dafür, dass unsere transtemporalen Identitätsbedingungen – sollten wir wesentlich Exemplare der biologischen Spezies *Homo sapiens* sein – *nicht* groß anders aussähen als die vieler anderer biologischer Lebewesen. Zu einer grundlegenden Art gehört demnach nicht zwangsläufig ein allein für diese Art spezifisches Bündel an transtemporalen Identitätsbedingungen. Wir Menschen könnten also unsere transtemporalen Identitätsbedingungen primär aufgrund der Tatsache haben, dass wir biologische Lebewesen (einer bestimmten Komplexität) sind – wie viele andere Lebewesen auch. An dieser Stelle soll keine Vorentscheidung darüber getroffen werden, *was nun genau* an der Tatsache, dass wir Exemplare der Art *Homo sapiens* sind, letztlich erklären könnte, welche Veränderungen wir überleben können und

(iii) So lange ein Einzelding x existiert, verändern sich die transtemporalen Identitätsbedingungen von x nicht und hat x keine konkurrierenden Bündel an transtemporalen Identitätsbedingungen.

Sollten wir also beispielsweise wesentlich Personen sein, hätten wir die transtemporalen Identitätsbedingungen von Personen. Wir existierten so lange, so lange wir *Personen* sind. Wären wir hingegen wesentlich lebendige Organismen der Spezies *Homo sapiens*, hätten wir die Identitätsbedingungen lebendiger (menschlicher) Organismen. Offen bleiben kann an dieser Stelle, welche Identitätsbedingungen nun Personen oder auch lebendige (menschliche) Organismen genau haben. Entscheidend für die folgenden Überlegungen ist jedoch die Annahme, dass wir *nicht zugleich* wesentlich Personen und wesentlich biologische Lebewesen sein können. Und man könnte sich fragen, warum dem überhaupt so sein sollte.

Nun, wenn ein Wesen nur dann eine Person ist, wenn es Selbstbewusstsein besitzt, ist ein Wesen, das wesentlich eine Person ist, nur so lange eine Person, so lange es Selbstbewusstsein aufweist.[22] Wären *wir* wesentlich Personen, würden wir demnach aufhören zu existieren, käme uns (unwiderruflich) die Fähigkeit abhanden, selbstbewusste Gedanken zu denken. Es dürfte ausgemacht sein, dass lebendige Organismen der biologischen Art *Homo sapiens* nicht auf diese Weise zu existieren aufhören. Gewiss können solche Organismen den Verlust von Selbstbewusstsein überleben. Und aller Voraussicht nach können biologische (menschliche) Organismen den Verlust *aller* mentalen Fähigkeiten überleben.[23] Was auch immer also genau die transtemporalen Identitätsbedingungen solcher Organismen sind – es steht zu erwarten, dass es nicht die Identitätsbedingungen sind, die sich aus der Annahme ergeben, wir seien wesentlich Personen. Es handelt sich also tatsächlich um *konkurrierende* Antwortkandidaten auf die Frage, was wir grundlegend (wesentlich) sind.

All diese Annahmen – insbesondere die dezidiert ontologischen Voraussetzungen im obigen Thesenbündel – sind zugegebenerweise begründungsbedürftig. Aber sie sind nicht wirklich exzentrisch. Vor allem aber sind sie in einer wichtigen Hinsicht neutral: Sie bilden ganz sicher keine Vorentscheidung zu-

---

welche nicht. (Es *könnte* etwas sein, was nur auf menschliche Organismen zutrifft, aber eben auch etwas, was wir mit anderen Tieren teilen.)

22 Es ist an dieser Stelle aber irrelevant, ob man Selbstbewusstsein oder irgendwelche anderen mentalen Fähigkeiten als kennzeichnend für Personen ansieht.

23 Hier besteht gewiss weiterer Diskussionsbedarf. Zwar kommt der Standardbegriff des biologischen (auch menschlichen) Organismus ohne Rekurs auf mentale Fähigkeiten aus. Ich möchte jedoch nicht ausschließen, dass es überzeugende Gründe für die Behauptung gibt, dieser Standardbegriff sei ein *zu karger* Begriff. Vielleicht sind zumindest *menschliche* Organismen – im Unterschied zu anderen biologischen Organismen – ja doch wesentlich repräsentierende oder auch empfindende Systeme? Die Idee wäre also, dem Gedanken nachzugehen, ob die Artausdrücke ‚Person' und ‚menschlicher Organismus' doch nicht so arg verschieden sind, wie in diesem Text unterstellt wird.

gunsten des Animalismus. Noch ist *offen*, ob wir wesentlich Personen oder wesentlich biologische Lebewesen sind.

Unterstellt ist für alles Weitere also, dass die Frage, welche Art von Einzelding wir wesentlich sind, und die Frage nach unserer transtemporalen Identität zusammengehören – auf ungefähr die Weise, die durch das Thesenbündel (i)-(iii) skizziert wurde. Der nächste Schritt im Überlegungsgang besteht nun darin, die beiden Annahmen (1) und (2) in diesen größeren (Problem-) Zusammenhang zu stellen. Geschehen soll dies mit Hilfe eines konkreten, wenn auch imaginierten ‚Falls'.

Stellen wir uns einen Mann namens Tim vor, der im Monat Mai all die üblichen mentalen Fähigkeiten hat, dank derer man eine Person ist. Tim ist also im Mai eine Person. Einige Wochen später verliert er jedoch aufgrund massiver Hirnschädigungen genau die mentalen Fähigkeiten, die für den Personenstatus ausschlaggebend sind. Dieser Patient ist im Juli somit keine Person. Was ist er statt dessen? Nehmen wir an, trotz Schädigungen im Bereich des Großhirns seien alle lebensnotwendigen Funktionen des Patienten intakt. Verloren gegangen sind *allein* die für den Personenstatus relevanten mentalen Fähigkeiten. In diesem Fall ist der Patient gewiss ein lebendiger biologischer (menschlicher) Organismus.

Und der Patient im Juli ist aller Voraussicht nach auch *wesentlich* ein solcher Organismus. Was sollte er auch sonst wesentlich sein?[24] Zumindest würden gewiss viele, die nicht zu der Ansicht neigen, dass Tim im Mai wesentlich ein biologischer (menschlicher) Organismus war, dieser Behauptung ohne größere Vorbehalte zustimmen. Es ist vergleichsweise unstrittig zu behaupten, dieser Patient – dieses Einzelding, das vor uns liegt – würde aufhören zu existieren, wäre er nicht länger ein lebendiger biologischer Organismus. Unterstellt werden kann also, dass der Patient im Juli ‚biologische Identitätsbedingungen' hat. Aber war Tim im Mai ein anderes Einzelding als dieser Organismus – vielleicht eine Person?

Der bislang skizzierte Fall sollte noch mit einigen Details angereichert werden. Angenommen, im Juli habe der Patient – dieser biologische Organismus – bestimmte körperliche Eigenschaften, die Tim auch im Mai hatte. (Eine natürlich sehr naheliegende Unterstellung.) Es genügt erst einmal, sich auf *eine* solche Eigenschaft zu beschränken – beispielsweise die Eigenschaft, eine bestimmte Narbe am rechten Unterarm zu haben. Diese Eigenschaft hat der Patient im Juli und auch Tim im Mai. Aber *wem oder was genau* – welcher Art von Einzelding – kam im Mai die Eigenschaft zu, diese Narbe zu haben?

---

24 Zugegeben: Ganz so einfach ist es letztlich nicht. Der Patient könnte zu diesem Zeitpunkt wesentlich eine bestimmte Zusammensetzung aus physikalischen Kleinstpartikeln sein – so, wie auch Tim im Mai wesentlich eine (dann aber andere) Zusammensetzung aus solchen Partikeln gewesen sein könnte. Aber wie eingangs gesagt, ignoriere ich diesen Antwortkandidaten.

Man darf sicherlich annehmen, dass der biologische Organismus, der im Juli existiert, auch schon im Mai existierte. Und naheliegenderweise kamen diesem Organismus auch im Mai bestimmte körperliche Eigenschaften zu. Insofern ist der Organismus im Mai gewiss ein *sehr* aussichtsreicher Kandidat für dasjenige Einzelding, dem im Mai die Eigenschaft zukam, die betreffende Narbe zu haben. Nehmen wir also – bis auf Weiteres – an, dass der Organismus im Mai die Eigenschaft hatte, eine Narbe am rechten Unterarm zu haben.

Welchem Einzelding kamen nun aber im Mai *mentale* Eigenschaften zu? Erinnern wir uns an Annahme (1): Sie besagt, dass es *ein und dasselbe Einzelding* ist, dem *sowohl* mentale *als auch* körperliche (nicht-mentale) Eigenschaften zukommen. *Wenn* nun der biologische Organismus im Mai dasjenige Einzelding sein sollte, das die betreffende Narbe hatte (und im Grunde auch eine Unmenge anderer körperlicher Eigenschaften), dann müsste eigentlich auch *genau dieser Organismus* im Mai dasjenige Einzelding sein, dem *mentale* Eigenschaften zukamen.

An dieser Stelle könnte man Annahme (2) bemühen und erwidern, im Mai habe eben die *Person* Tim die fragliche Narbe gehabt und auch mentale Eigenschaften besessen. Und unbestritten ist ja auch, dass Tim im Mai eine Person war. Allerdings stellte dieser Einwurf *nur dann* einen wirklichen *Einwand* dar, wenn unterstellt würde, dass die Person Tim im Mai *numerisch verschieden* von dem betreffenden Organismus war. Denn natürlich kann auch der Animalismus zugestehen, dass Tim im Mai insofern eine Person war, als dieser biologische Organismus im Mai bestimmte mentale Fähigkeiten hatte.

Ein mit einer solchen Unterstellung vorgebrachter Einwand sollte jedoch irritieren. Denn die Behauptung, dass im Mai die Person und *nicht* der Organismus sowohl die Narbe als auch mentale Eigenschaften hatte, läuft Gefahr, dem Organismus im Mai letztlich *alle* körperlichen Eigenschaften zu rauben. Denn was für die Narbe gilt, müsste auch für alle anderen körperlichen Eigenschaften gelten. Alle körperlichen Eigenschaften kämen dann also der Person, *nicht* aber dem Organismus zu. Diese Konsequenz wäre aber einigermaßen absurd – nicht zuletzt deshalb, weil sie die Annahme untergräbt, dass im Mai *überhaupt* ein biologischer Organismus existierte.

Oder sollte man tatsächlich bestreiten, dass im Mai ein biologischer Organismus existierte? Gab es vielleicht an der Raum-Zeit-Stelle, wo sich Tim im Mai aufhielt, nur die Person Tim, aber kein biologisches Einzelding – also den betreffenden Organismus? Diese Option ist gänzlich unattraktiv. Denn dann müsste man behaupten, dass in dem Moment, in dem Tim seine mentalen Fähigkeiten verlor, die ihn zu einer Person ‚machten', Tim nicht nur zu existieren aufhörte, sondern auch ein *neues Einzelding* – eben ein biologischer Organismus – zu existieren anfing. Ich denke, kaum jemand wollte ernsthaft behaupten, dass biologische Organismen auf diese Weise zu existieren anfangen.

Die plausibelste Deutung des skizzierten Falles scheint zu sein, zumindest vor dem Hintergrund von Annahme (1), dass es der biologische (menschliche) Organismus im Mai war, dem sowohl körperliche als auch mentale Eigenschaften zukamen. Natürlich hatte *Tim* im Mai körperliche und mentale Eigenschaften. Aber Tim und der biologische Organismus sind *ein und dasselbe Einzelding*. Und *genau dieser* biologische Organismus überlebte den Verlust mentaler Fähigkeiten. Und damit überlebte auch Tim. Tim war also offenbar nur *zeitweise*, in einer bestimmten, wenn auch sehr langen Phase seines Lebens, eine Person.

Tims Fall lässt sich natürlich verallgemeinern: Wir *sind* biologische (menschliche) Organismen und haben die transtemporalen Identitätsbedingungen biologischer (menschlicher) Organsimen. Gleichwohl kann jedoch Annahme (2) respektiert werden – und damit auch Strawsons Personenbegriff: Wir, zumindest viele von uns, sind in der Tat Personen; und es sind Personen, denen mentale und nicht-mentale Eigenschaften zukommen. Auch mag der Personenstatus von enormer sozialer und ethischer Bedeutung sein. Daran muss nicht gerüttelt werden. Aber, und das ist die Einschränkung, wir sind offenbar nicht in einem ontologisch grundlegenden Sinn Personen. Der Personenstatus ist in ontologischer Hinsicht ziemlich irrelevant.

Haben sich diese Überlegungen damit sehr weit von Strawsons Behandlung des Personenbegriffs entfernt? Vor allem: Beißt sich dieses Ergebnis mit Strawsons Behauptungen über die ontologische Relevanz des Personenbegriffs? Das ist schwer zu sagen. Eine Antwort auf diese Frage hängt insbesondere davon ab, welche ontologischen Implikationen die notorisch schwer zu interpretierende Behauptung Strawsons genau haben soll, der Personenbegriff sei ein *‚logisch' primitiver* Begriff.[25] Dass diese Behauptung *überhaupt* ontologische Implikationen haben soll, scheint ausgemacht zu sein. Darauf deuten zumindest etliche Bemerkungen Strawsons hin. So etwa die folgende:

> Given, then, that our scheme of things includes the scheme of a common spatio-temporal world of particulars, it appears that a central place among particulars must be accorded to material bodies and to persons. These must be the primary particulars.[26]

Es würde zu weit führen, hier in eine Strawson-Exegese einzusteigen. Angemerkt sei nur, dass diese Formulierungen zumindest so klingen, als sei der Personenbegriff für Strawson nicht lediglich ein sogenanntes Phasensortal – also ein ‚Artausdruck', der auf uns nur in einer (wenn auch meist langen) Phase unseres Lebens zutrifft. Aber womöglich ist Strawson ja nicht wirklich auf die Behauptung festgelegt, dass Personen insofern ‚primary particulars' sind, als Per-

25 Siehe Strawson (1959), S.101f. Strawsons Text selbst gibt keine eindeutige Antwort auf die Frage, auf welche ontologischen Behauptungen er sich festlegen möchte (oder auf welche er festgelegt ist).

26 Strawson (1959), S.246.

sonen grundlegend – eben wesentlich – Personen sind. Vereinbar sind die Kernthesen des Animalismus auf jeden Fall mit einer Lesart der Bemerkungen Strawsons über den Status des Personenbegriffs, die sich bei Peter Hacker findet:

> [...] Strawson argues, we must acknowledge the primitiveness of the concept of a person, i.e. of a type of entity to which both M- and P-predicates are applicable, a type of entity not reducible to a conjunction of two distinct entities to one of which M-predicates are ascribable and to the other of which P-predicates are ascribable.[27]

Wir, die wir Personen sind, bestehen sicherlich nicht aus *zwei* Einzeldingen. Es ist ein und dasselbe Einzelding, dem mentale und körperliche (nicht-mentale) Eigenschaften zukommen. Es ist jedoch der biologische (menschliche) Organismus, der wir jeweils sind, dem diese Eigenschaften allesamt zukommen. Es ist also jeweils ein solcher Organismus, der, zumindest eine gewisse Zeit lang, eine Person ist.[28]

Der Überlegungsgang, der den Animalismus motivieren sollte, ist damit in seinen Grundzügen abgeschlossen. Nicht wenige werden vermutlich denken, dass eine derart kontroverse ontologische These so leichtfüßig gewiss nicht überzeugend begründet werden kann. Das ist im Grunde auch richtig. Allerdings hatten die vorgetragenen Überlegungen ein bescheideneres Ziel: Verdeutlicht werden sollte, wie bruchlos sich die auf den ersten Blick vielleicht radikal anmutenden Kernthesen des Animalismus an einen wichtigen Ausschnitt unserer Interpretations- und Urteilspraxis anschließen lassen.

Sicherlich kann man mit guten philosophischen Gründen an einigen Stellen der Überlegungen kritisch einhaken. Für den Überlegungsgang besonders relevant ist natürlich eine buchstäbliche Lesart der Annahme (1) – also der These, dass wir ein und demselben Einzelding mentale als auch nicht-mentale (körperliche) Eigenschaften zuschreiben. Und gerade diese Lesart kann man angreifen.

In letzter Zeit haben einige Autoren Versionen eines sogenannten *Konstitutionsansatzes* vorgelegt. Die zentrale Idee dieser Theorien besagt, dass wir mit unserem jeweiligen Körper bzw. Organismus zwar nicht identisch sind, jedoch

---

27 Hacker (2002), S. 24.

28 Diese Konsequenz zieht Hacker nicht: „Not all animals are spoken of *having* a body and hence of *not* being identical with the body they *have*. Insects or crustaceans are not said to have bodies. It is only of ourselves and of higher animals, indeed, perhaps only of those who have a character, temperament and individuality exhibited in their behaviour, that we say that they have a body. […] That of which it can be said that it *has* a body, cannot, *in the same sense*, be said to *be* a body. For it is not my body that goes for a walk, intends to go to London, feels joy or sorrow, is an agent responsible for actions, and who has virtues and vices." (Hacker (2002), S.31 und S.39.) Die Redewendung „einen Körper haben" richtet einiges ontologische Unheil an. Eine Deutung dieser und verwandter Redewendungen habe ich in Reuter (2009) vorgeschlagen.

von ihm *konstituiert* werden.[29] Auch solche Theoretiker würden behaupten, ein und demselben Einzelding kämen sowohl mentale als auch körperliche Eigenschaften zu. Entscheidend ist jedoch der *Zusatz*, dass Personen mentale Eigenschaften *auf eine andere Weise* zukommen als körperliche (nicht-mentale) Eigenschaften. Etabliert werden soll eine Redeweise, wonach Personen mentale Eigenschaften *direkt* bzw. *nicht-derivativ* zukommen, körperliche Eigenschaften hingegen nur *indirekt* bzw. *derivativ*. So soll ich beispielsweise ein bestimmtes Gewicht *aufgrund der Tatsache* haben, dass der Körper, durch den ich konstituiert bin, ein bestimmtes Gewicht hat – wobei dieses Gewicht meinem Körper direkt und mir, der Person, indirekt (oder eben derivativ) zukommt.

Was passiert gemäß einem solchen Ansatz beispielsweise, wenn ich beim Fußballspielen in eine Pfütze falle? Vermutlich müsste man ontologisch korrekt sagen, dass *ich* – auf indirekte Weise – aufgrund der Tatsache in eine Pfütze falle, dass derjenige Körper in eine Pfütze fällt, zu dem ich in einer ‚intimen' Konstitutionsrelation stehe (er ist ‚mein Körper'). Ähnlich komplex scheint die ‚ontologische Struktur' auszufallen, wenn ich meine kleine Tochter trage: Wird *sie* – dasjenige Einzelding, auf das sie Bezug nimmt, wenn sie „ich" sagt – in einem indirekten Sinn getragen aufgrund der Tatsache, dass ihr Körper, der sie (die Person) konstituiert, in einem direkten Sinn getragen wird? Und was ist mit meiner Handlung des Tragens? Diese Handlung bestünde offensichtlich auch aus einem Beitrag meines Körpers – denn klarerweise bewegt *er* sich in einem physischen Sinne. Aber bewege *ich* mich nur indirekt dadurch, dass sich mein Körper bewegt? Kommt mir die physische (oder physikalisch beschreibbare) Tragebewegung nur indirekt zu?

Ich denke, es dürfte offenkundig sein, wie artifiziell diese Formulierungen anmuten.[30] Vielleicht kann man für eine derartige Rekonstruktion unserer Urteilspraxis gute philosophische Gründe auftreiben. Wichtig ist hier aber nur, dass es *tatsächlich guter* Gründe bedarf, um uns von diesem Stück philosophischer Theorie zu überzeugen. Eine buchstäbliche – und vielleicht philosophisch naive – Lesart der Annahme (1) hat schlicht den Vorzug, von einem wichtigen Ausschnitt unserer Urteils- und Interpretationspraxis abgesichert zu sein.

---

29 Dabei soll die Konstitutionsrelation ein sehr ‚intimes' Verhältnis zwischen einer Person und ihrem Körper markieren, das fast, aber eben nur fast, der Identitätsrelation gleichkommt. Die wichtigste Vertreterin dieses Ansatzes ist Lynne Baker (siehe Baker (2000)). Verwandte Ideen finden sich beispielsweise in Parfit (2008) und Shoemaker (2008).

30 Diese sehr verkürzten und etwas suggestiv vorgetragenen Anfragen an Bakers Theorie ergeben zugegebenermaßen nicht schon ein Argument gegen ihre Theorie. Eine detailliertere und (hoffentlich) substantiellere Kritik an ihrem Konstitutionsansatz habe ich in Reuter (2009) vorgelegt. Der Frage, welche spezifisch handlungstheoretischen Konsequenzen ein Konstitutionsansatz hat (insbesondere für die Rolle des Akteurs), werde ich an anderer Stelle ausführlicher nachgehen.

Die Behauptung, wir seien wesentlich biologische Lebewesen, steht auch nicht wirklich quer zu unserem *Selbstverständnis* als Personen. Weiterhin können wir sagen, dass wir vielerlei mentale Eigenschaften und Fähigkeiten haben, dank derer wir Personen sind und in sozialen und kulturellen Praktiken agieren. Unter diesen mentalen Fähigkeiten mögen auch einige sein, die für Menschen in dem Sinn *spezifisch* sind, dass Exemplare anderer Spezies diese Fähigkeiten nicht haben und auch nicht die Ausstattung besitzen, diese Fähigkeiten zu entwickeln. Vielleicht trifft es beispielsweise zu, dass nur Menschen Selbstbewusstsein (oder eine bestimmte komplexe Form von Selbstbewusstsein) haben bzw. entwickeln können. Sofern Selbstbewusstsein das entscheidende Merkmal von Personen sein sollte, könnte man dann auch sagen, dass – im Vergleich zu anderen biologischen Spezies – nur Menschen Personen sind oder zumindest im Zuge der Ontogenese werden können. All das ließe sich auch mit dem Animalismus behaupten und ausbuchstabieren. Zugestanden werden kann also, dass wir ziemlich *besondere* biologische Lebewesen sind. Hinzufügen würde ein Vertreter des Animalismus nur, dass diese besonderen Fähigkeiten keinen *ontologischen* Unterschied machen. Wir haben beispielsweise nicht deshalb, weil wir Selbstbewusstsein besitzen, andere transtemporale Identitätsbedingungen als biologische Lebewesen, denen diese Fähigkeit abgeht. Aber warum sollte es auch anders sein?

Vielleicht können diese letzten, sehr knappen Bemerkungen ebenfalls dazu beitragen, die Kernthese des Animalismus in ein anderes, günstigeres Licht zu rücken und die Argumentationslasten neu zu verteilen. Damit wäre auch schon viel erreicht. Auf den ersten Blick mag der Animalismus diejenige Position sein, die die Bürde kontraintuitiver Konsequenzen mit sich herumträgt. Ein zweiter oder dritter Blick lohnt sich jedoch. Eine Motivierung des Animalismus kann auf jeden Fall an einen wichtigen Ausschnitt unserer Urteils- und Interpretationspraxis anknüpfen. Zudem *harmoniert* er mit weiten Teilen unseres Selbstverständnisses als Personen. Vielleicht kann man sich mit dem Animalismus also doch schneller anfreunden, als anfangs gedacht.[31]

Literaturverzeichnis

Baker, L. R.: Persons and Bodies: A Constitution View. Cambridge UP, Cambridge, 2000

Fine, K.: „Essence and Modality". Philosophical Perspectives, 8, 1994. S. 1-16

31 Diese Arbeit ist im Rahmen meines *Dilthey-Fellowships* der VolkswagenStiftung entstanden. Für diese Förderung möchte ich mich bei der VolkswagenStiftung herzlich bedanken. Mein Dank gilt auch Jasper Liptow für viele hilfreiche Bemerkungen zu einer Vorfassung dieses Textes.

Geach, P.: „Identity". Review of Metaphysics, 21, 1967. S. 3-12

Hacker, P.: „Strawson's Concept of a Person". Proceedings of the Aristotelian Society, Vol. CII, 2002. S. 21-40

Kripke, S.: „Identity and Necessity". In: *Munitz, M. K. (Hrsg.):* Identity and Individuation. New York UP, New York, 1971. S. 135-164

Locke, J.: Versuch über den menschlichen Verstand. 2 Bd., Meiner, Hamburg, 1690/1988

McGinn, C.: Logical Properties, Blackwell, Oxford, 2000

Olson, E.: The Human Animal. Personal Identity Without Psychology. Oxford UP, Oxford, 1997

Olson, E.: „Temporal Parts and Timeless Parthood", Noûs, 40, 2006. S. 738-752

Olson, E.: What Are We? A Study in Personal Ontology. Oxford UP, Oxford, 2007

Parfit, D.: „Personal Identity", Philosophical Review, 80, 1971. S. 3-27

Parfit, D.: „Persons, Bodies, and Human Beings". In: *Hawthorne, J./ Sider, T./Zimmerman, T. (Hrsg):* Contemporary Debates in Metaphysics. Blackwell, Oxford, 2008. S.177-208

Reuter, G.: „Wem schreiben wir mentale Eigenschaften zu? Biologische Lebewesen als Subjekte von Erfahrungen". In: *Becker, A. und Detel, W. (Hrsg)*: Natürlicher Geist. Beiträge zu einer undogmatischen Anthropologie. Akademie-Verlag, Berlin, 2009. S. 65-98

Shoemaker, S.: „Persons, Animals, and Identity". Synthese, 162, 2008. S. 313-324

Snowdon, P.: „Persons, Animals, and Ourselves". In: *Gill, C. (Hrsg)*: The Person and the Human Mind. Clarendon Press, Oxford, 1990

Snowdon, P.: „Objections to Animalism". In: *Petrus, K. (Hrsg):* On Human Persons. Ontos-Verlag, Frankfurt/London, 2003. S. 47-66

Spitzley, T.: Facetten des „Ich". mentis, Paderborn, 2000

Strawson, P.: Individuals: An Essay in Descriptive Metaphysics. Methuen, London, 1959

Tugendhat, E.: Selbstbewußtsein und Selbstbestimmung. Suhrkamp, Frankfurt/M, 1979

Van Inwagen, P.: Material Beings. Cornell UP, Ithaca, 1990

Wiggins, D.: Sameness and Substance Renewed. Cambridge UP, Cambridge, 2001

Nonconceptualism: The Argument from Animal Perception

Eva Schmidt
eva.schmidt@mx.uni-saarland.de
Universität des Saarlandes, Saarbrücken

Abstract/Zusammenfassung

I discuss an argument for nonconceptualism based on animal and infant perception. Crudely put, some animals and infants who possess no concepts nonetheless have perceptual states with *nonconceptual* content. Perceptual experiences of adult humans have the same kind of content as the experiences of animals and infants; so the content of the perceptual experiences of adult humans is also nonconceptual.

I defend this argument against potential attacks from the conceptualist. I argue that there are indeed creatures that possess no concepts, but have perceptual experiences, and I attack McDowell's view that we share perceptual sensitivity with animals and infants, but not genuine perceptual contents.

In diesem Aufsatz erörtere ich ein Argument für den Nonkonzeptualismus, das auf der Wahrnehmung von Tieren und Kleinkindern aufbaut. Grob gesagt haben manche Tiere und Kleinkinder, die über keinerlei Begriffe verfügen, dennoch Wahrnehmungen mit *nichtbegrifflichem* Inhalt. Wahrnehmungserlebnisse von Erwachsenen haben dieselbe Art von Inhalt wie die Erlebnisse von Tieren und Kleinkindern; also ist der Inhalt der Wahrnehmungserlebnisse von Erwachsenen ebenfalls nichtbegrifflich.

Ich verteidige dieses Argument gegen potentielle Erwiderungen des Konzeptualisten. Ich argumentiere dabei, dass es tatsächlich Lebewesen gibt, die nicht über Begriffe verfügen, die aber Wahrnehmungserlebnisse haben, und ich greife McDowells Behauptung an, dass wir mit Tieren und Kleinkindern „perzeptuelle Sensitivität“ gemeinsam haben, aber nicht echte Wahrnehmungsinhalte.

Introduction

Conceptualists and nonconceptualists argue whether perceptual content is conceptual just like belief content, or whether it is nonconceptual. There are two answers to the question what it is for a content to be (non)conceptual. According to the state view, the content of a mental state is conceptual if, in order to undergo a mental state, the subject has to possess the concepts (read 'conceptual abilities') that characterize its content and nonconceptual if the subject does not need to possess these concepts. According to the content view, the content of a mental state is conceptual if it is constituted by concepts (as in, for instance, Fregean

senses) and nonconceptual if it is not constituted by concepts (cf. Byrne (2005)). I cannot discuss these different readings of 'conceptual' and 'nonconceptual' here, but will simply assume that the state view entails the content view. The fact that I have to possess and employ certain conceptual abilities in order to undergo a mental state entails that the content of this state is conceptually structured. If I need not possess or employ the respective conceptual abilities, then the content of my mental state is not structured by concepts; it is nonconceptual.

In this paper, I will defend the argument from animal and infant perception for nonconceptualism. First, I will present the argument and motivate its premises. I will then defend it against two potential conceptualist objections.

The Argument

Let me introduce the argument from animal and infant perception: It seems plausible enough that some animals and pre-linguistic children do not possess any concepts, but that they do have perceptual experiences. If this is true, the content of their perceptual states cannot be structured by concepts. Next, it can be argued that adult human perception and animal and infant perception have the same kind of content, or at least that there is a *core* content that they share. It follows that the content of adult human perception is at least partly nonconceptual.

1. There are animals and infants who do not possess any concepts, but have perceptual experiences with genuine content.
2. The content of their perceptual experiences is nonconceptual (by (1)).
3. This content and the content of adult human perception are partially identical.
4. Therefore, the content of adult human perception is – at least in part – nonconceptual. (by (2) and (3)).

The main proponents of this argument are Peacocke, Evans, and Bermúdez (cf. Peacocke (2001a, 2001b), Evans (1982), Bermúdez (1998, 2003a, 2003b)). Different steps of the argument are attacked by Byrne, Brewer and McDowell (cf. Byrne (2005), Brewer (2002), McDowell (1994)). Let us examine the premises of the argument in more detail before turning to possible conceptualist objections.

Let's start with premise (1). It seems clear enough that there are some animals that do not have any conceptual powers at all, for instance snails or amoeba. The same might be argued for very young infants, for instance, newborn babies. But premise (1) also claims that these animals and infants have perceptual experiences with genuine content. This claim is intuitively plausible for animals such as cats and dogs and for older infants; but with respect to these, the claim that they do not possess any conceptual abilities at all might seem questionable. The underlying problem is that there exists a tension between the two assumptions

that premise (1) combines: on the one hand, the relevant animals and infants lack certain demanding cognitive (viz. conceptual) powers, but, on the other hand, they have other relatively demanding mental (viz. perceptual) abilities.

To see how this tension can be resolved we need to get clear on what it is to possess a concept. For a subject to possess a concept is for her to have certain cognitive abilities: recognitional and inferential abilities as well as the ability to form certain thoughts. The subject has to be able to reidentify the corresponding objects and properties, she has to be able to draw inferences involving the concepts she possesses and she has to meet the Generality Constraint (cf. Evans (1982), p. 102). The Generality Constraint asserts that, in order to possess a concept *b* of an object, I have to be able to combine, in thought, my concept *b* with any other concept *F* of a property that I possess to form new thoughts *Fb*. In other words, to possess *b*, I must be able to know what it is for b to have all those properties F for which I possess concepts. It follows that I need to have a full understanding of what bs are. (And *vice versa* for possessing a concept of a property.)

What makes premise (1) plausible is the Generality Constraint. While it may be debatable whether, for example, a dog can reidentify its owner in certain contexts or whether it can draw limited inferences about its owner, it is out of the question to ascribe fully general thought to a dog. A dog is simply not able to entertain the thought that its owner is, for instance, tall, especially not in situations in which its owner is not present. The same is true for very young infants. Accepting the Generality Constraint as a condition for concept possession guarantees that animals and infants do not possess concepts.

How about the content of animal and infant perception? Bermúdez provides empirical evidence for the claim that animals and infants have perceptual experiences with genuine content (cf. Bermúdez (1998), pp. 62-66; Bermúdez (2003c), pp. 85-87). Let me summarize the studies he presents concerning human infants – similar things could be said with respect to animals. Recent research in developmental psychology disproves the older view that, for human infants, the world is almost completely undifferentiated until the end of the sensorimotor period, which is to say that their perceptual states have no content. Even three-month-old babies have certain expectations concerning the behavior of objects. They have certain principles by which they parse their visual fields. For instance, they show surprise when a solid object apparently moves through a solid surface. Their perceptual experiences must have some sort of content which explains these expectations. Yet at the age of three months, it is plausible that these infants do not meet the Generality Constraint.

Note that subscribing to the first premise requires the nonconceptualist to accept the so-called Autonomy Thesis, the claim that it is possible for a subject to undergo perceptual experiences with genuine content even if she possesses no concepts whatsoever (cf. Peacocke (1992), p. 90; Bermúdez (1998), p. 61). The

Autonomy Thesis is controversial even among nonconceptualists. It was originally rejected by Peacocke (cf. Peacocke (1992), pp. 90/91) and is still rejected by Tye for phenomenally conscious content. According to him, one precondition of perceptual experiences having genuine content is that there is a cognitive system (including states with conceptual content) that these perceptual states can have an impact on (cf. Tye (1995), p. 138).

There are some problems accepting the Autonomy Thesis that I cannot go into here. The alternative view, granting limited conceptual abilities to animals and infants, cannot support the argument, so I will stick with the Autonomy Thesis in this presentation.

I will assume the truth of premise (1) for now, ignoring the issues I just mentioned: There are animals and infants who do not meet the Generality Constraint and therefore possess no concepts, but who have perceptual experiences with genuine content.

On the understanding of 'nonconceptual' I introduced in the beginning, premise (2) follows from premise (1). If a subject possesses no conceptual abilities that she could exercise in undergoing her perceptual experiences, and if her experiences have genuine content, then this content must be nonconceptual.

The controversial claim involved in premise (3) is that not only do the respective animals and infants have perceptual experiences with genuine content, this content is supposedly of the same kind as the content of the perceptual experiences of adult humans. Peacocke tries to provide support for this claim by appeal to intuition: The denial of premise (3)

> entails that the following cannot be literally true: that the animal has a visual experience as of a surface at a certain orientation, and at a certain distance and direction from itself, in exactly the same sense in which an adult human can have a visual experience with that as part of its content. (Peacocke (2001a), p. 260)

In addition to a relatively weak appeal to intuition, there is a stronger point underlying Peacocke's argument. We normally use *empirical* methods to test whether an animal's perception is similar to that of an adult. The perceptual organs and brain structures underlying adult human perception and the perception of higher animals are very similar; this is normally taken to be evidence for how similar their perceptual states and their contents are. An example for the role of empirical research in this context is that scientists argue that dogs cannot perceive the differences between some colors that humans can experience based on behavioral tests with dogs and on the make-up of their eyes (cf. Lindsay (2000), pp. 128-132).

By contrast, the conceptualist argues that animals cannot have perceptual experiences with the same kind of content as humans just because there is an *a priori* connection between concept possession and the possibility of perceptual states with genuine content. (As we will see later, he claims that there can be no genuine content without concept possession.) Thereby, he implies that actual

similarities or differences between human and animal perception, which can be studied by empirical investigation, are completely irrelevant to the similarity of human and animal perceptual content. This is extremely implausible. The question whether animals, infants and adult humans have the same perceptual states with the same kind of content cannot be decided by *a priori* reasoning alone. If there are empirical studies showing that the brain structures and behavior involved in animal, infant and human perception are very similar, then our theories of perceptual content have to accommodate these results. That is to say, if there is sufficient empirical evidence for shared perceptual content, then premise (3) is true.

Let me concede that Bermúdez presents what might seem to be *empirical* evidence against premise (3) (cf. Bermúdez (2003c) pp. 84/85). He cites empirical research pertaining to both animals and human infants showing that they have different expectations of object behavior than adult humans and therefore different underlying principles of what counts as an object. Contrary to what Bermúdez suggests, this does not show that animal and infant perception has a different *kind* of content than adult human perception. Even if the visual field is parsed differently in animal and infant perception and in adult human perception, this is compatible with infant, animal and adult human perceptual content being of the same kind. For all the latter claim amounts to is that they are constituted by the same kind of non-conceptual elements, e.g. objects or properties, as opposed to, say, Fregean senses. Further, what the studies cited by Bermúdez show is that animals, infants and adult humans all perceive objects, even if they have slightly differing expectations concerning the objects' behavior.

To sum up, given the empirical evidence, we have strong reasons to believe that conclusion (4) is true. Animals and infants have perceptual states with non-conceptual content; adult human perception has, in part, the same kind of content; so adult human perception must have at least partly nonconceptual content.

Objections

Now, let us turn to objections against the argument. First, the conceptualist might argue that premise (1) is false. He might claim that animals and infants who have genuine perceptual experiences also possess concepts. Animals and infants who have genuine perceptual experiences have limited inferential and recognitional abilities. That is, the conceptualist can oppose my acceptance of the Generality Constraint.

A philosopher who abandons the Generality Constraint will hold that concept possession is a matter of degree. There is a whole spectrum of concept possessors, ranging from very sophisticated adult humans to infants and other animals with only limited conceptual capacities. If this is true, the argument from animal

and infant perception fails. For premise (1) is false: those animals and infants who clearly have perceptual content similar to ours will have conceptual abilities, however limited. In order to defend the argument, the nonconceptualist has to give a reason why concept possession is an all or nothing affair – as it is if we accept the Generality Constraint.

There are several possible lines of argument that could be put forth in defense of the Generality Constraint, such as the claim that it is presupposed both by inferential and recognitional abilities. Most of these arguments can easily be discounted by the conceptualist by pointing out that they either tend to equate thought with language or require full generality where partial generality is sufficient.

Let me present what I take to be the most promising defense of the Generality Constraint: I contend that there is no way to explain the ability of adult humans to draw inferences that does not involve full generality and therefore context-independence of concepts. The only way to give a deeper explanation of our ability to draw inferences is by assuming that concepts are indeed reusable in different thoughts and in different combinations.

What constitutes the inference from $Fa \,\&\, Gb$ and $Fa \rightarrow Ga$ to $Ga \,\&\, Gb$, for instance, is that the concept compounds Fa, Ga, and Gb show up in different places in the premises and the conclusion. Nothing can count as a concept unless it is reusable in this way. The concept G of a property, for instance, has to be combinable with different concepts (a, b) of different objects. But once a concept is able to show up in more than one place in this way, there can be no limits at all to the combinations, the premises or conclusions it can be used in. There might be some practical hindrances to full generality, such as problems with a subject's brain chemistry that prevent conclusions from being drawn or propositions from being contemplated (cf. Peacocke (1992), p. 43). But once we have a genuine conceptual capacity – an ability to draw certain inferences – there cannot be any principled limits to the thoughts it can be used to form.

This reply leaves the problem of what we should say of animals and infants who are apparently capable of reidentifying objects or of drawing limited inferences. There are several solutions to this problem on offer (that I cannot elaborate on here), such as the ones put forth by Bermúdez (cf. Bermúdez (2003c)) and Susan Hurley (cf. Hurley (2001)).

So, the conceptualist objection to premise (1) fails. At any rate, only philosophers with a very liberal view of concept possession would be willing to attack this premise. But the central proponents of conceptualism have rather high demands on what it takes to possess a concept. McDowell, for instance, thinks that without full rationality or the full-fledged ability to draw inferences and reassess her judgments, a subject is not a possessor of concepts. He tries to counter the argument by attacking the claim inherent in premises (1) and (3) that the animals and infants under consideration have experiences with genuine content. He at-

tempts to account for the perceptual similarities between animals and infants, on the one hand, and adult humans, on the other, by appeal to a perceptual sensitivity that we all have in common. He claims,

> [w]e do not need to say that we have what mere animals have, non-conceptual content, and we have something else as well, since we can conceptualize that content and they cannot. Instead we can say that we have what mere animals have, perceptual sensitivity to features of our environment, but we have it in a special form. Our perceptual sensitivity to our environment is taken up into the ambit of the faculty of spontaneity, which is what distinguishes us from them. (McDowell (1994), p. 64)

According to the nonconceptualist, there is only one possible explanation for the fact that adult humans, human infants and non-human animals all undergo the same kind of perceptual states: all of these states have the same kind of content. McDowell's alternative explanation is that adult humans have *perceptual sensitivity* in common with human infants and with non-human animals. But the perceptual sensitivity of adult humans is transformed by their conceptual abilities – instead of being a mere mechanism that enables subjects to react to their environments appropriately (but nothing more), their perceptual sensitivity produces mental states with a conceptual content.

As adult humans, we are able to critically reflect on our perceptual states and revise our beliefs in their light if necessary. Thanks to our conceptual abilities and our rationality we can build up an objective picture of the world. This is to say that our perceptual states have content; we can truly appreciate what is happening in the world. Without conceptual abilities, animals and human infants can do nothing more than react to their environments; they are so tied up in them that they cannot be said to have more than simple perceptual sensitivity (cf. McDowell (1994), pp. 114-123.)

According to McDowell, then, animal and human infant perception has no content, and *a fortiori* it does not have the same kind of content as adult human perception. My conclusion (4) – the content of adult human perception is partially nonconceptual – does not follow.

There are a few things that could be said about McDowell's notion of *perceptual sensitivity*. Let me present the central problem. There is a tension between two of McDowell's claims. On the one hand, adult humans share perceptual sensitivity with animals and infants. On the other hand, there is a stark contrast between both sides. While adult humans have perceptual experiences with genuine content, all animals and infants have is the ability to react appropriately to current needs. It seems to be nothing more than a terminological maneuver to call both of these things 'perceptual sensitivity'. The problem is aggravated by another claim of McDowell's – he emphasizes that perception "does not even make a notionally separable contribution to the co-operation" between perception and thought (McDowell (1994), p. 9). That is to say that, in the case of adult humans, we cannot even *conceptually* distinguish between perception and

thought and their contents. Perceiving is simply a different way of actualizing one's conceptual abilities. If this is true, perceiving (in human adults) cannot *also* be a kind of perceptual sensitivity just like the one that animals and infants have, for the perceptual sensitivity of animals and infants does not involve content, much less conceptual content.

So, McDowell fails to give a convincing conceptualist account of the similarities between adult human, animal, and infant perception. If the conceptualist wants to accommodate the intuition that adult humans, human infants, and animals have something in common with respect to perception, he has to concede that they must share the same kind of perceptual content.

Conclusion

Let me summarize my discussion of the argument from animal and infant perception. The nonconceptualist appeals to animals and infants to show that adult humans have perceptual states with nonconceptual content. His argument relies on the combination of the following claims: there are subjects of whom it is true that they have no conceptual abilities whatsoever and that they have perceptual experiences with genuine content. Moreover, their perceptual contents are partly identical with the contents of adult human perception. The weakest point of the nonconceptualist argument consists in the tension between these claims. To say that animals and infants possess no concepts is to say that they are very dissimilar from adult humans; it is to grant them only very limited mental capacities. To say that animals and infants have perceptual contents, and even stronger, contents that are just like those of adult humans, is to say that they are very similar to adult humans; it is to grant them very high-level mental capacities.

Correspondingly, the conceptualist can attempt to attack the argument by resolving the tension in one of two directions. McDowell's emphasis on the differences between adult human perception and infant and animal perception – his denial of the claim that infants and animals have perceptual experiences with genuine content – is not very promising exactly because, at the same time, he tries to maintain a semblance of commonality between animal, infant and adult human perception. To make the conceptualist view more consistent, he could give up on his notion of shared perceptual sensitivity, but would then be left with the implausible Cartesian view that animals and infants are mere automata.

The other conceptualist option is to abandon the demanding view of concept possession as tied to full-fledged rationality. He can argue that animals and infants resemble adult humans not just with respect to perception, but also with respect to concept possession. As I have shown, the only condition on concept possession that animals and infants clearly cannot meet is the Generality Constraint. Concept possession stands and falls with this constraint. The most com-

pelling argument for this claim is that our ability to draw inferences, which is a necessary condition for concept possession, entails full generality of thought. What enables adult humans to draw inferences is their ability to employ one concept in different combinations in premises and conclusions. But once a concept can be separated from its instances to play this role, there can be no limits to the contexts it can be used in, so it can be applied in a fully general way.

References

Bermúdez, J.: The Paradox of Self-Consciousness. Bradford Books, Cambridge (MA), 1998

Bermúdez, J.: "Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States". In: *Gunther, Y.* (Hrsg.): Essays on Nonconceptual Content. Bradford Books, Cambridge (MA), 2003a. S. 183-216

Bermúdez, J.: "Peacocke's Argument Against the Autonomy of Nonconceptual Content". In: *Gunther, Y. (Hrsg.):* Essays on Nonconceptual Content. Bradford Books, Cambridge (MA), 2003b. S. 293-308

Bermúdez, J.: Thinking without Words. Oxford University Press, New York, 2003c

Brewer, B.: Perception and Reason. Clarendon Press, Oxford, 2002

Byrne, A.: "Perception and Conceptual Content". In: *Steup, M. and Sosa, E. (Hrsg.)*: Contemporary Debates in Epistemology. Blackwell, Malden, 2005. S. 231-250

Evans, G.: The Varieties of Reference. Oxford University Press, New York, 1982

Hurley, S.: "Overintellectualizing the Mind". Philosophy and Phenomenological Research, 63, 2001. S. 423-431

Lindsay, S.: Handbook of Applied Dog Behavior and Training, Vol. 1: Adaptation and Learning. Wiley-Blackwell, Ames, IA, 2000.

McDowell, J.: Mind and World. Harvard University Press, Cambridge (MA), 1994

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

Peacocke, C.: A Study of Concepts. MIT Press, Cambridge (MA), 1992

Peacocke, C.: ”Does Perception Have a Nonconceptual Content?”. Journal of Philosophy, 98, 2001a. S. 239-264
Peacocke, C.: “Phenomenology and Nonconceptual Content”. Philosophy and Phenomenological Research, 62, 2001b. S. 609-615
Priest, S.: Theories of the Mind. Penguin Books, London, 1991

Strawson, P.: Individuals: An Essay in Descriptive Metaphysics. Methuen, London, 1959

Tye, M.: Ten Problems of Consciousness. MIT Press, Cambridge (MA), 1995

5 Metaphysik und Ontologie

Against Fundamentalist Essentialism (Including Scientific Essentialism)

Ralf Busse
ralf.busse@psk.uni-regensburg.de
Universität Regensburg, Regensburg

Abstract/Zusammenfassung

Scientific essentialists claim that fundamental physical features play their lawlike roles by strict metaphysical necessity. Scientific essentialism is a version of fundamentalist essentialism, as the necessity in question is claimed to be something fundamental. I argue that such essentialist have to assume a fundamental modal property underlying their assumed necessities, but that no fundamental property can play the role of a modality. The only way to escape this argument is to adopt a non-standard view of existence, which, I take it, is not what the typical essentialist has in mind.

Wissenschaftliche Essenzialisten behaupten, dass fundamentale physikalische Charakteristika ihre gesetzmäßige Rolle mit strikter metaphysischer Notwendigkeit spielen. Der Wissenschaftliche Essenzialismus ist eine Version des fundamentalistischen Essenzialismus, da behauptet wird, die in Anspruch genommene Notwendigkeit sei fundamental. Mein Argument lautet, dass solche Essenzialisten eine fundamentale modale Eigenschaft annehmen müssen, die den angenommenen Notwendigkeiten zugrunde liegt, dass aber keine fundamentale Eigenschaft die Rolle einer Modalität zu spielen vermag. Die einzige Möglichkeit, diesem Argument zu entgehen, besteht in der Annahme einer unorthodoxen Auffassung von Existenz. Das dürfte nicht im Sinne eines typischen Essenzialisten sein.

1. Actuality-based and fundamentalist essentialism

By essentialism I shall understand a view to the effect that something is a certain way by strict, metaphysical necessity, where this is not already a matter of logic or conceptual truth.[1] Three kinds of essentialism can be distinguished: an entity-feature essentialism is a claim to the effect that an entity (or several entities) has a certain feature by strict necessity; an entity-entity essentialism is a claim to the effect that two (or more) entities are such that, by strict necessity, if one exists so does the other; a feature-feature essentialism is a claim to the effect that two (or more) features (properties or relations) are such that when the one is instantiated so is the other.

1 Kit Fine (1994) famously argues against such a modal analysis of essentiality. See sec. 3 below.

Scientific essentialism is a feature-feature essentialism. It is the thesis that (some or all) fundamental physical properties and relations play their lawlike roles with respect to each other by strict metaphysical necessity.[2] For example, suppose that Charge is elementary charge, that Field is a particular electric field strength, and that Force is the particular force a particle that has Charge experiences in the presence of Field in accordance with the law $F = q \cdot E$. For simplicity, think of Charge, Field and Force as properties of particles, and assume that they are all irreducible, fundamental physical characteristics. Then the scientific essentialist's claim is that it is strictly impossible – and not just excluded by the laws of nature that happen to hold – that any particle has Charge and Field but not Force. Charge and Field are accompanied by Force by metaphysical necessity.

However, the thesis that Charge and Field by strict necessity lead to the presence of Force can be construed in two quite different ways. Someone might hold that certain actual patterns in the occurrence of fundamental physical features – the patterns we then call laws of nature – are so indispensable for our conceptual and epistemic access to these features that scenarios in which these very physical features fail to exhibit these patterns just do not make any sense to us as possible scenarios, as ways the actual world might be. Such a theorist will probably claim that there is no possible scenario according to which anything in the world has Charge and Field but not Force. But according to her this is only so because the three properties *in the actual world* exhibit a correlation that is indispensable for our access to them. Her claim is one of *actuality-based essentialism*: it is because the actual world exhibits a certain distinctive structure, a certain pattern of distribution of physical features, that the presence of Charge and Field combined with the absence of Force makes no sense to us as a possibility.

Scientific essentialism, by contrast, is a version of *fundamentalist essentialism*. According to the scientific essentialist, Charge and Field necessarily result in Force not because a certain important pattern of distribution occurs in actuality. Rather, the entailment of Force by Charge and Field is seen as an ultimate fact. And the contention is that it is *because* Charge and Field necessarily lead to Force that the correlation of the three properties counts as lawlike. This thesis is often formulated by saying that there are irreducible, fundamental dispositions: the fundamental property Charge inherently is the disposition of particles to experience Force in the presence of Field.[3]

2 For different versions of the position see Shoemaker (1980), Swoyer (1982), Bigelow/Ellis/Lierse (1992), Ellis/Lierse (1994), Ellis (2001), Heil (2003), Bird (2007). For an overview see Psillos (2002: ch. II).

3 See, for example, Ellis/Lierse (1994), Bird (2007).

2. Fundamental properties

By a fundamental property I do not mean every property that belongs to the level of description of (some future ultimate) fundamental physics. Only the truly basic characteristics at that level count as fundamental. For example, the disjunctive property of having some charge or other is not fundamental. Only the different absolutely determinate charges, masses, field strengths and so on are fundamental in the intended sense. Thus, a fundamental property is what David Lewis calls a perfectly natural property (1986: 59-69).

It will be helpful to identify properties with the classes of their instances. We know that classes of actual things cannot perfectly play the roles of properties. But for our purposes it will suffice to construe, say, the property of negative elementary charge as the class of all and only the negatively charged particles.

This does not already mean to reject universals and to embrace a thoroughgoing class nominalism. Even if properties are classes, universals can be invoked in order to distinguish the fundamental or perfectly natural classes among them. For example, there might exist the universal elementary charge, and the class of the elementarily charged things might count as perfectly natural because all and only the things in it instantiate a certain universal, namely, elementary charge.

Within this picture of so-called sparse universals, it is natural to say that two things that instantiate the same universal are thereby similar to each other in a certain basic way. To be sure, the things are not duplicates, as they may differ in which other universals they instantiate. But by sharing a certain universal they are nevertheless *perfectly* similar in a good sense, though in a weaker sense than that of being duplicates. They are not just more or less similar to each other like a red and an orange thing. Rather, they perfectly agree, if only in one basic respect of how they are.

The same can be said about all the things in the class of the elementarily charged particles that is identified with the property of being elementarily charged. They all instantiate the universal elementary charge and are therefore perfectly similar in one basic respect of how they are. What is more, we can assume that elementary charge is the only universal they all share. So the things in the class are perfectly similar in exactly one basic respect.

The theory of sparse universals is not the only account of fundamental or perfectly natural properties. Nominalists about perfect naturalness embrace the distinction of a minority of properties as fundamental, but intent to account for the distinction without assuming sparse universals. However, the universals theory of perfect naturalness just sketched may serve as a model for the nominalist[4].

4 In what follows, "nominalist" will be short for "nominalist about perfect naturalness". A nominalist in this sense may well assume abstract entities like numbers and sets. She may even embrace special abstract entities she identifies with properties and relations. The nominalist in this sense only rejects the assumption of a realm of sparse abstract entities

For an adherent of fundamental properties who rejects universals will nevertheless agree that a perfectly natural class of things is a class of things that are perfectly similar in exactly one basic respect. She only denies that perfect similarity in exactly one basic respect can be analyses as the sharing of exactly one universal. She insists that the notion of perfect similarity in exactly one basic respect of the things in a class can be grasped and used without an analysis in terms of universals.

This does not mean that the nominalist advances either perfect naturalness of classes or perfect similarity in one respect of the things in classes as a fundamental property or relation. In particular, the resemblance nominalist about perfect naturalness advances perfect similarity in one respect only as a primitive, undefined and unanalysed concept or notion, not as a fundamental relation between things. Her picture is this: all there fundamentally is to the world is the existence of a certain range of particulars and their fundamentally being certain ways, their having certain basic qualitative determinations. In a sense there is more to the world than the existing particulars, namely, their basic qualitative determinations. But these determinations are not further existing entities. It's just that the existing things fundamentally are certain ways.

This is the nominalists core conviction: the world's particulars fundamentally are certain ways without there occurring further entities like universals in or at them. A quark, for example, may be ontically unstructured. That is, it may be an ontic atom, a thing without further proper parts. In particular, no universals occur in it as parts. The nominalist insists that the quark can nevertheless exhibit a *qualitative* complexity. It is qualitatively determined in several basic ways, say, by having a certain mass and a certain electric charge. It just is these basic ways without incorporating a mass and a charge universal.

Probably the nominalist will admit that we do not know the thing's fundamental qualitative determinations by acquaintance. At best we know or can know them by description, namely, by finding out their lawlike roles. Notice that the universals theorist will probably say something very similar about the sparse universals. We are not acquainted with the universal elementary charge, but only postulate it as the role-player of a certainly lawlike role assumed by electrodynamics.[5] But the nominalist insists that we understand the general notion of two things' being *alike* in their basic qualitative determination. More specifically, the resemblance nominalist about perfect naturalness suggests that we can understand the general notion of the things in a class being perfectly similar in exactly one basic respect. Thus, all there fundamentally is to the world

---

like the universals considered above that serve to distinguish the fundamental properties or classes among the totality of properties and classes.

5 Does this entail quidditism? Not if quidditism is the position that there are possible scenarios in which the fundamental features fail to play the roles they actually play. For this may well be denied by a version of actuality-based essentialism.

according to the nominalist is the existence of certain things and their having certain basic qualitative determinations. But we can understand an unanalysed notion of perfect alikeness of qualitative determination. Using this notion, we as it were collect the world's particulars into classes of things that are all perfectly similar in exactly one basic respect. We have a general notion of a class of things being perfectly natural, or, equivalently, of a property being fundamental.

This nominalist account of perfect naturalness is in fact the view I endorse.[6] For the purpose of this paper, however, I do not need to defend it. I do not even need to defend the idea of perfectly natural or fundamental features in general. For I am concerned with the tenability of scientific essentialism. It is the scientific essentialist who has to assume fundamental characteristics. Her thesis is, after all, that fundamental physical properties and relations play their lawlike roles by metaphysical necessity. The outline of a universals theoretic and a nominalist account of fundamental features just sketched is my offer to the essentialist how to substantiate her underlying assumption of fundamental physical characteristics.[7]

Yet the notion of a fundamental property also enters the discussion of essentialism in a further way. It is not just that the essentialist assumes fundamental physical features that are by metaphysical necessity lawfully correlated in certain ways. She also claims that the metaphysical necessities in question are fundamental, ultimate facts. Scientific essentialism is a version of fundamentalist essentialism, not of actuality-based essentialism. According to the scientific essentialist, the metaphysical necessity of the lawlike correlations does not rest on anything further, but is itself fundamental. So the crucial question arises what the fundamentality of metaphysical necessity may consist in. It is here that the notion of a fundamental property will enter the discussion in a second way.

For the most part of what follows it makes no difference whether one construes the fundamentality of properties in a universals theoretic or in a nominalist manner. My formulations will appear nominalistically, but only because I shall most of the time not mention universals. The universals theorist can easily insert her analysis of fundamentality at each place. This will not affect the argument's substance.

Let me also point out in advance that the nominalist picture of the world adumbrated above is not completely irrelevant also for the theory of universals. The theorist of universals construes a particular's having of a certain basic qualitative determination as its instantiating a certain universal. So it might seem that she is at no point committed to the nominalist's view that entities can be qualitatively determined without there occurring abstract things like universals in or at them. But it will not do for the theorist of universals to say that all there funda-

6 See my (2009).

7 A third alternative is a theory of sparse tropes.

mentally is to the world is the existence of particulars and the existence of universals. For this does not suffice to determine which particular instantiates which universal. There must be something more to the world that glues the universals to the particulars.

Just adding more universals will not do, even if one calls them relational universals of instantiation. It seems that the theorist of universals has to embrace a fundamental nexus of instantiation that neither is a relational universal itself nor is distinguishes as fundamental by a relational universal. It seems unavoidable to resort to the nominalist's picture at least at this point: the particulars have a fundamental determination with respect to some of the universals, so that they can be said to instantiate these universals. But this fundamental determination is not again analysable as the obtaining of a universal between the particulars and their universals. Thus, the relationship of instantiation must and can be construed in a nominalist manner. It is a fundamental, perfectly natural class of pairs $\langle a, F \rangle$ of particulars and universals. The perfect naturalness of the class has to be taken as a conceptual primitive. It is not in turn analysed by reference to a relational universal of instantiation that obtains between the particulars and universals in the pairs.[8]

Finally, one should rather speak of a picture of reality than of a picture of the world. For the totality of what exists may go beyond what exists in the concrete world. Reality also comprises all the abstract entities like sets and numbers, if such there be. These abstract entities may have their own kinds of fundamental characteristics. In particular, the realm of the sets may be interconnected by the fundamental relationship of set-membership. So the dualist picture of existences as their fundamental monadic and relational determinations captures to reality in general, both its concrete and its abstract part. This will be important in what follows, for we shall have to deal with a fundamental property of abstract constructions.

8 Typical metaphysicians of universals account for instantiation by the assumption of states of affairs. Particular *a*'s instantiating universal *F* consists in there existing the state of affairs of a's having *F*, which has *a* and *F* as its constituents. But this view cannot be assumed here, as it arguably presupposes a version of fundamentalist essentialism. The state of affairs of *a*'s having *F* is not just the mereological sum of *a* and *F*. It is a substantial additional entity. But the theorist of universals typically wants it to be the case that the state of affairs necessitates its constituents: necessarily, if *a*'s having *F* exists, *a* exists and *F* exists. This amounts to a fundamentalist entity-entity essentialism concerning states of affairs and their constituents.

3. The argument

The distinguishing contention of a fundamentalist essentialist is that she claims certain *de re* necessities to be fundamental and ultimate. In particular, the scientific essentialist's claim is that it is an ultimate fact that fundamental physical features play such-and-such lawlike roles by strict necessity. The crucial question is how to construe the fundamentality of the alleged *de re* necessities.

Someone might object that the essentialist does not claim the *de re* necessities to be fundamental at all. Her view, so the objection, rather is that the necessities ground on the essences or natures of the things in question. On one reading this objection is beside the point, on another it by itself does not help the essentialist much. If one is content with capturing a thing's having of essential properties by ascribing a modal property to it, the objection is beside the point. For if a thing's being essentially *F* is precisely captured by saying that it is necessarily such that it is *F*, one cannot say that the *de re* necessity is not fundamental but ground on the thing's essence. The thing's having that essence then just is the obtaining of the *de re* necessity. On this reading to claim that the thing's essence is basic precisely is to say that the *de re* necessity is fundamental. Nor does it take us any farther to reify essences or natures. Let's grant that there are special universals, called the essences or natures of things. One still has to account for the fact that a thing's instantiating a universal of that kind *necessitates* it's behaving in a certain manner. On this route one ends up with *de re* necessities of the so-called essences. These necessities again have to be construed as fundamental. So it remains the case that according to the fundamentalist essentialist certain *de re* necessities are fundamental.

On another reading the objection has a point, but does not help very much by itself. The objection might be that the ascription of essential properties cannot be analysed in modal terms, that is, as a thing's necessarily being a certain way. Kit Fine (1994) famously argues that we have to embrace an essentiality operator and that the truths expressed with it ground the *de re* modal truths rather than the first being analysable by the latter. Thus, the proper formulation of an essentiality fact would be that a thing *a* is essentially such that *F*. From this it would follow that *a* is necessarily *F*, but the converse entailment would not hold. Socrates is necessarily a member of his unit set. But it is not part of his essence that he is a member of that set.

This move by itself does not help the essentialist much, because it only leads to a reformulation of the question that is at stake: how are we to construe the fundamentality of alleged essentiality facts? The question above was how the locution "*a* is necessarily such that *F*" can capture something fundamental about the world. The reformulated question would be how this might be done by the locution "*a* is essentially such that *F*". In substance the question has not changed.

I have no firm opinion about Fine's observation that essentiality is not analysable by *de re* necessity and his contention that the former is in fact prior the latter. I shall present my argument for a modal construal of essentiality. The main reason is that we are much better acquainted with different views about modality than with accounts of an alleged unanalysable essentiality. But the argument can be reformulated so that it also captures views on which an essentiality operator is basic.

This is not to deny that the "essentiality first" thesis may be part of a comprehensive metaphysical outlook that can in principle escape my argument against fundamentalist essentialism. I think that Fine is developing such a metaphysics. However, his metaphysics comprises a non-standard, non-quantificational view of ontology, of what it is for something to exist. I am convinced that this is no accident. In the final section of this paper I shall explain why I think that one can escape my argument only if one denies the standard quantificational, broadly Quinean view of existence. More specifically it is the argument's first premise that turns on the assumption of the standard view of existence. But this can better be explained after presenting and discussion the argument as a whole.

Here, then, is my argument against fundamentalist essentialism in outline. It is an argument with two premises:

Premise 1: Fundamentalist essentialism requires a fundamental modal property and not just a primitive modal concept.
Premise 2: No fundamental property can be a modality, because it is incapable of playing the distinctive doxastic role of a modality.
Conclusion: Fundamentalist essentialism is untenable.

I shall discuss the first and the second premises in turn, focussing on the example of scientific essentialism.

4. The first premise

Let's understand metaphysical possibility and necessity in terms of quantification over possible worlds (scenarios). Let's construe possible worlds in the spirit of what David Lewis (1986: 142-165) calls linguistic ersatzism: they are certain abstract entities, namely complete descriptions of or theories about the actual world. Thus, any possible world is a detailed theory that completely specifies which thing in the world has which fundamental property and which things stand in which fundamental relations. Most of these theories, to be sure, are false; only one of them, we can assume, the actualised world theory, is in fact true. But though in fact false all the other ersatz possible worlds specify a complete way the actual concrete world might be.

The scientific essentialist has to insist that there is no possible world according to which any particle in actuality has Charge and Field but fails to have Force. However, in and by themselves complete world theories that say that there are such particles are no more difficult to construe than world theories according to which all (Charge and Field) particles also have Force. So this essentialist has to provide something that excludes those theories that allow (Charge and Field) particles without Force as *possible* worlds.

The standard move is to introduce a primitive status of consistency which a world theory has to enjoy in order to qualify as a possible world or scenario. But what can this consistency status be? If we were concerned with mere actuality-based essentialism about lawlike roles of physical features, consistency may perhaps just be a primitive concept that we somehow understand and that is sensitive to the lawlike structure of the actual world. We may have a primitive notion, a primitive conception of what it is to be a possible scenario that respects those important structures in the actual distribution of fundamental physical features that we then call laws of nature.

But the scientific essentialist is a fundamentalist essentialist. For her, the strict connection between (Charge and Field) and Force does not rest upon important structures the concrete actual world happens to exhibit. Rather, this necessary connection is seen as brute or ultimate. Moreover, the abstract world theories that allow for (Charge and Field and not Force) particles have the same general build up, the same general structure as the theories that exclude such things. Nor can the three properties Charge, Field and Force have a relevant inner structure, for they are all fundamental, hence simple qualities. But then there is no structure, neither in actuality nor in the world constructions, to which a mere primitive concept of consistency could be sensitive. The having of the status of consistency by certain world theories in contrast to others has to be an additional fundamental fact about them. It has to consist in their having a certain *fundamental property C* of consistency. This establishes the first premise.

5. The second premise

Do not ask me what a fundamental property *C* of an abstract world theory, of a large set of interpreted sentences or the like, might be in the first place. For the theorist of universals it would be a class of world theories that all instantiate a certain universal. For the nominalist the theories in the class would be perfectly similar in one basic respect, where the similarity would be accepted as conceptually primitive. The theories in the class would somehow share a common basic qualitative determination. All this is worrisome, but let it pass. But certainly no fundamental property *C* of abstract world theories deserves to be called consistency; no fundamental property can be a modality.

To see this, consider David Lewis's (1994: 228, 239-240) argument from doxastic role against the assumption of a fundamental feature of objective chance. Suppose someone suggests that objective chance is a fundamental relation CH that holds between possible events and real numbers between 0 and 1. According to Lewis, any feature that deserves to be identified with objective chance has to satisfy the Principal Principle. This principle connects belief about chances to degrees of belief (credence), roughly as follows: when all one knows about a possible event *e* is that it has objective chance *r* to occur, one ought to believe precisely to degree *r* that *e* actually occurs. Lewis's point is that the belief that an event *e* stands in some funny fundamental relation CH to, say, the number 0.7 cannot in the least rationally motivate one's believing to degree 0.7 that *e* actually occurs. CH cannot play the doxastic role of chance.

Similarly, whatever deserves to be called (alethic) consistency has to satisfy the following constraint: the belief that all world theories that have this status say that *p* has to rationally motivate the belief that *p* is in fact the case. After all, truth according to all consistent world theories is necessity, and what is necessary is certainly in fact the case. But the belief that all world theories that have some funny fundamental property *C* say that no particle has Charge and Field but not Force cannot in the least motivate the belief that there are no such particles in the concrete actual world. The having or failing to have of some funny fundamental property by abstract world theories has nothing noticeable to do with the co-occurrence of fundamental physical features in the concrete world.

Perhaps the essentialist insists that the alleged property of consistency is not just "some funny fundamental property", but a fundamental property about which we have an explicit theory. In particular, the theory about *C* postulates that the actualised world theory, the world theory that describes the concrete world completely and correctly, has *C*.[9] Given this postulate, one can safely infer *p* from the information that all theories that have *C* say that *p*.

But this reply is a blind alley. For, taking everything together, the essentialist's analysis of necessity now reads as follows:

> Necessarily *p* := there is a fundamental property *C* (i) that happens to be had by the world theory that describes the actual world correctly and (ii) that is had by no world theory that does not say that *p*.

However, the right-hand side is but a roundabout way of saying that *p* is in fact the case and to add, irrelevantly, that the actual world theory happens to have *C* and that no non-*p* theories have *C*. This is certainly not what the scientific essentialist wants. Her point is certainly not to make the strange conjunctive statement that all particles with Charge and Field in fact also have Force and that, in addition, the correct world theory has a certain property *C* that is not had by any theory that allows for things with Charge and Field but without Force.

9 A reply along these lines has been suggested to me by André Fuhrmann and Hans Rott.

To illustrate this point, imagine the canonical copies of all the world theories filed in the University Library of Bremen. Imagine the copies are enclosed in envelopes of different colours. Suppose you learn that the actualised world theory is in a red envelope and that all theories in red envelopes say that it is now raining. You surely can infer from this information that it is now raining. But does the additional information about the envelope colours arouse the impression in you that the momentary rain is in any way necessary? Likewise, imagine you get the information that (i) there are in fact no particles with Charge and Field that lack Force and, in addition, (ii) that the actualised world story is in a red envelope and that no world theory in a red envelope says that there are such things. Although the extra information (ii) about envelope colours happens to entail the factual information (i) about the actual distribution of Charge, Field and Force in the world, it does nothing to raise the modal status of this information.

I do not intend to quarrel about words. Everybody is invited to henceforth use the locution "Necessarily *p*" as meaning that *p* is in fact the case, and, besides, that there is a fundamental property *C* that happens to be had by the actualised world story and that is only had by world stories according to which *p*. But a so-called fundamental necessity so defined will be of no use. In particular, scientific essentialists typically use their thesis that fundamental features play their lawlike roles necessarily to account for certain phenomena about laws of nature.

One is that laws of nature support counterfactuals. But suppose that Charge, Field and Force are in fact correlated according to $F = q \cdot E$ and, in addition, that a certain fundamental property is had by the actualised world theory and only by world theories that say that $F = q \cdot E$. It is impossible to infer from this a counterfactual conditional like "If this Charge-particle were in Field, it would experience Force". The inference can only be made after redefining the counterfactual conditional "If A were the case, B would be the case" in terms of possible world theories closest to actuality with the stipulation that a theory counts as possible only if it has the property *C* appealed to in the definition of necessity. But the explanatory task for a theorist of laws of nature is not to account for the support of arbitrarily redefined so-called counterfactual conditionals by laws, but of counterfactual conditionals as we ordinarily use them.

Likewise, many scientific essentialists intend to account for the reasonableness of induction on the basis of the assumed necessity of lawlike roles. Why are we justified to infer that $F = q \cdot E$ holds in every case after having examined a limited range of cases? The essentialist's idea is that we first infer by inference to the best explanation that the formula "$F = q \cdot E$" captures the essential roles of the quantities involved. From this we go on to infer that the correlation holds in every actual case. But given the definition of "Necessarily *p*" under consideration the statement that charge, field strength and force are necessarily correlated as $F = q \cdot E$ is the conjunction that, first, they are in fact so correlated, and, sec-

ondly, that the actualised world story has *C* and that all stories with *C* say that $F = q \cdot E$. In order to inductively establish this conjunctive statement, one first has to establish the first conjunct. The first conjunct, however, is the general statement that $F = q \cdot E$ in fact holds universally. But the inductive support of this universal statement is precisely what is at stake. As for the second conjunct, it is hard to see why a pattern of distribution of a property *C* over abstract world theories should be easier to justify by actually observed cases than the simple statement that in all actual cases $F = q \cdot E$.

The problem is that the actual distribution of physical characteristics and the distribution of the assumed property *C* over world theories are independent of each other. The minimum the essentialist needs in order to connect the two areas is the assumption that the actualised world theory does not just happen to have *C*, but that it *must* have *C*. Only on this assumption can she claim that the fact that all world theories that have *C* say that *p forces* actuality to be such that *p*. But in her assumption that the actualised world theory *must* have *C* the essentialist would rely on a further modal primitive, one that occurs outside her theory about property *C*. This further modality would again have to be or to rest on a fundamental property of which the essentialist would have to claim that it just does play the doxastic role of a modality. But this claim would be false, for the reasons given above. There is, then, no way to escape the second premise.

6. Lawlike roles as conditions of identity?

A popular locution for expressing essentialist claims is to say that the essential property belongs to the identity conditions of its bearer. So the scientific essentialist might say that the lawlike role in accordance with the formula "$F = q \cdot E$" just belongs to the conditions of identity of, say, electric charge. I can think of three different ways to understand this claim. On neither construal does the suggested locution help the essentialist to escape the argument under consideration.

One option is to spell out the thesis as saying that we simple do not call a physical feature "electric charge" when we consider it in counterfactual circumstances in which it does not play the role captured by "$F = q \cdot E$". This, however, amounts to a resort to actuality-based essentialism. The claim would be that the lawlike role the fundamental property of electric charge in fact plays is so important to our access to this property that we do not consider scenarios as possible in which that very property fails to play this role. This may perhaps be true. But it is no longer a version of fundamental essentialism about lawlike roles.

A second way to spell out the "conditions of identity" locution is to insist that it is instead an ultimate matter of fact that no fundamental property in a counterfactual scenario that fails to play the charge role is identical to the actual fundamental charge property. But this is equivalent to the thesis that it is a fundamen-

tal, ultimate fact that the actual charge property plays the charge role in all possible scenarios. So construed the "conditions of identity" locution is equivalent to the thesis of fundamentalist essentialism about lawlike roles, so that nothing is gained.

The third construal of the thesis that its lawlike role belongs to the conditions of identity of a property like electric charge does indeed lead to a new view. The idea would be that physical features like charge, field strength and force are but aspects of the total lawlike structure that obtains in the world. This view might be called holistic structuralism. The idea is that the truly fundamental physical feature in the world, the total lawlike structure, is a huge fundamental relation $R(x,y,z,\ldots)$ that connects all the things in the world. Particular physical quantities such as charge, field strength, and force are defined on the basis of this gigantic relation. One property may perhaps be that of being an x such that there are $y, z,\ldots$ such that $R(x,y,z,\ldots)$.

The first thing to note is that even if this view worked, this would not rescue scientific essentialism. Holistic structuralism is a view quite different form scientific essentialism. The latter position says that fundamental physical characteristics such as charge, field strength, and force play their lawlike roles by strict necessity. By contrast, according to holistic structuralism the particular physical features are not fundamental at all. Rather, they are derived as mere aspects of the truly fundamental lawlike structure.

The more substantial point is that it is hard to say why the assumed fundamental lawlike feature $R(x,y,z,\ldots)$ deserves to be called a structure. That there is such a holistic fundamental feature means that the total plurality of the physical things in the actual world and the total plurality of things in certain alternative worlds belong to the same perfectly natural class. That is to say, the actual plurality of things and the alternative plurality of things are perfectly similar to each other in exactly one respect. For example, according to the theorist of universals there would exist a universal that is had by the actual plurality and the merely possible pluralities. But this does not mean that the pluralities share a structure. They may share a single holistic qualitative determination. But for a structure a complexity, a manifold of aspects is needed. A fundamental qualitative determination does not already amount to a lawlike *behaviour* of its bearers, even if the determination pertains to total pluralities of things. The only possible way to arrive at such a behaviour seems to be to presuppose further fundamental characteristics – perhaps those of spacetime – that interact with the relation $R(x,y,z,\ldots)$. The idea would be that the obtaining of $R(x,y,z,\ldots)$ among the things in the world necessitates them to exhibit a certain pattern of occurrence in spacetime. But then we are back to a claim of fundamentalist essentialism, of a fundamental connection between different fundamental features, this time between $R(x,y,z,\ldots)$ and the relations of spacetime.

In sum, the slogan that the lawlike roles belong to the conditions of identity of physical characteristics does not express a solution to the problems of essentialism, but a way not to see the problems.

7. Essence and existence

There is no way to escape premise two of my argument: if a fundamental modality has to be a fundamental property in the sense explained in sec. 2, then there can be no fundamental modality. For no fundamental property can play the role of a modality. Everyone who wants to avoid the conclusion that fundamentalist essentialist is untenable has to try to circumvent premise one. The question then is what the fundamentality status of a modality should be if it is not that of being a fundamental property.

It is beside the point to object that no essentialist has ever suggested that at the basis of fundamental necessity lies anything like a fundamental consistency property of world theories. I do not claim that anybody has suggested such a view. I complain that fundamentalist essentialists have typically said almost nothing about the status of fundamentality of necessity, and I argue that the only way to go for them is to assume a fundamental, perfectly natural modal property. It is equally beside the point to repeat that necessity is fundamental and to insist that for that very reason little more can be said about it. For what I demand is an account of the status of fundamentality of the allegedly fundamental necessity. This must not be misconstrued as the demand of a reduction of necessity or consistency to something else. The universals theorist and the nominalist about perfect naturalness have an account of what it is for a property to be fundamental or perfectly natural. This does not mean that they suggest a reduction of the assumed fundamental properties – say, the charges, field strengths and forces – to something more basic.

One can of course reject the specific suggestion made above that the essentialist‘s fundamental property is best identified with a consistency property of abstract world theories. An essentialist may well reject such world theories and claim that necessity itself is already the perfectly natural characteristic. The most straightforward way to spell out this view would be to say that necessity is a perfectly natural property of propositions, whatever propositions may be. The argument for premise two above is easily adapted so as to also rule out this version. The same can be said about a corresponding position that takes into account Fine’s observation that what is truly fundamental is not modality but essentiality. Roughly, the essentiality operator in “object *a* is essentially such that *F*” might be construed as standing for a perfectly natural relation between objects and properties. The same problems arise that have been pointed out in the argument for premise one of my argument. For if “essentially” just stands for

some funny fundamental relation between objects and properties, why should the acceptance of "*a* is essentially *F*" rationally motive the acceptance of "*a* is *F*"? Object *a* may well stand in a peculiar fundamental relation to the property *F*-ness without actually being *F*.

If one wishes to escape the argument against fundamentalist essentialism presented in this paper, a much more drastic manoeuvre in one's metaphysics is required. The strategy of my argument is that I first, in the argument for the first premise, force the essentialist into the assumption of a certain fundamental features as the core element of her account of fundamental necessity and then, in the argument for the second premise, show that no such feature can play the role of a modality. In order to escape this reasoning the essentialist has to find a way to directly fundamentalise our notion of necessity (or essentiality, for that matter) which *includes* the logico-doxastic role played by necessity. Fine (2000) has found a general locution for such a direct fundamentalisation. He suggest that we can understand a reality operator "is it constitutive of reality that" and say, in our case, that for some *p* it is constitutive of reality that necessarily *p*. Thus, we use our modal operator "necessarily" as we ordinarily understand it, including its logico-doxastic role, and say in effect that some facts expressed by instances of the schema "necessarily *p*" are not derivative from more basic facts, but are themselves constitutive of reality. We thereby say that the necessity operator in general stands for something basic in reality.

I shall not question the intelligibility of Fine's reality operator[10] and the claims about reality expressed with it. My point is that such a direct fundamentalisation of necessity or essentiality is not available to the large majority of philosophers today. Recall the picture of reality sketched in sec. 2 about the notion of a fundamental or perfectly natural property. The picture is that when we ask what there fundamentally is to reality, two kinds of answers are required. What there fundamentally is to reality is, firstly, that there exist certain (maybe in some sense basic) things.[11] But there is more to say. The things are not just there, but they also fundamentally are certain ways. They have certain basic qualitative determinations, and they can be collected into natural classes in accordance with their basic determinations.

It is essential to realise that the things' having of qualitative determinations does not amount to there existing further things. The fact that the things are fundamentally certain ways does not enlarge the range of things there are. To be sure, the theorist of universals suggests that a particle's having of the fundamental physical feature of electric charge consists in its instantiating the universal

10 For doubts see Hofweber (2009).

11 One might say that only the existence of ontically simple, atomic things should count as fundamental. However, the fundamental qualitative determinations may well concern complex things. So it seems wise not to complicate the discussion with a requirement of ontic simplicity.

electric charge, which is a second thing in addition to the particle. But the theorist of universals needs her own kind of qualitative determination that does not consist in the addition of things. At the very least she needs to assume a fundamental two-place nexus of instantiation that connects the particulars to their universals. So in this way the two-dimensional picture of what there fundamentally is to reality is correct both for the nominalist and for the theorist of universals. It's only that according to the latter theorist there are two kinds of things, particulars and universals, while there is perhaps just a single kind of qualitative determination, namely, instantiation.

My position is that if the standard quantificational view of existence is correct, then the existence of things and their having of basic qualitative determinations is all there fundamentally is to reality. Everyone who wishes to add a further fundamental aspect to reality, like the directly fundamentalised necessity envisaged above by means of Fine's reality operator, has to dismiss the standard view of existence.

The standard view of existence is the Quinean one that existence is what is adequately expressed by first-order objectual quantification. The core claim is not that there is no first-order predicate of existence. There may well be such locutions. The essence of the standard view is rather that in order for something to exist it suffices that it is available as a referent of singular constants or variables. The point of this is best seen by comparison with an opposing, non-quantificational view. Fine, for example, agrees that there is such a thing as first-order objectual quantification. He even agrees that this kind of quantification is the means to express being. Thus, to assert that there are real numbers is to express and endorse the being of real numbers. Availability as a referent of a singular term is sufficient for being. But existence is something else. Even the radical nominalist, says Fine, should admit that there are real numbers. The intelligibility of the mathematician's discourse suffices to reveal that the real numbers are available as referents of singular terms. Somehow, we can say, reality features or supplies the real numbers as topics of rational discourse.[12] The nominalist should only deny that the real numbers exist. She should agree that there are real numbers, but deny that any facts about them are constitutive of reality.

The core conviction behind the standard view of existence is that there is no such contrast between being and existence. Being as expressed by first-order quantification already is existence. When we hit with our first-order quantificational apparatus at something to talk about, this topic of discourse exists. This can also be viewed in a different way: on the standard view, what one hits at with one's quantifiers, variables and singular constants is *ipso facto* one of the

12 The locutions "featuring or supplying" and "topics of discourse" are mine, not Fine's.

pieces of which reality consists. That is to say, reality just is everything there is taken together. It is the mereological sum of all there is.[13]

This is not so on the alternative view. On this view, reality is not everything there is taken together, for the topics of talk that have being but do not exist are certainly not pieces of reality. Nor is it adequate in the Finean framework to say that reality is everything that exists taken together. According to Fine's definition, to exist is to be a topic of discourse *x* such that for some way to be *W* it is constitutive of reality that *x* is *W*. The reality operator expresses a special status of complete truths, like a truth of the form that *x* is *W*. The status as existent of topic of discourse *x* is derivative on the status of being constitutive of reality of some truth that *x* is *W*. It does not follow from this derivative special status of existence that reality is adequately viewed as the mereological sum of all the topics of discourse that feature in truths that are constitutive of reality.

So it is the Quinean's distinctive view that reality just is all there is taken together, the sum of everything there is. This view, however, has considerable consequences for what one can assume to be fundamental about reality. Certainly, if reality just is everything there is taken together, one thing that can be fundamental about reality is that there are certain (maybe in some sense basic) things. But what is more, everything else that can be fundamental about reality must then be entity-bound. If reality just is the sum of everything there is, then what is to be fundamental about reality must concern one or more of the existents of which reality consists. There can be no free-floating fundamental truths, truths that do not directly concern objects that help to form reality. That is to say, what there fundamentally is to reality must either be the being or existence of some piece of reality or else *something fundamental about* one or more such existences.

In other words, in addition to the being or existence of certain things in it what can be fundamental about reality is only that one or more such existences fundamentally are certain ways. This is what in sec. 2 I have called a basic qualitative determination of things. A single thing may have a basic monadic determination. Two or more things may exhibit a basic relational determination with respect to each other. In particular, if the theory of universals is true, a particular and a universal can be relationally determined with respect to each other so that it is correct to say that the particular instantiates the universal. The existents in question may be few or many, small or large, categorically homogeneous or heterogeneous – all there can fundamentally be to them is that they fundamentally are certain ways.

13 Some philosophers do not believe in the mereological sum of the things of the world. Others think this sum is a substantially further object over and above its parts. Both can construe the taking together as the forming of a plural term for all the things there are, such as "all the things there are".

For simplicity, focus on the nominalist's view of the monadic determination of a single thing. It is a fundamental way the thing is. It is not, to be sure, an additional universal entity that occurs in or at the thing. It is just a way of the thing to be. There are some more things that can be said about the world on the basis of this thing's being the way it is. There may be other things that are qualitatively determined in perfectly the same way as the thing under consideration. Together with this thing they form a class of things that are perfectly alike in exactly one respect. They form a perfectly natural class, a fundamental property as envisaged by the nominalist about perfect naturalness. In addition there may perhaps be less straightforward kinds of similarity. One particle may be qualitatively determined in the way we call having a mass of 1g. Two other particles may have a mass of 2g and of 3g. No two of the particles need to be *perfectly* alike in any respect. But it seems that the 1g-thing is more similar to the 2g-thing than to the 3g-thing. So perhaps we do not, at the fundamental level, just have the perfectly natural properties (and likewise, the perfectly natural relations). There may also obtain similarity-like relationships between such properties. Having a mass of 1g is closer to having 2g than to having 3g. But this is all we have: the world's existing things, their basic qualitative determinations, perfectly natural classes that comprise things that are perfectly alike in their qualitative determinations, and perhaps similarity-like relationships between such classes.

The devastating consequences for fundamentalist essentialism can be seen in two different ways: firstly, what we have in addition to the existences of things is that a thing fundamentally is a certain way. "Is", not "must be". No necessities occur in the listing of what there fundamentally is to reality. Secondly, the only fundamental characteristics and structures that occur are perfectly natural classes of things and perfectly natural class of *n*-tuples of things. Thus, the only fundamental characteristics and structures available are fundamental properties and relations in the sense explained in sec. 2. But then the first thesis of the argument against fundamentalist essentialism stands: underlying the alleged fundamental necessity or essentiality can only be a fundamental, perfectly natural property (or relation). But, says the second premise, no fundamental property in the sense explained in sec. 2 can play the role of a modality. So given the orthodox quantificational view of existence fundamentalist essentialism is untenable.

The argument against fundamentalist essentialism and against scientific essentialism in particular draws, then, on an assumption that can in principle be denied. It depends on the standard quantificational view of what it is for something to exist. Nevertheless the result is forceful. Firstly, I take it that few philosophers tempted by a version of fundamentalist essentialism ever thought of dismissing a broadly Quinean view of ontology. Secondly, in order to establish fundamentalist essentialism it does not suffice to verbally deny the commitment to the standard view of existence. One has to develop a positive account of reali-

ty and the objects that exist in it that provides a tenable alternative to the view that reality is just the things there are taken together. I know of precisely one essentialist who decidedly develops such a metaphysics. This is Kit Fine, whom I have for this reason chosen as my paradigmatic anti-Quinean about existence.

Thirdly, any metaphysics that denies that reality is the things there are taken together confronts a serious worry. A view according to which reality does not consist of all the things there are or exist seems to be alarmingly close to something like Spinozism. The suggestion seems to be that the only truly existing thing is reality itself. All the different particular entities are not pieces or elements of this existing thing. They are just topics of discourse of different statuses (mere beings or existences) that are supplied by the one reality. To be sure, this is not fair to the letter of the anti-Quinean's doctrine. The anti-Quinean does say that there are many different existences, and even more different beings. Yet it is hard to suppress the worry that this theorist does not quite provide the manifold of existences one wants – true pieces or elements of reality –, but only aspects of a single huge One. This felt threat of Spinozism is my main reason to stick to the position that reality *consists* of everything there is and hence to deny fundamentalist essentialism and scientific essentialism in particular.

References

Bigelow, John/Ellis, Brian/Lierse, Caroline: "The World as One of a Kind". 1992. In: *J. W. Caroll:* Readings on Laws of Nature. University of Pittsburgh Press, Pittsburgh, 2004. S. 141-160

Bird, Alexander: Nature's Metaphysics. Oxford University Press, Oxford, 2007

Busse, Ralf: Properties in Nature. A Nominalist Account of Fundamental Properties. MS, Habilitation thesis, Regensburg, 2009

Ellis, Brian: Scientific Essentialism, Cambridge University Press, Cambridge, 2001

Ellis, Brian, Lierse, Caroline: "Dispositional Essentialism". Australasian Journal of Philosophy, 72, 1994. S. 27-45

Fine, Kit: "Essence and Modality". Philosophical Perspectives. Vol. 8, Logic and Language, 1994. S. 1-16.

Fine, Kit: "The Quest of Reality". Philosopher's Imprint. Vol. 1, no. 1., 2000

Fine, Kit: "The Question of Ontology". In: *Chalmers, D./Manley, D./Wasserman R. (Hrsg.):* Metametaphysics. Oxford University Press, Oxford, 2009. S. 157-177

Heil, John: From an Ontological Point of View. Oxford University Press, Oxford, 2003

Hofweber, Thomas: "Ambitious, Yet Modest Metaphysics". In: *D. Chalmers, D. Manley, R. Wasserman (eds.)*: In: *Chalmers, D./Manley, D./Wasserman R. (Hrsg.):* Metametaphysics. Oxford University Press, Oxford, 2009. S 260-289

Lewis, David: On the Plurality of Worlds. Blackwell, Oxford, 1986

Lewis, David: "Humean Supervenience Debugged". 1994. In: *D. Lewis*: Papers in Metaphysics and Epistemology. Cambridge University Press, Cambridge, 1999. S. 224-247

Psillos, Stathis: Causation and Explanation. McGill Queens University Press, Montreal, 2002

Shoemaker, Sydney: "Causality and Properties". 1980. In: *S. Shoemaker*: Identity, Cause, and Mind. Cambridge University Press, Cambridge, 1984. S. 206-233

Swoyer, Christ: "The Nature of Natural Laws". Australasian Journal of Philosophy, 60, 1982. S. 203-223

Ein Argument für die Realität der Zeit

Georg Friedrich
georg.friedrich@live.at

Abstract/Zusammenfassung

Es ist möglich einen Tisch wahrzunehmen, ohne dass es deshalb einen Tisch zu geben braucht, aber es ist nicht möglich die Zeit wahrzunehmen, ohne dass es die Zeit gibt. Der erste Teil der vorangegangenen Aussage scheint völlig unproblematisch. Skeptiker aller Zeiten haben sich bemüht Argumente vorzulegen, deren Ziel es war, zu beweisen, dass die Wahrnehmung irrtumsanfällig ist. Diese Argumente sind dermaßen gewichtig, dass keine erkenntnistheoretische Überlegung sie außer Acht lassen kann. Der zweite Teil der obigen Aussage ist weitaus weniger offensichtlich. Ganz im Gegenteil könnte man sagen, dass die Behauptung, die Wahrnehmung der Zeit impliziere die Realität der Zeit, der These der Irrtumsanfälligkeit der Wahrnehmung direkt entgegensteht. Die Behauptung, dass die Wahrnehmung der Zeit ihre Realität impliziert, bedarf einer gut fundierten Begründung. Die Frage ist also, warum man im speziellen Fall der Zeit von der Wahrnehmung zum Gegenstand kommen kann, in anderen Fällen aber nicht. Die Antwort ist in der Besonderheit der Zeitwahrnehmung zu suchen und zu finden.

Das Argument, das von der Wahrnehmung im Allgemeinen ausgeht und schließlich zur Realität der Zeit gelangt, ist denkbar einfach; es kann folgendermaßen kurz skizziert werden. Das Argument steht und fällt mit einer einzigen Annahme, nämlich, dass Veränderung unmöglich wäre, wenn es keine Zeit gäbe. Veränderung kann nur in der Zeit vor sich gehen, d. h. eine instantane Veränderung wäre eine Unmöglichkeit. Der nächste Schritt besteht darin festzustellen, dass man Veränderungen wahrnimmt. Der Punkt dabei ist, dass es völlig gleichgültig ist, wo diese Veränderungen stattfinden. Ob es sich also „nur“ um Veränderungen in der Wahrnehmung handelt, oder ob sich tatsächlich etwas außerhalb des Bewusstseins verändert, das wahrgenommen wird, spielt überhaupt keine Rolle. Zu meiner Wahrnehmung habe ich einen privilegierten Zugang. Wenn ich also Veränderung wahrnehme, dann verändert sich irgendetwas – sei es im Bewusstsein, sei es außerhalb des Bewusstseins. Die wahrgenommene Veränderung ist eine Veränderung. Eben wurde festgestellt, dass eine Veränderung nur dann vor sich gehen kann, wenn es Zeit gibt. Mit diesem letzten Schritt komme ich auch schon zur Konklusion, nämlich, dass es eine, wie auch immer geartete, Zeit geben muss.

Annäherung an das Argument

In diesem Beitrag möchte ich ein Argument präsentieren und diskutieren, welches zeigen soll, dass die Zeit existiert bzw. real ist. Der Zusatz „dass die Zeit real ist“ mag auf den ersten Blick vielleicht weniger Einwände hervorrufen, sei jedoch mit der ersten Formulierung synonym. Es ist zudem noch vorauszuschicken, dass das hier vorgestellte Argument der allgemeinen Tendenz entgegen-

läuft, die Zeit selbst auf vielfältige Weise zu reduzieren und zu eliminieren. Beispiele dazu finden sich in der Physik, welche die Zeit entweder zu einem bloßen Parameter macht oder überhaupt unberücksichtigt lässt. Auch in der Logik bietet sich ein ähnliches Bild, denn die logische Standardgrammatik berücksichtigt die Zeit nicht, wie Quine es ausgedrückt hat (vgl. Quine 1973, S. 39) und erst die verschiedenen zeitlogischen Systeme versuchen dieses Versäumnis zu kompensieren. In der Sozialwissenschaft wird die Zeit oft als Konstrukt aufgefasst. Die Zeit ist so etwas wie ein nützliches und notweniges Hilfsmittel um soziale Aktivitäten zu koordinieren, d. h. um sich in einer Gesellschaft zurechtzufinden und zu eben diesem Zweck wurde die Zeit erfunden (Vgl. Elias 1988). In der Psychologie wird die Zeit, so behauptet zumindest eine der möglichen Positionen, auf nichts anderes als die Aktivität von Nervenzellen zurückgeführt und in dieser Weise reduziert. Einerseits sucht man nach der Art und Weise der zeitlichen Ordnung von Ereignissen im Gehirn und andererseits nach einer neuralen Uhr, einer Art Zeit- bzw. Taktgeber, doch die Zeit selbst gerät auch hier ins Hintertreffen.

Über die dahinterliegenden Gründe kann man vielfach nur Vermutungen anstellen, doch man muss jedenfalls feststellen, dass die Zahl der Versuche die Irrealität der Zeit zu beweisen oder die Zeit auf etwas Anderes zurückzuführen Legion ist. Namentlich sind es m. E. die Probleme, die der Zeitbegriff mit sich bringt, welche die Annahme, dass die Zeit irreal ist, attraktiv erscheinen lassen, denn dadurch löst man viele Probleme und Paradoxien, die mit der Zeit bzw. mit dem Zeitbegriff zusammenhängen, mit einem Schlag. Wenn es keine Zeit gibt, dann sollte man sich auch nicht darüber wundern, dass man sich, wenn man doch über die Zeit spricht, in Widersprüche verstrickt. Trifft die Vermutung genannter These zu, so wären Vergangenheit, Gegenwart und Zukunft irreal. Einer der bekanntesten Vertreter dieser Position ist der britische Philosoph und Hegelianer McTaggart, der einen Beweis für die Irrealität der Zeit führte. Der Begriff der Zeit stellt sich nach seinem Beweis (vgl. McTaggart 1908) als widersprüchlich heraus, weshalb McTaggart ihn ablehnt. Eine seiner Schlussfolgerungen daraus ist, dass in Wirklichkeit nichts vergangen, nichts gegenwärtig und nichts zukünftig ist. Es gibt auch nichts, was früher oder später als etwas anderes ist. Und – so erklärt McTaggart – wann immer wir irgendetwas in der Zeit wahrnehmen, dann nehmen wir es so wahr, wie es in Wirklichkeit nicht ist. Soviel sei zur These der Irrealität der Zeit gesagt.

Wahrscheinlich ebenso oft kommt es vor, dass jemand nur glaubt, dass er glaubt, dass die Zeit irreal ist, was ein durchaus relevanter Unterschied ist. Viele Vertreter der These der Irrealität der Zeit sind nämlich nur der Meinung, dass so etwas wie eine objektive und „außerhalb“ liegende Zeit im Sinne Newtons irreal ist. So kommt beispielsweise Norbert Elias in seinen Überlegungen zum Schluss, dass die Zeit ein soziales Konstrukt ist, ein Orientierungsmittel in der sozialen Welt. An dieser Stelle muss man entgegnen, dass soziale Konstrukte

und Orientierungsmittel sicherlich nicht nichts sind; und damit käme der Zeit wieder eine gewisse Realität zu; die Zeit wäre immerhin ein soziales Konstrukt. Man könnte dafür argumentieren, dass es viele verschiedene Arten von Gegenständen gibt, und dass einige davon soziale Konstrukte sein könnten; und als solche könnten sie auch ontologisch abhängig von der Existenz einer sozialen Gruppe sein. Wesentlich dabei ist, dass soziale Konstrukte sich sicherlich in vielen Punkten von z. B. Begriffen, Hunden oder Häusern unterscheiden, aber allen diesen Dingen ist gemeinsam, dass sie existieren; das Sein eines Gegenstandes ist verschieden vom Wiesein (z. B. grün sein) oder Wassein (z. B. ein Pferd sein) dieses Gegenstandes. In diesem Sinne behaupte ich, dass wenn etwas zumindest eine Eigenschaft hat, dann existiert es. Wenn die Zeit in irgendeiner Weise existiert, dann ist sie nicht irreal. An diesem Punkt möchte ich einhacken und etwas länger verweilen bzw. gleich darauf zurückkommen, denn dies ist eine Grundlage des nun folgenden Arguments.

Wenn man nun das Gesagte kurz überdenkt, dann könnte man meinen, dass es ein allzu hoch gestecktes Ziel ist, die Existenz der Zeit beweisen zu wollen; und da die Zeit etwas ist, das von vielen Seiten her reduziert und eliminiert werden soll, ist dies ein zugegebenermaßen hohes, doch erreichbares Ziel.

Mit dem folgenden Gegensatz komme ich zurück zum Argument: Es ist möglich einen Tisch wahrzunehmen, ohne dass es deshalb einen Tisch zu geben braucht, aber es ist nicht möglich die Zeit wahrzunehmen, ohne dass es die Zeit gibt. Der erste Teil der vorangegangenen Aussage ist, wie ich meine, völlig unproblematisch. Skeptiker aller Zeiten haben sich bemüht Argumente vorzulegen, deren Ziel es war, zu beweisen, dass die Wahrnehmung überaus irrtumsanfällig ist. Dazu sei lediglich angemerkt, dass diese Argumente dermaßen gewichtig sind, dass keine erkenntnistheoretische Überlegung sie außer Acht lassen kann. Hingegen ist der zweite Teil der obigen Aussage weitaus weniger offensichtlich. Ganz im Gegenteil könnte man sagen, dass die Behauptung, die Wahrnehmung der Zeit impliziere die Realität der Zeit, der eben erwähnten These der Irrtumsanfälligkeit der Wahrnehmung direkt entgegensteht. Die Behauptung, dass die Wahrnehmung der Zeit ihre Realität impliziert, bedarf daher einer gut fundierten Begründung.

Eine solche Begründung zu geben ist der Gegenstand und das Ziel meines Beitrags. Die Frage ist also, warum man im speziellen Fall der Zeit von der Wahrnehmung zum Gegenstand kommen kann, in anderen Fällen aber nicht. Die Antwort ist in der Besonderheit der Zeitwahrnehmung zu suchen und zu finden. Die Wahrnehmung im Allgemeinen und die Zeitwahrnehmung im Speziellen gehen in der Zeit vor sich. Sie setzen die Zeit also in gewisser Weise voraus. Dazu nun etwas genauer.

Veränderung und Wahrnehmung

Die Argumentation beginnt bei der Wahrnehmung und gelangt schließlich zur Realität der Zeit; sie ist, wie die nun folgende Skizze zeigt, überraschend einfach, besteht aus nur zwei Prämissen von denen ausgehend man zur Konklusion gelangt, und zwar über eine einmalige Anwendung des *Modus Ponens*.

1. Wenn es Veränderung gibt, dann gibt es Zeit.
2. Es gibt Veränderung.
3. Es gibt Zeit. (aus 1 und 2, Modus Ponens)

Der Ausgangspunkt der Überlegung ist die Zeitwahrnehmung. Was ist die Zeitwahrnehmung? Dazu muss gleich gesagt werden, dass man, obwohl man von Zeitwahrnehmung spricht und diese auch Gegenstand zahlreicher, z. B. psychologischer Untersuchungen ist, die Zeit nur in indirekter Weise wahrnehmen kann. Man wird daher zu Recht fragen, was es genau ist, was wir von der Zeit wahrnehmen. Die Antwort ist, dass das Einzige, was wir von der Zeit wahrnehmen können, irgendeine Art von Veränderung ist. Die Zeit selbst ist nicht wahrnehmbar, aber die Veränderung ist ein Zeichen für die Zeit, so wie Rauch ein Zeichen für Feuer ist. Es ist zwar umstritten, ob Zeit Veränderung impliziert, d. h. ob auch dann die Zeit vergehen würde, wenn sich absolut nichts veränderte, jedoch scheint der umgekehrte Schluss, also von der Veränderung auf die Zeit, jederzeit möglich zu sein. Hier tritt die erste Prämisse hervor: Wenn es Veränderung gibt, dann gibt es Zeit.

Veränderung und die Realität der Zeit

Was ist Veränderung? Die auf diese Frage aufbauende Argumentation steht und fällt mit dieser einzigen Annahme, nämlich, dass Veränderung unmöglich wäre, wenn es keine Zeit gäbe. Ich glaube, dass die Annahme der ersten Prämisse nicht sonderlich problematisch ist. Sydney Shoemaker (Shoemaker 1969, 363) nennt sie sogar eine allgemein bekannte Wahrheit (*engl.* truism). Veränderung kann nur in der Zeit vor sich gehen, d. h. eine instantane Veränderung wäre eine Unmöglichkeit, wie sich aus den Begriffen der Veränderung und der Zeit zu ergeben scheint. Der Begriff der Veränderung hat im hier vorliegenden Argument eine entscheidende Rolle inne.

Eine Veränderung ist genau dann gegeben, wenn es einen Gegenstand a gibt, einen Gegenstand im weitesten Sinn, der zu einem Zeitpunkt t eine Eigenschaft F hat und, wenn derselbe Gegenstand a die Eigenschaft F zu einem anderen Zeitpunkt t_1 nicht mehr hat. Es ist nicht möglich zu bestimmen, was Veränderung ist, ohne auf die Zeit Bezug zu nehmen. Wenn man das Ganze etwas ungenau, aber dafür umso anschaulicher ausdrücken will, dann kann man sagen, dass im Zuge einer jeden Veränderung der Gegenstand a die Eigenschaft F hat und

auch nicht hat, und dass es nur das Vergehen der Zeit ist, das den Widerspruch vermeidet.

Der nächste Schritt der Argumentation besteht einfach darin, festzustellen, dass man Veränderungen wahrnimmt. Das ist eine empirische Feststellung; der wesentliche Punkt dabei ist, dass es völlig gleichgültig ist, wo diese Veränderungen stattfinden. Ob es sich also „nur" um eine Veränderung in der Wahrnehmung selbst handelt, oder ob sich tatsächlich etwas außerhalb des Bewusstseins verändert, das dann wahrgenommen wird, spielt überhaupt keine Rolle – so meine Überlegung. Da man zu seiner eigenen Wahrnehmung einen privilegierten Zugang hat, scheint es außerdem so zu sein, dass man Irrtümer in Bezug auf die eigenen Wahrnehmungen auszuschießen kann. Ich kann mich beispielsweise auch nicht irren, wenn ich glaube Schmerzen wahrzunehmen. Ob auch eine physische Ursache für die Schmerzempfindung vorhanden ist, ist natürlich eine andere Frage. Analoges gilt für die Wahrnehmung anderer Dinge.

Ich kann mich aber nicht irren, wenn ich glaube, dass ich eine Veränderung wahrnehme. Wenn ich eine Veränderung wahrnehme, dann verändert sich irgendetwas – sei es im Bewusstsein, sei es außerhalb des Bewusstseins. Auch der nächste Schritt der Argumentation scheint ganz unverfänglich zu sein: Eine wahrgenommene Veränderung ist eine Veränderung.

Am Anfang wurde festgestellt, dass eine Veränderung nur dann vor sich gehen kann, wenn es Zeit gibt. Mit diesem letzten Schritt komme ich auch schon zur Konklusion, nämlich dass es die Zeit geben muss, wenn irgendeine Veränderung wahrgenommen wird. Die Wahrnehmung bzw. genauer die Wahrnehmung der Veränderung beweist somit die Realität der Zeit.

Endbemerkungen

Am Ende angekommen müssen noch zwei Anmerkungen gemacht werden. Die erste Anmerkung bezieht Position zur Frage, ob diese Art der Argumentation auch dazu dienen könnte die Realität anderer Dinge zu beweisen. Die zweite Anmerkung geht der Frage nach, was nun durch eine solche Argumentation bewiesen wird und was nicht.

Erstens: Man könnte meinen, dass es naheliegend wäre die Realität des Raumes auf ähnlich einfache Art zu beweisen, wobei man allerdings bald feststellen wird, dass dies nicht gelingen kann. Der Grund dafür ist, dass der Raum nur der äußeren Wahrnehmung zugänglich ist; die äußere Wahrnehmung ist jedoch irrtumsanfällig. Diese Antwort deutet auch schon an, dass die an dieser Stelle geführte Argumentation nicht dazu verwendet werden kann die Realität irgendwelcher anderen Dinge zu beweisen.

Zweitens: Ich glaube, dass man einerseits sagen kann, dass man die Realität der Zeit beweisen kann, was auch das Ziel des Arguments ist. Andererseits muss

man aber eine tiefgreifende Einschränkung machen, die jedoch den Wert dieses Arguments nicht mindert. Die Einschränkung ist, dass aus keinem der Schritte der Argumentation irgendetwas über die konkreten Bestimmungen der Zeit gefolgert werden kann; ausgenommen ist selbstverständlich, und dies muss betont werden, die Realität der Zeit.

Die Bestimmungen der Zeit können außerordentlich vielfältig sein und sind Gegenstand von Kontroversen in der gegenwärtigen Philosophie der Zeit. Als Beispiele seien hier nur die Frage nach der Endlichkeit bzw. Unendlichkeit der Zeit, die Frage nach der Objektivität der Zeit oder etwa die Frage nach der genauen Struktur der Zeit genannt. Die letzte Frage betrifft die Bestimmungen der einzelnen Zeitstellen. Sind die Zeitstellen ausgedehnt oder unausgedehnt? Gibt es Verzweigungen oder Zyklen der Zeit? Diese und noch mehr Fragen wären in einer gesonderten Untersuchung zu klären und sollen hier weiters keine Rolle spielen. Nun wird man zu Recht fragen, was von der Zeit überbleibt, wenn die zuvor genannten Fragen und alle konkreten Bestimmungen ausgeklammert werden. Ich würde sagen, dass bleibt, was die Zeit an sich ist, nämlich ein System einer Ordnung und strukturierten Aufeinanderfolge von einzelnen Zuständen, die in einer gewissen Beziehung zueinander stehen, die ebenfalls noch näher zu definieren wäre. In dieser Annäherung an eine Definition bleibt einerseits viel Freiraum. Die Ordnung, die Strukturierung, die Art der Beziehung etc. sind nicht festgelegt. Andererseits wird durch sie schon ein Rahmen geschaffen, der beispielsweise das Phänomen der Veränderung aufnehmen kann. Wenn man von Veränderung spricht, dann spricht man nämlich von mindestens zwei Zuständen von irgendetwas, die voneinander verschieden sind. Man spricht von einem Vorher und von einem Nachher, die über die zeitliche Dimension miteinander in Verbindung stehen. Und diese zeitliche Dimension muss als real angesehen werden.

Literaturverzeichnis

Elias, Norbert: Über die Zeit. Suhrkamp, Frankfurt am Main, 1. Auflage, 1988

McTaggart, John: „The Unreality of Time“. Mind, Band 17, 1908. S. 457-474

Quine, Willard von Orman: Philosophie der Logik. W. Kohlhammer, Stuttgart, 1973

Shoemaker, Sydney: „Time Without Change“. Journal of Philosophy, Band 66, 1969. S. 363- 381

The Distinction between Genuine Properties and Mere Cambridge Properties[1]

Vera Hoffmann-Kolss
vera.hoffmann@uni-osnabrueck.de
Institut für Kognitionswissenschaft, Universität Osnabrück

Abstract/Zusammenfassung

In this paper, I aim to show how the metaphysically important distinction between genuine properties and mere Cambridge properties can be defined in an adequate way. Mere Cambridge properties are properties whose instantiation by an individual *x* depends only on the environment of *x*, while genuine properties are properties whose instantiation by *x* may also depend on what *x* itself is like. Examples of genuine properties are *being cubical* or *being the only cubical object*; examples of mere Cambridge properties are *being accompanied by a cube* or *being such that some other individual smiled at some time*. I first discuss the relationship between Geach's notion of a mere Cambridge change and the notion of a mere Cambridge property and show that despite the strong conceptual connections between these two notions, defining the distinction between genuine and mere Cambridge properties in terms of the notion of a mere Cambridge change yields problematic results. Then, I investigate causal criteria put forward by Shoemaker and Cleland and argue that each of them fails to give an adequate definition of mere Cambridge properties. Since an alternative criterion proposed by Francescotti encounters difficulties as well, I conclude that the most promising approach to defining the distinction between genuine and mere Cambridge properties relies on the distinction between intrinsic properties and extrinsic properties, where mere Cambridge properties are defined as properties whose instantiation is in a certain sense independent of the intrinsic properties of their bearers.

Ziel dieses Artikels ist es, zu zeigen, wie sich die in metaphysischer Hinsicht fundamentale Unterscheidung zwischen genuinen Eigenschaften und reinen Cambridge-Eigenschaften in adäquater Weise definieren lässt. Reine Cambridge-Eigenschaften sind Eigenschaften, deren Instantiierung durch ein Individuum *x* nur von der Umgebung von *x* abhängig ist, während die Instantiierung einer genuinen Eigenschaft durch *x* auch davon abhängig sein kann, wie *x* selbst beschaffen ist. *Ein Würfel zu sein* oder auch *der einzige Würfel zu sein* wären Beispiele für genuine Eigenschaften; *zusammen mit einem Würfel zu existieren* oder *derart beschaffen zu sein, dass ein anderes Individuum zu irgendeinem Zeitpunkt gelächelt hat* wären Beispiele für reine Cambridge-Eigenschaften. Zunächst erläutere ich den Zusammenhang zwischen dem von Geach eingeführten Begriff der reinen Cambridge-Änderung und dem Begriff der reinen Cambridge-Eigenschaft und zeige, dass es trotz der engen konzeptionellen Beziehungen zwischen diesen beiden Begriffen problematisch ist, die Unterscheidung zwischen genuinen Eigenschaften und reinen Cambridge-Eigenschaften allein mit Hilfe des Begriffs der reinen

1 This paper is an expanded and modified version of part I, section 5.2 of *The Metaphysics of Extrinsic Properties* (cf. Hoffmann-Kolss 2010).

Cambridge-Änderung zu definieren. Weiterhin diskutiere ich kausale Kriterien, die von Shoemaker und Cleland vorgeschlagen werden, und zeige auf, dass auch diese keine adäquaten Definitionen von reinen Cambridge-Eigenschaften liefern können. Da ein von Francescotti formuliertes alternatives Kriterium ebenfalls zu problematischen Resultaten führt, komme ich zu dem Schluss, dass der vielversprechendste Ansatz zur Definition der fraglichen Unterscheidung auf der Unterscheidung zwischen intrinsischen und extrinsischen Eigenschaften basiert, derart dass reine Cambridge-Eigenschaften Eigenschaften sind, deren Instantiierung in einem bestimmten Sinne unabhängig von den intrinsischen Eigenschaften ihrer Träger ist.

1. What are mere Cambridge properties?

Compare the property of *being accompanied by a cube* to the property of *being the only cubical object* and to the property of *being a cube*. The former two properties are extrinsic. Whether or not they are instantiated by some individual *x* depends not only on what *x* itself is like, but also on the environment of *x*. The property of *being a cube*, by contrast, is intrinsic. Its instantiation by some individual *x* is independent of the environment in which *x* is placed. Yet, there is also a second possibility to classify these three properties. Whether or not *x* instantiates the property of *being accompanied by a cube* never depends on what *x* is like, but only on the environment of *x*, whereas the instantiation of *being the only cubical object* partly depends on what *x* is like since only cubical individuals can have this property. Properties such as *being accompanied by a cube*, whose instantiation by *x* only depends on the features of the environment of *x*, are usually called *mere Cambridge properties*, whereas all other properties – including the property of *being the only cubical object* and the property of *being cubical* – are called *genuine properties*. Thus, extrinsic properties can be either genuine or mere Cambridge properties, while intrinsic properties are always genuine properties.

The conceptual distinction between genuine properties and mere Cambridge properties traces back to the distinction between real change and mere Cambridge change introduced by Geach in order to criticise the so-called *Cambridge criterion* of change. According to this criterion, an individual *x* changes if there is a property *P*, such that *x* has *P* at a time *t*, but lacks *P* at a later time *t*'. (Cf. Geach 1969, pp. 71-72.) Geach points out that this criterion yields a notion of change too weak to correspond to our ordinary intuitive concept of change. If, for instance, a person becomes shorter than her own son as he grows, she satisfies the proposed criterion of change. For there is a property, viz. the property of *being taller than one's own son*, which she has at a certain time, but lacks at a later time. Yet, intuitively, the person herself does not really change in virtue of her son's growing. Geach proposes to call changes of this kind 'mere Cambridge changes' as they satisfy the Cambridge criterion of change, but are not real changes in any intuitive sense. The intuitive relationship between the notion of a

mere Cambridge change and the notion of a mere Cambridge property is obvious: whenever some individual *x* loses or acquires a mere Cambridge property, this constitutes a mere Cambridge change of *x*.

The distinction between genuine and mere Cambridge properties is metaphysically important since it captures the intuition that there are entities which superficially look like properties, but should not be considered as real or genuine properties in the sense of characterizing their bearers in any relevant way. If, for instance, two individuals *x* and *y* both instantiate the mere Cambridge property of *being such that some other individual smiled at some time*, this does not provide any relevant information about what *x* and *y* have in common. Thus, a first step in devising theories, according to which properties are sparse, i.e., according to which not every entity that can be described by an arbitrary predicate is a real or genuine property, will typically consist in excluding mere Cambridge properties from the class of entities considered as properties in a metaphysically robust sense. Yet, even though it is hence rather uncontroversial that the distinction between genuine properties and mere Cambridge properties is metaphysically significant and intuitively well-founded, exact definitions of the distinction are rare and controversial. My aim in this paper is to investigate the most prominent criteria of the distinction currently available and to argue that the currently best account is a criterion based on the distinction between intrinsic and extrinsic properties.

I first discuss the relationship between Geach's notion of a mere Cambridge change and the notion of a mere Cambridge property and show that despite the strong conceptual connections between these two notions, it is problematic to define the distinction between genuine and mere Cambridge properties in terms of the notion of a mere Cambridge change alone (section 2). Then, I investigate the causal criteria put forward by Shoemaker and Cleland and argue that each of them fails to define mere Cambridge properties in an adequate way (section 3). Since an alternative criterion proposed by Francescotti encounters difficulties as well (section 4), I argue that the most promising approach to defining the distinction between genuine and mere Cambridge properties relies on the distinction between intrinsic properties and extrinsic properties (section 5).

2. A Geach-style approach

Geach does not employ the notions of intrinsicness and extrinsicness to specify the distinction between real change and mere Cambridge change. Yet, it appears to be in accordance with his conception to assume that an individual undergoes real change iff some of its intrinsic properties change, and a mere Cambridge change iff some of its extrinsic, but none of its intrinsic properties change. The person becoming shorter than her own son hence undergoes a mere Cambridge

change, since her intrinsic properties are not affected by the process of her son's growing, whereas the son undergoes real change, as his height, an intrinsic property of his, changes. Since instantiation of a mere Cambridge property by *x* does not depend on the intrinsic characteristics of *x*, it follows that *x*'s loosing or acquiring a mere Cambridge property will not entail a change of *x*'s intrinsic properties, i.e., it will not entail a genuine change of *x*.

It might seem natural to exploit this relationship in order to formulate a criterion of the distinction between genuine and mere Cambridge properties, such that *P* is defined as a mere Cambridge property iff (i) for all possible individuals *x* instantiating *P*, *x*'s losing *P*, does not imply that *x*'s intrinsic properties change and (ii) for all *x* not instantiating *P*, *x*'s acquiring *P* does not imply that *x*'s intrinsic properties change. However, this definition is too weak. Consider, for instance the complex property *Q* of *being either cubical and accompanied by a cube or spherical and not accompanied by a cube* (i.e. (*being cubical* $\wedge$ *accompanied by a cube*) $\vee$ (*spherical* $\wedge$ *not accompanied by a cube*)). Suppose that *x* instantiates *Q* in virtue of being cubical and accompanied by a cube. Then, even though *x*'s losing *Q* might be due to *x*'s changing shape, i.e., one of its extrinsic properties, *x* might also lose *Q* in virtue of not being accompanied by a cube any more. Thus, *x*'s losing *Q* does not *imply* that *x*'s intrinsic properties change. An analogous consideration holds in the case where *x* instantiates *Q* in virtue of being spherical and not accompanied by a cube. Thus, *Q* satisfies condition (i). To see that *Q* also fulfils condition (ii), suppose that *x* is non-cubical and accompanied by a cube and hence does not instantiate *Q*. Then, *x*'s acquiring *Q* does not imply that *x* undergoes any intrinsic changes, since *x* might acquire *Q* because the cube initially accompanying *x* vanishes. Again, an analogous consideration holds if *x*'s failing to instantiate *Q* is due to *x*'s being non-spherical and not accompanied by a cube. Therefore, the complex property *Q* satisfies both condition (i) and condition (ii) and has to be classified as a mere Cambridge property according to the proposed definition.

Yet intuitively, *Q* is clearly not a mere Cambridge property since whether or not some individual instantiates *Q* always depends partly on its intrinsic properties, notably its shape. Thus, there are at least some properties with respect to which the proposed Geach-style definition does not yield adequate results. Strictly speaking, the fact that the above definition fails, does not exclude that there are other, more adequate, ways of accounting the distinction between genuine and mere Cambridge properties in terms of Geach's notion of a mere Cambridge change. In view of the mentioned difficulty, it appears understandable, however, that the definitions of mere Cambridge properties proposed in the literature usually do not explicitly rely on Geach's conception, but on different considerations.

3. Causal approaches: Shoemaker's and Cleland's criteria

Shoemaker ties the criterion of whether a property is a mere Cambridge property to causal relevance. According to his causal theory of properties, only properties that contribute to the causal powers of their bearers count as genuine properties. The notion that a property contributes to the causal powers of its bearers, is captured by the notion that individuals have conditional powers, where an individual *x* has a certain power *Q* conditionally upon having the properties contained in some set **S** if *x* also has some property *P*, such that *x*'s having *P* and the properties contained in **S** is causally sufficient for *x*'s having *Q*, but *x*'s having the properties contained in **S** without having *P* is not causally sufficient for *x*'s having *Q*. For illustration, Shoemaker considers the power to cut wood, which an object has if it is knife-shaped, made of steel and knife-sized. If *x* is made of steel and knife-sized and also has the property of *being knife-shaped*, *x* has the power to cut wood. If, on the other hand, *x* is made of steel and knife-sized, but not knife shaped, this is not sufficient for *x*'s having the power to cut wood. Accordingly, *being knife-shaped* is the property *P* in Shoemaker's definition, while *being made of steel* and *being knife-sized* are members of **S** and *being able to cut wood* is the conditional power *Q*. Shoemaker identifies genuine properties with clusters of conditional powers, i.e., those conditional powers with which they endow their bearers. (Cf. Shoemaker 1980, pp. 212-213.) The property *P* of *being knife-shaped* hence is a genuine property on Shoemaker's account, as it bestows its bearers with certain conditional powers, for instance, the power to cut wood (which a knife-shaped individual has if it is knife-sized and made of steel), or the power to butter a sandwich (which a knife-shaped individual has if it is knife-sized and made of wood).

Shoemaker contends that mere Cambridge properties are the opposite of genuine properties, as they do not contribute to the causal powers of their bearers and therefore do not satisfy the causal criterion a property has to fulfil in order to count as genuine (cf. Shoemaker 1980, p. 219).[2] This contention is only justified if one assumes that the conditional causal powers of a thing are determined by its intrinsic properties. Then, it follows that if *x* has the mere Cambridge property of *being accompanied by a cube*, for instance, *x*'s having (or lacking)

2 Shoemaker's original definition is slightly more complicated since he also regards properties involving a reference to a time other than the time at which the property is instantiated, e.g. Goodman's property of *being either green and examined before 2000 AD or blue*, as mere Cambridge properties (cf. Shoemaker 1980, p. 208; 1988, p. 201). In the present context, I assume that properties should not be classified as mere Cambridge properties merely in virtue of involving a reference to a time other than the time at which they are instantiated, i.e., that such properties should not be classified as mere Cambridge properties *per se*. The discussion of whether or not this assumption is justified would go beyond the scope of this paper, however.

this property contributes nothing to *x*'s conditional causal powers, since the fact that *x* is (or is not) accompanied by a cube does not have any implications concerning its intrinsic properties.

The adequacy of the principle that an individual's causal powers are determined by its intrinsic properties can be called into question though. As Francescotti points out, the property of *being fifty miles south of a burning barn*, which Shoemaker regards as a mere Cambridge property (cf. Shoemaker 1980, p. 220-221), can contribute to the causal powers of the individuals instantiating it. For suppose that Jack instantiates this property and that a pilot flies over a burning barn and over Jack, thereby acquiring the belief that Jack is fifty miles south of a burning barn. Intuitively, Jack's having the property of *being fifty miles south of a burning barn* is at least partly causally responsible for the pilot's forming of the belief that Jack is fifty miles south of a burning barn. (Cf. Francescotti 1999b, p. 296.) If this intuition is accepted, Jack's having the property of *being fifty miles south of a burning barn* bestows him with the power to produce certain beliefs in other persons (conditionally upon his having certain other properties, e.g. his being outside and hence visible from a plane, etc.) and is consequently misclassified as a genuine property by Shoemaker's account.

Thus, Shoemaker's criterion either faces a number of counterexamples – as further cases of mere Cambridge properties which are misclassified as genuine on his account can easily be devised in analogy to Francescotti's example – or has to reject common intuitions concerning the causal relevance of properties. Both options are unsatisfactory. Shoemaker's approach can hence not be considered as yielding an adequate definition of the distinction between genuine and mere Cambridge properties.

A further approach can be extracted from an argument given by Cleland. Presupposing that there is a fundamental difference between rest and motion, she contends that real motion can be distinguished from mere Cambridge motion, i.e., motion which consists only in a change of position relatively to other objects, in virtue of the fact that real motion involves what she calls *operative tendencies to be elsewhere*, whereas mere Cambridge motion does not (cf. Cleland 1990, pp. 273-275). As she claims that the idea underlying her account of motion can be generalized such as to apply to other types of change as well (cf. Cleland 1990, pp. 278-279), Francescotti suggests that her theory may be taken to inspire a general criterion of the distinction between genuine and mere Cambridge properties. According to this criterion, a property *P* is a genuine property of *x* iff *x*'s having *P* consists in *x*'s undergoing some inner causal process (cf. Francescotti 1999b, p. 298). A property *P* is then classified as a genuine property *tout court* iff there is an individual *x*, such that *P* is a genuine property of *x*, and as a mere Cambridge property otherwise.

If one presupposes that the notion of an inner causal process can be characterized in an adequate way, this account appears to improve on Shoemaker's ac-

count at first sight. For even though the mere Cambridge property of *being fifty miles south of a burning barn* might bestow certain causal powers on their bearers, no individual undergoes an *inner* causal process in virtue of having this property. In general, however, it is at least questionable whether Cleland's analysis yields an adequate account of the distinction between genuine and mere Cambridge properties. The property of *being either non-cubical or accompanied by a cubical object*, for instance, should be classified as a genuine property. Yet, it is not clear what the inner causal process should be which an individual undergoes when instantiating this property. Considering its negation, i.e., the property of *being cubical and not accompanied by a cubical object*, does not solve this problem either since whether or not some individual *x* instantiates this property not only depends on *x*'s inner causal structure, but also on whether or not *x* coexists with a cubical object. Thus, *x*'s instantiating the property of *being cubical and not accompanied by a cubical object* can never be *equated* with *x*'s undergoing some inner causal process.

Therefore, neither Shoemaker's nor Cleland's causal criterion offers a satisfactory approach to defining the distinction between genuine properties and mere Cambridge properties. In the remaining two sections, I consider two alternative accounts which are not grounded on causal considerations.

4. Francescotti's relational approach

Francescotti proposes a non-causal definition of the distinction between genuine and mere Cambridge properties relying on the notion that mere Cambridge properties are a certain type of relational property. The key concept of his approach is the notion of an internal property, where *P* is an internal property of an individual *x* iff *P* is neither a universal nor an existential d-relational property of *x*. Here, *P* is an existential d-relational property of *x* iff *x*'s having *P* consists in *x*'s standing in some relation *R* to some member of a class of individuals **C**, which has a member to which *x* bears *R* and which is completely distinct from *x*. *P* is a universal d-relational property of *x* iff *x*'s having *P* consists in *x*'s standing in some relation *R* to all members of a class of individuals **C**, which has a member that is possibly completely distinct from *x* (cf. Francescotti 1999a, p. 602, 604; 1999b, p. 301).[3] The *consists-in* relation occurring in this definition is de-

3 Additionally, Francescotti devises the notion of an impure d-relational property, where *P* is an impure d-relational property of *x* iff there is a relation *R* and an individual *y*, such that *x*'s having *P* consists in *x*'s standing in *R* to *y* and *y* is completely distinct from *x* (cf. Francescotti 1999a, p. 601; 1999b, p. 301). However, since every property which satisfies this criterion also satisfies the criterion of being an existential d-relational property (think of the class **C** as containing exactly one individual, viz. *y*), I do not consider this third no-

fined in terms of identity of events: *x*'s having a certain property *P* consists in *x*'s having a property *Q* iff the event of *x*'s having *P* is identical to the event of *x*'s having *Q* (cf. Francescotti 1999a, p. 599; 1999b, p. 306, n. 12).

Being accompanied by a cube, for instance, is an existential d-relational property of every individual instantiating it. For if *x* has this property, then there is a class of individuals **C**, i.e., the class of all cubes, and a relation *R*, i.e., the dyadic relation of *being accompanied by*, such that *x*'s being accompanied by a cube consists in *x*'s standing in *R* to some member of **C**, and it is plausible to assume that there is a member of **C** to which *x* bears *R* and which is completely distinct from *x*. *Being bigger than all cubes*, by contrast, is a universal d-relational property of every individual instantiating it. For if *x* has this property, then there is a class of individuals **C**, viz. the class of all cubes, and a relation *R*, viz. the *bigger-than* relation, such that *x*'s being bigger than all cubes consists in *x*'s standing in *R* to every member of **C**, and it is, or course, possible that there is a member of **C** that is completely distinct from *x*. Thus, neither the property of *being accompanied by a cube*, nor the property of *being bigger than all cubes* are classified as internal according to this definition.

Francescotti then defines a property *P* as a mere Cambridge property of an individual *x* instantiating *P* iff for all properties $F_1, \ldots, F_n$, if *x*'s having *P* consists in *x*'s having $F_1, \ldots, F_n$, none of the properties $F_1, \ldots, F_n$ is an internal property of *x* (cf. Francescotti 1999b, p. 302). Since arguably, there is no internal property F_i, such that *x*'s having the property of *being accompanied by a cube* consists in *x*'s having F_i, *being accompanied by a cube* is correctly classified as a mere Cambridge property by this definition.

However, although apparently adequate in these and a number of other cases, Francescotti's criterion faces a serious objection. Given his proposal to conceive of the *consists-in* relation as grounded in event identity, a property *P* is classified as a universal d-relational property of *x* iff the instantiation of *P* by *x* is the same event as *x*'s standing in *R* to every member of a class **C**, at least one of whose elements is possibly completely distinct from *x*. The problem is that Francescotti does not specify a theory of identity of events which excludes a trivialization of his criterion. For suppose that *P* is an arbitrary qualitative property[4] of *x*'s and that *R* is the relation in which *x* stands to *y* iff *x* has *P* and *y* has exactly the same qualitative properties as *x*. Further suppose that **C** consists of all individuals that have the same qualitative properties as *x*. Then one can argue that *x*'s instantiating *P* consists in *x*'s standing in *R* to all members of **C**, viz. to all individuals having the same qualitative properties as *x*. Thus, *P* is a universal d-relational property of *x*. If this argument is generalized, it shows that Francescotti's criteri-

---

tion any further. (For a more detailed discussion of Francescotti's notion of d-relationality cf. Hoffmann-Kolss 2010, part I, section 3.4.2.)

4 Here as well as throughout the paper, 'qualitative' just means that properties like *being identical to x* are excluded.

on has to classify each qualitative property P instantiated by some individual x as a universal d-relational property of x. This in turn implies that no qualitative property can be an internal property of any individual instantiating it. But if there are no qualitative internal properties, genuine properties like *being cubical* or *being the only cubical object* cannot be adequately accounted for by Francescotti's criterion. For according to this criterion, P is a genuine property of x iff there are internal properties $F_1, \ldots, F_n$, such that x's having P consists in x's having $F_1, \ldots, F_n$. Yet, if no qualitative property is internal, it is entirely unclear how an individual's instantiating a qualitative property like *being cubical* can ever consist in its instantiating some internal property. Thus, even though at first sight, Francescotti's approach seems to improve on the causal accounts discussed in the previous section, it turns out to be inadequate upon closer examination.

5. Defining mere Cambridge properties in terms of intrinsic properties

Interestingly, the contemporary debate on the definition of mere Cambridge properties seems to entirely neglect the fact that a simple and adequate definition of this type of property has already been given by Lewis at the beginning of the 1980ies. Lewis's definition draws on the consideration that the set containing all possible individuals, i.e., all individuals inhabiting the actual or some possible world, can be divided into equivalence classes $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, … under the relation of duplication. This relation holds between the individuals x and y iff x and y are intrinsic duplicates, i.e., iff x and y have exactly the same qualitative intrinsic properties. Accordingly, two individuals belong to the same equivalence class iff they have exactly the same qualitative intrinsic properties. A property P is extrinsic iff it divides at least one of the classes $\mathbf{C}_i$, i.e., iff there are members of $\mathbf{C}_i$ that have P and members of $\mathbf{C}_i$ that lack P. Some extrinsic properties divide all of the classes $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, …. Lewis calls these *purely extrinsic properties*. (Cf. Lewis 1983, p. 356, n. 16.)

Lewis's notion of a purely extrinsic property corresponds exactly to the notion of a mere Cambridge property whose definition is at stake. For if P is a purely extrinsic property, there is no intrinsic property, such that an individual's having or lacking this intrinsic property would be sufficient or necessary for its having P. Thus, whether or not an individual has P always depends on the features of its environment, but never on what it is like itself. For illustration, compare again the property of *being accompanied by a cube* to the property of *being the only cubical object*. There is no intrinsic property whose possession or non-possession by x has any influence on whether x has the former property. However, if x is not cubical, x cannot have the latter property, i.e., be the only cubical

object. Thus, in contrast to *being accompanied by a cube*, the genuine property of *being the only cubical object* does not divide those classes of duplicates whose elements do not instantiate the property of *being cubical* and hence does not satisfy Lewis's condition of being a mere Cambridge property. Therefore, Lewis's definition is adequate in both cases.

One might reject Lewis's approach on the grounds that it relies on a controversial metaphysical assumption. The possibility to subdivide the set of all possible individuals into equivalence classes of individuals standing in the relation of duplication to each other rests on the presupposition that individuals are world-bound, i.e., that no individual inhabits more than one possible world. For if there are individuals inhabiting several possible worlds, it is plausible to assume that there is an individual x and a qualitative intrinsic property P, such that x has P at one world, but lacks P at a different world. But then it is not clear whether x should be assigned to equivalence classes containing individuals that instantiate P or to equivalence classes containing individuals that instantiate $\neg P$. If individuals are assumed to be world-bound, this difficulty does not occur.

However, it is possible to devise a definition which is very similar to Lewis's, but does not presuppose that individuals can be assigned to equivalence classes under the relation of duplication. Define a property P as a mere Cambridge property iff it satisfies the following two conditions: (i) for all x, if x instantiates P, it is possible that there is an intrinsic duplicate of x's which instantiates $\neg P$, (ii) for all x, if x instantiates $\neg P$, it is possible that there is an intrinsic duplicate of x's which instantiates P. Otherwise, P is a genuine property.

To see that this definition yields adequate results, consider first the genuine property of *being cubical*, which is also a qualitative intrinsic property. If some individual x has this property, there will hence be no individual that is an intrinsic duplicate of x's, but lacks this property. *Being cubical* therefore violates condition (i) and is correctly classified as genuine by the proposed definition. The genuine property of *being the only cubical object* does not raise any difficulties either. Suppose that x is spherical and consequently lacks this property. Then, since *being spherical* is a qualitative intrinsic property, all intrinsic duplicates of x's will be spherical and therefore lack the property of *being the only cubical object*. Accordingly, condition (ii) cannot be met by this property, which is correctly classified as genuine, too. Finally, consider the mere Cambridge property of *being accompanied by a cube*. Whenever x has or lacks this property, this does not have any implications concerning x's intrinsic properties. Thus, if x has this property, it is possible that there is an intrinsic duplicate of x's lacking it, while if x lacks this property, it is possible that there is an intrinsic duplicate of x's

having it. *Being accompanied by a cube* therefore satisfies both condition (i) and condition (ii) and is correctly categorized as a mere Cambridge property. [5]

A further objection to the proposed account could be that it merely postpones the problem of defining the distinction between genuine properties and mere Cambridge properties to the problem of defining the intrinsic/extrinsic distinction. On the one hand, this criticism is justified: if there is no adequate definition of the intrinsic/extrinsic distinction available, the proposed definition cannot be adequate either. On the other hand, each definition has to take some concepts as primitive. This also holds for the approaches discussed in the previous sections. The Geach-style account presented in section 2 would not work if the notion of a mere Cambridge change could not be taken for granted. Likewise, the causal accounts put forward by Shoemaker and Cleland heavily rely on a causal theory of properties. Francescotti's account presupposes the notion that the instantiation of a certain property by *x* may *consist in*, i.e., be the same event as, *x*'s having some other property. The account defended in this section is no exception in this respect. However, whereas all the other accounts turn out to be problematic even if the primitive notions on which they rely are accepted, the definition proposed in this section yields adequate results provided that the intrinsic/extrinsic distinction is taken for granted. This, I think, is reason enough to prefer it to the other accounts, while giving a definition of the intrinsic/extrinsic distinction would by far exceed the scope of the present investigation.

References

Cleland, C.E.: "The Difference between Real Change and Mere Cambridge Change". Philosophical Studies, 60, 1990. S. 257-280

5 Two further approaches to defining the distinction between genuine and mere Cambridge properties are worth mentioning. One is Denby's account of the intrinsic/extrinsic distinction, which contains a definition of so-called *pure extrinsic properties* that can plausibly be equated with mere Cambridge properties. Yet, I argue elsewhere that Denby's analysis has to be rejected in general (cf. Hoffmann-Kolss forthcoming). Therefore, Denby's account cannot be regarded as grounding an adequate definition of the distinction between genuine properties and mere Cambridge properties.
A further approach is discussed by Kremer who argues that the distinction between genuine and mere Cambridge properties can most suitably be accounted for in terms of relevance logic (cf. Kremer 1997, pp. 46-56). Yet, the conceptual and logical framework underlying Kremer's account is entirely different from the one underlying the present argument, which is chiefly based on classical logic and possible world semantics. Therefore, whether Kremer's approach represents a viable alternative to the approach of this paper would have to be discussed in a different context.

Francescotti, R.M.: "How to Define Intrinsic Properties". Noûs, 33 (4), 1999a. S. 590-609
Francescotti, R.M.: "Mere Cambridge Properties". American Philosophical Quarterly, 36 (4), 1999b. S. 295-308

Geach, P.T.: God and the Soul. Routledge and Kegan Paul, London, 1969

Hoffmann-Kolss, V.: The Metaphysics of Extrinsic Properties. ontos verlag, Frankfurt, 2010
Hoffmann-Kolss: "Denby on the Distinction between Intrinsic and Extrinsic Properties". Mind, forthcoming

Kremer, P.: "Dunn's Relevant Predication, Real Properties and Identity". Erkenntnis, 47 (1), 1997. S. 37-65

Lewis, D.: "New Work for a Theory of Universals". Australasian Journal of Philosophy, 61(4), 1983. S. 343-377

Shoemaker, S. (1980), 'Causality and Properties'. In: P. *Van Inwagen (Hrsg.),* Time and Cause. Essays Presented to Richard Taylor. Reidel. S. 109-136; quoted from: *S. Shoemaker*, Identity, Cause, and Mind. Cambridge University Press, Cambridge, 1984. S. 206-233
Shoemaker, S.: "On What There Are". Philosophical Topics, 16 (1), 1988. S. 201-223

6 Angewandte Ethik

Toleranz – unproblematisch, aber uninteressant?[1]

Christine Bratu

Abstract/Zusammenfassung

The concept of toleration is a much debated topic in current political philosophy. I want to focus on two issues of this debate that require further clarification: In the first part of my text I discuss one of the so called *paradoxa of toleration*. The purpose of this is to show that these paradoxa should not be understood as arguments against toleration as such, as some of the authors seem to imply. Rather they should be taken as counterexamples against certain explications of the concept, thus contributing to a better understanding of the notion of toleration.

In the second part I will try to illuminate the *relation between liberalism and toleration*. Many have assumed that there is a close link between these two concepts, but few have taken the trouble to spell out what it could be. I want to argue that tolerant behaviour is just what liberalism consists in: If you are a liberal, you take others to have the right to do whatever they care to undertake, as long as their actions do not infringe the same right of others. This implies that you have to let them do even the most foolish things they have chosen to do, as long as these actions do not prevent others from acting just as freely. But this form of restraining yourself could rightly be called toleration. Thus, the connection between liberalism and toleration is in fact so close (as the latter is simply the way the former manifests itself) that there is no point in arguing for or demanding toleration if you are liberal, since toleration is what you have to display anyways.

Über den Begriff der Toleranz wird aktuell viel debattiert. Mein Beitrag möchte zum einen zeigen, dass ein bestimmter Teil dieser Debatte kritisch betrachtet werden sollte, nämlich die so genannten *Paradoxien der Toleranz*. Diese werden scheinbar von vielen Autoren als Argumente dafür angeführt, Toleranz insgesamt als fruchtbare Idee für die politische Philosophie fallen zu lassen. Meiner Ansicht nach sollten die verschiedenen Paradoxien jedoch lediglich als Gegenbeispiele gegen bestimmte Explikationen des Toleranzbegriffs verstanden werden. Unter dieser Deutung stellen die Paradoxien also nicht Gründe gegen Toleranz per se, sondern lediglich für oder gegen bestimmte Auffassungen dieses Begriffs dar. In der Auseinandersetzung mit der nach meiner Einschätzung wichtigsten Paradoxie, der so genannten Paradoxie der Wahrheitsrelativierung, soll ein Begriff von Toleranz entwickelt werden, der auf die gängigen Gegenbeispiele antworten kann.

In einem zweiten Teil möchte ich das *Verhältnis von Toleranz und Liberalismus* näher beleuchten. Denn obwohl zwischen diesen häufig ein Zusammenhang vermutet wird, wird dieser nur selten genau dargestellt. Ich werde dafür argumentieren, dass liberales Handeln gerade darin besteht, sich tolerant zu verhalten. Denn liberal handelt man, wenn man andere Personen aufgrund ihres individuellen Rechts auf Freiheit in all ihren Überzeugungen und Handlungen gewähren lässt, sofern durch diese keine weiteren Personen in dem selben Recht ein-

1 Für viele wertvolle Anregungen danke ich Paulus Esterhazy, Johann Schulenburg und den Damen und Herren Sektionsteilnehmer, die sich auf der GAP7 in Bremen in Anschluss an meinen Vortrag an der Diskussion beteiligt haben.

geschränkt werden. Ein solches nicht-intervenierendes Verhalten ist aber, ausgehend von der im ersten Teil erarbeiteten Explikation, als tolerantes Handeln zu bezeichnen. Wenn sich liberale Überzeugungen in tolerantem Verhalten manifestieren, ist es also vor dem Hintergrund des Liberalismus überflüssig, Toleranz noch explizit zu fordern.

Einleitung

Rainer Forst hat sein für das hier zu diskutierende Thema maßgebliches Werk von 2003 *Toleranz im Konflikt*[2] genannt, und dieser Titel ist für das Nachdenken über Toleranz in zweierlei Weise programmatisch: Einerseits sind Konflikte der Hintergrund, vor dem Überlegungen zu Toleranz erst relevant werden. Denn gäbe es keine latenten Spannungen in einer Gesellschaft, so müsste man sich auch keine Gedanken darüber machen, wie sich diese friedlich aushalten lassen. Doch genau diesen Zweck, d.h. aufzuzeigen, wie ein friedliches Miteinander trotz gesellschaftlicher Spannungen möglich ist, soll Toleranz erfüllen. Toleranz ist daher insofern ein Konfliktbegriff, als sie ohne gegebene gesellschaftliche Konflikte überflüssig wäre. Andererseits hat Toleranz nicht nur gesellschaftliche Konflikte zum Gegenstand, sondern sie ist auch selbst Gegenstand von Konflikt in der philosophischen Debatte. Ein großer Teil des Streits dreht sich dabei um die Frage, was dafür spricht, tolerant zu sein; d.h. es wird darum gestritten, welche der Gründe, die für Toleranz angeführt werden, überzeugen. Darüber hinaus besteht aber noch ein weiterer Konflikt, welcher der Begründungsfrage vorgelagert zu sein scheint, nämlich der um die so genannten Paradoxien der Toleranz. Die Schwierigkeiten, die im Rahmen dieser Diskussion auftreten, muten dabei so schwerwiegend an, dass man daran zweifeln möchte, ob Toleranz insgesamt eine fruchtbare Idee für die praktische Philosophie sein kann.

Im ersten Teil meiner Ausführungen möchte ich anhand eines Beispiels deutlich machen, welcher argumentative Status den Paradoxien der Toleranz eigentlich zukommt – eine Klärung, die viele Autoren in der Debatte vernachlässigen. Dabei soll gezeigt werden, dass die Paradoxien nicht dafür sprechen, Toleranz per se fallen zu lassen, sondern lediglich Gegenbeispiele gegen bestimmte Explikationen dieses Begriffes sind. Die Paradoxien können Toleranz also nicht insgesamt problematisch erscheinen lassen. Im zweiten Teil, der die zuvor entwickelte Explikation übernimmt, möchte ich das Verhältnis von Liberalismus und Toleranz näher beleuchten. Dabei wird deutlich werden, dass die Forderung nach Toleranz vor dem Hintergrund des Liberalismus zwar begründet, aber zugleich in noch näher darzustellender Weise überflüssig ist. Für den Anhänger des Liberalismus ist es also uninteressant, Toleranz zu fordern.

2 Vgl. Forst 2003.

Zum argumentativen Status der Paradoxien der Toleranz

Ein Blick in die Forschungsliteratur zeigt, dass es mittlerweile Tradition hat, auf die Paradoxien von Toleranz hinzuweisen, und dass sich diese zudem vervielfältigen: In einem Aufsatz von 1981 stellt Karl Popper noch eine solche Paradoxie vor, John Horton nennt 1994 bereits drei und Forst führt schließlich im oben erwähnten Werk fünf an.[3] Dabei macht allerdings keiner der genannten Autoren explizit, welchen argumentativen Status die Paradoxien in ihren jeweiligen Ausführungen haben – d.h. ob diese für oder gegen etwas sprechen sollen und wenn ja, wofür oder wogegen. Doch die prominente Stellung der Paradoxien in der Diskussion legt nahe, sie als prima facie-Argumente gegen Toleranz zu verstehen. Denn in ihnen offenbaren sich Implikationen, die so problematisch sind, dass Toleranz insgesamt unhaltbar erscheint. Man fühlt sich an Platons *Politeia* erinnert: Auch hier decken die so genannten Paradoxien der gerechten Stadt die kontraintuitiven Implikationen von Sokrates' Staatsentwurfes auf. Und diese widersprechen der gängigen Meinung so sehr, dass Sokrates daran zweifelt, ob irgendjemand seine Theorie ernst nehmen, geschweige denn umsetzen können wird. Ein ähnlicher Zweifel scheint auch die Autoren der Toleranzdebatte umzutreiben: So spricht David Heyd von Toleranz als einer „elusive virtue"[4], Bernard Williams spitzt das noch zu einer „impossible virtue"[5] zu und Thomas Scanlon bleibt letztlich nur übrig, die „difficulty of tolerance"[6] zu konstatieren.

Die in den genannten Texten implizit mitschwingende Auffassung der Paradoxien der Toleranz, wonach diese prima facie-Argumente gegen Toleranz darstellen, halte ich für falsch. *Denn meines Erachtens treten die Paradoxien nur auf, wenn man ein noch nicht vollständig entwickeltes Verständnis von Toleranz zugrunde legt.* Doch vor dem Hintergrund eines umfassend und sinnvoll explizierten Toleranzbegriffs verschwinden sie. Um einen solchen zu erlangen, muss man in zwei Schritten vorgehen. Zuerst entwickelt man eine Grundvorstellung von Toleranz, indem man exemplarisch die eigenen begrifflichen Intuitionen befragt. Hiermit ist nichts anderes als das in der analytischen Philosophie übliche Vorgehen gemeint: Anhand von Beispielen und Gegenbeispielen wird getestet, welche Fälle kompetente Sprecher als Instanzen toleranten Handelns bezeichnen würden. Die Eigenschaften, die in diesen Fällen maßgeblich sind, werden als notwendige (und, wenn die Untersuchung erfolgreich abgeschlossen ist, hinreichende) Bedingungen für Toleranz aufgeführt. Dabei dienen sprachliche Intuitionen zwar als notwendiger Ausgangspunkt der Begriffsexplikation, nicht aber als deren unhintergehbare Bewährungsinstanz. Denn Zweck einer solchen Explikation ist nicht, nur festzuhalten, wie gesprochen wird; vielmehr darf der

3 Vgl. Popper 1987, Horton 1994 sowie Forst 2003, S. 30- 41.
4 Vgl. Heyd 1996.
5 Vgl. Williams 1996.
6 Vgl. Scanlon 2003.

Sprachgebrauch teilweise kritisiert werden, wenn eine leicht modifizierte Verwendung des fraglichen Begriffs zu neuen interessanten Ergebnissen führt. Eine Begriffsexplikation dieser Art strebt also ein *reflective equilibrium* zwischen dem Sprachgebrauch kompetenter Sprecher und wissenschaftlicher Fruchtbarkeit an.[7]

Eine solche Explikation gelangt meiner Ansicht nach zu folgendem Verständnis von Toleranz: In einer Situation A, in der zwei Akteure S_1 und S_2 bezüglich eines Sachverhaltes unterschiedliche Meinungen vertreten, ist S_1 genau dann tolerant gegenüber S_2 wenn S_1

(1) die Überzeugung von S_2 aus Gründen$_1$ für falsch hält
(2) glaubt, effektiv dagegen vorgehen zu können, dass S_2 ihrer Überzeugung durch Äußerungen oder Taten Ausdruck verleiht
(3) aufgrund von (1) und (2) davon überzeugt ist, dagegen vorgehen zu müssen, dass S_2 ihrer Überzeugung Ausdruck verleiht, und zu einem solchen Handeln auch motiviert ist
(4) aus Gründen$_2$ aber davon Abstand nimmt, S_2 daran zu hindern, ihrer Überzeugung Ausdruck zu verleihen
(5) dabei jedoch ihre ablehnende Haltung gegenüber den Überzeugungen von S_2 beibehält.[8]

Um davon sprechen zu können, dass S_1 tout court tolerant und dies nicht nur in der Situation A gegenüber S_2 gewesen ist, muss S_1 sich in Situationen, die A in relevanter Hinsicht ähnlich sind, verlässlich auf die explizierte Art und Weise verhalten. Toleranz ist also eine Tugend, da sie als Charaktereigenschaft nur derjenigen Person zugeschrieben werden kann, die habituell tolerant handelt.

Die Bedingungen (1) und (5) müssen gefordert werden, da man in Fällen, in denen S_1 keinen Anstoß an Überzeugungen oder Handlungen von S_2 nimmt, intuitiv nicht davon sprechen würde, dass S_1 diese toleriert, wenn sie nicht dagegen vorgeht. Nicht-intervenierendes Verhalten von S_1 wäre in solchen Fällen treffender als Indifferenz oder sogar als Akzeptanz zu beschreiben. Fordert man dagegen (2) nicht, könnte beispielsweise Sklaven Toleranz gegenüber ihren Herren attestiert werden, wenn sie sich mangels effektiver Widerstandsmöglichkeiten nicht gegen diese wehren. Wenn S_1 das Verhalten von S_2 aber nur deswegen

7 Auf diesen Begriff hat John Rawls das beschriebene Vorgehen gebracht (vgl. Rawls 1999, S. 18 (§4)). Dabei stützt sich Rawls allerdings auf die Ausführungen, die Nelson Goodman dazu gemacht hat, wie eine philosophische Explikation aussehen sollte: „While explication must respect the presystematic application of terms, it need not reflect the manner or order of their presystematic adoption; rather it must seek maximum coherence and articulation." (Goodman 1965, S. 47)

8 Die Kennzeichnung „Gründe1" und „Gründe2" soll deutlich machen, dass es sich um *inhaltlich verschiedene* Überzeugungen handelt: Gründe1 sprechen dafür, dass S1 dagegen vorgeht, dass S2 ihren Überzeugungen Ausdruck verleiht; Gründe2 sprechen dafür, dass S1 eine Intervention in das Verhalten von S2 unterlässt. Die gleiche Überzeugung kann nicht für und gegen eine Intervention in das Verhalten von S2 sprechen.

hinnimmt, weil es nicht in der Macht von S_1 steht, dieses zu ändern, ist es intuitiver, von „erdulden“ oder „ertragen“ zu sprechen als von „tolerieren“.[9] Ohne (3) wären auch Fälle praktischer Irrationalität Beispiele für tolerantes Verhalten. Doch es scheint umpassend, Toleranz auf Phlegma zu gründen. Schließlich hätte Toleranz ohne (4) nichts mehr damit zu tun, eine andere Person in ihrem Handeln gewähren lassen. Und wenngleich dies keine hinreichende Bedingung für Toleranz ist, so doch sicherlich eine notwendige. Ingesamt beinhaltet diese Explikation die für Toleranz typische Ablehnungskomponente (denn S_1 lehnt die Überzeugung von S_2 ab und wünscht zu verhindern, dass S_2 diese auslebt) ebenso wie die ebenfalls charakteristische Akzeptanzkomponente (da S_1 trotz ihrer ablehnenden Haltung nicht in das Handeln von S_2 interveniert) – selbst wenn diese Akzeptanz ihre Grenzen hat (nämlich dort, wo die Gründe$_1$, die für das Vorgehen gegen die Überzeugungen von S_2 sprechen, nicht mehr von den Gründen$_2$, die dazu drängen, S_2 in ihren Überzeugungen gewähren zu lassen, überwogen werden).[10]

Auf dieser Ebene einer ersten Explikation treten nun die Paradoxien der Toleranz auf. Dies möchte ich anhand der so genannten Paradoxie der Wahrheitsrelativierung[11] veranschaulichen: Grundlegend für diese ist die korrekte Feststellung, dass eine tolerante Handlung im soeben dargelegten Sinne impliziert, dass Gründe$_1$, die S_1 für ihre Ablehnung einer bestimmten Überzeugung und für das Vorgehen gegen diese hat, von Gründen$_2$ ausgestochen und daher nicht mehr handlungswirksam werden. Des Weiteren wird angenommen, dass die einzigen Gründe$_2$, die Gründe$_1$ auf diese Art ausstechen können, solche sind, die gegen die Korrektheit der Gründe$_1$ sprechen; d.h. Gründe$_2$ müssen Zweifel an Gründen$_1$ wecken. Ausgehend von diesen beiden Prämissen kommt man zu dem – in der Literatur als paradox bezeichneten – Resultat, dass beides zusammen nicht möglich ist: Toleranz gegenüber den Überzeugungen anderer und Anspruch auf Wahrheit bezüglich der eigenen Überzeugungen. Diese allgemeinen Bemerkungen lassen sich an einem konkreten Fall verdeutlichen: Wenn S_1 nicht dagegen vorgeht, dass sich S_2 gemessen an den Überzeugungen von S_1 bezüglich Höflichkeit schlecht benimmt, so geschieht dies nach der dargestellten Paradoxie nur deswegen, weil S_1 zudem über Gründe verfügt, die sie an der Stichhaltigkeit ihrer eigenen Überzeugungen über Höflichkeit und Etikette zweifeln lassen. Demnach würde S_1 das Verhalten von S_2 tolerieren, weil S_1 die Wahrheitsansprüche, die sie für ihre Überzeugungen über gutes Benehmen erhebt, relativiert.

Doch die Erwiderung auf diese Paradoxie liegt auf der Hand: Warum sollten Gründe$_2$, die S_1 dazu bewegen, nicht gegen die Position ihres Gegenübers vor-

9 Forst weist darauf hin, dass man während der Stoa und des frühen Christentums unter „Toleranz“ tatsächlich das Erdulden von Unabänderlichem verstand (vgl. Forst 2003, S. 54f.). Insofern hat sich unser aktuelles Begriffsverständnis von diesen Wurzeln entfernt.

10 Vgl. Forst 2003, S. 32- 37 sowie Forst 2007 und Cohen 2004.

11 Vgl. Forst 2003, S. 37.

zugehen, notwendigerweise solche sein, die Zweifel an Gründen$_1$ wecken? Andere Gründe$_2$ sind denkbar, die nicht die Relativierung der Wahrheitsansprüche, die S_1 mit ihren Gründen$_1$ verbindet, mit sich bringen: So könnte beispielsweise das Gebot, andere Personen in ihrer Freiheit zu respektieren, als guter Grund dafür dienen, nicht in das Handeln von S_2 einzugreifen. Erkennt S_1 die Pflicht, andere in ihrer Freiheit zu respektieren, als guten Grund an, so wird sie Abstand davon nehmen, in die Handlungen von S_2 zu intervenieren (außer in den Fällen, in denen S_2 durch ihre Handlungen selbst die Freiheit anderer missachtet) – und zwar ohne ipso facto daran zu zweifeln, dass sich S_2 gerade fürchterlich benimmt.

Diese Erwiderung offenbart zum einen, dass obige erste Explikation des Toleranzbegriffs unvollständig war und in einem weiteren Schritt ergänzt werden muss. Denn um ein korrektes Verständnis von Toleranz zu erlangen, muss auch expliziert werden, welche Überzeugungen als Grund, sowohl für die Ablehnung bestimmter Überzeugungen (also Gründe$_1$), vor allem aber dafür, nicht gegen diese vorzugehen (also Gründe$_2$), zulässig sind. Zum anderen wird aber deutlich, dass vor dem Hintergrund einer geeigneten begrifflichen Präzisierung die paradoxe Implikation, dass tolerantes Verhalten die Relativierung der eigenen Wahrheitsansprüche mit sich bringt, nicht bestehen bleiben muss. Insofern sollte man die Paradoxie der Wahrheitsrelativierung nicht als ein prima facie-Argument gegen Toleranz verstehen, das offenlegt, wie viel uns tolerantes Verhalten kosten würde, sondern lediglich als Argument dafür, das angenommene Verständnis des Begriffes entsprechend anzupassen und zu verfeinern.[12] Damit kommt den Paradoxien der Toleranz der selbe argumentative Status zu wie etwa den unzähligen Gegenbeispielen in der erkenntnistheoretischen Debatte um den Begriff des Wissens, d.h. sie dienen der genaueren Begriffsbestimmung. Jahrzehnte der *gettierology* haben kaum jemanden dazu veranlasst, Wissen für unmöglich zu halten und zum erkenntnistheoretischen Skeptiker zu werden. Ebenso sollten die Paradoxien der Toleranz, wenn man sie wie dargelegt als Gegenbeispiele gegen bestimmte Explikationen des Begriffs versteht, nicht als Argumente gegen Toleranz gewertet werden, sondern als Anregungen zu einer bestimmten begrifflichen Auffassung.

12 Das gilt auch für die weiteren Paradoxien der Toleranz, selbst wenn dies hier nicht ausgeführt werden kann. Tatsächlich stellt die Paradoxie der Wahrheitsrelativierung schon vor dem Hintergrund einer ersten Explikation kein echtes Problem dar. Denn nicht-intervenierendem Verhalten, das darauf zurückgeht, dass der Akteur an der Stichhaltigkeit seiner eigenen Überzeugungen zweifelt, fehlt die für Toleranz charakteristische Ablehnungskomponente. Dennoch erschien mir eine detaillierte Diskussion dieser Paradoxie wichtig, weil gerade sie Toleranz nachhaltig in Verruf zu bringen droht. Denn in vielen politischen Debatten schwingt die in dieser Paradoxie zum Ausdruck kommende Angst mit, durch tolerantes Verhalten die eigenen Überzeugungen und Werte implizit aufzugeben.

Die Ortlosigkeit von Toleranz im Liberalismus

Um ein sinnvolles, gegen die gängigen Paradoxien gewappnetes Verständnis von Toleranz zu gewinnen, muss man also ausführen, was dafür spricht, sich wie in obiger Explikation skizziert zu verhalten. So argumentiert John Stuart Mill in *On Liberty* utilitaristisch: Man soll andere Personen deswegen falsche Überzeugungen äußern und verfehlte Lebenswege einschlagen lassen, weil es die gesamtgesellschaftliche Nutzensumme letztlich am meisten befördert, wenn niemand in seiner Autonomie eingeschränkt wird. Nach Mill ist es also der Nutzen autonomer Entscheidungen, der als $Grund_2$ dafür spricht, von Interventionen in das Verhalten anderer Personen abzusehen.[13] Die klassisch liberale Argumentation für Toleranz beruft sich dagegen nicht auf deren vorteilhafte Folgen, sondern auf das individuelle Recht auf Freiheit: Dieses Recht kommt jeder Person zu und impliziert, dass ihre Handlungen respektiert werden müssen und nicht behindert werden dürfen, sofern durch sie kein anderes Individuum in seinem Recht auf Freiheit eingeschränkt wird. Auf die Frage, warum man eine Person in einer Überzeugung oder Handlung gewähren lassen sollte, die man begründeterweise für falsch hält, würde der Liberale demnach als $Grund_2$ anführen: Weil jede Person – unabhängig davon, ob die Überzeugungen, die sie äußert, oder die Handlungen, die sie tut, richtig oder falsch sind – das Recht hat, frei zu handeln, solange dadurch kein anderes Individuum in seinem gleichen Recht beschnitten wird. Sowohl Rawls als auch Forst führen das individuelle Recht auf Freiheit als $Grund_2$ für tolerantes Verhalten an, und auf dieser Überlegung basierte auch obige Abweisung der Paradoxie der Wahrheitsrelativierung.[14]

---

13 Vgl. Mill 2004. Es sollte allerdings zugestanden werden, dass Mills Toleranzbegründung nicht typisch utilitaristisch, weil nicht offensichtlich konsequentialistisch sind. Denn scheinbar argumentiert Mill deontologisch, indem er sich auf das so genannte Schadensprinzip beruft, wonach „der einzige Zweck, um dessentwillen man Zwang gegen den Willen eines Mitglieds einer zivilisierten Gemeinschaft rechtmäßig ausüben darf, der ist: die Schädigung anderer zu verhüten." (Mill 2004, S. 16) Doch diesem Prinzip kommt keine intrinsische Geltung zu, sondern es fasst lediglich die Einsicht zu einer Faustregel zusammen, dass es für die gesamtgesellschaftliche Nutzensumme immer förderlicher ist, Autonomie zuzulassen als sie einzuschränken. Der deontologisch anmutenden Argumentation liegt demnach doch eine Folgenabwägung zugrunde. Aber auch diese ist nicht typisch utilitaristisch. Denn dadurch, dass Mill davon ausgeht, dass ein bestimmter Handlungstyp (nämlich autonome Handlungen) immer die besten Folgen zeitigt, bricht er letztlich mit den empiristischen Wurzeln des Utilitarismus. Eigentlich müsste nämlich von Fall zu Fall geprüft werden, welche Handlungsalternative der gesamtgesellschaftlichen Nutzensumme am förderlichsten ist.

14 Nach Rawls lässt sich Toleranz als Handlungsregel für politische Akteure aus dem Urzustand deduzieren, d.h. die Parteien würden sich im Urzustand auf diese Handlungsregel einigen (vgl. Rawls 1999, S. 181). Der Urzustand ist aber nichts anderes als eine Operationalisierung der Überzeugung, dass Individuen frei und gleich behandelt werden sollten. Von diesem Recht auf Freiheit und Gleichheit spricht Forst nicht, stattdessen von einem

Hat man geklärt, welche Erwägungen als Gründe$_2$ für Toleranz sprechen, so gilt es zudem festzulegen, an wen sich diese Gründe richten, d.h. wer der Adressat der wie auch immer begründeten Aufforderung „Sei tolerant!“ ist. Implizit finden sich in der Literatur zwei Ansichten zu dieser Frage, die ich als die *politische* und die *moralische Auffassung von Toleranz* einführen möchte. Nach der moralischen Auffassung richtet sich die Aufforderung zur Toleranz an alle Individuen qua Individuen. Nach der politischen Auffassung sind dagegen lediglich politische Akteure Adressat dieser Aufforderung. Dabei umfasst die Menge der politischen Akteure nicht nur politische Institutionen (wie etwa die Legislative oder den obersten Gerichtshof), sondern auch alle Individuen qua Bürger. Ein Vertreter der politischen Auffassung von Toleranz ist beispielsweise Thomas Nagel. In *Equality and Partiality* argumentiert er dafür, dass politische Akteure es tolerieren müssen, wenn Menschen Lebensentwürfen anhängen, die sie für falsch halten, die aber niemandem schaden. Daher sollten die Bürger beispielsweise keine Partei wählen, die solche alternative Lebensentwürfe verbieten will.[15] Und auch John Lockes *Letter Concerning Toleration* ist eine Streitschrift für politische Toleranz, denn Locke versucht hier die staatliche Obrigkeit davon zu überzeugen, dass sie es zulassen sollte, wenn einige ihrer Bürger einer Minderheitenreligion angehören.[16] Die moralische Auffassung von Toleranz – also die Aufforderung, weder als Bürger (also etwa bei der Stimmabgabe in der Wahlkabine), noch als Privatperson (d.h. im täglichen Umgang miteinander) gegen bestimmte Formen des Fehlverhaltens anderer vorzugehen – findet sich dagegen u.a. bei Mill.

Vor dem Hintergrund dieser begrifflichen Klärungen möchte ich nun abschließend das Verhältnis von Toleranz und Liberalismus betrachten. Zwischen beiden wird häufig eine enge Verbindung vermutet, allerdings wird selten deutlich gemacht, wie man sich diese genau vorzustellen hat. Eine mögliche Art, die beiden zueinander in Beziehung zu setzen, wurde bereits dargestellt: Der Liberalismus kann durch seine Kernaussage, dass dem Einzelnen ein Recht auf Freiheit zukommt, einen Grund für tolerantes Verhalten liefern. Da der Liberalismus als argumentatives Fundament von Toleranz dienen kann, liegt es nahe, Toleranz als eine Tugend aufzufassen, die sich aus dem Liberalismus ableiten lässt und die der Anhänger des Liberalismus erwerben sollte. So verstanden ist Toleranz ein Habitus, der zum Liberalismus hinzutreten kann und für den liberale Autoren

---

Recht auf Rechtfertigung, nach dem das Individuum nur rechtfertigbaren Zwängen unterworfen werden darf (vgl. Forst 2003, S. 592). Vertreter des *Justificatory Liberalism* wie Gerald Gaus argumentieren aber dafür, Freiheit gerade in diesem Sinne zu verstehen: Nur Zwängen unterworfen zu sein, denen man begründeterweise zustimmen könnte (vgl. Gaus 1996, S. 165). Ein Recht auf Rechtfertigung ist demnach gleichbedeutend mit einem Recht auf Freiheit.

15 Vgl. Nagel 1995, S. 154- 168.

16 Vgl. Locke 1996.

argumentieren sollten. Autoren wie Locke und Rawls, die sich zu beiden Themen einflussreich geäußert haben, verleihen dieser Behauptung eine gewisse Plausibilität: Nach Rawls soll sein liberales Programm der Gerechtigkeit als Fairness auch „an account of certain political virtues – [...] such as the virtues of civility and tolerance“[17] beinhalten. Und Locke argumentiert in einem eigenständigen Werk für Toleranz und tut dies größtenteils mit anderen Argumenten als denen, die im *Second Treatise* für den Liberalismus im Allgemeinen angeführt werden.[18]

Im Folgenden möchte ich gegen diese Auffassung argumentieren und stattdessen die These vertreten: *Toleranz ist nicht eine Tugend, die zum Liberalismus hinzutreten kann und für die argumentiert werden muss – vor dessen Hintergrund ist es vielmehr überflüssig, Toleranz zu fordern.* Um diese These zu verteidigen, muss ich zuerst eine weitere Unterscheidung treffen. Wie bereits erwähnt kennzeichnet den Liberalismus, dass er dem Einzelnen ein Recht auf Freiheit zuspricht. Dabei wird dieses Recht von den verschiedenen Autoren unterschiedlich ausbuchstabiert in Abhängigkeit von ihrer jeweiligen Auffassung von Freiheit: Locke legt einen negativen Freiheitsbegriff zugrunde, so dass für ihn das individuelle Recht auf Freiheit darin besteht, dass keine Person in ihrer körperlichen Unversehrtheit verletzt oder ihres rechtmäßigen Eigentums beraubt werden darf.[19] Scanlon geht dagegen von einem positiven Freiheitsbegriff auf, so dass das Recht auf Freiheit in seiner Version ein Recht darauf ist, nur solchen Interventionen in das eigene Handeln zu unterliegen, die man nicht vernünftigerweise ablehnen könnte.[20] Doch um obige Behauptung zu verteidigen ist weniger entscheidend, wie die verschiedenen liberalen Autoren Freiheit (und im Anschluss daran das Recht auf Freiheit) inhaltlich genau fassen, sondern wem durch dieses Recht Pflichten erwachsen: Der *moralische Liberalismus* – wie ich ihn in Anlehnung an obige Binnendifferenzierung von Toleranz nennen möchte – postuliert für alle Akteure die Pflicht, Personen in ihrer Freiheit zu respektieren. Der *politische Liberalismus* beschränkt sich dagegen darauf, diese Pflicht politischen Akteuren aufzuerlegen. Während der moralische Liberalismus also fordert „Respektiere die Freiheit eines jeden, sowohl insofern du als politischer Akteur auftrittst als auch in Fragen des alltäglichen Miteinanders!“, lässt sich

17 Rawls 2005, S. 194.

18 Die Argumente, die Locke im *Letter Concerning Toleration* anführt, sind vor allem, dass man sich einerseits nicht sicher sein kann, ob man mit den eigenen religiösen Überzeugungen tatsächlich richtig liegt (vgl. Locke 1996, S. 33), und dass man authentischen Glauben ohnehin nicht erzwingen kann (vgl. Locke 1996, S. 15).

19 Vgl. Locke 2006, S. 203.

20 Vgl. Scanlon 2000, S. 153. Allerdings führt Scanlon dieses Recht nicht als das Recht auf Freiheit ein, sondern als Grundprinzip des Kontraktualismus. Doch wie schon in Fußnote 14 ausgeführt, kann man das Recht, nur rechtfertigbarem Zwang ausgesetzt zu sein, als Recht auf Freiheit verstehen. Damit wird der Kontraktualismus zu einer Variante des Liberalismus.

aus dem politischen Liberalismus lediglich die Forderung „Respektiere die Freiheit eines jeden, insofern du als politischer Akteur auftrittst!“ ableiten. Nach dieser Bestimmung gehört Rawls klarerweise zum Lager des politischen Liberalismus, und zwar nicht erst seit der Veröffentlichung des gleichnamigen Werkes. Denn schon die *Theory of Justice* verhandelt Fairness, also die Achtung vor der Freiheit und Gleichheit von Personen, als „first virtue of social institutions“[21]. Ebenso sollte Lockes *Second Treatise* als Werk des politischen Liberalismus verstanden werden, da liberale Verhaltensrichtlinien hier primär für den *body politick* aufgestellt werden. Scanlon hingegen ist ein Verteter des moralischen Liberalismus, denn er betrachtet nicht nur, was politische Akteure ihren Bürgern schulden, sondern sein Interesse gilt grundsätzlich dem, *what we owe to each other*.

Vor dem Hintergrund eines solchen moralischen Liberalismus lässt sich einfach für Toleranz argumentieren: Dass eine Person ein Recht auf Freiheit hat, spricht dafür, sie in ihren abwegigsten Handlungen gewähren zu lassen, solange diese niemand anderen im gleichen Recht einschränken. Gleichzeitig ist es aber überflüssig, von einem Liberalen dieser Art Toleranz zu fordern. Denn was sollte die Forderung nach moralischer Toleranz – also die Forderung „Sei tolerant, d.h. lass den anderen in Überzeugungen und Handlungen gewähren, die du für falsch hältst, sofern durch diese Überzeugungen und Handlungen keine weitere Person in ihrer Freiheit eingeschränkt wird!“ – noch für theoretische und praktische Arbeit leisten, die nicht schon durch die Forderung „Sei liberal, d.h. respektiere die Freiheit des anderen zu tun, was immer er will, sofern durch seine Handlungen keine weitere Person in ihrer Freiheit eingeschränkt wird!“ abdeckt ist? Ebenso findet sich für die Forderung nach politischer Toleranz im Rahmen des moralischen Liberalismus kein eigenständiger Ort. Denn politische Toleranz ist nur ein Sonderfall moralischer Toleranz, da erstere lediglich den Kreis derer einschränkt, an die das Toleranz-Gebot ergeht. Die Grundannahme des moralischen Liberalismus, dass sich Personen wechselseitig in ihrer Freiheit respektieren müssen, hat aber bereits die Forderung nach moralischer Toleranz überflüssig gemacht. Also muss auch die Forderung nach politischer Toleranz *a fortiori* in diesem Sinne ortlos werden. Die Forderung „Sei als politischer Akteur tolerant und wähle etwa keine Partei, die bestimmte Lebensentwürfe verbieten möchte, obwohl diese niemanden in seiner Freiheit einschränken!“ erschöpft sich demnach vor dem Hintergrund des moralischen Liberalismus ebenfalls in dem Hinweis „Sei liberal!“.

Die vorangegangenen Überlegungen lassen sich folgendermaßen zusammenfassen: Toleranz muss mit Gründen unterfüttert werden, um ein sinnvolles Konzept zu sein. Der moralische Liberalismus kann eine solche Toleranzbegründung leisten, denn dass der Einzelne ein Recht auf Freiheit hat spricht dafür, ihn in al-

21 Vgl. Rawls 1999, S. 3 (§1).

len Handlungen gewähren zu lassen, die die Freiheit anderer nicht einschränken, selbst wenn diese Handlungen in anderer Hinsicht falsch erscheinen. Doch vor dem Hintergrund eines moralischen Liberalismus wird die Forderung nach Toleranz überflüssig. Denn alles, was dadurch gefordert wird, wird bereits durch die Forderung, sich liberal zu verhalten und also die Freiheit des Individuums zu respektieren, abgedeckt. *Insgesamt wird deutlich, dass Toleranz nur eine Umschreibung dessen ist, was es bedeutet, im Sinne des moralischen Liberalismus zu handeln.* Wollte man beschreiben, wie ein Mensch handelt, der die Grundannahmen dieses Liberalismus beherzigt, könnte man sagen, dass er andere in all ihren Überzeugungen und Handlungen gewähren lässt, solange dadurch keine weiteren Personen in ihrer Freiheit eingeschränkt werden – d.h. also (nach obiger Explikation) dass er sich tolerant verhält. Damit ist aber obige These bestätigt: Toleranz kann nicht zum liberalen Handeln hinzutreten und muss insofern nicht erst explizit gefordert werden. Denn liberales Handeln besteht in Toleranz.

Vor dem Hintergrund eines rein politischen Liberalismus wird Toleranz nicht in vergleichbarer Art ortlos: Zwar geht die Forderung nach politischer Toleranz in analoger Weise in politischer Liberalität auf. Denn als politischer Akteur tolerant zu agieren und etwa nicht durch den Erlass von Gesetzen in die Überzeugungen und Handlungen seiner Mitbürger zu intervenieren, solange durch diese keine andere Person in ihrer Freiheit beschränkt wird, ist nichts anderes als sich qua politischer Akteur liberal zu verhalten. Doch die umfassenderen Ansprüche der moralischen Toleranz, nämlich auch als Privatperson tolerant zu sein, werden von dieser Form des Liberalismus nicht abgedeckt. Moralische Toleranz behält also vor dem Hintergrund eines politischen Liberalismus die Stellung einer eigenständigen, nicht auf das liberale Grundprinzip reduzierbaren Forderung.

Doch diese Eigenständigkeit ist teuer erkauft: Die Tatsache, dass sich der politische Liberalismus auf das Freiheitsrecht des Einzelnen gegenüber politischen Akteuren konzentriert, verbürgt zwar, dass die Forderung nach moralischer Toleranz nicht redundant wird. Dadurch kann diese Forderung aber auch nicht mehr mithilfe der argumentativen Ressourcen des politischen Liberalismus eingeholt werden. Denn dass jeder das Recht hat, nicht von politischen Akteuren behelligt zu werden, solange seine Handlungen nicht die Freiheit anderer gefährden, spricht nicht ohne weiteres dafür, dass jeder auch das Recht hat, von den Interventionen anderer Akteure verschont zu bleiben. Will der politische Liberale die Forderung nach moralischer Toleranz untermauern, kann er dafür also nicht auf seine Grundannahme, d.h. auf das Recht auf Freiheit gegenüber politischen Akteuren zurückgreifen, sondern muss sich gewissermaßen externer Ressourcen bedienen.

Kritische Geister fassen dieses Ergebnis als Grund dafür auf, den politischen Liberalismus insgesamt zu verwerfen: Warum sollte man dessen Programm politischer Ordnung akzeptieren, wenn sich vor seinem Hintergrund nicht einmal die Aufforderung, anderen gegenüber auch im täglichen Miteinander tolerant

aufzutreten, begründen lässt? Stellt die zugrundeliegende Einschränkung des Untersuchungsgegenstandes – also der Fokus auf politische Akteure, statt auf Akteure im Allgemeinen – nicht ohnehin eine Verkürzung des liberalen Gedankens dar? Mir scheint diese kritische Haltung berechtigt, und ich möchte mit einer kurzen Stellungnahme dazu schließen.

Zuerst ist darauf hinzuweisen, dass die genannten Kritikpunkte vom prominentesten Vertreter des politischen Liberalismus nicht geleugnet, aber auch nicht als Kritik verstanden werden würden. Denn für Rawls ist die dargestellte Einschränkung des Untersuchungsgegenstandes unumgänglich, da die Moderne zudem durch das fortschreitende Bewusstsein einer weiteren Einschränkung gekennzeichnet ist: Mittlerweile ist allgemein bekannt, dass die menschliche Vernunft hinsichtlich der Frage, was richtig ist, bestimmten Bürden des Urteilens unterliegt, und dass es daher einen berechtigten Pluralismus der Antworten auf diese Frage gibt. Angesichts dieses berechtigten Pluralismus ist kaum zu erwarten, dass es auf die umfassende Frage, was sich die Menschen qua Menschen in ihrem wechselseitigen Miteinander schulden, eine Antwort geben wird, die vor allen Betroffenen begründet ist. Beschränkt man sich aber auf ein Programm dessen, was politische Akteure ihren Bürgern schulden, steigen die Chancen auf eine allgemein begründete Antwort. Und lässt man zudem zu, dass die einzelnen Streitparteien die vorgeschlagene Theorie nicht aus den gleichen Gründen für angemessen halten, sondern ihre jeweils partikularen Gründe anführen – d.h. fordert man für eine *political conception* nicht, dass sie auf einer allgemein überzeugenden Begründung fußt, sondern lediglich, dass sie als freistehendes Modul von einem *overlapping consensus* der *comprehensive moral doctrines* getragen wird – so kann man hoffen, zumindest einen für westliche Demokratien überzeugenden Vorschlag zu machen. Nach Rawls ist die kritisierte Einschränkung also notwendig, um vor dem Hintergrund eines vernünftigen Pluralismus überhaupt für eine normative Theorie politischer Ordnung argumentieren zu können.[22] Dass die so entstehende normative Theorie moralische Toleranz nicht begründen kann, fällt demnach nicht schwer ins Gewicht: Unter den Bedingungen der Moderne darf man nichts anderes erwarten. Zudem kann darauf hoffen, dass die *comprehensive moral doctrines*, die den politischen Liberalismus stützen, argumentative Ressourcen bereitstellen, um die Forderung nach moralischer Toleranz zu untermauern.

Diese Erwiderung scheint mir aus zweierlei Gründen unbefriedigend: Zum einen bietet der politische Liberalismus nach Rawls (von einigen historischen Überlegungen abgesehen) kein stichhaltiges Argument dafür an, warum sich der vernünftige Pluralismus – der eine allgemein akzeptierbare Theorie dessen, was sich Menschen qua Menschen schulden, unmöglich macht – nicht auch auf die Frage erstrecken sollte, was politische Akteure ihren Bürgern schulden. Warum

22 Diese Zusammenhänge werden von Rawls ausführlich dargestellt in Rawls 2003, Part I.

sollte das Problem, wie sich der Staat und seine Bürger zu verhalten haben, weniger umstritten sein als die Frage, wie sich Individuen untereinander behandeln sollten? Diese Frage wird umso dringlicher, wenn man sich zum anderen vor Augen führt, dass auch Rawls mit seiner Theorie zumindest implizit das wechselseitige Miteinander zu regeln versucht. Denn sein Grund dafür, nur die politische Grundstruktur zu betrachten, ist neben der genannten Notwendigkeit, dem vernünftigen Pluralismus Rechnung zu tragen, der, dass „major institutions define men's rights and duties and influence their life prospects, what they can expect to be and how well they can hope to do. [...] its effects are so profound and present from the start".[23] D.h. Rawls ist sich der Auswirkungen, die von politischen Akteuren gestaltete Rahmenbedingungen haben, durchaus bewusst. Diese erstrecken sich auch in den Bereich dessen, was sich Individuen wechselseitig qua Individuen schulden. Wenn beispielsweise im Grundgesetz festgehalten wird, dass die Würde des Einzelnen nicht anzutasten ist, erwachsen daraus nicht nur dem Staat Pflichten gegenüber seinen Bürgern, sondern auch den Individuen in ihrem Umgang untereinander. Durch die Bestimmung dessen, was politische Akteure ihren Bürgern schulden, lassen sich also mittelbar auch Regeln für das wechselseitige Miteinander der Individuen qua Privatpersonen bestimmen – nämlich dadurch, dass politische Akteure den Rahmen für deren Umgang miteinander festlegen.

Implizit scheint also auch der politische Liberalismus davon auszugehen, dass Regeln des wechselseitigen Miteinanders wie beispielsweise die der moralischen Toleranz allgemein zu rechtfertigen sind. Damit nähert sich der politische Liberalismus implizit wieder dem moralischen an – wodurch die Forderung nach moralischer Toleranz zwar wieder begründbar, aber eben auch wieder in der dargestellten Weise überflüssig wird.

Literaturverzeichnis

Cohen, Andrew: „What Toleration Is". Ethics 1, 2004. S. 67-95

Forst, Rainer: Toleranz im Konflikt. Geschichte, Gehalt und Gegenwart eines umstrittenen Begriffs. Suhrkamp, Frankfurt a.M., 2003

Forst, Rainer: „Toleration". In: Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/toleration/. 2007

Gaus, Gerald: Justificatory Liberalism. An Essay on Epistemology and Political Theory. Oxford University Press, New York/ Oxford, 1996

23 Rawls 1999, S. 6f. (§2).

Goodman, Nelson: Fact, Fiction, and Forecast. Bobbs-Merrill, Indianapolis et al., 1965

Heyd, David: „Introduction". In: *Heyd, David (Hrsg.)*: Toleration – An Elusive Virtue. Princeton University Press, Princeton, 1996. S. 3-17.
Horton, John: „Three (Apparent) Paradoxes of Toleration". Synthesis Philosophica, 17, 1994. S. 7-20

Locke, John: Ein Brief über Toleranz. Übersetzt, eingeleitet und in Anmerkungen erläutert von Julius Ebbinghaus. Englisch – Deutsch. Meiner, Hamburg, 1996
Locke, John: Zwei Abhandlungen über die Regierung. Herausgegeben und eingeleitet von Walter Euchner. Suhrkamp, Frankfurt a.M., 2006

Mill, John Stuart: Über die Freiheit. Aus dem Englischen übersetzt von Bruno Lemke. Mit Anhang und Nachwort herausgegeben von Manfred Schlenke. Reclam, Stuttgart, 2004

Nagel, Thomas: Equality and Partiality. Oxford University Press, New York/ Oxford, 1995

Popper, Karl: „Toleration and Intellectual Responsibility". In: *Mendus, Susan/ Edwards, David (Hrsg.):* On Toleration. Oxford University Press, Oxford, 1987. S. 17-34

Rawls. John: A Theory of Justice. Revised Edition, Belknap Press, Harvard, 1999
Rawls, John: Justice as Fairness – A Restatement. Belknap Press, Harvard, 2003
Rawls, John: Political Liberalism. Expanded Edition. Columbia University Press, New York/ Chichester, 2005

Scanlon, Thomas: What We Owe to Each Other. Belknap Press, Harvard, 2000
Scanlon, Thomas: The Difficulty of Tolerance. Essays in Political Philosophy, Cambridge University Press, Cambridge, 2003

Williams, Bernard: „Toleration – An impossible virtue?". In: *Heyd, David (Hrsg.):* Toleration – An Elusive Virtue. Princeton University Press, Princeton, 1996. S. 18-27

Verantwortung ohne Grenzen? Zum Widerstreit zwischen globaler Gerechtigkeit und Freiheit

Henning Hahn
henning.hahn@uni-kassel.de
Institut für Philosophie, Universität Kassel

Abstract/Zusammenfassung

The controversy on global poverty and global justice has lead to discussing boundaries of our responsibilities for global justice. In this paper I will discuss some of the most prominent arguments for limiting these responsibilities. I will thereby criticize current types of arguments for failing to give a plausible justification of such boundaries. In the central part of the paper I discuss Samuel Scheffler's "agent-centered prerogative" in order to underline that we are by no means *justified* to give priority to our personal liberties. I will then conclude that these liberties are simply not generally justifiable, let alone towards those who are seriously deprived. That's why my paper closes with a dilemma: In the absence of a just global order we have to acknowledge that the normative perspectives of freedom and justice fall apart. This means above all that we have no elegant theory at hand for giving priority to one of these norms; hence it also means that our global responsibilities are not objectively limitable.

Innerhalb der gegenwärtigen Debatte um globale Armut und globale Gerechtigkeit wird die Frage nach den Grenzen moralischer Verantwortung neu gestellt. Dieser Beitrag enthält einen kursorischen Durchgang durch einige der wichtigsten Argumente zur Begrenzung globaler Verantwortung. Im Ergebnis zeigt die Erörterung, dass die unterschiedlichen Argumentationsmodelle letztlich ungeeignet sind, Grenzen der Verantwortung zu *rechtfertigen*. In der zentralen Auseinandersetzung mit Samuel Schefflers „agent-centered prerogative" wird vielmehr deutlich, dass es zwar aus der Sicht der verpflichteten Person nachvollziehbar ist, wenn sie sich Freiheiten gegenüber ihren Verantwortlichkeiten herausnimmt, dass diese Freiheiten unter dem Vorzeichen scharfer Ungerechtigkeiten aber niemals allgemein, vor allem nicht gegenüber denjenigen, die unter globaler Armut leiden, zu rechtfertigen sind. Daher schließt der Beitrag mit dem Dilemma, dass die normativen Logiken von Freiheit und Gerechtigkeit unter der Perspektive globaler Ungerechtigkeit im Widerstreit zueinander stehen. Dies bedeutet vor allem, dass wir über keine elegante Lösung für das Priorisierungsproblem verfügen, aber auch, dass unsere globale Verantwortung letztlich nicht aus objektiver Perspektive zu begrenzen ist.

> No plausible argument can justify the use of human resources involved in producing *patê de canard en croûte* against possible alternative beneficent ends to which these resources might be put (Susan Wolf)[24]

Mir geht es in diesem Beitrag vornehmlich darum, ein Problem in der Frage nach den Grenzen unserer Verantwortung für globale Gerechtigkeit genauer zu beleuchten. Eine entsprechende Debatte wird vor allem in Hinsicht auf globale Armut geführt. Ganz zu Anfang dieser Debatte hat Peter Singer auf einer umfänglichen individuellen Mitverantwortung insistiert, die unseren Lebensstil insgesamt in Frage stellt.[25] Singer stellt ein auf den ersten Blick unverfängliches, aber mit Blick auf die Reichweite individueller Verantwortung herausforderndes Moralprinzip auf. Er konstatiert dass, wenn wir a) etwas moralisch Schlechtes verhindern können, ohne b) etwas von vergleichbarer moralischer Wichtigkeit aufs Spiel zu setzen, wir es dann (c) moralisch gesehen auch tun sollten.[26] Und wie wir, so Singers suggestives Beispiel, ein ertrinkendes Kind aus einem Teich retten sollen, sollten wir auch nach geeigneten Wegen suchen, schwere Armut auf der Welt zu lindern, der täglich tausende Kinder zum Opfer fallen.

In seiner Grundstruktur ist Singers Moralprinzip *kein* genuin utilitaristischer Grundsatz, sondern eine formale Definition von Moral, die ebensogut von deontologischen Ansätzen geteilt wird. Utilitaristisch sind allenfalls die Konsequenzen, die Singer aus dem Prinzip zieht, wenn er jedem einzelnen Menschen in die Pflicht nimmt, sich bis zum Grenznutzen – also bis zu dem Punkt, an dem das eigene Leid Überhand nimmt – gegen globale Armut zu engagieren.

An dieser Stelle greifen die bekannten Überforderungsargumente gegen den Utilitarismus, die ich mir an späterer Stelle etwas genauer ansehen möchte. Zunächst werde ich einen kurzen Überblick über einige der wichtigsten Argumente geben, die gegen Singer und für eine Begrenzung unserer Verantwortung gegenüber globaler Armut vorgebracht werden. Dabei zeigt sich, dass Individuen in einer *nicht einzugrenzenden Mitverantwortung* für globale Armut stehen. Diese Verantwortung ist vom Einzelnen kaum zu bewältigen. Trotzdem werde ich dafür argumentieren, dass die eingeführten Überforderungsargumente unter der Voraussetzung globaler Ungerechtigkeit unzureichend sind, um eine Beschränkung unserer individuellen Mitverantwortung gegenüber globaler Armut allgemein zu *rechtfertigen.*

Das positive Resultat dieser kurzen Erörterung liegt in einer Explikation (oder zumindest in einer ersten Skizze) eines Widerstreits zwischen Gerechtigkeit und Freiheit in der globalen Arena. Gerechtigkeit und Freiheit, so die These, entfalten vor dem Hintergrund globaler Ungerechtigkeit widerstrebende normative Logiken. Das Primat der Freiheit, das in den diskutierten Überforderungsar-

24 Zitiert nach Thomas Nagel (1986, S. 191).
25 Peter Singer (1972).
26 Ebd., S. 231.

gumenten und Eingrenzungsversuchen mitschwingt, lässt sich dann nur noch perspektivisch, nämlich aus Sicht derer, deren Freiheit gefährdet ist, rechtfertigen. Aus der Perspektive global deprivierter Personen ist individuelle Freiheit aber kein gültiges Argument gegenüber einem gesamtgesellschaftlichen Primat der Gerechtigkeit.

1. Individuelle Verantwortung für globale Armut: Eingrenzungsversuche

Gesetzt, dass Singers Moralprinzip auf formal unkontroversen Voraussetzungen aufruht, lässt sich immer noch bezweifeln, dass es besonders strenge Formen moralischer Individualverantwortung nach sich zieht. Ich werde hier in aller Kürze einige der wichtigsten Strategien zur Beschränkung unserer Verantwortung gegenüber globaler Armut diskutieren, um zu zeigen, dass sie zwar geeignet sind, Singers utilitaristische Argumentation zu entkräften, dass sie aber gegenüber anderen Konzeptionen, insbesondere gegenüber der Konzeption Thomas Pogges, daran scheitern, eine Eingrenzung unserer individuellen Mitverantwortung für globale Armut zu rechtfertigen.

a) Überforderung: psychologisch, motivational, epistemisch

Bisweilen tritt der Vorwurf einer *psychologischen, motivationalen oder epistemischen Überforderung* als eigenständiges Argument auf. Die Verantwortung für globale Armut übersteigt die Fähigkeiten eines einzelnen Menschen. Ihm fehlt es an Wissen, um die Hintergründe globaler Armut zu ergründen, und an den Mitteln, das Problem zu lösen.

Gegen Singers Entgrenzung individueller Verantwortung wird daher geltend gemacht, dass die Zuschreibung moralischer Verantwortung nur unter der Voraussetzung Sinn macht, dass wir ein Problem durchdringen und etwas zur Problemlösung bewirken können – was im Fall globaler Armut und aus der Sicht einer individuellen Person nicht der Fall ist. Nun ist aber die Tatsache allein, dass Hintergründe und Ursachen eines Problems komplex und kontrovers sind, dass der Einzelne weitgehend ohnmächtig ist und dass die damit verbundenen Pflichten im Prinzip unabschließbar sind, kein Spezifikum des globalen Armutsproblems, sondern ein Merkmal vieler Verantwortungsverhältnisse. Wer sich seinen Kindern gegenüber verpflichtet fühlt, ihnen eine gute Ausbildung zu ermöglichen, trifft auf ebenso kontroverse wie komplexe Modelle kognitiver Lernpsychologie, er ist in hohem Maße auf die Kooperation anderer angewiesen und trägt eine strukturell unabschließbare Verantwortung. Ein anderes Beispiel ist die staatsbürgerliche Verantwortung. Nur weil ein einzelner Mensch aufgrund seiner psychologisch-motivationalen Ausstattung nicht in der Lage ist, seiner

umfänglichen Verantwortung abschließend gerecht zu werden, verfügen wir darum noch nicht über ein normativ hinreichendes Argument, um eine moralische Verantwortung kategorisch zu begrenzen.

b) Moralische Arbeitsteilung

Modelle moralischer Arbeitsteilung stellen einen vielversprechenden Ansatz bereit, um dem Überforderungsproblem Herr zu werden. Hervorzuheben sind hier solche Ansätze, die nationale und institutionelle Zuständigkeiten betonen. Globale Armut fiele demnach in den Verantwortungsbereich der jeweiligen Nation respektive in die Verantwortung zuständiger transnationaler Institutionen wie UNESCO, Weltbank oder auch Oxfam. Die Auffassung, dass globale Armut allein eine Frage nationaler Verantwortung oder eine Frage außerhalb unserer Verantwortung agierender globaler Institutionen ist, hat aber viel an Plausibilität verloren, seit Autoren wir Thomas Pogge herausgearbeitet haben, dass nationale Verantwortungsbereiche nicht unabhängig von globalen Regeln, Praktiken und Institutionen verstanden werden können, und dass diese globale Hintergrundstruktur wiederum nicht außerhalb des Verantwortungsbereichs individueller Akteure in demokratischen und relativ mächtigen Ländern liegt.[27]

Nicht nur stehen wir in der Mitverantwortung für die Unrechtseffekte einer globalen Grundstruktur, von der wir deutlich profitieren, sondern wir tragen auch dort eine Bringschuld für den Aufbau einer gerechtigkeitssichernden globalen Grundstruktur, wo diese ineffizient arbeitet oder gänzlich fehlt. Kurz, die potentielle Entlastungs- und Koordinierungsfunktion von Institutionen bewahrt uns nicht vor einer individuellen Mitverantwortung zum Aufbau und zur Kontrolle solcher Institutionen, wo diese fehlen oder ineffizient arbeiten. Allein die Menschenrechtsperformance der bereits bestehenden und zum Teil in unserem Namen operierenden Institutionen, so Pogges Punkt, legt uns bereits eine persönliche Mitverantwortung auf, die vor dem Hintergrund globaler Armut ebenso dringlich wie aus der Sicht des Einzelnen unabschließbar zu sein scheint.[28]

c) Schwache und starke Pflichten

Eine dritte Strategie, individuelle Verantwortung einzuschränken, besteht darin, zwischen schwachen und starken Pflichten zu unterscheiden. Auf diesem Wege wird zwar eine individuelle Verantwortung gegenüber globaler Armut eingeräumt, in der Regel sind aber andere Verantwortlichkeiten wichtiger – unter Umständen auch die Verantwortung gegenüber unseren je eigenen Projekten.

27 Vgl. Thomas Pogge (2002).
28 Ebd., S. 48-50.

Ein Weg, zwischen schwacher und starker Verantwortung zu unterscheiden, drückt sich in der Differenz zwischen *humanitären Pflichten* und *Pflichten der Gerechtigkeit* aus. Humanitäre Pflichten werden als mehr oder weniger freiwillige Wohlfahrtspflichten definiert, Pflichten der Gerechtigkeit erscheinen dagegen als Pflichten gegenüber den sozialen Regeln, deren Übertretung auch externe Sanktionen durch dritte Parteien, die an der allgemeinen Geltung dieser Regeln interessiert sind, rechtfertigt.[29] Ohne dies an dieser Stelle weiter ausführen zu können, scheint es aber durchaus angemessen und auch empirisch belegbar zu sein, globale Armut nicht zwangsläufig „nur" als eine humanitäre Katastrophe, sondern eben auch als Gerechtigkeitsproblem zu beschreiben, dass aus ungerechten sozialen Regeln resultiert, die zu reformieren wir eine in diesem Sinne starke Gerechtigkeitspflicht hätten.

Zusätzlich stützt sich die Unterscheidung zwischen starken und schwachen Pflichten auf die Unterscheidung zwischen negativen und positiven Pflichten,[30] bzw. auf der Unterscheidung von aushelfender (*remedial*) und kausaler (*outcome*) Verantwortlichkeit.[31] Eine starke Verantwortung bestände demzufolge nur dann, wenn ein Schädigungszusammenhang vorliegt, während die bloße Fähigkeit zu helfen nur schwache Pflichten generieren könnte. Stark ist eine Unterlassungspflicht vor allem deswegen, weil sie sich genau bestimmen lässt. Sie ist justiziabel und eignet sich daher besonders gut zur Festsetzung externer Sanktionen.

Aber auch hier gilt, dass es gute Gründe dafür gibt, das Set globaler Regeln, Institutionen und Praktiken bereits als einen Schädigungszusammenhang zu rekonstruieren, der negative Gerechtigkeitspflichten im engeren Sinne auf den Plan ruft. Und wichtiger noch: die Rede von starken (Gerechtigkeits-) und schwachen (humanitären) Pflichten macht vor dem Hintergrund schwerer Menschenrechtsverletzungen wenig Sinn. Humanitäre Pflichten können aufgrund der *Dringlichkeit* des auf dem Spiel stehenden moralischen Guts unter Umständen „stärker" sein als Gerechtigkeitspflichten.

Und selbst dann, wenn unsere Verantwortung gegenüber globaler Armut auf humanitären Pflichten beruhte und solche Pflichten an sich selbst „schwach" wären, so dass wir sie in der Abwägung mit unseren alltäglichen speziellen Pflichten wenig zu gewichten bräuchten, wären wir damit doch keineswegs *ent*pflichtet, sondern stünden nach wie vor in der Pflicht, abwägen und gegebenenfalls auch positiven Hilfspflichten nachkommen zu *müssen*. Das Überforderungsproblem bleibt innerhalb des hier entworfenen Abwägungsbildes insofern bestehen, dass wir unsere persönlichen Gewichtungen ständig rechtfertigen müssen und dass unsere in jedem Wortsinne anspruchsvolle Verantwortung gegenüber globaler Armut gleichsam in jedem Moment wieder in den Vordergrund

29 Vgl. David Miller (2007, S. 248); Henning Hahn (2009).
30 Pogge (2002, S. 70-73).
31 Miller (2007, S. 94-108).

rückt, in dem wir eine „stärkere“ Verantwortung abgearbeitet haben oder einer stärker einzufordernden negativen Pflicht nachgekommen sind. Dass eine Pflicht relativ schwach ist, heißt nicht, dass es keine Pflicht ist.

d) Fair-share-Argumente

Ein vierter und auf den ersten Blick überzeugender Versuch, dem Überforderungscharakter globaler Verantwortung Herr zu werden, bieten uns sogenannte *Fair-share-Argumente*. Demnach kann jede Person nur zu dem Anteil verpflichtet werden, der bei allgemeiner Normenbefolgung ausreichen würde, ein Übel zu beseitigen. Die moralische Verantwortung ließe sich unter der idealen Voraussetzung einer *full compliance* damit erheblich eingrenzen.[32] Von einer Person mehr zu fordern als zugleich alle anderen zu geben bereit sind, erscheint gegenüber der pflichtgemäß handelnden Person als ungerecht. Sie würde mit einer weiter reichenden Verantwortung dafür „bestraft“, dass sie pflichtgemäß handelt. Allerdings bleibt die Frage, ob es in einer nichtidealen Welt Sinn macht, bei der Bestimmung der eigenen Mitverantwortung von idealen Annahmen auszugehen. Wenn zwei Personen an einen Teich mit zwei ertrinkenden Kindern vorbeigehen, ist zwar zunächst jede Person nur für die Rettung eines Kindes zuständig. Wenn aber die zweite Person einfach weitergeht und ihrer Verantwortung nicht gerecht wird, ist damit nicht gleich zu rechtfertigen, dass der verbleibenden Person nicht auch die Verantwortung für das zweite Kind zufällt.

Die Verantwortung einer Person allein relational zur Verantwortung anderer Personen bestimmen zu wollen und nicht unabhängig davon auch ihr Potential als Retter und die *dringende* Not des Kindes einzubeziehen, scheint hier unhaltbar zu sein. *Dass andere Menschen ihrer Verantwortung nicht nachkommen, ist ja gerade auch deswegen so verwerflich, weil sie damit nicht nur das Opfer schädigen, sondern darüber hinaus auch die Erfüllungsbedingungen gemeinsamer Verantwortung für andere unfair gestalten.* Was einer sehr viel ausführlicheren Argumentation bedürfte, möchte ich hier nur noch einmal plakativ auf den Punkt bringen: Wie wir unserer Verantwortung nicht gerecht werden, wenn wir einem Unfallopfer fünf Milliliter Blut spenden, so werden wir auch unserer Verantwortung gegenüber globaler Armut nicht gerecht, wenn wir auf die faire Verteilung vor dem Hintergrund einer kontrafaktischen *full-compliance* Voraussetzung bestehen.

32 Eine starke Variante des *Fair-share*-Arguments diskutiert Liam Murphy, der sein “*limited principle of beneficence*” folgendermaßen definiert: “You are required to act such that you will produce as great an expected overall benefit, given what you have reason to believe, as you would acting in any other way available to you, but only to the extent that this does not make you less well-off than you would be if everyone from now on acted in accordance with this principle.” (2002, S. 5).

e) Schefflers agent-centered prerogative

Gegen Singer ist wiederholt eingeworfen worden, dass die utilitaristische Verantwortungskonzeption unverträglich mit personaler Integrität ist, mit der Freiheit also, seine eigenen sinnvollen Projekte unabhängig von den Bedürfnissen anderer verfolgen zu können.[33] Der Utilitarismus disqualifiziert sich dadurch, dass er eine, wie Samuel Scheffler es auf den Punkt bringt, „Versklavung" am Leid des Anderen erfordert und so die Integrität von Personen unterminiert.[34] Wenn einer Person eine unabschließbare Verantwortung für globale Gerechtigkeit aufgebürdet wird und sie zugleich ihre Freiheit als einen Wert an sich selbst ansehen können soll, bleibt die Frage, wie zwischen dem Wert der Freiheit und dem Wert der Gerechtigkeit zu gewichten ist.

Scheffler ist es dann auch, der in diesem Zusammenhang ein *agent-centered prerogative* eingeführt hat, das Vorrecht jeder Person, selbst darüber zu entscheiden, wie sehr sie Eingriffe in ihre eigene Freiheit gegenüber ihrer Verantwortung für andere gewichten darf:

> „I believe that a plausible agent-centered prerogative would allow each agent to assign a certain proportionately greater weight to his own interests than to the interests of other people. It would than allow the agent the non-optimal outcome of his choosing, provided only that the degree of its inferiority to each of the superior outcomes he could instead promote in no case exceeded … the degree of sacrifice necessary for him to promote the superior outcome." (20)

Ein akteurszentriertes Vorrecht zur Abwägung zwischen Gerechtigkeitspflichten und persönlicher Freiheit muss allerdings weiter eingegrenzt werden. Erstens endet die Freiheit des Einzelnen an der Freiheit des Anderen, was bedeutet, dass Schädigungen an anderen und die erforderlichen Kompensationsleistungen gar nicht unter dieses Vorrecht fallen können. Die Verantwortung gegenüber unserer Beteiligung an globalen Schädigungszusammenhängen, wie sie Thomas Pogge beschreibt, bliebe also von vornherein unberührt.[35]

Weiter ist zu hinterfragen, aus welcher Perspektive ein subjektives Vorrecht, Gerechtigkeit und Freiheit abzuwägen, gerechtfertigt ist. Da es sich bei Gerechtigkeitspflichten um Interessen der Gesellschaft handelt, könnte man Scheffler ein „gesellschaftszentriertes Vorrecht", ein *societal-centered prerogative* entgegenstellen, Gerechtigkeitspflichten ein besonderes Gewicht gegenüber der Freiheit des Einzelnen zu verleihen. In dieser gesellschaftlichen Perspektive kann es keine moralisch legitimierte Freiheit ohne Gerechtigkeit geben. Die Willkürfreiheit des Einzelnen ist nur allgemein akzeptabel, wenn gerechte Normen zumindest über einen Minimalsockel hinaus etabliert sind. Was unterschiedliche (libe-

33 Vgl. Bernard Williams (1979).
34 Samuel Scheffler (1982).
35 Diesen Punkt macht etwa Shelly Kagan (1984).

rale) staatsphilosophische Modelle miteinander zu vermitteln suchen, nämlich die Gewährleistung von sozialer Gerechtigkeit und individueller Freiheit in ein und demselben politischen Arrangement, fällt in der Diskussion um globale Gerechtigkeit aber auseinander. In der Abwesenheit institutioneller Gerechtigkeitsgarantoren lassen sich unsere persönliche Freiheit und unsere individuellen Gerechtigkeitspflichten zur Etablierung solcher Garantoren nicht in vergleichbarer Weise vermitteln und geraten in Widerstreit.

2. Ausblick: Zum Widerstreit zwischen globaler Gerechtigkeit und Freiheit

Im Ergebnis lässt sich der Widerstreit zwischen Gerechtigkeit und Freiheit nicht einfach nach einer Seite hin auflösen. Ohne staatsanaloge Institutionen verfügen wir auch über keine Vorfahrtsregeln für die Freiheit, wie Scheffler behauptet – jedenfalls über keine allgemein gerechtfertigten. Ebenso ist eine totale Heteronomisierung am Leid der Welt zu rechtfertigen, wie Singer zu behaupten scheint. Die Tatsache, dass wir einerseits umfängliche Pflichten gegenüber globaler Armut tragen – seien es nun negative Pflichten, positive Gerechtigkeitspflichten oder besonders schwerwiegende humanitäre Verpflichtungen – und dass diese Pflichten andererseits unabschließbar und belastend sind, lässt eine besondere Spannung zwischen Gerechtigkeit und Freiheit erkennen. Die Freiheit, unseren eigenen Projekten nachzugehen, ist nur vor dem Hintergrund mehr oder weniger erfüllter bzw. an Institutionen delegierter Gerechtigkeitspflichten vor anderen zu rechtfertigen. In Abwesenheit einer vergleichbaren Gerechtigkeitsstruktur in der globalen Arena fallen Freiheits- und Gerechtigkeitsansprüche aber perspektivisch auseinander: In unseren Augen, den Augen wohlhabender Europäer, ist eine totale Aufzehrung unserer freiheitlichen Integrität inakzeptabel, vor den Augen derer, die unter schwerer globaler Armut leiden, ist diese Freiheit, die wir uns gegenüber unseren Pflichten herausnehmen, aber niemals zu rechtfertigen. Deswegen, meine ich, ist es hier zumindest unangemessen, von einem Prärogativ, also einem Vorrecht zu sprechen. Ein Prärogativ ist der historische Rechtsausdruck für das Privileg eines Monarchen, am Parlament vorbei zu entscheiden. Gemeint ist das Privileg, aber auch die Bürde einer Person, ihre Freiheit gegenüber ihrer Verantwortung selbst gewichten zu können bzw. zu müssen. Aber die Rede von einem „Recht“ ist hier unangebracht. Es ist eine Entscheidung, die eine Person möglicherweise vor sich selbst begründen, nicht aber vor anderen als ihr moralisches Vorrecht rechtfertigen kann.

Literaturverzeichnis

Hahn, Henning: "The Global Consequence of Participatory Responsibility". Journal of Global Ethics, 5, 2009. S. 43-56
Hahn, Henning: Globale Gerechtigkeit. Eine philosophische Einführung, Campus, Frankfurt/M., 2009

Kagan, Shelly: „Does Consequentialism Demands Too Much?". Philosophy and Public Affairs, 13, 1984. S. 239-54

Miller, David: National Responsibility and Global Justice. OUP, Oxford, 2007
Murphy, Liam: Moral Demands in Nonideal Theory. OUP, Oxford, 2002

Nagel, Thomas: The View from Nowhere. OUP, Oxford, 1986

Pogge, Thomas: World Poverty and Human Rights. Polity Press, Cambridge UK, 2002

Scheffler, Samuel: The Rejection of Consequentialism. OUP, Oxford, 1982
Singer, Peter: "Famine, Affluence and Morality". Philosophy and Public Affairs 1, 1972. S. 229-43

Williams, Bernard: Kritik des Konsequentialismus. Suhrkamp, Frankfurt/M., 1979

Young, Iris Marion: "Responsibility and Global Justice: A Social Connection Model". Social Philosophy & Policy Foundation 23, 2006. S. 102-130

Menschen im permanenten vegetativen Zustand - ihr moralischer Status, ihre moralischen Rechte

Christoph Lumer
lumer@unisi.it
Università di Siena

Abstract/Zusammenfassung

In this contribution a reply to the question about the moral status and the rights of humans in a permanent vegetative state is sketched - on the basis of a systematic theory of the moral status of beings on various ontic levels. A background assumption of the theory presented is that the question of defining 'death' - which, among others, implies drawing a border between different ontic levels - and the question of attributing moral rights have to be distinguished. A theory of ontic levels, from the ethical viewpoint, provides only more precise descriptive categories, by which the ethical theory can establish more precisely which protective rights beings on which level shall have.
In the paper's first main part the ontic levels relevant for the question of moral status are differentiated - material object, sortal object, livings being, sentient being, person - and the boundaries between them determined. Humans traverse these levels during the different phases of their subsistence. All sortal objects arise, subsist and vanish. With respect to vanishing, biological death of living beings, mental death of sentient beings and personal death of persons can be distinguished. These concepts are defined.
In the paper's second main part a theory is sketched by which rights can be ascribed to humans on the various levels. This theory is applied in particular to people in a permanent vegetative state. Because the function of moral rights is to protect interests, first, it is examined interests of which kind can be attributed to beings of which ontic level. The next, now really ethical, step assumes on the methodological level that all ethics use certain sources of morals as basis for attributing rights, different sources in different methodological approaches: moral intuitions, the semantics of moral language, pure reason, (hypothetical) treatises etc. and also altruistic and intersubjectively converging motives (resistant to new information). The theory sketched is based on the last source: altruistic (in a wide sense) and intersubjectively converging motives. The following motives are identified as being of this kind: sympathy, respect for other beings, interest in cooperation, human solidarity. On the basis of these motives certain rights are attributed to living beings with certain interests. According to this scheme, humans in a postpersonal and postmental permanent vegetative state have only a weak positive and negative prima facie right to live, which though is nearly always trumped by other interests of society.

In diesem Beitrag wird eine Antwort auf die Frage nach dem moralischen Status und den Rechten von Menschen im permanenten vegetativen Zustand skizziert, und zwar auf der Basis einer systematischen Theorie des moralischen Status von Wesen verschiedener ontischer Stufen. Eine Hintergrundannahme der vorgestellten Theorie ist, daß zwischen Todesdefinition - die u.a. eine Grenzziehung zwischen verschiedenen ontischen Stufen impliziert - und der Fra-

ge der Zusprechung von moralischen Rechten unterschieden werden muß Eine Theorie ontischer Stufen liefert rein ethisch nur eine Präzisierung der Beschreibung, mit deren Hilfe die ethische Theorie genauer festlegen kann, welche Schutzrechte Wesen in welchem dieser Stadien haben sollen.
Im ersten Hauptteil werden die für die Problematik des moralischen Status relevanten ontischen Stufen differenziert - materieller Gegenstand, Sortal, Lebewesen, fühlendes Wesen, Person - und die Grenzen zwischen ihnen bestimmt. Menschen durchlaufen diese Stufen in den verschiedenen Stadien ihres Bestehens. Alle Sortale entstehen, bestehen und vergehen. Beim Vergehen kann man unterscheiden zwischen dem biologischen Tod des Lebewesens, dem mentalen Tod des fühlenden Wesens und dem personalen Tod der Person. Diese Begriffe werden definiert.
Im zweiten Hauptteil wird eine Theorie skizziert, mit der Menschen in den verschiedenen Stadien Rechten zugeschrieben werden können. Diese Theorie wird insbesondere auf Menschen im permanenten vegetativen Zustand angewendet. Da die Funktion moralischer Rechte ist, Interessen zu schützen, wird zunächst untersucht, Wesen welcher ontischer Stufe welche Interessen zugesprochen werden können. Der nächste eigentlich ethische Schritt geht methodisch davon aus, daß alle Ethiken gewisse Quellen der Moral als Basis für die Zusprechung von Rechten verwenden, je nach methodischem Ansatz unterschiedliche: moralische Intuitionen, Bedeutung der Moralsprache, reine Vernunft, (hypothetische) Verträge etc. sowie aufklärungsstabile altruistische und intersubjektiv konvergierende Motive. Die skizzierte Theorie stützt sich auf die letztgenannte Quelle: i.w.S. altruistische und intersubjektiv konvergierende Motive. Als solche Motive werden ausgemacht: Empathie, Achtung vor anderen, Kooperationsinteresse, zwischenmenschliche Solidarität. Auf der Basis dieser Motive werden dann Lebewesen mit bestimmten Interessen bestimmte Rechte zugeschrieben. Menschen im postpersonalen und postmentalen permanenten vegetativen Stadium haben danach nur ein schwaches positives und negatives Prima-facie-Lebensrecht, das aber durch gesellschaftliche Interessen nahezu regelmäßig übertrumpft wird.

Einleitung

In der Diskussion über den dauerhaften vegetativen Zustand von Menschen wird häufig angenommen, die moralischen Fragen seien - mehr oder weniger - automatisch beantwortet, wenn geklärt sei, ob diese Menschen tot sind bzw. als tot zu gelten haben oder nicht. Diese Annahme erscheint mir und manchen anderen Ethikern [z.B. Singer 1995; Stoecker 2003] methodisch völlig verfehlt. Positiver gesagt: Mit der Antwort auf die Frage, ob ein Mensch im permanenten vegetativen Zustand tot ist oder nicht, ist moralisch noch lange nichts entschieden. Denn zum einen haben Menschen nicht bis zum Tod alle Rechte und verlieren sie dann. Der Besitz von moralischen oder auch juridischen Rechte ist keine Sache des Alles oder Nichts. Sondern Wesen in unterschiedlichen Stadien und Zuständen haben sehr unterschiedliche Rechte. Zum anderen ist die Festlegung, wann ein Mensch tot ist, eine biologisch-medizinische und auch philosophisch metaphysische Frage; eine ganz andere, nämlich erst wirklich moralische Frage ist, in welchem Stadium wer welche Rechte haben *soll*.

Dies bedeutet für die Ethik: Wir benötigen zuerst eine i.w.S. *metaphysische*, nämlich ontologische und auch personentheoretische, Theorie, wie welche ontischen Stadien beim Menschen zu definieren sind. Eine solche Theorie liefert rein ethisch nur eine Präzisierung der Beschreibung und eine nützliche Klassifikation von Gegenständen. Anschließend benötigen wir eine ethische Theorie darüber, welche Schutzrechte Wesen, insbesondere Menschen in welchem dieser Stadien haben sollen.

Dieser Beitrag ist deshalb wie folgt aufgebaut: In einem ersten Teil werden die ontischen Stadien von Gegenständen, insbesondere Menschen, nach metaphysischen, ontologischen Kriterien und nach Kriterien der Personentheorie definiert. Im zweiten Teil wird diskutiert, welche Rechte u.ä. Wesen dieser ontischen Stufen aus moralischen Gründen zustehen sollen.

1. Ontologie der Lebewesen - Stufen des Organischen

In der Ontologie werden u.a. innerhalb der Hauptgruppen diverser Gegenstände - wie Abstrakta, Substanzen, Tropen, Ereignisse, mentale Phänomene - auch Hierarchien immer komplexerer Gegenstände innerhalb solcher Hauptgruppen thematisiert. Für unsere Frage interessiert eine *Hierarchie der materiellen Gegenstände* [zusammengefaßt in Tabelle 1]:

Materielle Gegenstände sind Substanzen und haben eine Masse. Zu den materiellen Gegenständen gehören auch Elementarteilchen oder Wolken.

Sortale sind größere materielle Gegenstände mit einer gewissen Kohäsion ihrer Teile. Sie haben Formen oder Funktionen, über die sie definiert sind. Sie können sich aber verändern. Die Sortale sind definiert durch ein raumzeitliches Kontinuum dieser kohäsiven Teile. Für das Fortbestehen ein und desselben Sortals in der Zeit ist wesentlich, 1. daß jeweils der größte Teil der Substanz von einer Zeitschicht zur nächsten erhalten bleibt, 2. daß die Form und Funktion erhalten bleibt. Wenn diese Bedingungen nicht mehr erfüllt sind, hat sich das Sortal als solches aufgelöst, es besteht nicht mehr. Sortale haben empirisch immer einen zeitlichen Anfang, ab dem sie bestehen, und ein zeitliches Ende, ab dem sie nicht mehr bestehen.

Lebewesen sind Sortale besonderer Art. Sie sind durch Lebensfunktionen gekennzeichnet. Zu den Lebensfunktionen gehören: Selbsterhaltung (Aufrechterhaltung eines Sollzustandes) durch Metabolismus, Entwicklung, Reproduktion, Beziehungen mit der Umgebung. Schon Einzeller sind in diesem Sinne Lebewesen.

Fühlende Wesen sind Lebewesen, die die Fähigkeit zur subjektiven Empfindung haben. Dazu gehören ein wahrnehmendes Bewußtsein und die Schmerz- und Lustempfindlichkeit.

Personen sind fühlende Wesen mit personalen Eigenschaften. Die wichtigsten personalen Eigenschaften sind: Selbstbewußtsein; rationale Entscheidungsfähigkeit, d.h. Fähigkeit, Handlungsalternativen zu imaginieren, diese zu bewerten und zwischen ihnen zu entscheiden; Urteilsfähigkeit; Kritikfähigkeit.

Gegenstandstyp	*Definiens*	*Beispiele*
materieller Gegenstand	Substanz und Masse	Elementarteilchen, Wolken
Sortale	materielle Gegenstände mit raumzeitlich kohäsiven Teilen, die eine bestimmte Form haben oder eine bestimmte Funktion erfüllen	Berge, Stühle, Maschinen, Leichen
Lebewesen	Sortal mit Lebensfunktionen: Selbsterhaltung, Stoffwechsel, Entwicklung, Fortpflanzung	Einzeller, Insekten, Reptilien
fühlendes / Wesen	Lebewesen mit der Fähigkeit, Lust oder Schmerz zu empfinden	Säugetiere
Person	Lebewesen mit Personenqualitäten: Selbstbewußtsein, Fähigkeit, rational zu entscheiden (Vermögen, Handlungsoptionen zu imaginieren, zu bewerten und zu wählen), Urteilskraft, Kritikfähigkeit	normale erwachsene Menschen

Tabelle 1: Hierarchien von Gegenständen in der Ontologie

Alle Sortale entstehen in der Zeit, bestehen eine Zeit lang und vergehen dann wieder. Für diese Ereignisse und Zustände gibt es je nach Stufe der Entwicklung z.T. unterschiedliche Bezeichnungen [vgl. Tabelle 2].[1] Die Bezeichnungen für Sortale generell können selbstverständlich insbesondere auch für Sortale höherer Stufe verwendet werden. Der Blick auf die Tabelle macht aber deutlich, daß für den Anfang, das Bestehen und das Ende verschiedener (hierarchischer) Stufen der Sortale zum großen Teil dieselben Ausdrücke verwendet werden. In unserem Zusammenhang problematisch ist insbesondere die fehlende terminologische Differenzierung des Vergehens verschiedener Stufen von Sortalen.

Gegenstandstyp	*Anfang*	*Bestehen*	*Ende*
Sortal	entstehen	bestehen	vergehen, Auflösung
Lebewesen	entstehen, Befruchtung	leben	sterben, Tod
fühlende Wesen	entstehen	bestehen, mentales Leben	vergehen, mentaler Tod
Person	entstehen	bestehen, personales Leben	Tod der Person, personaler Tod

Tabelle 2: Bezeichnungen für Anfang, Ende und Bestehen der Sortale

Der Alltagsausdruck "Tod" ist relativ eindeutig biologisch. Denn alltagssprachlich können auch niedere Tiere und Pflanzen tot oder lebendig sein. Aus mehreren Gründen gibt es in der Philosophie aber die Tendenz, daneben auch noch die Ausdrücke "personaler Tod" und "mentaler Tod" einzuführen. 1. Es

1 Der Ausdruck "existieren" wird häufig auch für das Bestehen oder Leben verwendet. Philosophisch ist "Existenz" jedoch in der Hauptsache der Existenzquantor, also ein logischer Operator und kein Prädikat. Existenz im logischen Sinne ist zeitlos; 'Sokrates existiert'. 'Bestehen' ist hingegen ein Prädikat mit Zeitvariable.

fehlt einfach ein Ausdruck der Alltagssprache, der dies bezeichnet. 2. Dies als "*Tod*" zu bezeichnen wird der Bedeutung, die dem Tod kulturell beigemessen wird, gerechter als die Bezeichnungen "Vergehen", "Ende" o.ä. 3. Im juristischen Bereich ist der Tod die Schwelle, mit der ein Wesen die entscheidende Schwelle zur nahezu völligen Rechtlosigkeit übertritt. Aus rein sachlichen Gründen spricht aber vieles dafür, diesen Schritt nicht mit dem biologischen, sondern mit dem mentalen Tod anzusetzen. (Durch die Bezeichnung "mentaler *Tod*" wird also der sachliche Streit um den Verlust der Schutzrechte auch linguistisch geführt.) 4. Den alltagssprachlichen biologischen Todesbegriff aufheben, beseitigen zu wollen oder besser: den Ausdruck "Tod" nicht mehr für den biologischen Tod verwenden zu wollen ist vermutlich weder realistisch durchführbar noch sachlich angemessen. Dieser Ausdruck ist zu stark verankert; und die Tatsache, daß man auch bei niederen Tieren und Pflanzen von "Tod" spricht, ist eine zu deutliche Erinnerung, daß es hier um den biologischen Tod geht. Sachlich ist der Ausdruck "Tod" also ursprünglich für den biologischen Tod reserviert. Aber vor der Entstehung der Intensivmedizin bestand auch gar nicht die Notwendigkeit, die drei Todesarten zu differenzieren, weil sie zwingend aneinandergekoppelt waren und in kurzem Abstand nacheinander erfolgten. Wenn man die biologische Konnotation des Ausdrucks "Tod" akzeptiert, hat man nur die Möglichkeit, die Analogie der drei Vergehensarten auszunutzen und den Ausdruck "Tod" - aber zusammen mit einer Spezifizierung - noch für diese analogen Phänomene zu verwenden, also vom "mentalen" und "personalen Tod" zu sprechen.

Entstehung und Vergehen eines Sortals sind - abgesehen von der raumzeitlichen Identität - an die sie definierende Funktion gebunden, an das Einsetzen der Funktion und das Aufhören der Funktion. Zusätzlich können die Funktionen aber z.T. für eine Zeit suspendiert sein. Wenn derselbe materielle Gegenstand später die Funktion wieder aufnimmt, hat er weiter bestanden. Für das Vergehen des Sortals ist deshalb das endgültige Aufhören der Funktion entscheidend. Entsprechend kann man 'Tod' wie folgt definieren:

Biologischer Tod := Ende aller vitalen Funktionen bei Tieren, Pflanzen oder Teilen von ihnen ohne Möglichkeit der Wiederherstellung (genauer: alle vitalen Funktionen hören auf, und der Körper tritt in einen Zustand ein, in dem eine Wiederherstellung der vitalen Funktionen nicht mehr möglich ist).[2]

Empfindungstod := Ende aller Empfindungsvorgänge und -funktionen bei Tieren ohne Möglichkeit der Wiederherstellung.

personaler Tod := Ende aller personalen Funktionen bei Tieren ohne Möglichkeit der Wiederherstellung.

Neben der Todes*definition* gibt es sekundäre, operationale Todeskriterien. Diese legen nach dem medizinischen Wissen relativ einfach zu beobachtende, operati-

2 Zur Diskussion der Kriterien Irreversibilität s. Youngner 2009, 294-297.

onale Bedingungen fest, die als sicheres Anzeichen des Erfülltseins der definitorischen Bedingungen gelten können. Der *Ganzhirntod*, d.h. der Ausfall des gesamten Hirns, inklusive Hirnstamm, ist das übliche medizinische Kriterium für den biologischen Tod.[3] Der *Teilhirntod*, d.h. das Ende der Funktionen des Großhirns, wird in der ethischen Diskussion, vergröbernd, oft als Kriterium für den mentalen Tod angeführt. Dies ist vermutlich nicht genau richtig; denn wahrscheinlich führt schon das Absterben gewisser Teile des Großhirns zum Verlust jeglichen Bewußtseins. Aber für die hiesige Diskussion genügt das Aussetzen des Großhirns als grobes Kriterium für den mentalen Tod: Genauere Differenzierungen sind noch zu unsicher, um darauf Entscheidungen über das Aufhören von Rechten aufzubauen. Deshalb kann man das sicherere Kriterium 'mentaler Tod' verwenden, das auf jeden Fall den mentalen Tod impliziert. Ein klares medizinisch operationales Kriterium für den personalen Tod hingegen fehlt bislang.

Nun durchlaufen Mensch in ihrem Leben diverse ontische Stadien; mit der Entwicklung werden sie zu ontisch komplexeren Wesen und entwickeln sich nachher wieder zurück. Die Stadien sind [vgl. auch Tabelle 3]:

1a. Bloßes Sortal: Ein eventuelles Anfangsstadium eines Menschen als bloßes Sortal gibt es nicht. Weder die Samenzelle noch das Ei sind schon Menschen. Beide Teile tragen zur besonderen Lebensfunktion bei. Man kann aber nicht rückwärtig von der Zygote aus sagen, daß diese schon mit dem Ei oder mit der Samenzelle identisch war. Der Sprung ist zu groß; und rückwärtig kann man zwei Identifikationen vornehmen. Das widerspricht aber der Idee der Identität.

2a. Lebewesen: Menschen entstehen, beginnen mit der Befruchtung als Lebewesen oder organisches Wesen.

3a. Fühlendes Wesen: Ab dem vierten Entwicklungsmonat, also beim Übergang vom Embryonal- zum Fötalstadium, entwickelt sich mit der Entfaltung der Neokortex etc. die Fähigkeit zur subjektiven Empfindung: Die Sinneswahrnehmungen entwickeln sich, ebenso die Schmerz- und Lustempfindlichkeit. Der Mensch wird von einem bloßen Lebewesen zu einem fühlenden Wesen.

4. Person: Personale Fähigkeiten entwickeln sich beim Menschen ungefähr ab einem Jahr nach der Geburt.

Diese aufbauende Entwicklung wird invertiert zum Ende des Lebens. Im Normalfall verlieren Menschen mit dem Tod gleichzeitig die Personalität, ihre Fähigkeit als fühlende Wesen und das Leben eines Organismus; bzw. diese drei Verluste spielen sich innerhalb von wenigen Minuten nacheinander ab. Aber es gibt auch Fälle, in denen die einzelnen Regressionsphasen deutlich verlängert sind, u.U. auf viele Jahre. Dies sind die im vorliegenden Zusammenhang problematischen Fälle.

---

3 Der Herzschlag ist zwar auf der Rückenmarksebene verankert, die Atmung hingegen im Stammhirn. Deshalb müssen Hirntote, denen später Organe entnommen werden sollen, künstlich beatmet werden. Wenn die künstliche Beatmung ausgesetzt wird, hört nach kurzer Zeit wegen mangelnder Sauerstoffversorgung der Herzmuskeln auch der Herzschlag auf.

3b. Fühlendes Wesen: Bei sehr starker Debilität verlieren Menschen alle ihre personalen Funktionen. (Von Geburt an sehr stark debile Menschen entwickeln sich nie zu Personen.) Sie können sich u.U. noch bewegen, Nahrung aufnehmen etc., aber nicht mehr rational entscheiden, wissen nicht mehr, wer sie sind etc.

2b. Lebewesen: Vor allem beim apallischen Syndrom, dem permanenten vegetativen Zustand, verlieren Menschen auch dauerhaft ihr Bewußtsein und ihre Empfindungsfähigkeit.

1b. Sortale: Nach dem biologischen Tod ist ein Mensch nur noch als Leiche vorhanden. Mit der Verwesung bleiben irgendwann nur noch Teile der Leiche übrig. Dies ist das endgültige Ende des Menschen. Es gibt dann nur noch *Reste* des Menschen, seine Gebeine etc., aber nicht mehr *den* Menschen.

Nummer	*ontol. Art des Stadiums*	*Stadium*	*Anfang*
1.a	**bloßes Sortal**	---	---
2.a	**Lebewesen**	Embryo	Befruchtung
3.a	**fühlendes Wesen**	Fötus, Säugling	Funktionieren der Großhirnrinde
4	**Person**	Kinder, normale Erwachsene	Entstehen von Selbstbewußtsein, kritischer und rationaler Fähigkeiten
3.b	**fühlendes Wesen**	starke Debilität	Aufhören der rationalen Fähigkeiten
2.b	**Lebewesen**	permanenter vegetativer Zustand	Funktionsausfall der Großhirnrinde
1.b	**Sortal**	Leiche	Funktionsausfall des gesamten Gehirns

Tabelle 3: Ontische Stadien beim Menschen

2. Die moralischen Rechte von Lebewesen einer bestimmten Entwicklungsstufe

Die soeben entwickelten Differenzierungen müssen nun angewendet werden. In diesem Teil des Beitrags geht es darum, wann Menschen welche Rechte haben. Wie schon in der Einleitung erwähnt, ist die Zusprechung von Rechten keine Frage des Alles oder Nichts in dem Sinne, daß ein Wesen alle moralischen Rechte hat oder eben keine. Menschen sind nicht die einzigen Wesen, die moralische Rechte haben. Und sie haben nicht immer alle moralischen Rechte auf einmal. Vielmehr sind moralische Rechte u.a. an den Stand der ontischen Entwicklung gebunden. Hier geht es nun darum, auf welcher Grundlage Menschen oder andere Lebewesen unter welchen Bedingungen welche moralischen Rechte haben.

Eine präzisierende Einschränkung ist aber noch erforderlich: Hier geht es nur um den *moralischen Status* eines Wesens und ob es *selbst* moralische Rechte hat. Wenn ein Wesen *überhaupt* keine moralischen Rechte mehr hat oder bestimmte moralische Rechte nicht hat, bedeutet dies nicht, daß dieses Wesen moralisch wertlos ist oder nicht aus moralischen Gründen geschützt werden sollte. Aber dabei geht es dann nicht mehr um die Rechte und Interessen dieses Wesens, sondern um die Interessen anderer. Es könnte also insbesondere sein, daß ein Mensch im permanenten vegetativen Zustand aufgrund seines moralischen Status für sich selbst kein Recht auf bestimmte Behandlungen mehr hat, daß aber Angehörige diese Behandlungen für ihn wünschen. Dann kann es sein, daß der Patient diese Behandlung auch aus moralischen Gründen erhalten sollte, obwohl er aufgrund seines moralischen Status kein Recht auf diese Behandlungen hat.

Moralische Rechte verleihen wir Wesen, die selbständig sind, einen Eigenwert und Interessen haben. Die Verleihung moralischer Rechte hat gerade die Funktion, die Interessen des Rechtsträgers zu schützen. In unserer üblichen moralischen Praxis verleihen wir Wesen vieler Arten Rechte, und wir betrachten sie als moralisch schutzwürdig auf der Basis ihrer Interessen. Wenn sie keine Interessen haben, sind sie auch nicht für sich schutzwürdig und haben sie keine moralischen Rechte.

Als Voraussetzung einer Bestimmung der moralischen Rechte verschiedener Arten von Wesen, müssen wir deshalb untersuchen, welche Wesen welche Arten von Interessen haben können.

Sortale - funktionale Werte: Man kann bei Sortalen in einem sehr weiten Sinn Wünschbarkeiten für dieses Sortal konstruieren: *x* ist gut / schlecht für *s*; z.B.: der Ölwechsel ist gut für den Motor. Diese Wünschbarkeit besteht in der Funktion des Wertgegenstandes für die Aufrechterhaltung und Förderung der Form und Funktion des Sortals. Die bloße Funktion von bloßen Sortalen ist aber in der Regel eine Funktion für etwas anderes, eine Funktionalität. Deshalb kann etwas zwar - in einem schwachen Sinn - gut für sie sein; aber wegen der Funktionalität ist die Funktionserhaltung primär instrumentell gut und nicht intrinsisch gut. Bloße Sortale sind nicht eigenständig, sie bilden kein eigenes Zentrum gegenüber der Welt; sie sind in diesem Sinne nicht subjektiv, und das für sie Gute ist nicht subjektiv gut. Da dies (subjektives Zentrum gegenüber der Welt) aber für die Existenz von Interessen vorausgesetzt wird, haben Sortale als solche noch keine Interessen.

Lebewesen - Lebensinteressen: Auch für Lebewesen kann man Wünschbarkeiten konstruieren. Die Wünschbarkeiten bestehen in der Funktion für die Aufrechterhaltung und Förderung der Lebensfunktionen des Lebewesens, insbesondere: Ist etwas seinem Weiterleben zu- oder abträglich? Bei Lebewesen nehmen wir schon eine gewisse Eigenständigkeit an, die sich in der "Sorge" des Lebewesens um sich selbst äußert. Deshalb sprechen wir ihnen auch Interessen zu. Inhaltlich sind diese Interessen also Lebensinteressen, Interessen an der Aufrechterhaltung und Förderung ihrer Lebensfunktionen. Lebewesen als solche können diese Interessen allerdings nicht selbst re-

präsentieren, es sind keine subjektiven Interessen und Wünschbarkeiten. Nur wir können ihnen diese Interessen zuschreiben.

Fühlende Wesen - hedonische Interessen: Fühlende Wesen haben Schmerz- und Lustempfindungen. Lust und Schmerz sind subjektive, dem jeweiligen Subjekt zugängliche Zustände. Sie sind für das Subjekt per se desiderativ bedeutsam. Es mag angenehme Zustände, wünscht ihr Andauern und ihre Wiederholung und arbeitet, wenn es funktionsfähig ist, auf dieses Andauern und Wiederholen hin; umgekehrt bei unangenehmen Zuständen. Diese subjektive desiderative Bedeutsamkeit und das entsprechende Streben der fühlenden Wesen sind die Basis dafür, ihnen subjektive Wünschbarkeiten zuzuschreiben. Der Inhalt dieser Wünschbarkeiten ist hedonisch. Und die Wünschbarkeit eines Gegenstands für *s* besteht in seinem Gesamtbeitrag zur hedonischen Bilanz von *s*. Das fühlende Wesen kann auf jeden Fall angenehme Zustände begrüßen und unangenehme ablehnen. Es muß aber bei weitem nicht für beliebige Gegenstände den Beitrag zur eigenen hedonischen Bilanz ermitteln können. Aus der subjektiven Perspektive des fühlenden Wesens sind die hedonischen Interessen vorrangig vor den vitalen Interessen - auch wenn die hedonischen Interessen evolutionär als Mittel zur Förderung der Überlebensfähigkeit entstanden sind. Bei anderweitig nicht beendigbarem Leiden geht die Beendigung des Leidens der Aufrechterhaltung der Lebensfunktionen vor.

Personen - präferentielle Interessen: Personen haben mit ihrer Handlungsfähigkeit auch Präferenzen, d.h. intrinsische motivationale Bewertungen von Ereignissen. Es ist ein wertendes Sich-zu-etwas-Verhalten mit Blick auf die handelnde Gestaltung der Welt. Aus diesen Präferenzen und intrinsischen Bewertungen können präferentielle subjektive Wünschbarkeiten konstruiert werden. Als Ausdruck einer aktionalen Subjektivität geben wir diesen motivationalen, präferentiellen Interessen noch eine größere Bedeutung als den hedonischen Interessen. (Dies schließt allerdings nicht aus, daß die motivationalen, präferentiellen Interessen selbst hedonische Inhalte haben.)

Bislang wurde hier nur betrachtet, wie Interessen und Wünschbarkeiten für bestimmte Wesen konstruiert werden können. Es fehlt noch die Beantwortung der moralischen Frage nach den moralischen Wünschbarkeiten und Rechten. Mit dieser Frage sind natürlich sofort große methodische Probleme der Moralbegründung verbunden und entsprechende Debatten und divergierende Positionen. Aber nach allen Positionen beruht die Moral jeweils auf einer Quelle der Moral - wobei hier unter "*Quelle der Moral*" eine Basis verstanden wird, die auch den Inhalt der Moral vorgibt. Auch die Verleihung moralischer Rechte und alle moralischen Bewertungen beruhen auf *Quellen* der Moral. Über die Quellen der Moral gibt es sehr unterschiedliche Ansichten in der Ethik. Als Quellen der Moral werden beispielsweise angenommen: unsere Moralsprache, unsere moralischen Intuitionen, Gottes Wille, die Grundbedingungen für Kommunikation überhaupt, Kooperationsinteressen, moralnahe Motive usw.

An dieser Stelle können die verschiedenen Konzeptionen der Quellen der Moral nicht diskutiert werden. Vielmehr rekurriere ich hier auf eine anderenorts von mir ausgearbeitete Theorie [Lumer <2000>/2009, insbes. Kap. 1-2]. Diese Theorie begründet zunächst formale Adäquatheitsbedingungen für eine gute oder die richtige Quelle der Moral: Diese Quelle muß kohärent, motivational

wirksam, rational und aufklärungsstabil, intersubjektiv einheitlich und damit konsensstiftend sowie intuitiv moralisch sein. Im zweiten Schritt werden dann mögliche Quellen der Moral daraufhin untersucht, ob sie diese Adäquatheitsbedingungen erfüllen. Als Ergebnis dieses Schritts erwiesen sich bestimmte unserer - rational ausgearbeiteten - altruistischen oder intersubjektiv konvergierenden Motive als die adäquate Quelle der Moral. Die wichtigsten derartigen Motive und damit Quellen der Moral sind [s. Lumer 1999]:

1. die Empathie, das Mitgefühl;
2. die Achtung vor anderen (im Sinne zunächst eines Gefühls, nach dem der andere in einem gewissen Sinne hoch entwickelt und empfindlich ist, z.T. sogar Autonomie besitzt und für sich sorgt, und deshalb als in sich wertvoll betrachtet wird, sodann im Sinne eines Motivs, dieses Wertvolle schützen zu wollen);
3. Kooperationsinteressen zum gegenseitigen Vorteil;
4. zwischenmenschliche Solidarität.

Welche moralischen Wünschbarkeiten und Rechte ergeben sich aus diesen Quellen? [Zusammenfassender Überblick s. Tabelle 4.]

Lebewesen als solchen bringen wir normalerweise Achtung entgegen: Hier ist etwas, das eigene Interessen hat, das eine höhere Entwicklungsstufe über das Anorganische hinaus entwickelt hat; wir erkennen dies an. Dies gilt aber nur prima facie, wenn dem keine menschlichen Interessen entgegenstehen. Im Verhalten äußert sich dies so, daß wir versuchen, die Interessen der Lebewesen nicht zu verletzen, wenn dies nicht unseren eigenen Interessen stark zuwiderläuft: Wir treten nicht auf Regenwürmer, zerhacken Schnecken nicht. Wir töten aber Mücken, um uns vor ihren Stichen zu schützen; wir fällen Bäume, wenn wir ihr Holz benötigen oder sie uns im Weg stehen. Aber wir schützen meistens die einen Tiere nicht vor den Angriffen der anderen. Eine rein willkürliche Verletzung oder Tötung solcher Lebewesen, nur zum Spaß ist unmoralisch. Wir gestehen also Lebewesen als solchen minimale, Prima-facie-Schutzrechte zu, ein passives Lebensrecht und ein Recht auf körperliche Unversehrtheit - wie gesagt, wenn dem keine stärkeren menschlichen Interessen zuwiderlaufen. - Motivationale Grundlage dieses schwachen Schutzes ist das Achtungsgefühl.

Von *fühlenden, empfindungsfähigen Wesen* erkennen wir die hedonischen Interessen an. Dies ist auch im Tierschutzgesetz geregelt: Fühlende Tiere dürfen nicht gequält werden, auch für unsere eigenen Interessen nicht - es sei denn, dem stehen sehr hohe menschliche Interessen entgegen. (Da bloß fühlende Wesen noch keine Vorstellung von ihrer Zukunft haben, um die sie sich sorgen könnten, sprechen viele Theoretiker den fühlenden Wesen kein Lebensrecht zu. Allerdings haben ja auch fühlende Wesen - ein allerdings nicht subjektiv repräsentiertes - hedonisches und vitales Interesse am Weiterleben. Deshalb gestehen viele Tierschützer allen fühlenden Wesen auch ein Lebensrecht zu. Sie sind z.B. ethische Vegetarier.) - Die Grundlage dieses Schutzes der fühlenden Wesen ist die Empathie und die Achtung.

Personen haben Autonomie, und wir erkennen normalerweise ihre Autonomie an. D.h. moralisch gewähren wir ihnen alle möglichen Freiheitsrechte, ihre Präferenzen umzusetzen. Und wir gewähren ihnen Schutzrechte, damit sie ihre Präferenzen in Ruhe umsetzen und sich selbst entwickeln können. Zudem gewähren wir ihnen positive Sozialrechte, damit sie die Ressourcen haben, einen Teil ihrer präferentiellen Interes-

sen umzusetzen und auf jeden Fall ihre hedonistischen und vitalen Interessen wahren können. Außerdem lassen wir sie an der Verteilungsgerechtigkeit und an politischer Macht teilhaben. (Wir geben den hedonischen Interessen dieser Menschen aber ein besonderes Gewicht. Zu diesen hedonischen Interessen gehört bei hoch entwickelten Wesen mit Selbstbewußtsein, Wissen und Sorge um die Zukunft auch, daß wir vermeiden, ihre Ängste mit Blick auf die Zukunft zu wecken.) - Grundlage dieser Schutzrechte sind die Empathie, aber auch die Achtung vor der Autonomie, z.T. aber auch die eigene Angst vor den zerstörerischen Fähigkeiten des anderen sowie das Interesse an Kooperation.

Mit *Menschen* allgemein verbindet uns über die bisher genannten Quellen der Moral hinaus noch die zwischenmenschliche Solidarität. Diese Solidarität ist nicht an spezielle Stufen des Organischen gebunden, sondern richtet sich allein nach der Speziszugehörigkeit. Dies führt dazu, daß wenn Menschen und nichtmenschliche Tiere auf der gleichen organischen Stufe stehen, Menschen immer noch ein paar Schutzrechte mehr haben.

Art des Wesens	*Interessen*	*Basis der moralischen Rechte beim moralischen Subjekt*	*moralische Rechte*
Lebewesen	Lebensinteressen	Achtung	• (schwaches) negatives prima facie Lebensrecht und Recht auf körperliche Unversehrtheit
fühlendes Wesen	Lebensinteressen und hedonistische Interessen	Achtung; Empathie / Mitgefühle	• (prima facie oder endgültiges?) negatives Lebensrecht und Recht auf körperliche Unversehrtheit; • Recht, nicht gequält zu werden.
Person	Lebensinteressen, hedonische und präferentielle Interessen	Achtung vor der Autonomie; Empathie / Mitgefühl; Kooperationsinteresse	• positives und negatives Lebensrecht und Recht auf körperliche Unversehrtheit; • Recht, nicht gequält zu werden; • Menschenrechte usw.
Menschen	Lebensinteressen, hedonische und präferentielle Interessen	menschliche Solidarität	• Verstärkung der Rechte im Vergleich zu den Rechten der Wesen gleicher ontischer Art.

Tabelle 4: Moralische Rechte verschiedener Wesen

Die soeben skizzierten moralischen Rechte der Wesen auf diversen ontischen Stufen gelten selbstverständlich auch für Menschen. Bei Menschen sind aber noch zwei Besonderheiten zu berücksichtigen, die zu einer komplizierteren Gestaltung ihrer Rechte in den verschiedenen Stadien führen. Zum einen führt die

Überlagerung des ontischen Status mit dem Menschsein zu einer Verstärkung der Rechte gegenüber nichtmenschlichen Tieren. Zum anderen gibt es Asymmetrien zwischen den Rechten bei der Aufwärts- und bei der Abwärtsentwicklung: Bei der Aufwärtsentwicklung müssen den Menschen mehr Rechte zugestanden werden, als dies ihrem aktuellen ontischen Status entspricht, weil sie vermutlich noch in höhere Stadien gelangen werden, die einen stärkeren Schutz verlangen; Schäden, die auf niedrigeren ontischen Stufen angerichtet werden, können Schäden in höheren ontischen Stadien bedeuten und dortige Schutzrechte verletzen. So darf z.B. auch ein Embryo bei Drogensucht der Mutter nicht durch deren Drogeneinnahme in seiner Entwicklung geschädigt werden: Das Interesse der Mutter an der Drogeneinnahme könnte zunächst einmal das Interesse des Embryos an einer ungestörten Entwicklung überwiegen. Wenn wir aber berücksichtigen, daß der drogengeschädigte Embryo später ein fühlendes Wesen und dann eine Person sein wird, dessen körperliche Integrität durch die Drogeneinnahme verletzt ist, dann ist die Drogeneinnahme aber auch in der Embryonalphase verboten.

Welche Rechte haben Menschen also auf den einzelnen ontischen Stufen? [Überblick s. Tabelle 5.]

2a: Embryonalstadium: Menschlichen Embryonen dürfen wir keine Schäden zufügen, die Auswirkungen auf den vollständig entwickelten Menschen haben könnten. Dieses Verbot gilt auch für alle Folgestadien. Wenn kein massives Interesse der Mutter, der Umgebung oder der Gesellschaft dagegensteht, gibt es zudem das Recht des Embryos auf ungehinderte Entfaltung zu einer Person, also ein sehr schwaches (prima facie) negatives und positives Lebensrecht. Das positive Lebensrecht schließt auch z.B. eine für die Embryonalentwicklung hinreichende Ernährung der Mutter ein und überhaupt eine bestimmte Lebensweise der Mutter. Bei massiven entgegenstehenden Interessen ist aber Abtreibung erlaubt; die Schutzrechte eines fühlenden Wesens oder gar einer Person werden dadurch nicht verletzt, denn der abgetriebene Embryo erreicht diese ontischen Stufen gar nicht. Außerdem hat der menschliche Embryo ein Recht auf medizinische Versorgung.

3a: Fötal- und Säuglingsstadium: Föten und Säuglingen dürfen, als fühlenden Wesen, keine unnötigen Leiden zugefügt werden. Dies gilt auch für alle weiteren Stadien, in denen der Mensch ein fühlendes Wesen ist. Daneben haben Föten und Säuglinge ein negatives Recht auf Leben, körperliche Unversehrtheit ... und ein positives Recht auf Ernährung, Fürsorge, Förderung der eigenen Anlagen, Wohnung, Pflege. Diese Rechte beruhen auf der Achtung und Empathie.

4: Personenstadium: Im Personenstadium haben Menschen alle Menschenrechte, inklusive der positiven Sozialrechte, sowie politische Rechte. Die Quelle für diese Rechte sind - über die bisher schon genannten, vor allem die Empathie, hinaus - die Achtung vor der Autonomie und dem Interesse an gegenseitiger Kooperation.

3b: Postpersonales Stadium des fühlenden Wesens: Menschen im postpersonalen Stadium des fühlenden Wesens dürfen keine unnötigen Leiden zugefügt werden. Sie haben ein positives Recht auf Ernährung, Fürsorge, Wohnung und Pflege. Außerdem haben sie ein negatives Recht auf Leben, körperliche Unversehrtheit ... Das Recht

auf Förderung der eigenen Anlagen entfällt; hier ist nichts mehr zu fördern, sondern nur noch zu erhalten.

2b: Postpersonales vegetatives Stadium: Menschen im postpersonalen vegetativen Stadium haben ein Recht auf würdige Behandlung, aber nur ein schwaches positives und negatives Prima-facie-Lebensrecht, das jedoch durch gesellschaftliche Interessen regelmäßig übertrumpft wird. - Dies wird gleich noch ausführlicher diskutiert.

1b: Leiche: Die menschliche Leiche schließlich hat ein Recht auf würdige Behandlung.

Nr.	*Art des Stadiums*	*Stadium*	*moralische Rechte (Auswahl)*
2.a	Lebewesen	**Embryo**	• bedingtes (es gibt kein wichtiges konkurrierendes Interesse der Mutter oder anderer Personen) negatives und positives Lebensrecht, Recht auf körperliche Unversehrtheit und Entwicklung; • positives Recht auf medizinische Versorgung
3.a	fühlende Wesen	**Fötus, Säugling**	• negatives und positives Lebensrecht, Recht auf körperliche Unversehrtheit und Entwicklung; • positives Recht auf Pflege und Förderung der eigenen Potentiale; • Recht, nicht unnötig leiden zu müssen; • …
4	Person	**Kinder, normale Erwachsene**	• Alle Menschenrechte; • (Erwachsene auch politische Rechte etc.)
3.b	fühlende Wesen	**schwere Debilität**	• negatives und positives Lebensrecht, Recht auf körperliche Unversehrtheit; • positives Recht auf Pflege; • Recht, nicht unnötig leiden zu müssen; • …
2.b	Lebewesen	**permanenter vegetativer Zustand**	• im Prinzip bedingtes (es gibt kein wichtiges konkurrierendes Interesse der Gesellschaft) negatives und positives Lebensrecht und Recht auf körperliche Unversehrtheit – die Bedingung ist aber nicht erfüllt; • Recht auf würdige Behandlung
1.b	Sortal	**Leiche**	• Recht auf würdige Behandlung

Tabelle 5: Rechte von Menschen auf den diversen ontischen Entwicklungsstufen

Welche Rechte haben nun Menschen im permanenten vegetativen Zustand? Die folgenden Überlegungen setzen voraus, daß es sich wirklich um einen *permanenten* vegetativen Zustand handelt, aus dem also eine Rückkehr zum Bewußtsein nicht möglich ist. Die Diskussion über die medizinischen operationalen Kriterien dafür, wann dies der Fall ist, kann hier nicht dargestellt oder gar fortgeführt werden.

Das Recht, nicht in einer Weise geschädigt zu werden, die Auswirkungen auf den vollständig entwickelten Menschen hat, entfällt bei Menschen im permanenten vegetativen Zustand; es ist gegenstandslos geworden.

Das Analogon zum präpersonalen vegetativen Stadium, dem Embryonalstadium wäre ein schwaches negatives und positives Lebensrecht. *Tieren* auf dem gleichen ontischen Niveau gestehen wir aber höchstens ein schwaches negatives Lebensrecht zu, kein positives Lebensrecht. Das positive Recht kommt bei *menschlichen Embryonen* nur hinzu wegen der Aussicht auf eine Entwicklung zur Person. Dieser Grund entfällt im postpersonalen vegetativen Stadium. Ein anderer Grund, Menschen im postpersonalen vegetativen Stadium doch auch dieses positive Lebensrecht zu gewähren, ist eventuell die speziesistische Solidarität. Diese Frage kann hier aber offenbleiben, denn schon dem Embryo steht nur ein *schwaches* (prima facie) negatives und positives Lebensrecht zu, das durch massive entgegenstehende Rechte der Mutter, der engen Umgebung oder der Gesellschaft übertrumpft werden kann. Ein solches entgegenstehendes Interesse ist bei Menschen im vegetativen Zustand aber *regelmäßig* gegeben: Die Versorgung des Menschen im vegetativen Zustand, um sein positives Lebensrecht zu gewährleisten, ist ziemlich aufwendig und u.U. äußerst langwährend.[4] Diese Ressourcen fehlen an anderer Stelle bei Menschen, denen sie sehr viel mehr Nutzen bringen können. (Weitere Disanalogien zum Embryonalstadium sind: Dort dauert die Entwicklung zum Fötalstadium nur drei Monate; vor allem aber erwartet man sich eine weitere Entwicklung und die Hervorbringung eines personalen Wesens.) Deshalb kann es unter diesen Bedingungen kein positives Lebens*recht* des Menschen im postpersonalen vegetativen Zustand geben. D.h., es gibt für niemanden eine *Pflicht*, für die Lebenserhaltung dieser Menschen zu sorgen.

Das Fehlen eines Lebensrechts schließt selbstverständlich nicht aus, daß sich insbesondere Angehörige entschließen, freiwillig für die positive Erhaltung des Menschen im vegetativen Zustand zu sorgen. Aber das ist ein privates Interesse, eine Liebhaberei, gewissermaßen ein Hobby; es ist keine moralische Pflicht. Und die Angehörigen haben dabei keinen Anspruch auf gesellschaftliche Unterstützung - wie man es bei seinen sonstigen privaten Interessen auch nicht hat.

Da sich Menschen im postpersonalen vegetativen Stadium nicht mehr selbst erhalten können, bedeutet dieses fehlende positive Lebensrecht in vielen Fällen ein Todesurteil. (Es wäre zu erwägen, ob man, um die Nerven der Umstehenden zu schonen und um weitere Ressourcen (Krankenhausbetten) zu sparen, dem

4 Die Kosten für die Pflege von Patienten im permanenten vegetativen Zustand in einer Einrichtung mit geschultem Personal betragen nach US-Angaben aus dem Jahr 1994, umgerechnet in heutige Euro ca. 126.000-180.000 Euro/Jahr. Allerdings werden Patienten mit apallischem Syndrom in Deutschland nach den ersten Monaten nach Hause entlassen. Dort bedürfen sie allerdings ebenfalls intensiver Pflege, medizinischer Behandlung, so daß die Angehörigen auf jeden Fall durch medizinisches Pflegepersonal unterstützt werden müssen. Dadurch wird die Behandlung zwar finanziell deutlich billiger als in einem Krankenhaus oder Pflegeheim, aber immer noch sehr teuer und welfaristisch gesehen nur wenig billiger, weil ja eben die Angehörigen nur finanziell kostenfrei eingespannt werden.

Kranken nicht den Gnadenstoß gibt, ihn also - z.B. durch KCl-Injektion - aktiv tötet.) Diese starken Thesen widersprechen der Intuition vieler Menschen. Aber diese Intuitionen müssen sich daraufhin befragen lassen, ob sie nicht einfach auf falschen Annahmen und dem trügenden Schein der im Wachkoma noch funktionierenden Reaktionen beruhen. Nach den neurophysiologischen Befunden fehlt jedoch einfach die Substanz [5] für ein stärker schützenswertes Gut: nämlich Gehirnaktivität, die Personalität und Empfindungsfähigkeit ermöglicht.

Literaturverzeichnis

Birnbacher, Dieter: Bioethik zwischen Natur und Interesse. Mit einer Einleitung von Andreas Kuhlmann. Berlin, Akademie Verlag, 2002. Nachdruck als Taschenbuch, Suhrkamp, Frankfurt/M., 2006

Feldman, Fred: "Death". In: *Craig, Edward (Hrsg.):* Routledge Encyclopedia of Philosophy. Bd. 2. Routledge, London/New York, 1998. S. 817-823

Lumer, Christoph: „Quellen der Moral. Plädoyer für einen prudentiellen Altruismus". Conceptus, 32, 1999. S. 185-216

Lumer, Christoph: Rationaler Altruismus. Eine prudentielle Theorie der Rationalität und des Altruismus. 2000. Mentis, Paderborn, 2., durchgesehene und ergänzte Auflage, 2009

McMahan, Jeff: The Ethics of Killing. Problems at the Margins of Life. Oxford [etc.], U.P. Oxford, xiii, 2002

Quante, Michael: Personales Leben und menschlicher Tod. Personale Identität als Prinzip der biomedizinischen Ethik. Suhrkamp, Frankfurt/M., 2002

Singer, Peter: "Is the Sanctity of Life Terminally Ill?". Bioethics, 9, 1995. S. 307-343. Wiederabdruck in: *Kuhse, Helga/ Singer, Peter (Hrsg.):* Bioethics. An Anthology. Blackwell, Oxford, 2. Auflage, 1999/2006. S. 292-301

Stoecker, Ralf: Der Hirntod. Ein medizinethisches Problem und seine moralphilosophische Transformation. Alber, Freiburg/München, 1999

5 Z.T. fehlt die Substanz im ganz wörtlichen Sinn: Im permanenten vegetativen Zustand atrophiert die Hirnsubstanz langfristig. Das Gehirn Terri Schiavos beispielsweise wog bei ihrem (biologischen) Tod nach 15 Jahren Wachkoma nur die Hälfte eines normalen Gehirns.

Stoecker, Ralf: „Sind hirntote Menschen wirklich tot?“. In: *Düwell, Marcus/Steigleder, Klaus (Hrsg.):* Bioethik. Eine Einführung. Suhrkamp, Frankfurt/M., 2003. S. 298-305

Youngner, Stuart: “The Definition of Death”. In: *Steinbock, Bonnie (Hrsg.):* The Oxford Handbook of Bioethics. Oxford U.P., Oxford [etc.], 2. Auflage, 2007/2009. S. 285-303

Strafe, Rache und retributive Gerechtigkeit

Julius Schälike
julius.schaelike@googlemail.com
Universität Konstanz

Abstract/Zusammenfassung

Präventionalistische Strafbegründungen gelten u.a. deshalb als problematisch, weil sie über kein plausibles Kriterium der Strafzumessung zu verfügen scheinen und den Delinquenten instrumentalisieren. Retributivistische Begründungen versprechen Abhilfe. In diesem Aufsatz werden eine Reihe retributivistischer Begründungsvorschläge geprüft und verworfen. Anschließend wird ein eigener Vorschlag entwickelt. Er rekurriert auf eine genuine Präferenz, Verbrechern Leid zuzufügen, die mit dem Wunsch nach Rache verwandt ist. Während der Retributivismus üblicherweise als deontologische Theorie gilt, erweist retributive Gerechtigkeit sich als ein intrinsisches Gut, das sich in einen konsequentialistischen Theorierahmen neben dem Gut der Wohlfahrt, auf das die Abschreckungstheorie abzielt, einfügt – ein Gut, dessen Gewicht allerdings gering ist.

1. Einleitung

Angesichts furchtbarer Verbrechen ist die Intuition besonders stark, dass der Täter bestraft werden muss. Die Vorstellung, dass er einen geruhsamen Lebensabend verbringen kann, ohne je belangt zu werden, ist schwer zu ertragen. Es erscheint vollkommen angebracht, dass großer Aufwand getrieben wird, ihn aufzuspüren, damit er sich vor Gericht verantworten muss und angemessen bestraft wird. Warum wollen wir, dass Verbrecher bestraft werden? Was steckt hinter solchen Intuitionen? Lassen sie sich rechtfertigen?

Diese Fragen sind von erheblicher ethischer Relevanz. Das wird deutlich, wenn man sich klar macht, was Strafe ist. Wer jemanden bestraft, der fügt ihm absichtlich Leid zu, weil er gegen eine Norm verstoßen hat. Unter normalen Umständen ist es moralisch falsch, jemandem absichtlich Leid zuzufügen. Man benötigt dazu einen guten Grund. Solche Gründe gibt es. So fügt es dem Bürger sicher ein Leid zu, dass er Steuern zahlen muss. Es gibt jedoch gute Gründe, ihm dies zuzumuten. Die Steuermittel werden vom Staat benötigt, um wichtige gesellschaftliche Aufgaben zu erfüllen. Wohl begründet ist es auch, in Notwehr einen Angreifer abzuwehren, selbst wenn diesem dadurch Schaden zugefügt wird, sofern der Schaden nicht unverhältnismäßig ist. Ist eine entsprechende Rechtfertigung von Strafe möglich?

Die unterschiedlichen Ansätze der Rechtfertigung von Strafe lassen sich anhand verschiedener Kriterien unterscheiden. Als Kriterium können beispielsweise temporale Aspekte dienen. Manche Begründungen orientieren sich an der Vergangenheit; ihnen zufolge verdient jemand Strafe allein aufgrund der Taten, die er verübt hat. Hiervon lassen sich zukunftsbezogene Begründungen abgrenzen; diese mögen der Vergangenheit durchaus Bedeutung beimessen, doch reicht der Umstand, dass etwas geschehen ist, allein nicht aus, um Strafe zu rechtfertigen. Darüber hinaus müssen durch das Strafen in der Zukunft Zwecke realisiert werden – Zwecke, die vom Zufügen der Strafe selbst unterschieden sind. Zu solchen Zwecken zählt die Prävention von Straftaten durch Abschreckung und die Besserung des Täters.

Geläufiger ist freilich die Unterscheidung zwischen konsequentialistischen und retributivistischen Strafbegründungen. Der Retributivismus wird dabei als eine *deontologische* Theorie verstanden. Dies ist jedoch, wie sich zeigen wird, unglücklich, da man retributive Gründe durchaus als konsequentialistische Gründe verstehen kann – etwa indem man in der Bestrafung Schuldiger einen *intrinsischen Wert* erblickt. Solch ein Konzept wäre konsequentialistisch, aber aus nicht-kontingenten Gründen vergangenheitsbezogen (da die Frage, wer schuldig ist, von der Vergangenheit abhängt). Ich werde daher mit der temporalen Klassifikation arbeiten und die retributivistischen Begründungen nicht den konsequentialistischen, sondern den präventionalistischen gegenüberstellen.[1]

In diesem Aufsatz werde ich eine Reihe von retributivistischen Begründungsversuchen diskutieren. Ich werde versuchen, den Retributivismus in einer unüblichen, konsequentialistischen Weise zu rekonstruieren und als Begründungskonzept zu rehabilitieren. Ich werde jedoch argumentieren, dass das Gewicht der mit diesem Ansatz verbundenen Gründe zu gering ist, um eine nennenswerte Rolle hinsichtlich der Frage der Legitimität harter Maßnahmen gegen Verbrecher zu spielen. Als weitaus gewichtiger – und ausschlaggebend – erweisen sich Gründe, die aus der Wohlfahrtsmoral resultieren und die für ein nicht-retributives Präventionsstrafrecht sprechen.

2. Präventionstheorien

Was retributive Theorien attraktiv macht, wird deutlich, wenn man sich die Probleme vor Augen führt, die sich mit dem stärksten Konkurrenzkonzept verbinden: der Präventionstheorie. Ihrer utilitaristischen Variante zufolge ist Strafe ge-

1 Dies ist allerdings nicht unproblematisch. Die zukunftsbezogenen Begründungen besitzen zwar alle einen präventiven Aspekt, insofern etwa jemand, der durch Strafe ein besserer Mensch wird, weniger Verbrechen verüben wird, doch muss die Besserung nicht unbedingt allein als *Präventionsmittel* gelten, sondern kann selbst als Ziel bzw. als Mittel für andere Zwecke geschätzt werden.

rechtfertigt, wenn sie mehr Nutzen als Schaden verursacht und es keine Alternative gibt, die noch mehr Nutzen generiert – wenn sie den Nutzen also maximiert. Diese Rechtfertigung wird aus verschiedenen Gründen kritisiert:[2]

1. Der Abschreckungskonzeption wird vorgeworfen, sie behandle Personen nicht gebührlich als autonome moralische Subjekte, die moralischer Argumentation zugänglich sind, sondern, so Hegel, wie Hunde, die abgerichtet werden.[3]
2. Eine andere Kritik ist kantianischer Provenienz. Sie lautet, der Bestrafte werde auf dem Altar der Nützlichkeit geopfert. Er werde instrumentalisiert, damit andere sicherer leben könnten. Mit Kant sei jedoch zu fordern, dass Menschen nie allein als Mittel, sondern stets auch Zweck behandelt werden. Als autonomes Wesen verdiene er einen Respekt, der mit seiner Instrumentalisierung nicht vereinbar sei.
3. Die Abschreckungskonzeption macht keine grundsätzliche Unterscheidung zwischen Schuldigen und Unschuldigen. Wenn es nützlich ist, ist es ihr zufolge legitim, auch Unschuldigen, die von hinreichend vielen für Täter gehalten werden, absichtlich Leid zuzufügen.
4. Das Problem der Opferung Unschuldiger ist nur der extremste Ausdruck eines generellen Problems: das der Taxierung des Strafmaßes. Maßstab ist für einen Utilitaristen der maximal erzielbare Nutzen, die Schuld des Täters hingegen ist grundsätzlich bedeutungslos. Wenn die drakonische Bestrafung von Bagatelldelikten die Nutzensumme maximiert, so ist sie unter präventiven Gesichtspunkten legitim.[4]

3. Retributivismus I: Rechtsverwirkung und Zustimmung

Retributivistische Ansätze versprechen hier Abhilfe. Ich werde nun die vielleicht wichtigsten einer kurzen Prüfung unterziehen.

Der *Rechtsverwirkungstheorie* zufolge verliert jemand Rechte, die ihm eigentlich zukommen, wenn er dem Verlust *zustimmt*. Wer weiß, dass bestimmte Taten mit Strafe bedroht sind, und sie dennoch vollzieht, der hat implizit zugestimmt, dass der Staat ihn bestrafen darf. Er hat sein Recht auf Freiheit bzw. auf einen Teil seines Geldes aufgegeben.[5]

Ist Zustimmung wirklich ausreichend, um einen Täter legitim bestrafen zu können? Angenommen, ein Gangster hat einige Menschen in seine Gewalt gebracht. Er teilt ihnen mit, ihnen werde nichts geschehen, wenn sie seine Befehle befolgten. Versuchten sie jedoch, zu fliehen oder sich sonst wie zu widersetzen, werde er sie töten. In einem oberflächlichen Sinne stimmen diejenigen, die nicht gehorchen, ihrer Bestrafung zu. Dennoch ist dies sicherlich keine moralisch ak-

2 Cf. auch Hallich 2011.

3 Hegel 1821, 190 (§99, Zusatz).

4 Wie Hallich darlegt, wird der Tendenz zu drakonischer Bestrafung von Bagatelldelikten allerdings dadurch entgegengewirkt, dass ein solches Strafrecht Angst und Schrecken verbreitet, cf. Hallich 2011.

5 Cf. Morris 1968, 479; Goldman 1979.

zeptable Weise, mit Menschen umzugehen. Die oberflächliche Zustimmung allein berechtigt nicht dazu, jemandem absichtlich Leid zuzufügen.[6]

Vielleicht lässt sich der Ansatz retten. Die Opfer der Entführer wollen nicht, dass Menschen entführt werden. Sie lehnen das normative System ab, das ihnen aufgezwungen wird. Das Strafrecht hingegen stellt ein normatives System dar, das im Interesse aller ist, auch der Verbrecher. Auch sie stimmen diesem System zu, denn sie wollen nicht Opfer von Taten werden wie der, die sie verüben. Insofern ist ihnen das System nicht aufgezwungen. Zustimmung macht, so die präzisierte These, Strafe nur dann legitim, wenn sie das gesamte normative System einschließt.[7]

Dieser Rettungsversuch scheitert jedoch. Es stimmt nämlich nicht, dass ein Verbrecher der Norm, gegen die er verstößt, zustimmen muss. Ein Vergewaltiger etwa könnte es vorziehen, in einem Normensystem zu leben, das Vergewaltigung nicht unter Strafe stellt. Insbesondere wenn man bestimmte Formen von Vergewaltigung definiert als sexuellen Übergriff eines Mannes gegenüber einer Frau, zeigt sich, dass die Norm, die diese Formen von Vergewaltigung verbietet, den Männern gar nicht bzw. nur indirekt – über ihre Anteilnahme am Wohl von Frauen – nützt, so dass sie ihr nicht aus Eigeninteresse zustimmen müssen.[8] Aber auch wenn man die Norm allgemeiner fasst, so dass sie jede Form von sexueller Gewalt verbietet, könnte der Täter sich weigern, ihr zuzustimmen, da sie ihm angesichts der geringen Wahrscheinlichkeit, Opfer zu werden, mehr schadet als nützt. Bei einem Sado-Masochisten versagt die Theorie vollends, da für ihn die Norm nicht einmal *pro tanto* zustimmungswürdig ist.

Der Rechtsverwirkungstheoretiker könnte sich angesichts dieser Schwierigkeiten von der Idee verabschieden, Rechtsverwirkung an Zustimmung zu knüpfen. Er könnte darauf verweisen, dass Rechte Pflichten implizieren, dieselben Rechte anderer zu achten, und behaupten, daraus folge, dass Rechte hinfällig werden, wenn den Pflichten nicht nachgekommen werde. Diese Behauptung ist jedoch inakzeptabel. Ein Richter, der unfaire Verhandlungen führt, hat klarerweise das Recht auf eine faire Verhandlung; wer das Recht anderer auf freie Rede verletzt, verliert doch selbst dieses Recht nicht.[9] Auf diese Weise lässt sich daher nicht begründen, dass der Staat rechtsbrüchige Bürger durch Strafe schädigen darf. Aber selbst wenn doch irgendwie gezeigt werden könnte, dass jemand, der ein Recht bricht, eigene Rechte verwirkt, so dass er bestraft werden *dürfte*, bliebe doch unklar, warum er bestraft werden *sollte*. Auf diese Frage gibt die Rechtsverwirkungstheorie keine Antwort.

6 Cf. Burgh 1982, 199.
7 Cf. Morris 1968, §2.
8 Cf. Burgh 1982, 201.
9 Burgh 1982, 198.

4. Retributivismus II: Relationale Gerechtigkeit (Fairness)

Einem anderen einflussreichen Vorschlag zufolge verdient der Täter Strafe, weil er auf unfaire Weise Vorteile gegenüber den rechtstreuen Bürgern erlangt hat. Strafe gleicht diesen unverdienten Vorteil wieder aus, stellt somit Gerechtigkeit her.[10]

Worin nun besteht der unfaire Vorteil, und wie hängt er mit den vollzogenen Handlungen zusammen? Herbert Morris[11] schlägt eine kontraktualistische Interpretation moralischer Normen vor: Moralische Normen sind das Produkt eines Handels zwischen den Subjekten, auf sie einigt man sich, da man, obgleich man etwas aufgibt, etwas anderes, Wertvolleres gewinnt. Man verzichtet etwa darauf, andere zu töten, gewinnt jedoch die Sicherheit, selbst nicht getötet zu werden. Aus der allgemeinen Befolgung des moralischen Normensystems erwachsen allen Beteiligten Kooperationsgewinne. Die Normen müssen aus unterschiedlichen Gründen von Sanktionen begleitet werden. Zum einen schafft dies Anreize für normenkonformes Verhalten – dies ist eine präventionistische Begründung. Zum anderen jedoch – und dies ist in unserem Kontext entscheidend – sorgen Strafen für Gerechtigkeit. Sie annullieren den unfairen Vorteil, der aus der Normverletzung erwächst, und stellen so das Gleichgewicht von Kosten und Nutzen wieder her.

Worin aber genau besteht der unfair erlangte Vorteil des Verbrechers? Laut Morris sind hier zwei Aspekte zu beachten. Zum einen profitiert der Täter von der Selbstbeschränkung, die sich die anderen, Gesetzestreuen, auferlegen, er genießt die entstehenden Schutzräume (*spheres of noninterference*). Zum anderen zahlt er nicht den Preis, der darin besteht, sich in gleicher Weise zu beschränken, also die „*burdens of self-restraint*“ zu tragen. Die entstandene Ungerechtigkeit lässt sich entweder dadurch ausgleichen, dass man den Nutzen, der aus den Schutzräumen gezogen wurde, annulliert, oder dadurch, dass der Preis nachträglich entrichtet wird.[12]

In welcher Währung aber sind Kosten und Nutzen zu messen? Plausibel ist der Verweis auf Präferenzenbefriedigung bzw. -frustration. Beim Gesetzestreuen bleiben einige Präferenzen unbefriedigt, die der Verbrecher hat befriedigen können, diejenigen nämlich, deren Befriedigung die Verletzung der Normen impliziert hätte. Auch der Verbrecher war andererseits in der Lage, Präferenzen zu realisieren, die ohne die Zurückhaltung der anderen frustriert worden wären. Der Ausgleich bestünde darin, Präferenzen des Verbrechers zu frustrieren, die entweder von ähnlichem Gewicht sind wie die, die er durch seine Tat hat befriedi-

10 Vertreter dieses Ansatzes sind Morris 1968, Murphy 1973, Sher 1987, Kap. 5, Dagger 1993 und Wolf 2003.

11 Morris 1968.

12 Burgh 1982, 203.

gen können, oder wie die, deren Befriedigung durch die Schutzräume ermöglicht wurden.

Dieser Vorschlag ist konsistent, hat aber problematische Implikationen. Betrachten wir zunächst den Nutzen, den der Täter aus den Schutzräumen zieht. Es stimmt einfach nicht, dass alle im gleichen Maße davon profitieren, dass alle anderen Menschen moralischen Normen folgen. Manche benötigen manche Schutzräume gar nicht, da sie nichts haben, was zu schützen wäre. Wer besitzlos ist, profitiert nicht vom Eigentumsrecht und vom Verbot der Unterschlagung. Manche Männer profitieren möglicherweise nicht einmal indirekt vom Verbot der Vergewaltigung von Frauen.[13]

Wie steht es um die Kosten, die mit der Normentreue einher gehen? Hier ergibt sich ein ähnliches Bild: Die allermeisten Menschen haben gar nicht den Wunsch, andere Menschen zu töten oder zu vergewaltigen. Vielen fehlen die Fähigkeiten, bestimmte Verbrechen zu begehen, etwa sich in fremde Computernetze einzuhacken.[14] Gesetzestreue stellt für sie somit gar keine Bürde dar: Das, was der Verbrecher hat, ist nichts, was sie auch gerne hätten, und auch hätten haben können, wenn sie versucht hätten, die Normen zu brechen.

Dies aber bedeutet, dass eine unfaire Kosten-Nutzen-Verteilung nicht nur dann entstehen kann, wenn Normen gebrochen werden, sondern auch dann, wenn alle sich an die Regeln halten. Ein Normverstoß kann unter diesem Umständen sogar dazu führen, dass die Unfairness *reduziert* wird. Eine Bestrafung des Täters ließe sich dann nicht im Rekurs auf Fairness begründen. Dies ist insofern problematisch, als es intuitiv auch dann angemessen erscheint, etwa einen Vergewaltiger zu bestrafen, wenn er von der Norm, gegen die er verstieß, gar nicht profitierte bzw. durch sie sogar benachteiligt wurde.

Der auf Fairness bezogene Ansatz interpretiert alle Verbrechen als Form des Trittbrettfahrens. Trittbrettfahren ist unfair, und Unfairness ist sicher moralisch zu kritisieren, aber nicht alle Verbrechen implizieren Unfairness, und auch bei denen, die einen Fairnessaspekt besitzen, ist dieser Aspekt oftmals nicht das, weshalb sie in erster Linie unmoralisch sind und nach Strafe rufen. Wer einen Mord oder eine Vergewaltigung als unfair kritisiert, der hat sicherlich nicht den Punkt getroffen, der diese Verbrechen moralisch so empörend macht. Empörend sind sie zweifellos primär deshalb, weil hier absichtlich Schaden beim Opfer verursacht wird. Der Fairness-Ansatz ist blind für dieses moralische Übel.

13 Dann nämlich, wenn ihnen das Wohl aller Frauen gleichgültig ist (cf. Burgh 1982, 205).
14 Cf. Boonin 2008, 123.

5. Retributivismus III: Expression moralischer Missbilligung

Eine andere retributive Begründungskonzeption hebt auf die expressive Dimension von Verbrechen und Strafe ab. Sie basiert darauf, dass Strafe nicht immer, aber doch oftmals damit einhergeht, dass der Strafende durch die Sanktion seine moralische Missbilligung ausdrückt. Wer einen Strafzettel für Parken ohne Parkschein erhält, muss sich nicht unbedingt als moralisch gerügt vorkommen. Bei schweren, moralisch empörenden Verbrechen hingegen bringt die Strafe immer auch die Ablehnung der Tat und des Täters zum Ausdruck.[15] Mit dieser expressiven Qualität von Strafe geht eine kommunikative Leistung einher: Dem Täter im Besonderen und der Gesellschaft im Allgemeinen wird mitgeteilt, dass die strafende Instanz bestimmte Taten missbilligt.[16] Wie lässt sich daraus eine Legitimation von Strafe ableiten?

Dies sieht man, wenn man sich fragt, was genau an einem Verbrechen eigentlich missbilligenswert ist. Einem Vorschlag von Jean Hampton[17] zufolge ist es die falsche moralische Botschaft, die sich in der Tat ausdrückt. Der Täter behauptet implizit, er sei mehr wert als sein Opfer und dürfe es sich deshalb unterwerfen. Die Idee nun ist, dass der Täter deshalb bestraft werden muss, damit der falschen Botschaft widersprochen wird.

Das Problem besteht jedoch darin, dass es alles andere als harmlos ist, Missbilligung auszudrücken und falsche Botschaften richtigzustellen, indem man sich des Mittels der Strafe bedient. Zu klären ist deshalb, ob es nicht weniger bedenkliche Weisen gibt, diese Funktion zu erfüllen.

Das scheint durchaus der Fall. Joel Feinberg weist darauf hin, dass sich Gerichtsverfahren so gestalten ließen, dass sie nicht mit der Verhängung von Strafen, sondern mit feierlichen Proklamationen moralischer Entrüstung beendet werden. Denkbar sind elaborierte öffentliche Rituale, die sich bewährter Mittel bedienen, die der Religion, der Musik und dem Theater entlehnt sind.[18] Kann es angesichts dieser Alternativen zulässig sein, das Mittel zu verwenden, absichtlich Leid zuzufügen?

Jean Hampton hält Strafe für unverzichtbar. Nur so könne die falsche Botschaft wirklich annulliert werden. Doch es fragt sich, ob eine falsche Botschaft auf diese Weise tatsächlich aus der Welt geschafft wird. Ungeschehen machen kann man ihre Äußerung ja nicht mehr. Bestenfalls kann erreicht werden, dass alle einsehen, dass die Botschaft falsch ist. Aber warum sollte dies gerade durch Strafe erreicht werden, und nicht durch moralische Argumentation sowie die von Feinberg skizzierten nichtpunitiven Maßnahmen? Was wäre, wenn ohnehin

15 Cf. Feinbergs Unterscheidung zwischen *punishment* und *penalty*, Feinberg 1965, 95ff.
16 Cf. Duff 2001, 27ff.; 79ff.
17 Hampton 1992.
18 Feinberg 1965, Kap. V.

niemand außer dem Täter an die falsche Botschaft geglaubt hätte, der Täter jedoch unbelehrbar wäre? Entfiele dann der Strafgrund?

Der vielleicht gravierendste Einwand gegen den expressiven Retributivismus ist folgender: Es ist unplausibel, dass es in erster Linie die Falschheit seiner Botschaft ist, die einen Verbrecher so strafwürdig macht. Die Botschaft kann auch ohne Verbrechen ausgedrückt werden. Klarerweise ist es jedoch viel schlimmer, eine Vergewaltigung zu begehen, als zu sagen, Vergewaltigung sei in Ordnung. Was eine Vergewaltigung so schlimm macht, ist nicht primär die Botschaft, die sich in ihr ausdrückt, sondern die Art und Weise, *wie* sie sich ausdrückt: durch die vorsätzliche oder fahrlässige Erzeugung von Leid.[19]

6. Retributivismus IV: Non-relationale (vergeltende) Gerechtigkeit und Rache

Angesichts dieser zahlreichen Probleme ist das Urteil des Strafrechtlers Claus Roxin nachvollziehbar, die retributive Strafbegründung sei „heute [...] nicht mehr haltbar."[20] Es drängt sich der Verdacht auf, dass die retributiven Begründungen von Strafe, die in der Diskussion anzutreffen sind, nichts als Rationalisierungen bestimmter übelwollender Motive sind, zu denen offen zu stehen vielen[21] Teilnehmern an der Debatte peinlich ist. Bemäntelt sich mit diesen Begründungen einfach ein archaischer Racheimpuls? James Fitzjames Stephen hat die Ansicht vertreten, das Strafrecht verhalte sich zum Racheaffekt wie die Ehe zum Sexualtrieb.[22] Das Strafrecht wäre demnach eine „zivilisierte" Art und Weise, übelwollende Strebungen zu kanalisieren und zu ihrem Ziel gelangen zu lassen. Aber sollte man solche Affekte nicht eher unterdrücken, als Institutionen schaffen, sie auszuleben?

Zunächst einmal ist festzuhalten, dass es nicht immer *direkt* der Wunsch nach Rache ist, der sich in den retributiven Intuitionen ausdrückt. Rache setzt voraus, dass man selbst oder jemand, zu dem man in einer persönlichen Beziehung steht, der Geschädigte ist. Man meint aber auch, dass Täter Strafe verdienen, die Subjekte geschädigt haben, zu denen die persönliche Beziehung fehlt.[23] Es ließe sich

19 Cf. Dolinko 1991, 552.

20 Roxin 1994, 42.

21 Nicht allen - den Zusammenhang zwischen dem Rache- und dem Strafbedürfnis erkennen *und billigen* u.a. Smith 1759, Stephen 1863, Oldenquist 1988 und Hershenov 1999.

22 Stephen 1863, 99.

23 Cf. Nozick 1981, 367. Nozick zählt vier weitere Merkmale auf, durch die sich Rache von retributiver Vergeltung unterscheidet (366-368): (i) Rache folgt keinen generellen Prinzipien, (ii) sie impliziert Freude am Leid des anderen, (iii) sie reagiert nicht notwendig auf eine Rechtsverletzung (*wrong*), sondern möglicherweise lediglich auf eine Schädigung (*injury/harm/slight*), (iv) sie hat kein internes Maß, während Vergeltung nach Maßgabe

im Stile David Humes[24] und John Stuart Mills[25] eine Genealogie des Sinnes für retributive Gerechtigkeit zeichnen, in der die Disposition zu Racheaffekten, gemeinsam mit der Einstellung der Sympathie, eine unpersönliche Haltung des Übelwollens gegenüber Menschen hervorbringt, die unmoralisch handeln.[26] Wenn Hans mit Klaus sympathisiert, und Peter schädigt Klaus, so reagiert Hans affektiv ähnlich, wie wenn Peter ihn selbst geschädigt hätte. Diese durch Sympathie zu einer unpersönlichen Haltung transformierte Disposition, sich zu „rächen", wäre das, was man den *Sinn für vergeltende Gerechtigkeit* – genauer gesagt: die *Präferenz* für vergeltende Gerechtigkeit – nennen könnte. Unklar ist, ob dieser Sinn sich aus den Motiven, aus denen er entstand, immer noch speist, oder ob er mittlerweile eine eigenständige Einstellung bildet. Klar scheint jedoch, dass er ein besseres Prestige genießt als seine genetischen Ursprünge. Was ihn problematisch macht, ist nicht, dass er der Moral entgegengesetzt ist. Vielmehr bildet er selbst eine Quelle moralischer Normativität, der Normativität vergeltender Gerechtigkeit nämlich. Retributive Gerechtigkeit erschöpft sich nicht in einer Spielart relationaler, fairnessbezogener Gerechtigkeit. Zusätzlich liegt dem Retributivismus ein non-relationaler, „vergeltender Gerechtigkeitssinn" *sui generis* zugrunde. Es geht hier nicht um die gerechte Verteilung von Kosten und Nutzen, sondern um eine universelle, unpersönliche Disposition, die dem Streben nach Rache verwandt ist.

Wie ergeben sich aus dem vergeltenden Gerechtigkeitssinn konkrete praktische Direktiven? Welche Strafe verdient ein Rechtsbrecher? Die Schwierigkeiten, die sich hier zeigen, sind vielfältig und gravierend; sie lassen es zweifelhaft erscheinen, dass der Retributivismus überhaupt anwendbar ist. Wenn es den vergeltenden Gerechtigkeitssinn tatsächlich gibt, so muss das Maß der angemessenen Vergeltung in diesem Sinn liegen. Es gilt, den Gehalt der diesen Sinn konstituierenden Präferenzen auszubuchstabieren. Als plausibel wird traditionell die biblische *lex talionis* erachtet: Auge um Auge, Zahn um Zahn. Doch in vielen Fällen greift dieses Prinzip nicht, etwa dann, wenn ein Blinder einen Sehenden blendet oder ein Kinderloser das Kind eines anderen entführt. Bei anderen Verbrechen fehlen die Opfer, insofern niemand einen messbaren Schaden erleidet, etwa bei Steuerbetrug oder Insiderhandel. Zudem muss sicherlich der *mens rea* Rechnung getragen werden: Wenn A erblindet, weil B fahrlässig gehandelt hat,

---

des Gewichts der Rechtsverletzung taxiert wird. Walker zweifelt die Merkmale (iii) und (iv) an (Walker 1995, 581f.).

24 Hume, *Treatise*, III. 2. ii.

25 Mill 1871, Kap. 5.

26 Cf. Mackie 1982; Singer 2005, 336. Analog lässt sich die Genealogie des Sinnes für distributive Gerechtigkeit beschreiben, die diesen als Derivat der Präferenz des Neides fasst, cf. Crisp 2003, Schälike 2009.

verdient B eine mildere Strafe, als wenn er die Tat absichtlich vollzogen hätte.[27] Unklar ist auch, wie fehlgeschlagene Versuche zu behandeln sind. Klar scheint zunächst nur, dass Taten, die moralisch kritikwürdiger sind, schwerere Strafen verdienen. Wie jedoch gelangt man von diesem Proportionalitätsgebot zu einem Prinzip, das konkreten Taten konkrete Strafen zuordnet?

Benötigt werden zwei Skalen, die einerseits die Verbrechen, andererseits die Strafen nach Maßgabe ihrer Gewichtigkeit ordnen. Verbunden werden die Skalen an den Punkten, von denen folgendes gilt: Erstens korrespondieren diesen Punkten Taten, für die der Täter die volle Verantwortung trägt, so dass in Bezug auf die *mens rea* keine Diskontierung erforderlich ist; zweitens haben diese Taten ein konkretes Opfer, das dem Täter in relevanter Hinsicht vergleichbar ist. Gemäß der *lex talionis* verdient es der Täter, dass an ihm exakt dieselbe Tat als Strafe vollzogen wird. Dass der Täter vom Opfer in relevanter Weise unterschieden ist, könnte etwa daran liegen, dass das Opfer durch die Tat stärker beeinträchtigt wird, als es der Täter durch die entsprechende Strafe wäre. Beispielsweise wird das Opfer am Bein so schwer verletzt, dass es seinen Beruf als Fußballspieler nicht mehr ausüben kann. Der Täter hingegen hat einen Beruf, für den voll funktionsfähige Beine nicht vorausgesetzt sind. Dann müsste die Strafe ihn in anderer Weise, die für ihn ähnlich hinderlich ist, schädigen. Ebenso wäre ein reicher Täter, der einem armen Opfer 1.000€ stiehlt, nicht ausreichend bestraft, wenn man ihm 1.000€ nähme. Der Täter soll in dem Maße in seiner Lebensqualität beeinträchtigt werden, in dem er die Lebensqualität seines Opfers reduziert hat. Entsprechend müsste auch bei Taten verfahren werden, bei denen sich die *lex talionis* ohnehin nicht direkt anwenden lässt, weil der Täter das, was er beim Opfer zerstört, selbst gar nicht besitzt.

Offenkundig sind die epistemischen Probleme, die sich mit dem Versuch verbinden, Differenzen der Lebensqualität zu quantifizieren, erheblich. Entsprechend schwierig ist es, Strafen auf einer Skala zu ordnen. Welche Geldstrafe ist welcher Gefängnisstrafe gleichwertig?[28] An seine Grenze gelangt der Ansatz, der bei der Taxierung der Strafe auf die verursachte Einbuße an Lebensqualität abstellt, bei Taten, die keine identifizierbaren Opfer haben, etwa Steuerhinterziehung. Um sie einzubeziehen, muss die Bezugsbasis erweitert werden: Schädigung der Lebensqualität ist nur eine Form unmoralischen Verhaltens, Unfairness durch Trittbrettfahren eine andere. Wenn retributive Strafe sich am Grade der moralischen Falschheit bemisst, so ist es erforderlich, Taten, die Opfer haben, und Taten, die sich in Unfairness erschöpfen, in eine Rangordnung zu bringen. Diese Aufgabe dürfte sich als äußerst schwierig erweisen. Grundsätzlich unlösbar erscheint sie mir nicht. Ist sie bewältigt, können allen Taten angemessene Vergeltungsstrafen zugeordnet werden. Wäre etwa ein bestimmter Akt der Steu-

27 Cf. Shafer-Landau 2000, 193. Einen entsprechenden Vorschlag macht Nozick 1981, 363-397.

28 Cf. Shafer-Landau 2000, 202f.

erhinterziehung moralisch ebenso schlimm wie die Schädigung eines Menschen, der dem Täter in relevanter Hinsicht gleicht, durch Freiheitsentzug einer bestimmten Dauer, so könnte die Tat durch eine solche Freiheitsstrafe vergolten werden.

Eine Komplikation entsteht, wenn ein Normenverstoß in einem Umfeld geschieht, das selbst moralisch defizient ist. Verdient jemand, der einen Raub begeht, Strafe, wenn er selbst massiv unter Benachteiligung litt? Vergeltende Strafe ist verdient nach Maßgabe der moralischen Schuld. Wenn die Umstände eine Tat, die in einem fairen Umfeld moralisch falsch wäre, moralisch rechtfertigen, so verdient der Täter keine Strafe. Auch wenn die Umstände die Tat nicht gänzlich rechtfertigen, so können sie doch die Schuld mindern, so dass die Strafe milder sein muss.

Muss das Leid, das der Täter verdient, ihm *mit der Intention* auferlegt werden, ihn für die Tat zu strafen? Oder erfüllen auch Naturereignisse die retributiven Präferenzen? Angenommen, jemand hat einen anderen gefoltert. Später erleidet er einen Autounfall, der ihn in vergleichbarer Weise schädigt. Stellt dies vergeltende Gerechtigkeit her? Was, wenn der Unfall sich *vor* der Folterung ereignet hätte? Hätte die Person dann gleichsam ein Guthaben auf ihrem „Retributionskonto" erworben, das sie zum straffreien Begehen von Verbrechen berechtigt? Ja wäre sogar etwas Gutes daran, wenn jemand, der so ein Guthaben erworben hat, Verbrechen verübt, da dies das Konto ausgleicht? Das wäre sicherlich sehr unplausibel. Dennoch lässt sich nicht leugnen, dass es intuitiv weniger dringlich erscheint, jemanden, der unmoralischer Taten schuldig ist, zu bestrafen, wenn er gleichsam „vom Schicksal bereits bestraft wurde". Dies scheint mir aber nicht daran zu liegen, dass dem Täter bereits vergeltende Gerechtigkeit widerfahren ist, sondern daran, dass sich in die Beurteilung der Frage, wie mit ihm zu verfahren ist, einerseits Mitleid, andererseits andere, relationale[29] Gerechtigkeitsaspekte mischen, die die Gründe, die für eine weitere Schädigung sprechen, abschwächen. Vergeltende Gerechtigkeit kann – anders als distributive Gerechtigkeit (Fairness) – nur durch *Strafe* hergestellt werden: Das Leid muss *absichtlich* und *für einen Normenverstoß* zugefügt werden, verbunden mit dem Wunsch, dass der Delinquent *weiß*, warum ihm das Leid widerfährt.[30] Vergeltende Gerechtigkeit ist nicht-relational. Daher fehlt der Vergleichsmaßstab, an dem sich zeigen könnte, wann der Täter – bezogen etwa auf den Durchschnitt der Lebensqualität aller anderen – genug gelitten hat. Das Maß wird allein durch die Tat vorgegeben, ohne Bezug auf ein „Normalmaß" an Lebensqualität oder

29 So geht es dem Täter vielleicht durch den Unfall so viel schlechter als den anderen, dass er einen Ausgleich verdiente.

30 Dies verbindet Vergeltung mit Rache. Nicht von ungefähr ist es „privaten" Rächern ein Anliegen, ihrem Opfer, bevor sie ihm die Strafe zufügen, den Grund wissen zu lassen („Das ist für x!"). Eine detaillierte, 9-teilige Analyse vergeltender Retribution liefert Nozick 1981, 369.

die relative Situierung in einer Vergleichsklasse, so dass Abweichungen von solchen Relata nicht in Rechnung zu stellen sind.

Welche Rolle spielt der *Vollzug* der unmoralischen Tat für das retributive Vergelten? Man könnte denken, der Normenverstoß liege bereits dann vor, wenn jemand moralisch kritikwürdige Charakterzüge oder Willenshaltungen besitzt. Unmoralische Handlungen, so könnte man argumentieren, sind ja lediglich kontingente *Äußerungsformen* solcher Laster; daher sind sie für die Bewertung des Subjekts irrelevant.[31] Da der Verstoß gegen Willens- bzw. Charakternormen grundlegender erscheint als der gegen Handlungsnormen, könnte man meinen, jemand verdiente Strafe nach Maßgabe der Taten, die er verüben *würde*, böte sich ihm die Gelegenheit dazu. Tatsächlich reagiert der Sinn für retributive Gerechtigkeit jedoch nicht auf Willens- oder Charakterfehler, sondern auf fehlerhaftes *Handeln*.[32] So wenig Taten, die jemand potentiell verüben würde, Rachegelüste hervorrufen, so wenig stimulieren sie den Sinn für retributive Gerechtigkeit.[33]

Entfällt der Strafgrund, wenn der Täter aufrichtig bereut? Intuitiv spricht vieles dafür.[34] Dies erscheint auch gerechtfertigt: Reue kann die Tat zwar nicht ungeschehen machen, doch ihre Beziehung zum Täter lockern. In gewisser Weise ist ein reuiger Täter nicht mehr derselbe, der er war, als er die Tat verübte. Der Makel der Tat überträgt sich deshalb nicht mehr auf ihn, gleitet an ihm ab. Ihm kann – ja muss – verziehen werden.[35]

7. Das Gewicht retributiver Gerechtigkeit

Retributive Gerechtigkeit besitzt zweifellos moralisches Gewicht, doch erschöpft sich weder die Gerechtigkeit, noch die Moral in ihr. Weitere relevante Aspekte sind distributive Gerechtigkeit und das gute Leben. Gerade mit diesem letzten Aspekt, den man mit dem Begriff „Wohlfahrtsmoral“ (*welfarism*) bezeichnet und der im Zentrum die Norm enthält, Glück zu steigern bzw. Leid zu verringern, konfligiert die Herbeiführung retributiv-vergeltender Gerechtigkeit

31 So etwa Kant, GMS 394.

32 In Nozicks Terminologie: „flouting of values“ legitimiert retributiv-vergeltende Strafe, nicht der Umstand, dass das Subjekt „is anti-linked with values“ (Nozick 1981, 382-384).

33 Es erscheint sogar semantisch verquer, dass man sich für hypothetische Taten rächt; ähnlich falsch erscheint es, dass hypothetische Taten faktische Ungerechtigkeit erzeugen.

34 Cf. Nozick 1981, 372.

35 Cf. Schälike 2010, Kap. 15. – Die Reue muss allerdings sehr tief gehen, sie muss einen Wandel indizieren, für den gilt: käme der Täter noch einmal in eine vergleichbare Tatsituation, würde er die Tat auch dann nicht noch einmal verüben, wenn er meinte, dass er keine Strafe zu befürchten hätte. – Schadenersatzansprüche werden hingegen *nicht* gegenstandslos, da sie nicht an moralische Schuld, sondern allein an kausale Verantwortung geknüpft sind.

häufig, da Letztere intrinsisch darauf abzielt, Leid zuzufügen, während Erstere genau dies verbietet. Diese Spannung entsteht innerhalb der Moral selbst.

Aber ist es nicht verwunderlich, dass unsere moralischen Einstellungen so anfällig für Konflikte sind? Dass wir die retributiven Einstellungen sowie die Dispositionen für Groll und Rache entwickelt haben, hat durchaus einen guten Sinn. Sie tragen dazu bei, antimoralische Impulse zu zügeln. Wer mit dem Gedanken spielt, andere zu schädigen, wird ihre retributiven Reaktionen in Rechnung stellen müssen, und dies wirkt abschreckend. Abschreckung jedoch arbeitet der Wohlfahrtsmoral in die Hände. So zeigt sich, dass beide Aspekte der Moral, retributive Gerechtigkeit und Wohlfahrt, in der Tendenz durchaus im Einklang miteinander stehen.[36] Im Einzelfall jedoch kann der retributive Impuls das Wohlfahrtsstreben durchkreuzen, dann etwa, wenn er Taten betrifft, die zu vollziehen sich potentielle Täter gar nicht abschrecken lassen. Auch was die *Dosierung* der Strafe betrifft, können Diskrepanzen entstehen. Retributiv gesehen, verdienen schwerere Taten härtere Strafen, aber das kann unter Wohlfahrtsaspekten anders sein.

Wie ist mit solchen Spannungen innerhalb der Moral umzugehen? Grundsätzlich scheinen Gerechtigkeitsansprüche wesentlich schwächer zu sein als die Ansprüche der Wohlfahrtsmoral; dies gilt nicht nur für vergeltende, sondern auch für relational-distributive Gerechtigkeit.[37] Das sieht man, wenn man über konkrete Fälle nachdenkt. Direkt Betroffene fordern zwar häufig „Auge um Auge“, so etwa jüngst eine iranische Frau, die von einem verschmähten Verehrer mit Schwefelsäure so schwer im Gesicht verätzt wurde, dass sie erblindete.[38] Und in der Tat: Eine gerechte Vergeltung wäre das. Dennoch sollte dies – gesetzt, die abschreckende Wirkung legitimiert es nicht – so wenig umgesetzt werden wie die Todesstrafe für einen Mörder. Dass direkt Betroffene es anders sehen, liegt daran, dass sich bei ihnen retributive Gerechtigkeitsforderungen mit direkten Rachegelüsten überlagern, die ihre motivationale Kraft potenzieren. Zudem beeinträchtigen die involvierten heftigen Affekte die sachgerechte Beurteilung der Situation im Lichte der relevanten Präferenzen.[39] Das Gewicht retributiver Gerechtigkeit als solcher erweist sich in der distanzierten Betrachtung nicht direkt Betroffener, die sich zuvor klar gemacht haben, dass die gängigen Begründungsversuche für den Retributivismus scheitern, so dass sie allein ihre

36 Cf. Mackie 1982.

37 Cf. Schälike 2009. Ähnlich meint Nozick, retributive Vergeltung sei, verglichen mit den Auswirkungen von Strafe auf das Wohl, „of lesser value and not as desirable. Yet still, it is of some considerable value“ (Nozick 1981, 375f.).

38 DER SPIEGEL 52 (2008), 83.

39 Der Gerechtigkeitssinn ist nicht in Affekten fundiert, sondern in Präferenzen. Diese Präferenzen sind zwar psychisch mit Affekten verbunden und fundieren diese auch, doch zeigt sich das normative Gewicht der Präferenzen nicht in der Stärke der Affekte, sondern in der Stärke der Präferenzen bei ruhiger und informierter Betrachtung.

intrinsische Präferenz zugunsten retributiver Gerechtigkeit artikulieren und gewichten. Im Konfliktfall zwischen retributiver Gerechtigkeit und Wohlfahrt müssten sicherlich regelmäßig die Gerechtigkeitsansprüche weichen.[40]

Im Lichte einer Wohlfahrtsmoral lässt sich ein Präventionsstrafrecht begründen. Abschreckung greift sehr gut bei leichteren und mittelschweren Delikten, deren Täter normalerweise zweckrational handeln und somit für Anreize – positive wie negative – empfänglich sind. An seine Grenze stößt Abschreckung bei Delikten wie Mord und Vergewaltigung, deren Täter sich durch Androhung von Sanktionen kaum beeinflussen lassen. Solche Delikte dürften somit unter Wohlfahrtsaspekten nicht bestraft werden. Hingegen könnte es unter diesen Aspekten nützlich sein, Bagatelldelikte sehr hart zu bestrafen.[41]

Das vergeltende Gerechtigkeitskonzept, das das Strafmaß an die Schwere der Schuld bindet, generiert hier abweichende Gründe. Um zu bestimmen, welche Gründe im Konfliktfall den Ausschlag geben, muss eine Abwägung durchgeführt werden. Lässt sich eine umfassende Theorie formulieren, die den unterschiedlichen Gründen ihren Platz zuweist?

Die Antwort kann ich hier nur skizzieren. Soviel immerhin zeichnet sich ab: Die umfassende Theorie ist konsequentialistisch. Dies ist insofern überraschend, als sie retributive Elemente enthält und der Retributivismus gemeinhin als nichtkonsequentialistische, deontologische Theorie gilt. Tatsächlich sollte er jedoch konsequentialistisch verstanden werden.[42] Der intrinsische Wert, der ihm zugrunde liegt, ist retributive Gerechtigkeit – das Zufügen verdienter Strafen, genauer: Das Minimieren der Anzahl von Verbrechen, die ohne verdiente Strafe bleiben.[43] Auch die abschreckungsbezogenen Aspekte sind konsequentialisti-

---

40 Dass der Wert retributiver Vergeltung zwar gering, aber durchaus messbar ist, zeigen psychologische Experimente, bei denen die Teilnehmer bereit waren, andere für unmoralische Handlungen zu bestrafen, obgleich sie sich dadurch Nachteile einhandelten (etwa zahlten sie einen bestimmten Betrag, wenn sie dadurch bewirken konnten, dass der Täter ebenfalls finanzielle Einbussen erlitt, cf. Fehr/Gächter 2002). Diese Experimente sind allerdings insofern nur bedingt aussagekräftig, als die Strafenden zugleich die Opfer der Normverstöße waren, so dass hier eher das Gewicht von Rache gemessen wurde.

41 Diese vermeintlichen Schwäche des Präventionskonzept wird jedoch durch verschiedene empirische Faktoren weitgehend kompensiert: Die verbreitete Angst, dass man selbst oder Nahestehende wegen Bagatelldelikten – ggf. aufgrund von Justizirrtümern – drakonisch bestraft werden, senkt die Nutzenbilanz, cf. Hallich 2011.

42 Die Möglichkeit eines konsequentialistischen Retributivismus anerkennen G. E. Moore 1962, §128; Ross 1965, 58; M. Moore 1993. Sie wird bestritten von Dolinko 1997. Dolinkos Kritik trifft jedoch nicht den Retributivismus als Element einer hybriden Theorie, in der die Verhängung einer Strafe nicht schon dann als Pflicht gilt, wenn sie retributiv gerecht wäre. – Ein so verstandener Retributivismus könnte unter besonderen Umständen durchaus die Bestrafung Unschuldiger rechtfertigen: dann nämlich, wenn dies zur Folge hätte, dass insgesamt mehr Schuldige ihre verdiente Strafe bekommen.

43 Diese komplizierte Formulierung ist genauer, weil die erste, einfachere Formulierung die Interpretation zulassen würde, dass die *Summe* verdienter Strafen *zu maximieren* wäre.

scher Natur, doch liegt ihnen ein anderer Wert zugrunde: das Wohl der Subjekte. Der Wert der Wohlfahrt wiegt jedoch deutlich schwerer als der der retributiven Gerechtigkeit. Nun lassen sich folgende Extremfälle unterscheiden:

(i) Abschreckung greift nicht, die Gerechtigkeit fordert jedoch eine hohe Strafe (etwa bei einem sadistischen Mord). Da hier das Gewicht der retributiven Gründe gering ist, ist Strafe unter dem Strich unbegründet.[44]

(ii) Abschreckung generiert Gründe für eine harte Maßnahme, obgleich diese ungerecht wäre (etwa bei folgenloser Fahrlässigkeit im Straßenverkehr). Dieser Fall ist spiegelbildlich zu (i): Da ihr intrinsisches Gewicht im Vergleich zu den Wohlfahrtsgründen sehr schwach ist, können die Gerechtigkeitsgründe sich nicht durchsetzen.

Welche Relevanz hat eine Auseinandersetzung mit dem Retributivismus überhaupt, wenn retributive Gerechtigkeit nur von derart geringer praktischer Bedeutung ist? Ich denke, dass sie durchaus eine praktische Bedeutung besitzt – allerdings nicht auf der Ebene der *Begründung* harter Maßnahmen, wohl aber auf der Ebene der *Intuitionen* und deshalb auch auf der Ebene der Erklärung der *faktischen* Praxis. Zwar haben sie kaum Einfluss darauf, welche Maßnahmen legitim *sind*, wohl aber darauf, welche wir für legitim *halten*. Unreflektierte Intuitionen aber haben in moralischen Fragen nicht das letzte Wort, sondern eine eher heuristische Funktion.[45] Da sich gezeigt hat, dass sich hinter den Intuitionen, die sich aus dem Interesse an retributiver Gerechtigkeit speisen, keine *gewichtigen* normativen Quellen verbergen, gilt es, diese Intuitionen als irreführend zu durchschauen und ihnen nur den geringen Einfluss auf die Maßnahmen gegenüber Verbrechern einzuräumen, der ihnen zusteht. Zudem müssen auch die genannten, auf retributiven Intuitionen basierenden Einwände gegen ein konsequentes Präventionsstrafrecht außer Betracht bleiben. Die relevanten staatlichen Institutionen sollten sich daher allein auf das Ziel der Prävention ausrichten. Al-

---

Dies würde die absurde Implikation haben, dass zunächst die Anzahl der Verbrechen zu maximieren wäre, denn verdiente Strafen setzen Verbrechen voraus.

44 Stellt man aber in Rechnung, dass die Meinung, jemand entgehe seiner gerechten Strafe, wiederum das Wohl beeinträchtigt, schlägt die Gerechtigkeit in der Gesamtkalkulation allerdings doppelt – intrinsisch und extrinsisch – zu Buche. Das Gewicht, mit dem sie in der letzteren indirekten, wohlfahrtsbezogenen Weise in die Abwägung eingeht, hängt nun auch ab von der Anzahl derer, die unter der Ungerechtigkeit leiden. Da diese Anzahl im Zeitalter der Massenmedien sehr groß sein kann, kann das extrinsische Gewicht der Gerechtigkeit (das sie durch ihren Einfluss auf das Wohl besitzt) ihr intrinsisches Gewicht schnell deutlich überwiegen. Dies ändert aber nichts an der Rolle, die die Gerechtigkeit unter dem Strich spielt, da sich auch die Wohlfahrtsaspekte, die im Mitleid fundiert sind und in die andere Richtung weisen, aufaddieren (Dafür, dass die Wohlfahrtsmoral im Mitleid fundiert ist, argumentiere ich in Schälike 2009a.).

45 Cf. Singer 2005.

lenfalls im Blick auf die *Akzeptanz* des Strafrechts könnten sich Abweichungen in Richtung auf retributive Gerechtigkeit als angeraten erweisen.[46]

Literaturverzeichnis

Boonin, D.: The Problem of Punishment. Cambridge University Press, Cambridge, 2008

Burgh, W.: „Do The Guilty Deserve Punishment?". The Journal of Philosophy, 79, 1982. S. 193-210

Cottingham, J.: „Varieties of Retribution". The Philosophical Quarterly, 16, 1979. S. 238-246

Crisp, R.: „Equality, Priority, and Compassion". Ethics, 113, 2003. S. 745-763

Dagger, R.: „Playing Fair With Punishment". Ethics, 103, 1993. S. 473-488

Dolinko, D.: „Some Thoughts About Retributivism". Ethics, 101, 1991. S. 537-559

Dolinko, D.: „Retributivism, Consequentialism, and the Intrinsic Goodness of Punishment". Law and Philosophy, 16, 1997. S. 507-528

Duff, A./D. Garland (Hrsg.): A Reader on Punishment. Oxford University Press, Oxford, 1994

Duff, R. A.: Punishment, Communication, and Community. Oxford University Press, Oxford, 2001

Fehr, E./Gächter: „Altruistic Punishment in Humans". Nature, 415, 2002. S. 137-140

Feinberg, J.: „Justice and Personal Desert". 1963. In: *Feinberg, J.*: 1970. S. 55-94

Feinberg, J.: „The Expressive Function of Punishment", 1965. In: *Feinberg, J.*: 1970. S. 95-118

Feinberg, J.: Doing and Deserving. Essays in the Theory of Responsibility. Princeton University Press, Princeton, 1970

Goldman, A.: „The Paradox of Punishment". In: *Simmons, A. J., et al. (Hrsg.):* Punishment: A Philosophy and Public Affairs Reader. Princeton University Press, Princeton, 1979. S. 30-46

46 Dieser Text ist eine gekürzte Version von Schälike 2011. Ich danke Vuko Andrić und Oliver Hallich für wertvolle Hinweise.

Hallich, O.: „Konsequentialistische Straftheorien“. In*: Gesang, B. / Schälike, J. (Hrsg.):* Die großen Kontroversen der Rechtsphilosophie. Mentis, Paderborn, 2011

Hampton, J.: „The Moral Education Theory of Punishment“. 1984. In: *Simmons et al. (Hrsg.):* Punishment. Princeton University Press, Princeton, 1995. S. 112-142

Hampton, J.: „An Expressive Theory of Retribution”. In: *Cragg, W. (Hrsg.):* Retributivism and Its Critics. Franz Steiner Verlag, Stuttgart, 1992. S. 1-25

Hegel, G. W. F.: Grundlinien der Philosophie des Rechts. 1821. Werke 7. Suhrkamp, Frankfurt/M., 1970

Hershenov, D. B.: „Restitution and Revenge“. The Journal of Philosophy, 96, 1999. S. 79-94

Hood, R.: „Capital Punishment“, In: *Tonry, M. (Hrsg.):* The Handbook of Crime and Punishment. Oxford University Press, Oxford, 1998. S. 739-776

Hume, D.: A Treatise of Human Nature. hrsg. von *P. H. Nidditch*. Oxford, Oxford University Press, ^{2}1978

Kant, I. (GMS): Grundlegung zur Metaphysik der Sitten. Akademie-Ausgabe.

Kleinig, J.: Punishment and Desert. Martinus Nijhoff, Den Haag, 1973

Mackie, J. L.: „Morality and the Retributive Emotions“. Criminal Justice Ethics, 1, 1982. S. 3-10

Mill, J. St.: Utilitarianism. Longmans, Green, Reader, & Dyer, London, 1871

Moore, G. E.: Principia Ethica. Cambridge University Press, Cambridge, 1962

Moore, M.: „Justifying Retributivism“. Israel Law Review, 27, 1993. S. 15-49

Morris, H.: „Persons and Punishment“. The Monist, 52, 1968. S. 475-501

Morris, H.: „A Paternalistic Theory of Punishment“. 1981. In: *Duff, A./D. Garland (Hrsg.):* A Reader on Punishment. Oxford University Press, Oxford, 1994. S. 92-111.

Murphy, J.: „Marxism and Retribution“. 1973. In: *Duff, A./D. Garland (Hrsg.):* A Reader on Punishment. Oxford University Press, Oxford, 1994. S. 44-70.

Murphy, J.: Retribution, Justice, and Therapy. Reidel, Boston, 1979

Nozick, R.: Philosophical Explanations. Harvard University Press, Cambridge (MA), 1981

Oldenquist, A.: „An Explanation of Retribution“. The Journal of Philosophy, 85, 1988. S. 464-478

Parfit, D.: „Gleichheit und Vorrangigkeit“. 1998. In: *A. Krebs (Hrsg.):* Gleichheit oder Gerechtigkeit. Texte der neuen Egalitarismuskritik. Suhrkamp, Frankfurt/M., 2000. S. 81-106

Ross, W. D.: The Right and the Good. Oxford University Press, Oxford, 1965

Roxin, C.: Strafrecht. Allgemeiner Teil, Bd. 1. Grundlagen. Der Aufbau der Verbrechenslehre. Beck, München, 1994

Schälike, J.: „Levelling-up-Egalitarismus. Gerechtigkeit, Gleichheit und Neid“. Zeitschrift für philosophische Forschung, 63, 2009

Schälike, J.: „Moral und Interesse. Vom interessenfundierten Konzept praktischer Rationalität zum normativen Universalismus“. Philosophisches Jahrbuch, 115, 2009a

Schälike, J.: Spielräume und Spuren des Willens. Eine Theorie der Freiheit und der moralischen Verantwortung. Mentis, Paderborn (i. V./2010)

Schälike, J.: „Retributivistische Straftheorien“. In: *B. Gesang/J. Schälike (Hrsg.)*: Die großen Kontroversen der Rechtsphilosophie. Mentis, Paderborn, 2011

Shafer-Landau, R.: „Retributivism and Desert“. Pacific Philosophical Qarterly, 81, 2000. S. 189-214

Sher, G.: Desert. Princeton University Press, Princeton, 1987

Singer, P.: „Ethics and Intuitions“. The Journal of Ethics, 9, 2005. S. 331-352

Smith, A.: The Theory of the Moral Sentiments. 1759. Liberty Press, Indianapolis, 1982

Stephen, J. F., General View of the Criminal Law of England. Macmillan&Co, London, 1863

Strawson, P.: „Freedom and Resentment“. 1962. In: *Watson, G. (Hrsg.):* Free Will. Second Edition. Oxford University Press, Oxford, 2003. S. 72-93

Walker, N.: „Nozick‘s Revenge“. Philosophy, 70, 1995. S. 581-586

Walker, N.: „Even More Varieties of Retribution“. Philosophy, 74, 1999. S. 595-605

Wolf, J.-C.: „Strafe als Wiederherstellung eines Gleichgewichts“. Jahrbuch für Recht und Ethik, 11, 2003. S. 199-216

Die Abhängigkeit zwischen Chancengleichheit und Freiheit

Ivo Wallimann-Helmer
ivowall@access.uzh.ch
Ethikzentrum der Universität Zürich, Schweiz

Abstract/Zusammenfassung

Commonly, in liberal thought equal opportunity is understood as principle of distribution. However, such a point of view misses the close conceptual relation between equal opportunity and liberty. This paper's aim is to show, why there is such a close conceptual relation between the two ideals. From this follows that within liberalism equal opportunity and liberty can only be defended together if they conceptually correlate.

In a first step the conceptual structure of both ideals is in focus. This discussion shows why equal opportunity must be conceived as an egalitarian conceptualization of claims of liberty. In a second step this paper discusses the potential conflicts between equal opportunity and liberty. Defending these conflicts has further consequences for conceptualizing liberty. Most commonly, in liberalism liberty guarantees conditions to realize an autonomous life. Therefore, it is necessary to show, why equal opportunity can serve this goal. Hence, in a third step this paper sketches two arguments for this purpose.

Chancengleichheit wird im Liberalismus gemeinhin als ein egalitäres Prinzip der Verteilungsgerechtigkeit aufgefasst. Deshalb wird in der Debatte das enge konzeptionelle Abhängigkeitsverhältnis zwischen Chancengleichheit und Freiheit zu wenig berücksichtigt. Dieser Aufsatz zeigt, weshalb Chancengleichheit und Freiheit aufgrund ihrer konzeptionellen Struktur voneinander abhängen. Dies hat zur Konsequenz, dass Chancengleichheit und Freiheit im Liberalismus nur dann widerspruchsfrei verteidigt werden können, wenn die beiden Ideale einander auch konzeptionell entsprechen.

In einem ersten Schritt wird die konzeptionelle Struktur der beiden Ideale näher untersucht. Diese Diskussion zeigt, dass Chancengleichheit immer eine egalitäre Explikation eines Freiheitsanspruches darstellt. In einem zweiten Schritt stehen die zwischen Chancengleichheit und Freiheit behaupteten Konflikte im Zentrum, deren Aufrechterhaltung Folgen für das im Liberalismus verteidigte Freiheitsverständnis hat. Da Freiheit – vereinfacht gesprochen – die Bedingungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens sicherstellen soll, muss sich aufgrund der Diskussion in diesem Aufsatz zeigen lassen, weshalb Chancengleichheit zur Sicherstellung der dazu notwendigen Bedingungen beiträgt. Abschliessend skizziert dieser Aufsatz deshalb in einem dritten Schritt zwei Argumente hierfür.

Chancengleichheit wird im Liberalismus meist als ein egalitäres Prinzip der Verteilungsgerechtigkeit aufgefasst. Bei einer solch einseitigen Betrachtung des Ideals bleibt allerdings das Verhältnis zu Freiheit zu wenig berücksichtigt, dem für die Verteidigung von Chancengleichheit im Liberalismus zentrale Bedeutung zukommt: Wie sich im Folgenden zeigen wird, besteht zwischen Chancen-

gleichheit und Freiheit ein viel engeres konzeptionelles Abhängigkeitsverhältnis als gemeinhin angenommen wird.

Chancengleichheit und Freiheit werden einander in der Debatte meist nur dann gegenübergestellt, wenn es darum geht, den Konflikt zwischen beiden Forderungen aufzuzeigen. Zudem wird behauptet, Chancengleichheit könne zugunsten von Freiheit aufgegeben werden, weil das Ideal von seiner konzeptionellen Struktur her mit derjenigen von Freiheit identisch sei. Beides ruft aber nach einer Aufgabe von Chancengleichheit als liberalem Ideal.

Im Folgenden soll gezeigt werden, weshalb Chancengleichheit als liberales Ideal nicht aufgegeben werden sollte, auch wenn es sich auf Freiheit reduzieren lässt und mit dieser in Konflikt stehen mag. Hierzu diskutiere ich in einem ersten Schritt die konzeptionelle Abhängigkeit zwischen Chancengleichheit und Freiheit (1.). Danach zeige ich, dass der behauptete Konflikt zwischen Chancengleichheit und Freiheit nicht nur zwischen diesen beiden Idealen sondern auch zwischen verschiedenen Freiheitsansprüchen besteht (2.). Vor dem Hintergrund dieser Diskussion lassen sich zwei Argumente zugunsten von Chancengleichheit anführen, die deren zentrale Bedeutung zur Sicherstellung der Bedingungen für die Verwirklichung eines selbstbestimmten, menschlichen Lebens zeigen (3.).

1. Lässt sich Chancengleichheit auf Freiheit reduzieren?

Richards behauptete, Chancengleichheit solle als liberales Ideal zugunsten von Freiheit oder anderer liberaler Forderungen aufgegeben werden, weil dessen Forderungen damit besser abgebildet werden können.[1] Im Folgenden soll die Berechtigung der Behauptung näher untersucht werden, Chancengleichheit lasse sich auf Freiheit reduzieren. Es wird sich zeigen, dass Chancengleichheit aufgrund der konzeptionellen Struktur des Chancenbegriffes unter bestimmten Umständen eine egalitäre Explikation von Freiheit darstellt.

Dieser These steht zunächst eine begriffliche Feststellung entgegen: Das Kompositum „Chancengleichheit“ legt nahe, das Ideal als eine Forderung der Gleichheit auszulegen, weil der Begriff „Gleichheit“ Teil dessen Bezeichnung ist. Während eine Forderung der Gleichheit Ansprüche relativ zu bestimmten Eigenschaften von Personen zusichert, werden Freiheitsansprüche im Liberalismus als absoluter Standard aufgefasst, die Personen unabhängig von ihren individuellen Eigenschaften zukommen. Da die beiden Ideale demzufolge zwei unterschiedlichen Begründungsstandards folgen, scheinen sie nur schon deswegen nicht aufeinander reduzierbar zu sein.

1 Richards, 1997, S. 261

Untersucht man allerdings die Struktur des Chancenbegriffes, dann zeigt sich, dass Chancengleichheit und Freiheit von ihrer konzeptionellen Struktur her identisch sind. Vor diesem Hintergrund lässt sich zeigen, unter welchen Bedingungen Chancengleichheit eine egalitäre Explikation von Freiheit darstellt und insofern auf diese reduzierbar ist. Nimmt man diese Feststellung ernst, dann führt dies zu einem engen konzeptionellen Abhängigkeitsverhältnis zwischen den beiden Idealen.

1.1 Chancengleichheit: Reale und realistische Gelegenheiten

Spricht man davon, dass jemandem eine Chance auf eine begehrte Kaderposition zukommt, dann sagt man von dieser Person, sie habe eine mehr oder weniger günstige Gelegenheit, in diese aufzusteigen. Allgemein gesprochen lässt sich deshalb eine Chance als eine dreistellige Relation zwischen einer Person X, einer mehr oder weniger günstigen Gelegenheit Y und einem Gut Z beschreiben.[2]

Chancen als Gelegenheiten stellen weder Garantien noch hypothetische Möglichkeiten dar. Vielmehr stehen sie für etwas dazwischen: Chancen sind keine Garantien, weil Chancen nicht sicherstellen, dass eine Person ein Gut Z auch tatsächlich erreicht, sobald sie es begehrt. Vielmehr kann eine Person bei der Realisierung einer Chance auch erfolglos sein. Wären Chancen als hypothetische Möglichkeiten aufzufassen, dann hätten alle eine Chance auf alle möglichen Güter, sofern ein hypothetischer Weltzustand denkbar ist, in dem sie eine Chance auf das Gut hätten, auf welches sie unter den gegebenen Umständen keine Chance haben.[3]

Chancen geben deshalb an, unter welchen Bedingungen Gelegenheiten für eine Person real sind. Ebenso geben sie an, wie realistisch der Erfolg im Wettbewerb um ein Gut für eine Person ist. Chancengleichheit als liberale Forderung gibt deshalb entweder an, unter welchen Bedingungen die Chancen für eine Person real sind, oder bestimmt, unter welchen Bedingungen die Chancen einer Person in hinreichendem Masse realistisch sind. Erläutern lässt sich diese Unterscheidung zwischen realen und in hinreichendem Masse realistischen Gelegenheiten mit folgendem Beispiel:[4]

In einer Gesellschaft, in der die begehrten Kaderpositionen nur mit Angehörigen bestimmter, reicher Familien besetzt werden, besteht weder eine reale noch eine realistische Gelegenheit für die restliche Bevölkerung in diese aufzusteigen. Denn unter den gegebenen Umständen ist es der restlichen Bevölkerung nicht möglich, sich um Kaderpositionen zu bewerben. Es besteht für sie demzu-

2 Westen, 1985, S. 838 & 1990, S. 169; siehe aber auch Arneson, 1989, S. 85; Campbell, 1975; O'Neill, 1993, S. 144; Hansson, 2004, S. 309; Meyer, 2007.

3 Westen, 1990, S. 166ff.

4 Dies stellt eine Variation des berühmten Beispiels von Williams Kriegergesellschaft dar (1978, S. 244).

folge keine real existierende Gelegenheit, im entsprechenden Wettbewerb erfolgreich zu sein. Ebenso ist unter diesen Umständen der Aufstieg in eine solche Position für den restlichen Teil der Bevölkerung unrealistisch, weil ihre Erfolgschancen aufgrund des Ausschlusses vom Wettbewerb gegen Null tendieren.

Erst wenn demzufolge durchgesetzt wird, dass der Wettbewerb um die Kaderpositionen allen offen steht, besteht auch für Angehörige der restlichen Bevölkerung eine reale Gelegenheit, in eine solche aufzusteigen. Denn durch eine solche Reformation des Wettbewerbs wird ermöglicht, dass sich alle Mitglieder einer solchen Gesellschaft um Kaderpositionen bewerben können. Wie realistisch das Erlangen einer solchen Kaderposition für den vormals nicht privilegierten Bevölkerungsteil allerdings ist, hängt davon ab, wie gut die Chancen der Angehörigen dieses Bevölkerungsteils sind, im Wettbewerb um Kaderpositionen überhaupt erfolgreich zu sein.

Der normative Gehalt eines Verständnisses von Chancengleichheit ergibt sich vor dem Hintergrund dieser Unterscheidung zwischen realen und realistischen Gelegenheiten: Betont ein Verständnis von Chancengleichheit den Aspekt realer Gelegenheiten, dann bestimmt das Ideal, unter welchen Bedingungen günstige Gelegenheiten als real gelten. Fokussiert ein Verständnis von Chancengleichheit demgegenüber auf den Aspekt realistischer Gelegenheiten, dann legt das Ideal fest, wie gross die Aussichten auf Erfolg für die Einzelnen sein müssen bzw. unter welchen Bedingungen ungleiche Erfolgsaussichten legitimiert werden können, damit die Gelegenheiten in hinreichendem Masse realistisch sind. Dabei werden reale Gelegenheiten durch eine angemessene Verteilungs*prozedur* sichergestellt. In hinreichendem Masse realistische Gelegenheiten stellt ein Verteilungs*zustand* her:

Betont ein Verständnis von Chancengleichheit den Aspekt der Verteilungs*prozedur*, dann bedeutet Chancengleichheit zu fordern, dass die Zugangsbedingungen zu den begehrten Kaderpositionen für alle Mitglieder real sein müssen. Damit wird festgelegt, welche Hindernisse in der Verteilungsprozedur von Kaderpositionen gerechtfertigterweise als überwindbar gelten und deshalb Teil derselben sein können. In Verbindung mit einer Verteilungsprozedur steht Chancengleichheit demzufolge für die Forderung, dass die Gelegenheiten Y zum Erlangen eines Gutes Z für alle Personen X nur als überwindbar gerechtfertigte Hindernisse umfassen dürfen. Dieses Verständnis von Chancengleichheit werde ich im Folgenden als prozedurale Chancengleichheit bezeichnen. So wie die soziale Reform im Beispiel vorgestellt wurde, wird damit prozedurale Chancengleichheit durchgesetzt. Denn durch die Reform wird sichergestellt, dass sich alle Mitglieder der Beispielgesellschaft unter den gleichen Bedingungen um die begehrten Kaderpositionen bewerben können.

Wird Chancengleichheit als eine Forderung nach einem bestimmten Verteilungs*zustand* aufgefasst, dann bedeutet deren Sicherstellung, für alle in hinreichendem Masse realistische Gelegenheiten im Zugang zu einem Gut zu ermög-

lichen. In unserer Beispielgesellschaft bedeutete dies, dass allen Mitgliedern der Erwerb von Fähigkeiten möglich sein muss, die für den Erfolg im Wettbewerb um Kaderpositionen relevant sind. Der normative Gehalt eines solchen Verständnisses von Chancengleichheit liegt deshalb in der Bestimmung einer gerechtfertigten Verteilung erworbener Fähigkeiten, um faire Startbedingungen für einen Wettbewerb sicherzustellen. Insofern bestimmt ein solches Verständnis von Chancengleichheit, unter welchen Bedingungen die Gelegenheiten in einem Wettbewerb für alle in hinreichendem Mass realistisch und unter welchen Bedingungen ungleiche Erfolgsaussichten legitim sind. Ein Verständnis von Chancengleichheit, das einen Verteilungszustand fordert, werde ich im Folgenden substantielle Chancengleichheit nennen. Da substantielle Chancengleichheit die Ermöglichungsbedingungen zum Erwerb von Fähigkeiten sicherstellt, betrifft dieses Verständnis die Eigenschaften der Personen X der dreistelligen Relation des Chancenbegriffes.

Unabhängig davon, welches Verständnis von Chancengleichheit verteidigt wird, ist eine Gelegenheit für eine Person unter gegebenen Umständen nur dann real, wenn ihr der Zugang zu einem Gut nicht prinzipiell verwehrt bleibt. Der Zugang zu einem Gut ist einer Person dann nicht prinzipiell verwehrt, wenn ihr der Wettbewerb um ein Gut nicht durch unüberwindbare Hindernisse verschlossen ist. Um in einem Wettbewerb aber erfolgreich sein zu können, muss eine Person auch über Fähigkeiten verfügen, die für den Erfolg im Wettbewerb relevant sind. Auch hier gilt unabhängig vom Verständnis von Chancengleichheit, dass der Zugang zu einem Gut für eine Person umso realistischer sein muss, über umso mehr relevante Fähigkeiten sie für einen Erfolg im Wettbewerb verfügt.

Diese Analyse der konzeptionellen Struktur von Chancengleichheit lässt den Anwendungsbereich des Ideals offen. Denn sie gibt nicht vor, dass Chancengleichheit einzig für den Wettbewerb um Kaderpositionen oder – wie in der Debatte häufig angenommen wird – für den Wettbewerb um soziale Positionen und Ausbildungsplätze relevant sein soll. Genauso wäre es möglich, das Ideal im Kontext der Medizin, der Politik oder für den Strassenverkehr einzusetzen, um anzugeben, wann für Personen der Zugang zu einem Gut real bzw. in hinreichendem Mass realistisch ist. Wie sich im Folgenden zeigen wird, lässt sich Chancengleichheit aufgrund dieser konzeptionellen Unterbestimmtheit auf Freiheit reduzieren. Denn sofern der Anwendungsbereich von Chancengleichheit mit demjenigen von Freiheit deckungsgleich ist, stellt das Ideal nichts anderes als eine egalitäre Explikation der Freiheitsforderung dar.

1.2 Chancengleichheit als Explikation von Freiheit

Chancengleichheit gilt im Liberalismus als ein egalitäres Ideal der Verteilungsgerechtigkeit. Aus diesem Grund bedeutet dessen Sicherstellung, allen die glei-

chen Gelegenheiten zu eröffnen, sofern sie in relevanter Hinsicht gleich sind: Prozedurale Chancengleichheit stellt dabei sicher, dass die Zugangsprozedur zu einem Gut oder einem Bündel von Gütern für alle Bewerberinnen und Bewerber die gleichen Hindernisse umfasst. Substantielle Chancengleichheit sichert demgegenüber die gleichen Startbedingungen für den Wettbewerb um ein Gut oder ein Bündel an Gütern, indem der Erwerb von dafür relevanten Fähigkeiten ermöglicht wird.

Dieser Verweis auf relevante Eigenschaften von Personen zur Sicherung von Chancengleichheit gilt für Freiheit nach gängiger Ansicht nicht. Denn Freiheitsansprüche sollen für alle Mitglieder einer Gesellschaft unabhängig von ihren individuellen Eigenschaften gesichert sein. Freiheitsansprüche lassen sich deshalb gemäss gängiger Ansicht nicht mit Verweis auf eine relevante Hinsicht rechtfertigen. Aus diesem Grund sichert Freiheit für alle Mitglieder einer Gesellschaft unabhängig von ihren individuellen Eigenschaften die Befriedigung bestimmter Ansprüche. Da Freiheit - etwas vereinfacht gesprochen - die Bedingungen für ein selbstbestimmtes, menschliches Leben ermöglichen soll, sichert dieses Ideal für alle unabhängig von ihren individuellen Eigenschaften entsprechende Bedingungen.

Im Anschluss an Berlins Unterscheidung zwischen negativer und positiver Freiheit, lassen sich grundsätzlich zwei Formen der Sicherung der Bedingungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens unterscheiden:[5] Negative Freiheit stellt sicher, dass alle frei von Beschränkungen ihres Handelns sind, die sie an der Verwirklichung eines selbstbestimmten, menschlichen Lebens auf ungerechtfertigte Weise hindern könnten. Mit positiver Freiheit werden diese Bedingungen dadurch sichergestellt, dass alle zur Verwirklichung eines selbstbestimmten, menschlichen Lebens befähigt werden. Dies bedeutet sicherzustellen, dass allen der Erwerb von Fähigkeiten möglich ist und sie über die notwendigen materiellen Ressourcen verfügen, die beide zur Verwirklichung eines solchen Lebens als notwendig erachtet werden.

Diese beiden Auslegungen von Freiheit lassen sich wie der Chancenbegriff in einer dreistelligen Relation fassen: Freiheit besteht in einer Relation zwischen einer Person X, die eine bestimmte Handlungsmöglichkeit Y hat, etwas Z zu tun oder zu werden.[6] Dabei steht wie bei Chancen Freiheit weder für eine Garantie noch für eine rein hypothetische Möglichkeit. Freiheit stellt keine Garantien sicher, weil Personen in der Verwirklichung ihres selbstbestimmten, menschlichen Lebens trotz der durch Freiheit sichergestellten Bedingungen erfolglos sein können. Gleichzeitig ist eine Person nur dann frei etwas zu tun oder zu werden, wenn es ihr unter den gegebenen Umständen tatsächlich möglich ist, das zu tun oder zu werden, was sie tun oder werden will. Auch Freiheit stellt deshalb keine

5 Berlin, 2005, S. 166ff.

6 McCallum, 1991, S. 102, übersetzt nach Pauer-Studer, 2000, S. 10; ähnlich Feinberg, 1973, S. 11

rein hypothetischen, sondern unter gegebenen Umständen reale Handlungsmöglichkeiten zur Verwirklichung eines selbstbestimmten, menschlichen Lebens sicher.

Im Rahmen dieser dreistelligen Relation betont negative Freiheit die Beschaffenheit der Handlungsmöglichkeiten Y. Positive Freiheit fokussiert demgegenüber die Fähigkeiten einer Person X sowie deren Verfügungsgewalt über materielle Ressourcen. Folgt man meiner Terminologie aus dem letzten Abschnitt, dann stellt negative Freiheit sicher, dass die Verwirklichung eines selbstbestimmten, menschlichen Lebens für alle real ist. Positive Freiheit garantiert den Erwerb von Fähigkeiten und die Verfügungsgewalt über Ressourcen, damit die Verwirklichung eines solchen Lebens für alle in hinreichendem Masse realistisch ist.

Da eine Gelegenheit, ein Gut zu erreichen, nichts anderes als eine Handlungsmöglichkeit darstellt und Handlungsbeschränkungen für Hindernisse stehen, stellt Freiheit wie Chancengleichheit reale und realistische Gelegenheiten sicher, wobei diese im Fall von Freiheit die Verwirklichung eines selbstbestimmten, menschlichen Lebens ermöglichen sollen. Dabei gilt auch für Freiheit, dass diese Gelegenheiten unter gegebenen Umständen nur dann real sind, wenn einer Person die Verwirklichung eines selbstbestimmten, menschlichen Lebens nicht prinzipiell verwehrt ist. Realistisch sind diese Gelegenheiten dann, wenn Personen zumindest über ein minimales Mass an Fähigkeiten und materiellen Ressourcen zur Verwirklichung eines solchen Lebens verfügen.

Gemäss dieser Analyse der konzeptionellen Struktur von Freiheit, stellt Chancengleichheit dann eine egalitäre Explikation der liberalen Freiheitsforderung dar, wenn die Komponenten der dreistelligen Relation des Chancenbegriffes identisch mit denjenigen von Freiheit sind: Fallen die relevanten Eigenschaften von Personen zur Rechtfertigung des Anspruchs auf gleiche Erfolgswahrscheinlichkeiten mit denjenigen zusammen, die Ansprüche positiver Freiheit begründen, dann stellt substantielle Chancengleichheit eine egalitäre Explikation positiver Freiheit dar. Sind die durch prozedurale Chancengleichheit sicherzustellenden Verteilungsprozeduren mit der durch negative Freiheit sicherzustellenden Abwesenheit von Beschränkungen deckungsgleich, dann expliziert prozedurale Chancengleichheit negative Freiheit egalitär. In beiden Fällen lässt sich ein entsprechendes Verständnis von Chancengleichheit auf Freiheit reduzieren, wenn der Anwendungsbereich von Chancengleichheit die Bedingungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens sicherstellen soll. Denn dann sind die Anwendungsbereiche beider Ideale deckungsgleich.

1.3 Abhängigkeit I: Konzeptionelle Widerspruchsfreiheit

Gegen diese Analyse der konzeptionellen Struktur von Chancengleichheit und Freiheit kann man folgendes Argument vorbringen: Vor dem Hintergrund des

Beispiels der Gesellschaft mit den begehrten Kaderpositionen kann man behaupten, Chancengleichheit könne keine egalitäre Explikation der liberalen Freiheitsforderung darstellen, weil das Ideal nur im Wettbewerb um soziale Positionen und um Ausbildungsplätze Anwendung finde. Diesem Einwand kann man in zweierlei Hinsicht begegnen: Zum einen kann man die im Rahmen des Glücksegalitarismus vertretenen Verständnisse von Chancengleichheit ins Feld führen. Zum andern lässt sich vorbringen, dass die Begründungslast bei denjenigen liegt, die eine entsprechende Auslegung der realen Gelegenheiten vertreten.

Zur ersten Entgegnung: Obwohl Chancengleichheit in der Debatte oft nur Anwendung auf den Wettbewerb um soziale Positionen und Ausbildungsplätze findet, vertreten Glücksegalitaristen wie Arneson und Cohen einen weiteren Anwendungsbereich von Chancengleichheit, der meiner Meinung nach mit demjenigen von Freiheit identisch ist: Arneson ist der Meinung, Chancengleichheit müsse Gleichheit zur Erlangung von Wohlergehen sicherstellen.[7] Cohen behauptet demgegenüber, das Ideal ermögliche den Zugang zu Vorteilen im Allgemeinen.[8]

Diese beiden Auslegungen von Chancengleichheit stellen eine egalitäre Explikation von Freiheit dar, weil meiner Meinung nach die Verwirklichung eines selbstbestimmten, menschlichen Lebens allgemein gesprochen in nichts anderem besteht, als Wohlergehen oder Vorteile im Allgemeinen zu erlangen. Ein solches Verständnis von Chancengleichheit muss deshalb zeigen können, weshalb dem Ideal im Liberalismus eine gegenüber Freiheit eigenständige Funktion zukommt, auch wenn der Anwendungsbereich der beiden Forderungen zusammenfällt.

Folgt man den Vorschlägen von Arneson und Cohen, dann steht Chancengleichheit für substantielle Chancengleichheit. Denn gleiche Aussichten auf Wohlergehen oder Vorteile im Allgemeinen lassen sich nur sichern, wenn Personen dazu entsprechend befähigt werden. Aufgrund ihrer Bestimmung des Anwendungsbereichs explizieren sie damit positive Freiheit. Positive Freiheit stellt demzufolge gemäss Arneson und Cohen sicher, dass es allen möglich ist, über diejenigen Fähigkeiten zu verfügen, die notwendig sind, um Wohlergehen oder Vorteile im Allgemeinen zu erlangen. Dies schliesst die Sicherung einer bestimmten Verteilung materieller Ressourcen ein. Denn auch wenn mit einem Verständnis substantieller Chancengleichheit häufig nicht die Behauptung einhergeht, dass Personen neben einem bestimmten Fähigkeitenerwerb Anspruch auf materielle Ressourcen haben, ist die Sicherstellung substantieller Chancengleichheit ohne eine Umverteilung materieller Ressourcen kaum denkbar: Eine solche ist notwendig, um die entsprechenden Ausbildungsinstitutionen zu schaffen. Dabei unterscheidet sich Arneson's und Cohen's egalitäre Explikation posi-

7 Arneson, 1989, S. 83f.
8 Cohen, 1989, S. 916f.

tiver Freiheit von Freiheit im üblichen Verständnis einzig darin, dass sie den individuellen Freiheitsanspruch relativ zu Selbstverschulden festlegt.

Die Festlegung von Freiheitsansprüchen relativ zu Selbstverschulden mag erstaunen, weil man in der Debatte üblicherweise der Meinung ist, Freiheitsansprüche stünden allen unabhängig von ihren selbst verschuldeten Entscheidungen als absoluter Anspruch zu. Dieser Einwand stellt allerdings kein Einwand gegen die hier vorgetragene Analyse der konzeptionellen Struktur von Chancengleichheit und Freiheit dar, sondern zeigt vielmehr einen Einwand gegen die Konzeptualisierung von Chancengleichheit, wie sie Arneson und Cohen vorschlagen. Denn dieser Einwand zeigt nur, dass für die Rechtfertigung von Ansprüchen positiver Freiheit Selbstverschulden als relevante Eigenschaft von Personen keine Begründungslast tragen sollte.

Die herausgearbeitete konzeptionelle Struktur von Chancengleichheit und Freiheit wird auch durch die zweite Entgegnung bestätigt: Die Reform der sozialen Strukturen in unserem Beispiel soll sicherstellen, dass alle eine reale Gelegenheit haben, in Kaderpositionen aufzusteigen. Ermöglicht eine solche Reform einzig, dass der Zugang zu diesen Positionen niemandem durch die Verteilungsprozedur grundsätzlich verwehrt bleibt, dann widerspricht eine Beschränkung dieser Reform sozialer Strukturen auf den Wettbewerb um Kaderpositionen dem Ideal moralischer Gleichheit. Denn eine solcherart beschränkte Reform stellt zwar sicher, dass alle im Rahmen dieses Wettbewerbs als Gleiche behandelt werden, garantiert aber nicht, dass ungerechtfertigte Beschränkungen der Handlungsmöglichkeiten von Personen in anderen Bereichen dieser Gesellschaft bestehen bleiben.

Eine Beschränkung des Anwendungsbereiches prozeduraler Chancengleichheit auf den Wettbewerb um Kaderpositionen ist deshalb vor dem Hintergrund des Ideals moralischer Gleichheit in hohem Grad rechtfertigungsbedürftig. Denn sofern prozedurale Chancengleichheit wie negative Freiheit nur sicherstellt, dass Personen bestimmte Gelegenheiten nicht grundsätzlich verwehrt bleiben, gibt es vor dem Hintergrund des Ideals moralischer Gleichheit keinen Grund, eine solche Reform auf den Wettbewerb um Kaderpositionen zu beschränken.

Um dieser Schwierigkeit zu entgehen, sind zwei Strategien möglich: Die erste Strategie besteht darin, den Anwendungsbereich von Chancengleichheit auf den Wettbewerb um Kaderpositionen zu beschränken und gleichzeitig negativer Freiheit die Funktion zuzusprechen, ausserhalb dieses Wettbewerbs sicherzustellen, dass niemandem die Verwirklichung eines selbstbestimmten, menschlichen Lebens ungerechtfertigterweise verwehrt bleibt. Die zweite Strategie besteht darin, den Anwendungsbereich prozeduraler Chancengleichheit auf alle Bereiche einer Gesellschaft auszuweiten.

Sofern aber negative Freiheit genauso wie prozedurale Chancengleichheit sicherstellt, dass niemandem Gelegenheiten durch unüberwindbare Hindernisse grundsätzlich verwehrt bleiben, stellt prozedurale Chancengleichheit im Fall der

zweiten Strategie allerdings nichts anderes dar, als eine egalitäre Explikation negativer Freiheit für den Wettbewerb um soziale Positionen. Im Fall der ersten Strategie bleibt unklar, weshalb Chancengleichheit mit denselben normativen Forderungen in einem Gesellschaftsbereich Bedingungen sicherstellen soll, die durch Freiheit für den restlichen Bereich der Gesellschaft gesichert werden.

Sowohl für substantielle als auch für prozedurale Chancengleichheit gilt demzufolge, dass das Ideal immer dann eine egalitäre Explikation von Freiheit darstellt, wenn dessen Anwendungsbereich gegenüber Freiheit nicht spezifiziert wird. Wird der Anwendungsbereich von Chancengleichheit zwar gegenüber Freiheit spezifiziert, ist aber die für das entsprechende Verständnis von Chancengleichheit relevante Komponente des Chancenbegriffes mit der entsprechenden Komponente des Freiheitsbegriffes identisch, dann ist eine solche Beschränkung in hohem Mass rechtfertigungsbedürftig.

Diese Rechtfertigungslast für eine Spezifikation des Anwendungsbereichs von Chancengleichheit führt zu folgendem, konzeptionellen Abhängigkeitsverhältnis zwischen den beiden Idealen: Sofern Chancengleichheit aufgrund einer mangelnden Spezifikation ihres Anwendungsbereiches eine egalitäre Explikation von Freiheit darstellt, bedeutet eine jegliche Einschränkung desselben, Chancengleichheit einen Anwendungsbereich innerhalb desjenigen von Freiheit zuzuweisen. Aus diesem Grund muss sich für die Verteidigung von Chancengleichheit immer zeigen lassen, weshalb die Durchsetzung des Ideals für die Verwirklichung eines selbstbestimmten, menschlichen Lebens von Bedeutung ist.

Chancengleichheit kann im Liberalismus demzufolge nur dann widerspruchsfrei verteidigt werden, wenn das Ideal zu den normativen Anforderungen von Freiheit nicht in Widerspruch steht: Stellt negative Freiheit die Abwesenheit bestimmter Beschränkungen sicher, dann müssen diese auch in einem durch prozedurale Chancengleichheit geregelten Wettbewerb abwesend sein. Sichert positive Freiheit einen bestimmten Standard an Fähigkeiten und Ressourcen, dann stünde ein Verständnis substantieller Chancengleichheit dazu in Widerspruch, wenn es eine Verletzung dieses Standards aufgrund von Selbstverschulden zulässt.

Zentral für die Verteidigung von Chancengleichheit im Liberalismus ist aufgrund dieser Anforderungen im Weiteren, inwiefern Freiheit innerhalb eines solchen überhaupt gesichert werden soll: Stellt zum Beispiel positive Freiheit im Liberalismus keine sicherzustellende Forderung dar, dann stünde substantielle Chancengleichheit zur Freiheitsforderung eines solchen Liberalismus in Widerspruch, weil damit die Sicherung eines Fähigkeitenerwerbs einhergeht, der gerade nicht sichergestellt werden soll. Das Gleiche gilt für das Verhältnis zwischen prozeduraler Chancengleichheit und negativer Freiheit: Auch hier ist zentral, ob im Liberalismus negative Freiheit überhaupt sichergestellt werden soll. Denn es wäre auch hier ein Widerspruch, wenn im Liberalismus negative Freiheit kein

Ideal darstellt, Chancengleichheit aber durch prozedurale Bedingungen genau dasjenige leistet, was negative Freiheit leisten könnte.

Zusätzlich gilt aufgrund dieser Überlegungen, dass Chancengleichheit im Liberalismus dann nicht auf ein prozedurales bzw. substantielles Verständnis beschränkt werden kann, wenn im Liberalismus McCallum folgend ein Freiheitsverständnis verteidigt wird, das sowohl negative als auch positive Freiheit umfasst. Denn auch in diesem Fall bestünde ein Widerspruch zwischen einer Konzeptualisierung von Freiheit und Chancengleichheit, sofern letzteres Ideal nur einen der beiden Aspekte von Freiheit auf egalitäre Weise in einem gegenüber Freiheit begrenzten Anwendungsbereich expliziert.[9]

2. Besteht ein Konflikt zwischen Chancengleichheit und Freiheit?

Gegen die Behauptung eines engen konzeptionellen Abhängigkeitsverhältnisses zwischen Chancengleichheit und Freiheit kann man einwenden, dass Chancengleichheit mit Freiheit in Konflikt geraten könne. Dieser Konflikt besteht entweder mit der Freiheit der Familie oder mit Freiheitsansprüchen von Unternehmerinnen. Deshalb kann das konzeptionelle Abhängigkeitsverhältnis zwischen diesen beiden Ideale nicht zutreffend sein. Im Folgenden wird sich aber zeigen, dass gerade dieses Konfliktpotential zwischen Chancengleichheit und Freiheit die Analyse der konzeptionellen Struktur der beiden Ideale bestätigt:

Es wird sich zeigen, dass der Konflikt zwischen Chancengleichheit und der Freiheit der Familie nur aufrecht erhalten werden kann, wenn Freiheit im Liberalismus als eine Verbindung zwischen negativer und positiver Freiheit verteidigt wird. Denn dieser Konflikt löst sich in einen Konflikt zwischen zwei Freiheitsansprüchen negativer und positiver Natur auf. Auch für den Konflikt zwischen Chancengleichheit und der Freiheit der Unternehmer gilt, dass er sich in einen Konflikt zwischen zwei Freiheitsansprüchen auflösen lässt. Diese sind aber beide negativer Natur.

2.1 Kein ausschliesslicher Konflikt zwischen Chancengleichheit und Freiheit

Der Konflikt zwischen Chancengleichheit und der Freiheit der Familie entsteht, sofern das Ideal als substantielle Chancengleichheit aufgefasst wird:[10] Soll es allen unabhängig von ihren sozialen und ökonomischen Umständen möglich sein, ihre natürlichen Anlagen zu entwickeln, dann sind starke Eingriffe in die Frei-

9 Ein solches Verständnis von Chancengleichheit verteidigt Mason am Explizitesten (2001), es findet sich aber auch bei Roemer (1998, S. 86) und Rawls (2001, S. 43f.) angedeutet.

10 Vergl. hierzu: Rawls, 1979, S. 92; Hayek, 1980/81, S. 119f.; Fishkin, 1983, S. 64ff.

heit der Familie notwendig. Denn die Sozialstrukturen der Familie und ihre ökonomischen Umstände prägen auf der einen Seite die Entwicklungsmotivation und ökonomische Nachteile beeinträchtigen auf der anderen Seite die Entwicklungsmöglichkeiten Heranwachsender. Um für alle die gleichen Startbedingungen im Sinne in hinreichendem Masse realen Gelegenheiten zu schaffen, müssen deshalb soziale Unterschiede und ökonomische Nachteile möglichst ausgeglichen werden. Dies kann gemäss dem Einwand nur dadurch geschehen, dass die unterschiedlichen Sozialstrukturen in Familien durch staatliche Eingriffe einander angeglichen und ökonomische Unterschiede ausgeglichen werden. Beides bedingt Eingriffe in die Freiheit der Familie, die sogar zur Auflösung dieser Institution führen können.

Der Anspruch von Familien auf Freiheit stellt eine Forderung negativer Freiheit dar. Es besteht demzufolge gemäss dieser ersten Konfliktbehauptung ein Konflikt zwischen einem negativen Freiheitsanspruch und substantieller Chancengleichheit. Anerkennt man allerdings, dass auch mit positiver Freiheit ein bestimmter Standard an Fähigkeitenerwerb zur Verwirklichung eines selbstbestimmten, menschlichen Lebens sichergestellt wird, dann besteht dieser Konflikt nicht nur zwischen einem Anspruch negativer Freiheit und substantieller Chancengleichheit sondern auch zwischen einem negativen und einem positiven Freiheitsanspruch. Denn um allen Heranwachsenden den Erwerb eines bestimmten Masses an Fähigkeiten ermöglichen zu können, muss sichergestellt sein, dass sie in sozialen Strukturen aufwachsen und ökonomisch Bedingungen vorfinden, die ihnen den Erwerb entsprechender Fähigkeiten erlauben. Auch dies bedingt einen Eingriff in die Freiheit der Familie und damit eine Verletzung eines Anspruchs negativer Freiheit.

Dasselbe gilt für den zweiten Konflikt, der sich nicht nur als ein Konflikt zwischen Chancengleichheit und Freiheit, sondern auch als ein Konflikt zwischen den Freiheitsansprüchen von Unternehmerinnen bzw. Unternehmern und denjenigen von Bewerberinnen und Bewerbern darstellen lässt: Aus libertärer Perspektive lässt sich gegen die Forderung prozeduraler Chancengleichheit einwenden, diese stehe mit dem negativen Freiheitsanspruch der Unternehmerinnen in Konflikt, über Privateigentum unabhängig von externem Zwang verfügen zu können. Denn sofern angenommen wird, Arbeitsplätze als Teil privater Unternehmen stünden genauso im Besitz der Unternehmerinnen wie das Unternehmen als Ganzes, dann wird durch die Anforderungen prozeduraler Chancengleichheit an die Vergabeverfahren für Arbeitsstellen die Verfügungsgewalt über Privateigentum eingeschränkt.[11] Wenn demzufolge die Kaderpositionen in unserem Beispiel in Privatbesitz sind, dann bedeutet eine Änderung der Vergabeprozedur für dieselben, Vorschriften bezüglich des freien Verfügens über Privateigentum zu

11 Cavanagh, 2002, S. 49ff.; ähnlich Arneson, 1996, S. 88; Thompson, 1993, S. 34

machen, was eine Verletzung der negativen Freiheit der Unternehmerinnen bedeutet.[12]

Die ungerechte Gesellschaftsstruktur in unserem Beispiel verletzt vor dem Hintergrund des Ideals moralischer Gleichheit nicht nur Forderungen prozeduraler Chancengleichheit sondern auch negativer Freiheit. Negative Freiheit besteht vor der Reform der sozialen Strukturen nicht für alle Mitglieder als absoluter Standard, sondern sichert für die Mitglieder bestimmter reicher Familien einen grösseren Handlungsspielraum als für den restlichen Teil der Bevölkerung, weil diesem keine negative Freiheit im Wettbewerb um Kaderpositionen zugestanden wird. Wird die Reform durchgesetzt, dann bedeutet dies für den Wettbewerb um Kaderpositionen deshalb, nicht nur prozedurale Chancengleichheit durchzusetzen, sondern auch für alle dieselben Bedingungen negativer Freiheit sicherzustellen. Aus diesem Grund entsteht durch die Reform der sozialen Strukturen neben dem Konflikt zwischen prozeduraler Chancengleichheit und dem negativen Freiheitsanspruch der Unternehmerinnen auch ein Konflikt zwischen dem negativen Freiheitsanspruch aller Mitglieder der Beispielgesellschaft und dem negative Freiheitsanspruch der Unternehmerinnen. Es stehen sich demzufolge zusätzlich zwei negative Freiheitsansprüche konfliktträchtig gegenüber.

2.2 Abwägung zwischen Freiheitsansprüchen

Folgt man der bisherigen Argumentation, dann ist es ein kleiner Schritt, auch den Konflikt zwischen Chancengleichheit und Freiheit als einen zwischen zwei Freiheitsansprüchen aufzufassen: Besteht ein positiver Freiheitsanspruch aller Mitglieder einer Gesellschaft, ihre Fähigkeiten bis zu einem gewissen Grad zu entwickeln, dann liesse sich dieser mittels substantieller Chancengleichheit egalitär explizieren. Kommt allen Mitgliedern ein Anspruch zu, ihre Fähigkeiten ohne ungerechtfertigte Beschränkungen auszuüben, dann könnte dieser mittels prozeduraler Chancengleichheit egalitär expliziert werden.

Im Fall substantieller Chancengleichheit besteht der Freiheitsanspruch darin, seine natürlichen Anlagen ohne ungerechtfertigte Einschränkung durch soziale und ökonomische Nachteile entwickeln zu können. Dieser Freiheitsanspruch steht deshalb für einen Anspruch positiver Freiheit. Wie oben gesehen, kann die Durchsetzung dieses Freiheitsanspruchs zur Folge haben, dass zur Sicherstellung der entsprechenden Entwicklungsbedingungen so stark in die Privatsphäre der Familie eingegriffen werden muss, dass deren Freiheitsanspruch aufgehoben wird. Eine Abwägung zwischen der Freiheit der Familie zugunsten der Freiheit, seine Fähigkeiten ohne soziale und ökonomische Benachteiligungen entwickeln zu können, kann deshalb zur Folge haben, einen Freiheitsanspruch gegenüber

12 Gemäss Cavanagh lässt sich sogar zeigen, dass die Forderung, einzig die bestqualifizierte Bewerberin bzw. den bestqualifizierten Bewerber einzustellen nur ein Gebot der Klugheit, nicht aber der Gerechtigkeit darstellt (2002, S. 68f.).

dem anderen für nichtig zu erklären. Sollen allerdings beide Freiheitsansprüche aufrecht erhalten bleiben, dann muss die vollumfängliche Sicherstellung beider aufgegeben werden.

Im Fall prozeduraler Chancengleichheit gilt dasselbe. Im Gegensatz zum Freiheitsanspruch substantieller Chancengleichheit besteht der entsprechende Freiheitsanspruch allerdings nicht im Erwerb von Fähigkeiten unter angemessenen sozialen und ökonomischen Bedingungen sondern in der freien Ausübung erworbener Fähigkeiten. Prozedurale Chancengleichheit steht deshalb für einen Anspruch negativer Freiheit. Denn wenn Einzelne nicht zum Wettbewerb um Kaderpositionen zugelassen oder ihre Erfolgschancen durch eine einseitig benachteiligende Vergabeprozedur geschmälert werden, dann wird ihr negativer Freiheitsanspruch verletzt, ihre Fähigkeiten ohne ungerechtfertigte Beschränkungen ausüben zu können. Auch in diesem Fall hat die Abwägung zwischen der negativen Freiheit der Unternehmerinnen und dem negativen Freiheitsanspruch, seine Fähigkeiten ohne ungerechtfertigte Beschränkungen ausüben zu können, dieselben Folgen wie im Fall substantieller Chancengleichheit: Entweder muss der Freiheitsanspruch der Unternehmerinnen zugunsten des Freiheitsanspruchs prozeduraler Chancengleichheit aufgegeben werden oder es ist nicht möglich, beide Ansprüche vollumfänglich sicherzustellen.

Die Auseinandersetzung mit den Konflikten zwischen Chancengleichheit und Freiheit bestätigt deshalb die behauptete konzeptionelle Abhängigkeit zwischen den beiden Idealen: Da eine jede Spezifikation des Anwendungsbereiches von Chancengleichheit gegenüber Freiheit eine egalitäre Explikation eines Teilbereiches derselben durch Chancengleichheit darstellt, steht jede Forderung nach Chancengleichheit für eine Freiheitsforderung. Aus diesem Grund bestehen auch die beiden genannten Konflikte zwischen Chancengleichheit und Freiheit für Konflikte zwischen verschiedenen Freiheitsansprüchen. Eine jegliche Beilegung dieser Konflikte bedeutet demzufolge eine Abwägung zwischen verschiedenen Freiheitsansprüchen.

2.3 Abhängigkeit II: Freiheitsverständnis und Chancengleichheit

Chancengleichheit lässt sich demzufolge immer auf Freiheit reduzieren, da das Ideal einen Freiheitsanspruch unter anderen explizieren kann: Im Fall substantieller Chancengleichheit den Anspruch, ein bestimmtes Mass an Fähigkeiten zu erwerben, im Fall prozeduraler Chancengleichheit den Anspruch, in der Ausübung seiner Fähigkeiten nicht ungerechtfertigterweise behindert zu werden. Sollen die beiden behaupteten Konflikte zwischen Chancengleichheit und Freiheit aufrecht erhalten werden, dann folgt hieraus, dass Freiheit im Liberalismus nicht als Freiheit *tout cours* aufgefasst werden kann. Vielmehr muss Freiheit für ein System von Freiheiten stehen, die miteinander in Konflikt geraten können.

Dabei bestehen diese Konflikte nicht nur zwischen dem Gesamtsystem von Freiheiten, die gemäss Rawls mit demjenigen aller anderer vereinbar sein müssen, sondern auch zwischen den einzelnen Freiheitsansprüchen als Teil dieses Systems.[13] Die Behauptung von Chancengleichheit und der beiden Konflikte hat demzufolge zur Konsequenz, dass das Ideal nur in Verbindung mit einem Freiheitsverständnis verteidigt werden kann, das Freiheit als ein System von Freiheiten auffasst, die als Gesamtsystem die Verwirklichung eines selbstbestimmten, menschlichen Lebens ermöglichen.

Werden die beiden Konflikte zwischen Chancengleichheit und Freiheit aufrecht erhalten, ergeben sich weitere Folgen für das Freiheitsverständnis im Liberalismus: Im Fall des ersten Konflikts wird ein negativer Freiheitsanspruch der Familie auf Privatsphäre und gleichzeitig ein positiver Freiheitsanspruch Heranwachsender behauptet, ihre Fähigkeiten ohne soziale oder ökonomische Beeinträchtigung zu entwickeln. Diese beiden Freiheitsansprüche können im Rahmen des Liberalismus deshalb nur aufrecht erhalten werden, wenn ein Freiheitsverständnis behauptet wird, das sowohl negative als auch positive Freiheit umfasst. Denn die beiden konfligierenden Ansprüche sind zum einen negativer und zum anderen positiver Natur.

Die Aufrechterhaltung des Konflikts zwischen prozeduraler Chancengleichheit und der Freiheit der Unternehmerinnen setzt demgegenüber einzig die Behauptung negativer Freiheit voraus. Denn es stehen sich in diesem Konflikt zwei Ansprüche negativer Freiheit gegenüber. Sollen aber die beiden Freiheitsansprüche nicht zusammenfallen und derjenige prozeduraler Chancengleichheit gegenüber dem ersten spezifiziert werden, dann bedeutet dies, dass negative Freiheitsansprüche im Liberalismus unterschiedlich restriktiv sein können: Im freien Markt gilt üblicherweise die Forderung, dass im Sinne negativer Freiheit niemandem die Teilnahme grundsätzlich verwehrt bleiben soll.

Für den Wettbewerb um Kaderpositionen in unserem Beispiel bedeutet dies im Sinne prozeduraler Chancengleichheit aber nicht, dass damit die Forderung einhergeht, diese einzig an die Bestqualifizierten zu vergeben. Denn so lange für diesen Wettbewerb keine Bedingungen gelten, die bestimmte Bevölkerungsgruppen ausschliessen, ist damit der Forderung negativer Freiheit genüge getan. Soll aber in einem solchen Wettbewerb prozedurale Chancengleichheit sicherstellen, dass nur die Bestqualifizierten in Kaderpositionen aufsteigen können, dann bedeutet dies, für diesen Wettbewerb grössere Restriktionen für legitim zu erklären, als sie im restlichen Markt gelten.

Im Fall beider Konflikte gilt allerdings, dass sich eine egalitäre Explikation einer Freiheitsforderung im Sinne von Chancengleichheit nur rechtfertigen lässt, wenn die zentrale Bedeutung eines solchen Anspruchs für ein selbstbestimmtes, menschliches Leben nachgewiesen werden kann. Denn lässt sich dies nicht zei-

13 Rawls, 1979, S. 115ff.

gen, dann kann man gegen die Überlegungen in diesem Abschnitt einwenden, Chancengleichheit stelle als Freiheitsforderung nichts anderes sicher als die anderen Freiheitsansprüche des Systems an Freiheiten, was aber nicht notwendig eine egalitäre Explikation derselben rechtfertigt, wie dies Chancengleichheit leistet.

Welche Freiheitsansprüche ein System von Freiheiten im Liberalismus sicherstellen soll und weshalb Chancengleichheit als Explikation eines Teils dieser Freiheitsforderungen von zentraler Bedeutung ist, hängt davon ab, wie die Bedingungen und Anforderungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens aufgefasst werden. In unserer Beispielgesellschaft müsste sich zum Beispiel zeigen lassen, weshalb der Aufstieg in eine Kaderposition für die Verwirklichung eines solchen Lebens von so zentraler Bedeutung ist, dass dies durch eine egalitäre Explikation eines Freiheitsanspruches im Sinne von Chancengleichheit sichergestellt werden muss.

3. Zwei Argumente für Chancengleichheit

Unabhängig davon, welche Bedingungen und Anforderungen für die Verwirklichung eines selbstbestimmten, menschlichen Lebens als notwendig erachtet werden, ergeben sich aus der bisherigen Diskussion des konzeptionellen Abhängigkeitsverhältnisses zwischen Chancengleichheit und Freiheit zwei Argumente für die Verteidigung von ersterem als egalitärer Explikation eines Freiheitsanspruchs:

i) Ist man der Meinung, für die Verwirklichung eines selbstbestimmten, menschlichen Lebens sei einzig die Sicherung eines Systems negativer Freiheiten notwendig und lässt sich ein Bereich wie die Kaderpositionen als zentral für ein solches ausscheiden, dann sind für diesen Bereich auch grössere Restriktionen rechtfertigbar. Denn wenn Kaderpositionen tatsächlich zentral für die Verwirklichung eines selbstbestimmten, menschlichen Lebens sind, dann sollte für den Wettbewerb um diese auch sichergestellt sein, dass einzig die Befähigung dazu über Erfolg oder Misserfolg entscheidet und keine weiteren Überlegungen.

Dies gilt insbesondere auch für den Wettbewerb um unter Umstände knappe Ausbildungsplätze zum Erwerb dieser Fähigkeiten. Denn sowohl im Fall der Kaderpositionen als auch im Fall von Ausbildungsplätzen bedeutete die Anwendung von Vergabeprozeduren, die nicht einzig die Befähigung von Personen berücksichtigen, diejenigen in der Verwirklichung eines selbstbestimmten, menschlichen Lebens zu behindern, die über anderweitige Eigenschaften nicht verfügen. Vor dem Ideal moralischer Gleichheit lassen sich aber solche zusätzlichen Bedingungen für eine Vergabeprozedur nicht rechtfertigen, weil diese den negativen Freiheitsanspruch ungerechtfertigterweise einschränken, in Ausübung erworbener Fähigkeiten ein selbstbestimmtes, menschliches Leben zu verwirkli-

chen. Die Verteidigung restriktiverer Anforderungen für die Vergabe sozialer Positionen (bzw. Kaderpositionen in unserem Beispiel) und von Ausbildungsplätzen stellt deshalb eine zentrale Bedingung zur Verwirklichung eines solchen Lebens dar.

ii) Hängt die Verwirklichung eines selbstbestimmten, menschlichen Lebens zentralerweise davon ab, dass man ein bestimmtes Mass an Fähigkeiten erwerben kann, dann scheint es plausibel, deren Erwerb unabhängig von sozialen oder ökonomischen Nachteilen im Sinne eines positiven Freiheitsanspruches zu ermöglichen. Wird der Erwerb solcher Fähigkeiten allerdings mittels eines absoluten Standards positiver Freiheit sichergestellt, dann bedeutet dies, allen Mitgliedern einer Gesellschaft nur bis zu einem gewissen Schwellenwert den Erwerb der relevanten Fähigkeiten zu ermöglichen.

Sind aber für die Kaderpositionen höherer Hierarchiestufen in unserem Beispiel zusätzliche Qualifikationen notwendig, dann sind diejenigen im Erwerb dieser Qualifikationen benachteiligt, denen es an sozialen und ökonomischen Ressourcen für eine entsprechende Ausbildung fehlt. Denn solcherart Benachteiligte werden unter diesen Umständen zwar unterstützt, die Fähigkeiten des Schwellenwerts zu erwerben, es ist ihnen aber aufgrund ihres Nachteils erschwert oder vielleicht gar unmöglich, darüber hinausgehende Qualifikationen zu erlangen.

Aus diesem Grund ist es plausibel, diesen Freiheitsanspruch im Sinne substantieller Chancengleichheit egalitär zu explizieren: Eine solche Explikation dieses Freiheitsanspruchs erlaubt eine zusätzliche Unterstützung derjenigen zu rechtfertigen, die sozial und ökonomisch benachteiligt sind. Gleichzeitig wird es dadurch möglich, einen Anspruch auf Unterstützung nur denjenigen zu gewähren, die aufgrund ihrer natürlichen Anlagen und ihres Engagements tatsächlich eine solche Unterstützung verdienen. Dabei kann man der Meinung sein, der Anspruch auf den Erwerb dieser Fähigkeiten sei einzig durch substantielle Chancengleichheit zu explizieren oder man kann der Meinung sein, ein Schwellenwert positiver Freiheit sei zwangsläufig sicherzustellen und Chancengleichheit komme erst oberhalb desselben zum Zug.

Je nachdem welches Freiheitsverständnis man verteidigt und je nachdem, welche Bedingungen und Anforderungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens notwendig sind, erhält demzufolge Chancengleichheit eine andere Funktion: Prozedurale Chancengleichheit stellt sicher, dass die Vergabeprozeduren für einen bestimmten Anwendungsbereich ein selbstbestimmtes, menschliches Leben in Ausübung erworbener Fähigkeiten ermöglichen. Substantielle Chancengleichheit stellt demgegenüber sicher, dass es auch sozial und ökonomisch Benachteiligten möglich ist, Fähigkeiten zu erwerben, deren Erwerb durch einen absoluten Schwellenwert positiver Freiheit nicht sichergestellt wird.

4. Fazit

In diesem Aufsatz wurde weder ein spezifisches Verständnis von Chancengleichheit vorgeschlagen, noch wurde ein bestimmtes Freiheitsverständnis, geschweige denn eine bestimmte Auslegung des Liberalismus verteidigt. Die Diskussion des Verhältnisses zwischen Chancengleichheit und Freiheit zeigte vielmehr, welches enge, konzeptionelle Abhängigkeitsverhältnis zwischen diesen beiden Idealen besteht und welche Bedingungen berücksichtigt werden müssen, soll Chancengleichheit im Liberalismus widerspruchsfrei verteidigt werden.

Es zeigte sich, dass Chancengleichheit eine egalitäre Explikation einer Freiheitsforderung darstellt: Wird Freiheit nicht als ein System von Freiheiten aufgefasst, dann kann Chancengleichheit die Bedingungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens egalitär explizieren. Stellt die liberale Freiheitsforderung ein System von Freiheiten sicher, dann gilt dasselbe für eine spezifische Freiheitsforderung, die sich als zentral für die Verwirklichung eines solchen Lebens herausstellen lässt. Für die Verteidigung eines bestimmten Verständnisses von Chancengleichheit ist von Bedeutung, welches Freiheitsverständnis im Liberalismus verteidigt wird, denn nur in Abhängigkeit davon lässt sich entweder prozedurale oder substantielle Chancengleichheit als liberale Forderung behaupten.

Aus diesem Grund stellt die wichtigste Bedingung für die Verteidigung von Chancengleichheit im Liberalismus die Klärung des Freiheitsverständnisses dar, in Verbindung mit dem das Ideal verteidigt wird: Wird negative Freiheit verteidigt, dann ist nur ein prozedurales Verständnis von Chancengleichheit damit vereinbar. Wird Freiheit als positive Freiheit ausgelegt, dann lässt sich in Verbindung damit nur substantielle Chancengleichheit behaupten. Folgt man McCallum Freiheitsverständnis, dann muss Chancengleichheit sowohl substantielle als auch prozedurale Aspekte umfassen.

Soll allerdings Chancengleichheit nicht einzig die liberale Freiheitsforderung egalitär explizieren, sondern eine gegenüber Freiheit eigenständige Funktion zukommen, dann muss der Anwendungsbereich des Ideals gegenüber Freiheit spezifiziert werden. Eine solche Spezifikation bedeutet aber, dass Chancengleichheit eine Funktion innerhalb des Anwendungsbereiches von Freiheit zukommt. Da Freiheit die Bedingungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens sicherstellt, muss sich deshalb zeigen lassen, weshalb ein bestimmter Bereich eines selbstbestimmten, menschlichen Lebens von so zentraler Bedeutung ist, dass dieser mittels eines gegenüber Freiheit gesonderten Ideals gesichert werden muss.

Unabhängig davon sprechen zwei Argumente für die Verteidigung von Chancengleichheit im Liberalismus: Für prozedurale Chancengleichheit spricht, dass die Verwirklichung eines selbstbestimmten, menschlichen Lebens den Erwerb und die Ausübung von Fähigkeiten beinhaltet. Deshalb scheint es plausi-

bel, das Ideal in denjenigen Bereichen einer Gesellschaft anzuwenden, die beides ermöglichen und als zentral für die Verwirklichung eines selbstbestimmten, menschlichen Lebens gelten. Für substantielle Chancengleichheit spricht demgegenüber, dass nur mittels einer egalitären Explikation des Freiheitsanspruchs auf einen bestimmten Fähigkeitenerwerb keine Beeinträchtigung der Erfolgschancen sozial oder ökonomisch Benachteiligter entsteht. Denn Freiheit als absoluter Standard sichert nur den Erwerb eines bestimmten Schwellenwerts an Fähigkeiten, rechtfertigt aber keine Unterstützung für den Erwerb darüber hinausgehender Qualifikationen.

Aufgrund der Diskussion in diesem Aufsatz stellt Chancengleichheit zwar zwangsläufig eine egalitäre Explikation einer Freiheitsforderung dar. Die beiden zuletzt genannten Argumente zeigen aber, weshalb das Ideal nicht zugunsten von Freiheit aufgegeben werden sollte: Nur in Verbindung mit Chancengleichheit können durch Freiheit die angemessenen Bedingungen zur Verwirklichung eines selbstbestimmten, menschlichen Lebens sichergestellt werden.

Literaturverzeichnis

Arneson, Richard J.: „Equality and Equal Opportunity for Welfare“. Philosophical Studies, 56, 1989. S. 77-93

Arneson, Richard J.: “Against Rawlsian Equality of Opportunity”. Philosophical Studies, 93, 1996. S. 77-112

Berlin, Isaiah: Liberty. Oxford University Press, Oxford, 2005

Campbell, T. D: “Equality of Opportunity”. Proceedings of the Aristotelian Society, 75, 1975. S. 51-68

Cavanagh, Matt: Against Equality of Opportunity. Oxford University Press, Oxford, 2002

Cohen, G. A.: “Equality of What? On Welfare, Goods, and Capabilities”. In: *Nussbaum, M. & Sen A. (Hrsg.)*. 1993. S. 9-29

Cohen, G. A.: „On the Currency of Egalitarian Justice“. Ethics, 99, 1989. S. 906-944

Fishkin, James: Justice, Equal Opportunity, and The Family. Yale University Press, New Haven, 1983

Feinberg, Joel: Social Philosophy. Prentice-Hall, Englewood Cliffs, 1973

Hansson, S. O.: “What are opportunities and why should they be equal?”. Social Choice and Welfare, 22, 2004. S. 305-316

Hayek, Friedrich, A.: Law Legislation and Liberty. Volume 2: The Mirage of Social Justice, Chicago (The University of Chicago Press), 1978; deutsch: Recht, Gesetzgebung und Freiheit. Band 2: Die Illusion der sozialen Gerechtigkeit. Verlag Moderne Industrie, München, 1980-1981

Mason, Andrew: "Equality of Opportunity, Old and New". Ethics, 111, 2001. S. 760-781

McCallum, Gerald C.: "Negative and Positive Freedom". In: *Miller, David (Hrsg.*): Liberty. Oxford University Press, Oxford, 1991. S. 100-122

Meyer, Kirsten: "Warum sollten Chancen gleich sein? Chancengleichheit und Egalitarismuskritik". In: *Bohse, H. & Walter S. (Hrsg.):* Ausgewählte Beiträge zu den Sektionen der GAP.6. Mentis, Paderborn, 2007. S. 764-779

Miller, David (Hrsg.): Liberty. Oxford University Press, Oxford, 1991

O'Neill, Onora: "Wie wissen wir, wann Chancen gleich sind?". In: *Rössler, Beate (Hrsg.):* Quotierung und Gerechtigkeit. Eine moralphilosophische Kontroverse. Campus, Frankfurt/New York, 1993. S. 144-157

Pauer-Studer, H.: Autonom Leben. Reflexionen über Freiheit und Gleichheit. Suhrkamp, Frankfurt/M., 2000

Rawls, John: Eine Theorie der Gerechtigkeit. Suhrkamp, Frankfurt/M., 1979

Rawls, John: Justice as Fairness: a Restatement. Harvard University Press, Cambridge, 2001; deutsch: Gerechtigkeit als Fairness. Ein Neuentwurf. Suhrkamp, Frankfurt/M., 2003

Richards, Janet R.: "Equality of Opportunity". Ratio X, 1997. S. 253-279

Roemer, John E.: Equality of Opportunity. Harvard University Press, Cambridge, 1998

Rössler, Beate (Hrsg.): Quotierung und Gerechtigkeit. Eine moralphilosophische Kontroverse. Campus, Frankfurt/New York, 1993

Thompson, Judith Jarvis: „Bevorzugung auf dem Arbeitsmarkt": In: *Rössler, Beate (Hrsg.):* Quotierung und Gerechtigkeit. Eine moralphilosophische Kontroverse. Campus, Frankfurt/New York, 1993. S. 29-48

Westen, Peter: "The Concept of Equal Opportunity". Ethics, 95, 1985

Westen, Peter: Speaking of Equality. Princeton University Press, Princeton, 1990

Williams, Bernard: „The Idea of Equality". In: *Williams, Bernard:*: Problems of the Self. Cambridge University Press, Cambridge, 1973. S. 230-249; deutsch: „Der Gleichheitsgedanke". In: *Williams, Bernard:* Probleme des Selbst. Reclam, Stuttgart, 1978. S. 366-397

Wilt Chamberlain und organische Gerechtigkeitsprinzipien

Fabian Wendt
fabian.wendt@uni-hamburg.de
Universität Hamburg

Abstract/Zusammenfassung

In 1974, Robert Nozick presents the Wilt Chamberlain-argument in his book *Anarchy, State, and Utopia*. It is designed to show that all "structured" principles of justice are inacceptable because "liberty upsets patterns". Most philosophers think that Nozick's argument is flawed. In this paper I will try to argue that the Wilt Chamberlain-argument is in fact a powerful argument; it is, admittedly, not successful against all kinds of structured principles of justice, but indeed successful against an important subclass, namely "organic" principles of justice including egalitarian ones.

1974 präsentiert Robert Nozick in *Anarchy, State, and Utopia* das Wilt Chamberlain-Argument. Es soll zeigen, dass alle „strukturellen" Gerechtigkeitsprinzipien inakzeptabel sind, weil „Freiheit Strukturen sprengt". Die weitaus meisten Philosophen sind der Meinung, dass Nozicks Argument fehlerhaft ist. In diesem Aufsatz dagegen soll gezeigt werden, dass das Wilt Chamberlain-Argument tatsächlich ein schlagkräftiges und eigenständiges Argument ist; es ist zwar nicht gegen jede Form struktureller Gerechtigkeitsprinzipien erfolgreich, wohl aber gegen eine bedeutende Unterart, nämlich „organische" Gerechtigkeitsprinzipien und damit auch gegen egalitäre Gerechtigkeitsprinzipien.

In einem der bekanntesten Abschnitte von *Anarchy, State, and Utopia* präsentiert Robert Nozick die Geschichte vom Basketballspieler Wilt Chamberlain. Diese Geschichte stellt das von Will Kymlicka so genannte „intuitive Argument" für den von Nozick vertretenen Libertarismus dar (Kymlicka 2002, 105). Es soll zeigen, dass alle „strukturellen" Grundsätze der Gerechtigkeit mit unseren Intuitionen über die Legitimität von freiwilligen Tausch- und Schenkungshandlungen inkompatibel sind. Nozicks libertäre „Anspruchstheorie", in der es keine strukturellen Grundsätze gibt, bliebe dann als auf der Hand liegende Alternative.

Das Wilt Chamberlain-Argument ist irritierend, weil es zugleich sehr einfach und sehr folgenreich zu sein scheint. Viele Philosophen sind deswegen sicher, dass das Argument fehlerhaft ist. Ich möchte in diesem Aufsatz zu zeigen versuchen, dass die Geschichte von Wilt Chamberlain zwar tatsächlich nicht gegen die Akzeptabilität jeder Form struktureller Gerechtigkeitsprinzipien spricht. Deswegen stellt es allein auch noch kein hinreichend überzeugendes Argument

für den Libertarismus dar. Doch ich möchte auch zu zeigen versuchen, dass das Wilt Chamberlain-Argument ein überzeugendes Argument gegen eine bestimmte Subklasse struktureller Gerechtigkeitsprinzipien darstellt, nämlich gegen *organische* Gerechtigkeitsprinzipien. Eine besonders wichtige Unterart organischer Gerechtigkeitsprinzipien sind egalitaristische Gerechtigkeitsprinzipien. Das Wilt Chamberlain-Argument könnte dann als Unterstützung der Egalitarismuskritik von Harry Frankfurt und anderen willkommen geheißen werden.

1. Bevor die Wilt-Chamberlain-Geschichte erzählt werden kann, muss die Unterscheidung zwischen strukturellen und nicht-strukturellen Gerechtigkeitstheorien ein wenig erläutert werden. Nozick unterscheidet nicht nur strukturelle und nicht-strukturelle Grundsätze, sondern auch historische und nicht-historische Grundsätze. Nozicks eigene Gerechtigkeitstheorie soll nicht-strukturell und historisch sein. Er kontrastiert sie – vor der Unterscheidung zwischen strukturellen und nicht-strukturellen Grundsätzen – zunächst mit unhistorischen Gerechtigkeitstheorien, die er auch Endzustands-Prinzipien nennt (Nozick 1974, 153ff.). Bei unhistorischen oder Endzustands-Prinzipien kann man die Gerechtigkeit einer Verteilung von Gütern ohne Wissen über ihr Zustandekommen beurteilen. Ein Blick auf die gegenwärtige Verteilung genügt. Ein Beispiel für ein solches Prinzip, das auch Nozick angibt, wäre ein utilitaristischer Grundsatz, demzufolge von zwei Verteilungen diejenige die gerechtere ist, die eine höhere Nutzensumme aufweist.

Danach erst führt Nozick die Unterscheidung zwischen strukturellen und nicht-strukturellen Prinzipien ein. Er sagt: „Let us call a principle of distribution *patterned* if it specifies that a distribution is to vary along with some natural dimension, weighted sum of natural dimensions, or lexicographic ordering of natural dimensions.“ (ebd., 156). Strukturelle Grundsätze, so Nozick, können entweder historisch oder unhistorisch sein, je nachdem, ob die angegebene natürliche Dimension auf vergangene Handlungen Bezug nimmt oder nicht. Eine natürliche Dimension, die auf vergangene Handlungen Bezug nimmt, wäre z.B. der Verdienst. Ein Gerechtigkeitsprinzip, das „jedem nach seinem Verdienst“ Güter zukommen lassen will, ist dementsprechend ein struktureller und zugleich historischer Gerechtigkeitsgrundsatz (ebd.). Ein struktureller Gerechtigkeitsgrundsatz, der unhistorisch ist, so Nozick weiter, gibt dagegen eine unhistorische natürliche Dimension an, die für die Verteilung materieller Güter als relevant erachtet wird, z.B. „jedem nach seinen Bedürfnissen“ (ebd.).

Sind nun die von Nozick eingangs diskutierten Endzustands-Prinzipien strukturell oder nicht-strukturell? *Können* Endzustands-Prinzipien überhaupt nicht-strukturell sein? O'Neill ist in der Tat der Auffassung, dass die genannten Endzustands-Prinzipien nicht-strukturell sind: „End-result principles and patterned principles differ in that the former usually do not determine the size of individu-

als' shares, and the latter do." (O'Neill 1981, 307[14]). Endzustands-Prinzipien geben also nach O'Neill nicht an, *wer genau welchen* Anteil bekommen soll, sie geben nur eine anonyme Struktur an. Ich denke, dass dies in Nozicks Kategorien nicht relevant ist. Auch Endzustands-Prinzipien geben eine *Struktur* an und sind deshalb den strukturellen Gerechtigkeitsgrundsätzen zuzuordnen.[15] Wir haben es also bei Nozick mit drei, nicht vier, Typen von Gerechtigkeitstheorien zu tun: Es gibt unhistorische strukturelle Prinzipien (die Endzustands-Prinzipien), historische strukturelle Prinzipien und drittens nicht-strukturelle Prinzipien. Letztere sind stets historisch. Alle Grundsätze, die nicht-strukturell sind, sind historisch, unter den Grundsätzen, die strukturell sind, sind manche historisch, manche unhistorisch.

Nozicks eigene Theorie ist, wie gesagt, nicht-strukturell und also auch historisch. Er nennt sie *entitlement theory*, deutsch „Anspruchstheorie". Sie besagt, im Kern, dass eine Güterverteilung gerecht ist, wenn jeder auf seine Güter Anspruch hat. Und man hat auf seine Güter Anspruch, wenn man sie sich gemäß einem Grundsatz der gerechten Aneignung angeeignet hat oder gemäß einem Grundsatz des gerechten Transfers übertragen bekommen hat. Obwohl Nozick John Lockes Theorie der gerechten Aneignung in vielen Punkten kritisiert und mehr Fragen als Antworten hinterlässt, scheint er zumindest etwas Ähnliches wie diese Theorie als Grundsatz der gerechten Aneignung akzeptieren zu wollen (Nozick 1974, 174ff.).[16] Der Grundsatz des gerechten Transfers scheint bei Nozick im Kern zu besagen, dass freiwilliger Tausch und freiwillige Schenkungen ein legitimes Mittel der Übertragung von Gütern sind. Letztlich läuft die Anspruchstheorie auf die Akzeptanz der klassisch-liberalen Rechte einer jeden Person auf Leben, Freiheit und Eigentum hinaus. Jeder ist Eigentümer seines Körpers, jeder kann sich externe Güter aneignen und diese dann tauschen oder verschenken. Alle anderen haben die Pflicht, seinen Körper und seine externen Güter nicht anzutasten.

Die Anspruchstheorie kennt jedenfalls keine strukturellen Prinzipien; sie gibt keine natürliche Dimension an, nach der Güter verteilt sein müssen, damit die Verteilung gerecht ist. Es ist irrelevant, wie die Güterverteilung in einer Gesellschaft strukturiert ist, solange sie durch legitime Mittel – erste Aneignung, freiwillige Übertragung – zustande kam. Zwei mögliche Welten, in denen das Eigentum völlig verschieden verteilt ist, können beide gerecht sein, wenn in beiden Welten Eigentum nur durch legitime erste Aneignung oder freiwillige Tausch- oder Schenkungsakte erworben wurde. Umgekehrt kann von zwei möglichen Welten, in denen das Eigentum genau gleich verteilt ist, die eine gerecht,

14 Vgl. auch Knoll 2008, 89

15 Entsprechend schreibt Nozick auch von „[…] end-state *and other* patterned conceptions of justice […]" (Nozick 1974, 172, Hervorhebung d. Verf.).

16 Für einen alternativen Grundsatz der gerechten Aneignung siehe Mack 1990, dazu auch Wendt 2009, 126ff.

die andere aber ungerecht sein, wenn in der einen diese Verteilung durch unzulässige Mittel zustande kam.

2. In dem Wilt-Chamberlain-Abschnitt (Nozick 1974, 160ff.) möchte Nozick nun zeigen, dass strukturelle Prinzipien unakzeptabel sind, weil sie fundamentalen moralischen Intuitionen widerstreiten. Der Leser soll sich zunächst vorstellen, dass in einem Gemeinwesen ein strukturelles Gerechtigkeitsprinzip seiner Wahl verwirklicht ist. Das strukturelle Gerechtigkeitsprinzip behauptet, dass eine bestimmte Struktur in der Verteilung von Eigentum, D1 genannt, verwirklicht sein muss, damit die Verteilung gerecht ist. Wenn die bevorzugte strukturelle Gerechtigkeitstheorie etwa das Differenzprinzip von Rawls ist, so kommen alle Ungleichheiten in Einkommen und Vermögen den Schlechtestgestellten zugute (Rawls 1971, 60). Diese Struktur D1 ist verwirklicht und Ausgangspunkt der Geschichte von Wilt Chamberlain.

Nun kommt der berühmte Basketballspieler Wilt Chamberlain in dieses Gemeinwesen und unterzeichnet einen Vertrag beim örtlichen Basketballteam. Er soll fünfundzwanzig Cent von jedem Heimspielticket bekommen. Die Zuschauer besuchen in Scharen die Heimspiele und geben je fünfundzwanzig Cent in einen extra zu diesem Zweck bereitgestellten Topf für Wilt Chamberlain. Dieser verdient auf diese Weise mit der Zeit viel Geld und wird reich. Wir haben – offenkundig – eine neue Eigentumsverteilung, die wir D2 nennen können, vor uns. Der strukturelle Gerechtigkeitsgrundsatz ist nicht mehr erfüllt. Nozick fragt:

> Is this new distribution [...] unjust? If so, why? There is *no* question about whether each of the people was entitled to the control over the resources they held in D_1; because that was the distribution (your favorite) that (for the purposes of argument) we assumed acceptable. Each of these persons *chose* to give twenty-five cents of their money to Chamberlain. (Nozick 1974, 161).

Was ist also mit der Geschichte gezeigt? Dass Freiheit – freiwilliges Tauschen und Schenken – strukturellen Gerechtigkeitsgrundsätzen widerstreitet. „Liberty upsets patterns“ heißt der Titel des Kapitels in *Anarchy, State, and Utopia*. Und freiwilliges Tauschen und Schenken ist intuitiv moralisch unproblematisch. Strukturelle Gerechtigkeitsgrundsätze müssten fordern, freiwillige Tausch- und Schenkungsakte rückgängig zu machen um die geforderte Struktur wiederherzustellen.

> The general point illustrated by the Wilt Chamberlain example [...] is that no end-state principle or distributional patterned principle of justice could be continuously realized without continuous interference with people's lives. [...] To maintain a pattern one must either continually interfere to stop people from transferring resources as they wish to, or

> continually (or periodically) interfere to take from some persons resources that others for some reason chose to transfer to them. (ebd., 163).[17]

Eine adäquate Gerechtigkeitstheorie, so Nozick, ist deswegen nicht-strukturell: Jeder Güterverteilung, die aus freiwilligen Tausch- oder Schenkungsvorgängen von (legitim erworbenem) Eigentum resultiert, ist gerecht. Die libertäre Konsequenz ist, dass ein Steuereinzug zum Zwecke der Umverteilung im Namen eines strukturellen Gerechtigkeitsprinzips – oder auch zu anderen Zwecken – ungerecht ist. Weshalb nach Nozick Besteuerung für die Zwecke des Minimalstaats dennoch legitim sein soll, kann hier nicht erläutert werden.

3. Wie reagieren nun Philosophen, die materielle Umverteilung für legitim halten, auf das Wilt-Chamberlain-Beispiel? Häufig wird behauptet, Nozick setze voraus, was er beweisen möchte, nämlich „absolute Eigentumsrechte". Der Fehler des Arguments liege darin, dass Nozick unterstellt, dass von der vom Leser bevorzugten Gerechtigkeitstheorie in D1 absolute Eigentumsrechte verteilt würden. So schreibt etwa Will Kymlicka: „The best response to his offer to specify D1 is to refuse to specify any distribution at all. For if Nozick insists on treating D1 as endowing absolute rights, then we may not believe there is a fair initial distribution of such rights." (Kymlicka 1990, 103).[18] Doch stimmt das? Setzt Nozick bereits absolut geltende Eigentumsrechte voraus? Edward Feser schreibt dazu:

> Some have alleged that Nozick's argument begs the question in that it presupposes that those given shares under D1 are „entitled" to them in Nozick's strong sense of having an *absolute property right* in them – a presupposition that Nozick's opponent would, of course, reject […]. But the argument assumes no such thing. It need assume, not that you can do *anything* you like with what you're given under D1, but only that you can freely do at least *something* with it. […] But as long as individuals can do *something* with their shares, surely they can in principle do something that breaks the pattern." (Feser 2004, 71f.).

Diese Antwort von Feser scheint mir schlicht korrekt zu sein. Natürlich läuft das Argument von Nozick darauf *hinaus*, dass jeder Eingriff in Eigentumsrechte im Namen eines strukturellen Gerechtigkeitsprinzips als illegitim zu erachten ist. Doch dies ist tatsächlich eine Schlussfolgerung, keine Voraussetzung im Wilt Chamberlain-Argument. Damit freiwilliges Geben Strukturen sprengen kann, reicht es, dass man zumindest ein paar Dinge mit dem unter D1 verteilten Eigentum machen darf, darunter auch zumindest bestimmte Formen von freiwilliger

17 An anderer Stelle schreibt Nozick, dass strukturelle Grundsätze die Konsequenz hätten, „loving behavior" zu verbieten (vgl. Nozick 1974, 167). Für eine formalere Darstellung des Wilt Chamberlain-Arguments vgl. Wendt 2011.

18 Ähnlich auch Kersting 2000, 317, Koller 1987, 160ff., Nagel 1981, 201, O'Neill 1981, 309, Scanlon 1981, 110f.

Schenkung oder freiwilligem Tausch.[19] Um zu verhindern, dass Strukturen gesprengt werden, müsste man schon behaupten, dass die Menschen unter D1 ihre Ressourcen *nur für sich selbst* gebrauchen dürfen und niemandem schenken dürfen (Nozick 1974, 167) – und wer würde das fordern wollen?

In einer zweiten Replik auf das Wilt Chamberlain-Argument wird bezweifelt, dass alle freiwilligen Tauschhandlungen moralisch zulässig sind. Peter Koller etwa schreibt:

> Die moralische Zulässigkeit sozialer Umstände läßt sich eben nicht allein daran festmachen, daß sie durch freiwillige Transaktionen zustande gekommen sind, sondern es kommt auch darauf an, ob die *Randbedingungen*, unter denen diese Transaktionen stattgefunden haben, ihrerseits moralisch zulässig waren. (Koller 1987, 164).

Dies würde nun allerdings auch Nozick nicht bestreiten. „Randbedingungen" sind für Nozick allerdings schlicht die Lockeschen Rechte. Zum Beispiel ist der freiwillige Tausch, den ein Hehler und sein Kunde vornehmen, natürlich aus Sicht der Nozickschen Anspruchstheorie problematisch, weil der Hehler eben keinen Anspruch auf das Gut hatte. Dazu kommt, dass Nozicks Grundsatz der gerechten Aneignung ein Lockesches Proviso enthalten soll, das vermeintlich freiwilligen Tausch in einer Monopolsituation als Rechte verletzend charakterisieren lässt (Nozick 1974, 178ff.). Klar ist für Nozick auch, dass Betrug gegen Lockesche Rechte verstößt bzw. dass ein auf Betrug basierender Vertrag kein freiwilliger Vertrag ist.[20]

Manches, was Marxisten als Ausbeutung bezeichnen würden, ist nach Nozick allerdings in der Tat freiwilliger Tausch: Jemand, der nur unter wenigen und wenig attraktiven Optionen wählen kann, ist dennoch jemand, der noch freiwillig eine der beiden Optionen wählt, solange niemand seine Lockeschen Rechte verletzt (oder droht, seine Lockeschen Rechte zu verletzen). Und dies auch dann, wenn die Handlungen anderer Personen dazu geführt haben, dass diese Person nur unter solchen Bedingungen wählen kann. In diesem Sinne ist niemand gezwungen, für einen sehr niedrigen Lohn bei einem Arbeitgeber zu arbeiten, sondern tut es freiwillig.

Nozick verdeutlicht die Plausibilität dieses Freiwilligkeitsbegriffs an einem Beispiel (das weniger bekannt ist als das Wilt Chamberlain-Beispiel): Nehmen wir an, wir haben 26 Männer namens A, B, C usw. und 26 Frauen namens A', B', C' usw. vor uns. Alle wollen heiraten. Alle Männer haben die gleiche Präferenzordnung: Sie ziehen die Frau A' der Frau B' vor, die Frau B' der Frau C' usw. Das gleiche gilt für die Frauen: Sie ziehen allesamt den Mann A dem Mann B vor, den Mann B dem Mann C usw. So kommt es dazu, dass A und A' heiraten werden. B und B' würden zwar lieber A bzw. A' geheiratet haben, werden nun aber miteinander vorlieb nehmen und ebenfalls heiraten. Dass am Ende Z

19 „What was it for if not to do something with?" (Nozick 1974, 161).

20 Vgl. z.B. Nozick 1974, ix, dazu Wolff 1991, 85f.

und Z' nur noch die Wahl zwischen Nicht-Heiraten und Z- bzw. Z'-Heiraten haben, bedeutet jedoch kaum, dass sie einander nicht mehr „freiwillig" heiraten werden (ebd., 263). Sie wählten freiwillig, obwohl sie wenige und wenig attraktive Optionen hatten und diese Tatsache durch die Handlungen der anderen Personen herbeigeführt wurde.

Doch müssen wir hier nicht entscheiden, ob dieser Begriff der Freiwilligkeit tatsächlich überzeugend ist.[21] In der Geschichte von Wilt Chamberlain jedenfalls wird kein besonders strittiger Freiwilligkeitsbegriff vorausgesetzt. Keine besonderen Randbedingungen stehen der Charakterisierung der Schenkungshandlungen von 25 Cent als „freiwillig" entgegen. Wir dürfen in der Geschichte davon ausgehen, dass die Zuschauer, die in die Spiele von Wilt Chamberlain strömen, *wirklich* freiwillig – gemäß des Lesers Lieblingstheorie der Freiwilligkeit – in die Spiele von Wilt Chamberlain gehen. Niemand sprach Drohungen aus für den Fall, dass sie es nicht tun, sie wurden auch nicht „manipuliert" – sie gehen schlicht und einfach in das Spiel, weil es sie interessiert oder weil es ihnen Spaß macht. Trotzdem werden Strukturen gesprengt…

4. Dass also freiwillige Tausch- und Schenkungshandlungen Strukturen sprengen können, sollte hinreichend einsichtig sein. Darüber, dass (wirklich) freiwillige Tausch- und Schenkungshandlungen moralisch unproblematisch sind, sollte auch Einigkeit herrschen. Doch wie relevant ist dies nun? Reicht es, um zu zeigen, dass alle strukturellen Grundsätze inakzeptabel sind und wir deswegen Nozicks Anspruchstheorie akzeptieren müssen[22]? Ist nicht ein Kompromiss möglich? Kann es nicht Gerechtigkeitsgrundsätze geben, „die strukturelle Kriterien mit Prinzipien einer (begrenzten) Handlungsfreiheit der Individuen in irgendeiner Weise kombinieren"? (Koller 1987, 174[23]). Solche Gerechtigkeitsgrundsätze würden, so Koller, nicht mehr einen „ständigen Eingriff in das Leben der Menschen" legitimieren (Koller 1987, 176).

---

21 Für eine Kritik an Nozicks Begriff der Freiwilligkeit siehe Cohen 1977, 34ff., dazu auch Wolff 1991, 83ff.

22 Den Linkslibertarismus werde ich als interessante nicht-strukturelle Alternative zu Nozicks Anspruchstheorie der Gerechtigkeit hier ignorieren müssen. Hillel Steiner *akzeptiert* die Schlagkraft des Wilt Chamberlain-Arguments gegen strukturelle Gerechtigkeitstheorien (vgl. insbesondere Steiner 1977), meint jedoch anders als Nozick und andere „klassische" Libertäre, dass es ein Naturrecht auf einen gleichen Anteil äußerer Güter gibt (vgl. Steiner 1977, Steiner 1994, 262ff., zur Diskussion vgl. Cohen 1986, 102ff., Mack 2009). Wenn jede Person diesen ihren gleichen Anteil an äußeren Gütern zugeteilt bekommen hat, ist nach Steiner jedoch tatsächlich jede Güterverteilung, die durch freiwilligen Tausch dieser Güter hergestellt wird, gerecht. Eine solche „starting gate"-Theorie ist, wie gesagt, eine Alternative *innerhalb* des Lagers nicht-struktureller Gerechtigkeitstheorien.

23 Vgl. auch Kersting 2000, 314, Knoll 2008, 239f.

Es gibt zwei Möglichkeiten, strukturelle Grundsätze mit der Intuition bezüglich der Legitimität von freiwilligen Tausch- und Schenkungshandlungen zu kombinieren. Die eine ist, den Widerstreit zwischen dem Prinzip der Tauschfreiheit, wie ich es von nun an nennen möchte, und dem strukturellen Prinzip einfach zu akzeptieren und also eine pluralistische politische Philosophie mit echten Prinzipienkonflikten zu vertreten. Die andere Möglichkeit besteht darin, strukturelle Grundsätze auf einer anderen Ebene zu verorten und zu behaupten, dass sich strukturelle Grundsätze und das Prinzip der Tauschfreiheit in Wirklichkeit nicht widerstreiten.

Rawls scheint den zweiten Weg zu wählen, wenn er zu Nozicks Vorwurf schreibt: „The objection that the difference principle enjoins continuous corrections of particular distributions and capricious interference with private transactions is based on a misunderstanding." Es geht nach Rawls schließlich nicht darum, jede einzelne Tauschhandlung zu beobachten und gegebenenfalls einzugreifen, wenn durch sie eine Struktur gesprengt wurde, sondern um ein Regelsystem in Gesetzesform, das zu einer bestimmten Verteilungsstruktur führen soll. Über das Steuersystem und Gesetze bzgl. öffentlichen Schulen, gesetzlichen Krankenkassen etc. soll das Differenzprinzip verwirklicht werden. In der Tat scheint das auf den ersten Blick plausibel. Wenn es eine Mehrwert- und Einkommenssteuer gibt, dann lässt man die Leute doch tauschen, wie sie wollen, nur nimmt man systematisch einen bestimmten Anteil davon weg. In einem gewissen Rahmen lässt man also freien Tausch tatsächlich geschehen.

Doch wie wirkt nun die Struktur? Bei der weiteren Klärung des Missverständnisses, dem Nozick erlegen war, schreibt Rawls:

> [E]ven if everyone acts fairly as defined by the rules that it is both reasonable and practicable to impose on individuals, the upshot of many separate transactions will eventually undermine background justice. [...] Thus, even in a well-ordered society adjustments in the basic structure are always necessary. (Rawls 1993, 283f.).

David Schmidtz kommentiert dies etwas süffisant: „The clarification makes it hard to see what Nozick misunderstood." (Schmidtz 2006, 200). Auch mir scheint Nozicks zentraler Punkt unangetastet zu bleiben: Es wird eine bestimmte Struktur gefordert, die durch ein bestimmtes Regelsystem erreicht werden soll und dieses Regelsystem muss geändert werden, wenn die Struktur nicht erreicht wird. Das Wilt-Chamberlain-Beispiel lässt uns stutzig werden, warum man dies im Namen der Gerechtigkeit verlangen sollte, da es dem intuitiv akzeptierten Prinzip der Tauschfreiheit widerstreitet.

Die Tatsache, dass die Eingriffe in freiwillige Tausch- und Schenkungshandlungen zum Zweck der Realisierung eines strukturellen Gerechtigkeitsgrundsatzes über Steuern passieren, ändert nicht ihren grundsätzlichen Charakter. Der einzige Unterschied ist, dass mit dem Mittel der Besteuerung systematisch und präventiv statt am Einzelfall orientiert und ex post eingegriffen wird. Es sind Eingriffe im Namen eines solchen Grundsatzes, die eine andere Struktur herbei-

führen sollen, als sie sich ohne solche Eingriffe ergeben hätte. Es bleibt also die Tatsache bestehen, dass es einen Widerstreit zwischen dem Prinzip der Tauschfreiheit und dem strukturellen Grundsatz gibt. Die Ebenen-Verlagerung ändert nichts daran.

5. Wenn man am Prinzip der Tauschfreiheit festhalten und gleichzeitig einen strukturellen Grundsatz vertreten möchte, so wird man diesen Widerstreit akzeptieren müssen. Nun ist es so, dass manche strukturellen Prinzipien mit dem Prinzip der Tauschfreiheit verträglicher sind als andere. Ich möchte behaupten, dass der Konflikt zwischen strukturellem Prinzip und dem Prinzip der Tauschfreiheit bei einer bestimmten Unterklasse struktureller Prinzipien derart gravierend wäre, dass diese Unterklasse als nicht-akzeptabel ausscheidet. Gegen sie ist das Wilt Chamberlain-Argument tatsächlich erfolgreich. Die Unterklasse, die ich meine, ist die *organischer* Prinzipien der Gerechtigkeit.

Nozick führt die Unterscheidung zwischen organischen und nicht-organischen strukturellen Prinzipien an anderer Stelle ein, nicht im Kontext des Wilt Chamberlain-Arguments. Er schreibt: „Let us say that a principle of distribution is *organic* if an unjust distribution, according to this principle, can be gotten from one the principle deems just, by deleting (in imagination) some people and their distributive shares.“ (Nozick 1974, 209). Das gedankliche Streichen einer Person kann dazu führen, dass ein organisches Prinzip in einer vor der Streichung gerechten Gesellschaft nicht mehr erfüllt ist oder dazu, dass ein organisches Prinzip in einer vorher ungerechten Gesellschaft nach der Streichung doch erfüllt ist. Dies dürfte genau bei solchen Prinzipien der Fall sein, die relationale Begriffe benutzen. Solchen Prinzipien kommt es darauf an, wie viel jemand im Vergleich zu anderen hat, deswegen muss stets die Gesellschaft als Ganze im Auge behalten werden und das gedankliche Streichen einer Person eine ungerechte Verteilung plötzlich zu einer gerechten machen (und umgekehrt). Die beliebtesten organischen Prinzipien sind natürlich solche Prinzipien, die in irgendeiner Form auf Gleichheit abzielen, also egalitaristische Prinzipien.

Es ist nicht behauptet worden, dass bei solchen Grundsätzen das gedankliche Streichen einer Person *immer* die Bewertung einer Verteilung als „gerecht“ oder „ungerecht“ in ihr Gegenteil verkehren wird. Zum Beispiel ist ein Grundsatz, der eine egalitäre Verteilung fordert, nicht verletzt, wenn von Personen A, B und C, die alle gleich viel haben, eine Person gedanklich gestrichen wird. Aber wenn C vorher mehr hatte, führt sein gedankliches Streichen dazu, dass die Verteilung gerecht wird. Der Test besteht also in der Überlegung, ob das gedankliche Streichen einer Person dazu führen *kann*, dass sich die Beurteilung einer Verteilung als gerecht oder ungerecht in ihr Gegenteil verkehrt. Dies ist offenkundig bei egalitaristischen Prinzipien der Fall.

Es gibt bekanntlich eine große Bandbreite egalitaristischer Prinzipien: Es kann um Gleichheit von Ressourcen, Wohlfahrt oder Chancen gehen. Und meistens werden egalitaristische Prinzipien um andere Prinzipien ergänzt. Selten wird Gleichheit als der einzige zu verwirklichende Wert gesehen, schließlich ließe sich Gleichheit auch dadurch schaffen, dass man alle Menschen gleich arm macht oder alle Menschen tötet (Krebs 2002, 566f.). Eine deswegen wohl unumgängliche Abmilderung des Egalitarismus erhält man, wenn man zusätzlich Wohlfahrt einen intrinsischen Wert zuspricht. So kann man auch das Differenzprinzip von Rawls lesen: Ungleichheit wird hingenommen, wenn dies einer Verbesserung der Lebenssituation *aller*, insbesondere der Schlechtestgestellten dient. Zweitens wird bekanntlich oft versucht, egalitäre Prinzipien mit der Idee von Verdienst oder Verantwortung zu kombinieren und Ungleichheiten zuzulassen, wenn sie verdient sind oder eigenverantwortet sind. Der Egalitarismus wirkt dann nur noch bei unverschuldeten oder unverdienten Ungleichheiten. Alle diese Varianten von egalitären Prinzipien bleiben jedoch organisch. So erfüllt etwa Rawls' Differenzprinzip weiterhin Nozicks Kriterium: „The difference principle *is* organic. If the least well-off group and their holdings are deleted from a situation, there is no guarantee that the resulting situation and distribution will maximize the position of the new least well-off group." (Nozick 1974, 209). Genauso kann das Streichen der bestgestellten Person, deren Besitz nicht zum Vorteil der Schlechtestgestellten war, dazu führen, dass eine ungerechte Verteilung zu einer gerechten Verteilung wird, obwohl sich die Situation der Schlechtestgestellten, nicht-relational betrachtet, nicht verändert hat.

Tausch- und Schenkungshandlungen *verändern* fast immer die Verteilungsstruktur einer Gesellschaft. Sobald jemand ein begehrtes Produkt geschaffen oder eine besonders geschätzte Fähigkeit hat (zum Beispiel Basketballspielen), wird durch freiwillige Tauschhandlungen die Verteilung von Geld zu seinen Gunsten verändert. Und die Verteilung von Geld dürfte unbestritten auch relevant sein für die Wohlfahrt und Chancen von Individuen (also das, was nach egalitären Prinzipien gleich sein sollte). Dies gilt natürlich ebenso für den Fall von freiwilligen Schenkungen von Eltern an ihre Kinder, Erbschaften usw. Die Tatsache, dass Tausch- und Schenkungshandlungen fast immer die Verteilungsstruktur verändern, spricht stark gegen die Vereinbarkeit des Prinzips der Tauschfreiheit mit einem organischen Gerechtigkeitsprinzip. Denn organische Prinzipien sind *besonders empfindlich* gegenüber Veränderungen.

Worin besteht genau diese besondere Empfindlichkeit? Denken wir uns irgendein organisches Prinzip, zum Beispiel ein egalitäres Prinzip, und nehmen an, dass es erfüllt ist: Wir haben eine Verteilung D1. Nun passiert eine kleine Veränderung hin zu einer Verteilung D2: Person Z bekommt etwas von Person Y. Alle anderen Personen, A bis X, behalten, was sie unter D1 hatten. Trotzdem kann es sein, dass sie nun nicht mehr ihren „gerechten Anteil" haben. Dies ist etwas, das bei nicht-organischen Prinzipien nicht passieren kann. Wenn eine

Person A in einer Verteilung D1 einen gerechten Anteil hat und in einer Verteilung D2 hat sie *dasselbe*, dann ist dies aus der Perspektive eines nicht-organischen Prinzips eben wieder ein gerechter Anteil. Bei organischen Prinzipien dagegen kann ein und dasselbe in einer Verteilung D1 ein gerechter Anteil, in einer Verteilung D2 aber ein ungerechter Anteil sein.

Dies ist schon durch die Bestimmung der Organizität eines Gerechtigkeitsprinzips klar: Die gedankliche Streichung einer Person lässt allen das, was sie vorher hatten, und trotzdem kann der Anteil von einem vorher gerechten nun zu einem ungerechten Anteil geworden sein (oder umgekehrt).

Ein ernst genommenes egalitäres Prinzip, davon kann man nach aller empirischen Evidenz ausgehen, wird deswegen ständig und intensiv, spätestens nach einigen wenigen Tausch- und Schenkungshandlungen, dem Prinzip der Tauschfreiheit widerstreiten. Es scheint mir aus diesem Grund wenig plausibel, das Prinzip der Tauschfreiheit mit einem egalitären strukturellen Prinzip verbinden zu wollen. Man kann kaum überzeugend behaupten, Tauschfreiheit grundsätzlich für unproblematisch zu halten, wenn man dies nur unter dem Vorbehalt tut, dass durch freiwilligen Tausch keine Ungleichheiten in Wohlfahrt oder Chancen (oder etwas anderem) produziert werden. Zu deutlich ist, dass Tauschfreiheit sofort und unvermeidlich Ungleichheiten in Wohlfahrt und Chancen produzieren wird. Das Bekenntnis zur Tauschfreiheit kann hier nur ein Lippenbekenntnis sein, da es keinen relevanten Raum mehr für legitimen freien Tausch gibt. Die Vereinigung eines organischen Prinzips mit dem Prinzip der Tauschfreiheit kann in Wirklichkeit keine Vereinigung sein, sondern nur eine Aufgabe des Prinzips der Tauschfreiheit.

6. Nicht-organische Prinzipien dagegen sind grundsätzlich verträglicher mit dem Prinzip der Tauschfreiheit. Sie lassen dem freiwilligen Tausch und der freiwilligen Schenkung Raum. Manche nicht-organischen Prinzipien dürften derart harmlos sein, dass sie so gut wie nie durch freiwillige Tausch- und Schenkungshandlungen verletzt werden.[24] Zum Beispiel ist es empirisch sehr unwahrscheinlich, dass ein struktureller Grundsatz, der fordert, dass alle Menschen, die einen Rolls Royce besitzen, auch einen Herd besitzen sollen, durch freiwillige Tausch- und Schenkungsakte häufig umgeworfen würde.

Aber es sind nicht nur abwegige strukturelle Grundsätze wie dieser, die dem Prinzip der Tauschfreiheit einigen Raum lassen. Es gibt auch für eine Gerechtigkeitstheorie plausible nicht-organische Grundsätze. Will Kymlicka etwa schreibt in einer ersten Reaktion auf die Geschichte von Wilt Chamberlain, dass seine Intuitionen ihm immer noch sagen würden, dass man Wilt Chamberlain besteuern darf, wenn dadurch verhindert werden kann, dass Menschen in absolu-

24 Auch Nozick sieht dies (vgl. Nozick 1974, 164).

te Armut gelangen oder verhungern (Kymlicka 2002, 106). Ein entsprechendes strukturelles Gerechtigkeitsprinzip würde verlangen, dass Menschen nicht unter ein bestimmtes Wohlfahrtsniveau fallen dürfen. Harry Frankfurt, zum Beispiel, vertritt ein solches *Suffizienzprinzip*: „Economic equality is not as such of particular importance. With respect to the distribution of economic assets, what *is* important from the point of view of morality is not that everyone should have *the same* but that each should have *enough*." (Frankfurt 1987, 134).

Das Suffizienzprinzip ist natürlich ein strukturelles Prinzip. Eine bestimmte Struktur muss erfüllt sein, damit eine Verteilung als gerecht gelten kann: Alle müssen ein bestimmtes Wohlfahrtsniveau haben. Doch das Suffizienzprinzip ist nicht organisch; es ist ohne relationale Begriffe formuliert und das gedankliche Streichen einer Person kann deswegen nichts an der Gerechtigkeit oder Ungerechtigkeit einer Verteilung ändern. Es ist deswegen auch nicht unplausibel, dass dieses Prinzip eher selten verletzt wird und den freiwilligen Tausch- und Schenkungshandlungen breiten Raum lässt. Die Verträglichkeit mit dem Prinzip der Tauschfreiheit ist natürlich eine graduelle Sache: Je niedriger das Wohlfahrtsniveau in einem Suffizienzprinzip bestimmt wird, desto mehr Raum wird dem Prinzip der Tauschfreiheit gelassen. Ein nicht-organisches Prinzip wie das Suffizienzprinzip hat aber jedenfalls den Vorteil, dass nie eine Tauschhandlung von D1 zu D2 aus *dem* Grund rückgängig gemacht werden muss, dass jemand, der in D1 und D2 *dasselbe* hat, nun nicht mehr „seinen gerechten Anteil" hat, weil sich die Verteilungsstruktur im gesellschaftlichen „Organismus" verändert hat. Dies macht die klar bessere Vereinbarkeit eines nicht-organischen Prinzips mit dem Prinzip der Tauschfreiheit aus.

Ich halte es für möglich und nicht unplausibel, das Prinzip der Tauschfreiheit mit einem nicht-organischen strukturellen Prinzip kombinieren zu wollen. Jedenfalls spricht die Geschichte von Wilt Chamberlain nicht dagegen. Auf die Konfliktträchtigkeit einer solchen Kombination von Tauschfreiheitsprinzip und strukturellem Prinzip macht die Geschichte natürlich gleichwohl aufmerksam.

7. Zusammengefasst: Die Geschichte von Wilt Chamberlain liefert noch kein Argument gegen jede Form struktureller Prinzipien. Hierfür reicht der in der Wilt Chamberlain-Geschichte enthaltene Appell an unsere Intuitionen bezüglich der Legitimität von freiwilligem Tausch und freiwilliger Schenkung noch nicht aus. Das Wilt Chamberlain-Argument ist aber dennoch kraftvoll: Es zeigt den Konflikt von strukturellen Gerechtigkeitsprinzipien mit dem Prinzip der Tauschfreiheit auf. Und es zeigt, dass dieser Konflikt im Falle von *organischen*, z.B. egalitaristischen, strukturellen Prinzipien derart gravierend sein dürfte, dass solche Prinzipien abgelehnt werden sollten. Es stellt somit ein beeindruckend einfaches, eigenständiges Argument gegen den Egalitarismus dar.

Literaturverzeichnis

Cohen, Gerald: „Robert Nozick and Wilt Chamberlain: How Patterns Preserve Liberty". 1977. In: *Cohen, G.*: Self-Ownership, Freedom, and Equality. Cambridge University Press, Cambridge, 1995. S. 19-37

Cohen, Gerald: „Are Freedom and Equality Compatible?". 1986. In: *Cohen, G.*: Self-Ownership, Freedom, and Equality. Cambridge University Press, Cambridge, 1995. S. 92-115

Feser, Edward: On Nozick. Wadsworth, Pasadena, 2004

Frankfurt, Harry: „Equality as a Moral Ideal". 1987. In: *Frankfurt, H.*: The Importance of What We Care About. Cambridge University Press, Cambridge, 1988. S. 134-158

Kersting, Wolfgang: Theorien der sozialen Gerechtigkeit. Metzler, Stuttgart, 2000

Knoll, Bodo: Minimalstaat: Eine Auseinandersetzung mit Robert Nozicks Argumenten. Mohr, Tübingen, 2008

Koller, Peter: Neue Theorien des Sozialkontrakts. Duncker & Humblot, Berlin, 1987

Krebs, Angelika: „Gleichheit oder Gerechtigkeit: Die Kritik am Egalitarismus". 2002. www.gap-im-netz.de/gap4Konf/Proceedings4/pdf/6%20Pol1%20Krebs.pdf, zuletzt aufgerufen am 8.7.2010

Kymlicka, Will: Contemporary Political Philosophy (First Edition). Oxford University Press, Oxford, 1990

Kymlicka, Will: Contemporary Political Philosophy (Second Edition). Oxford University Press, Oxford, 2002

Mack, Eric: „Self-Ownership and the Right of Property". The Monist, 73, 1990. S. 519-543

Mack, Eric: „What is Left in Left-Libertarianism?". 2009. In: *De Wijze, S./Kramer, M./Carter, I. (Hrsg.):* Hillel Steiner and the Anatomy of Justice. Routledge, London, 2009. S. 101-131

Nagel, Thomas: „Libertarianism without Foundations". In: *Paul, J. (Hrsg.)*: Reading Nozick: Essays on Anarchy, State, and Utopia. Rowman & Littlefield, Totowa, 1981. S. 191-205

Nozick, Robert: Anarchy, State, and Utopia. Basic Books, New York, 1974

O'Neill, Onora: „Nozick's Entitlements". In: *Paul, J. (Hrsg.)*: Reading Nozick: Essays on Anarchy, State, and Utopia. Rowman & Littlefield, Totowa, 1981. S. 305-322

Rawls, John: A Theory of Justice. Belknap Press, Cambridge (MA), 1971
Rawls, John: Political Liberalism. Columbia University Press, New York, 1993

Scanlon, Thomas: „Nozick on Rights, Liberty, and Property". In: *Paul, J. (Hrsg.)*: Reading Nozick: Essays on Anarchy, State, and Utopia. Rowman & Littlefield, Totowa, 1981. S. 107-129
Schmidtz, David: Elements of Justice. Cambridge University Press, Cambridge, 2006
Steiner, Hillel: „The Natural Right to the Means of Production". Philosophical Quarterly 27, 1977. S. 41-49
Steiner, Hillel: An Essay on Rights. Blackwell, Oxford, 1994

Wendt, Fabian: Libertäre politische Philosophie. Mentis, Paderborn, 2009
Wendt, Fabian: „Nozick's Wilt Chamberlain-Argument". In: *Bruce, M./Barbone S. (Hrsg.)*: Just the Arguments: 100 of the Most Important Arguments in Western Philosophy. Blackwell, Oxford, 2011. S. 254-257
Wolff, Jonathan: Robert Nozick: Property, Justice and the Minimal State. Polity Press, Cambridge, 1991

7 Normative Ethik

An Application of Parity in Decision-Making

Urs Allenspach
urs.allenspach@env.ethz.ch
ETH Zurich, Dept. of Environmental Sciences

Abstract/Zusammenfassung

This article discusses an application of a non-traditional evaluative relation of parity. Parity is presented as a solution to a problem in decision-making, along with an orthodox solution of transitive closure. The decision problem is that, although adopting a random process is a rational procedure when choosing a single option from a plurality of options, none of which is inferior to any other, the result may be less than ideal for those needing to make several such consecutive decisions. Parity and transitive closure block that outcome in different ways. It is argued that parity is the more plausible solution.

Dieser Artikel stellt eine Anwendung einer Form der nicht-traditionellen Vergleichsrelation Parität vor. Das Konzept der Parität bietet eine alternative Lösung für das Entscheidungsproblem, in dem Zufallsverfahren sub-optimale Resultate bei mehreren aufeinanderfolgenden Entscheidungen zwischen unvergleichbaren Optionen herbeiführen. Es wird argumentiert, dass die Paritätslösung plausibler sei als die herkömmliche Lösung, die von der Idee der transitiven Hülle Gebrauch macht.

1. Parity as a solution to a decision problem

How can we rationally decide among options, none of which is worse than another? - This question seems quite innocent: If none of the options is worse than any other, a decision in favour of any seems acceptable. In other words, no decision made by any random process can be criticised as irrational.
Such an assessment may be correct in singular cases. It is wrong, however, if we perceive these decisions as parts of a chain.

To arrive at this conclusion, consider the following example, which shows the typical structure of examples occurring in the philosophical literature on comparability and incomparability, and on commensurability and incommensurability.[1]

This paper is based on the author's research within the projects COST E45 and ClimPol of ETH Domain. An earlier version of it has been published in German in Studia Philosophica, 68 2009, 65–83.

1 Cf. De Sousa (1974, pp. 544-545), Parfit (1986, pp. 430-431), Griffin (1986, p. 81), Chang (2002, p. 668), and Carlson (2006, pp. 19-20).

Assume you and a friend are thinking about buying tickets to the opera. You want to be seated side by side. A call to the ticket counter reveals the following offer: A first pair of tickets, A, is located in the first row of the balcony. The others, B, are 7th row stalls. Both pairs would be excellent places, as their identical (exorbitant) price reflects. Although excellent, these seats are also disparate. When it comes to sight-lines, acoustics or legroom, there are differences – some in favour of A and some in favour of B. These differences make it hard to compare the tickets in terms of which would be the better purchase. Yet, you may come up with the somewhat minimalistic judgement that no pair is worse than the other. Consequently, you flip a coin which lands in favour of B. Instead of immediately buying them over the internet, you take a short walk to the counter. When arriving there, you are immediately informed that the A-tickets have just been sold out. However, the booking clerk offers you some A^--tickets he had overlooked before. Thus, instead of tickets A and B, you are offered tickets A^- and B, A^- referring to seats equal to A in all respects, but the row 8.

As it turns out, you made your decision on mistaken grounds. If the difficulty in comparing seats stems from the difference in the intensity of properties, then you would have to repeat your former judgement: Neither is A^- worse than B, nor B worse than A^-. According to the argumentation above, you have reason to flip the coin again. Let us assume the coin flip tells you to select A^-.[2]

Each decision, the first to go for B, if offered A and B, and the second to pick A^- if offered A^- and B, is rational per se. In an appropriate chain of offers, however, they cannot both be rational since they have resulted in an option worse than necessary.

This form of irrationality calls to mind the money-pump argument. Yet, in contrast to its famous cousin, no form of cycle is involved. Turning now to another feature: For some years, there has been an on-going debate as to which evaluative relations were involved in comparison. In particular, Chang[3] doubted whether the traditional list consisting of betterness, worseness and equality were complete. She argued that there is conceptual room for additional relations like parity, the existence of which she supports using not only examples but also several arguments which cannot be discussed here.

What is relevant in the following discussion is that in order to apply parity, it must be shown that there are options which can be compared in some respect and that none of the traditional relations adequately expresses that relation.

---

2 If you think flipping the coin again is pointless, since one has already made a decision on A^- and B-like seats, then think about how much less pointless you would consider adopting the random process again in case you had been offered a seat A^+ in the 6th row. Thus, it is not convincing to argue that flipping the coin again would be redundant. For those who deem the flipping of the coin again a mistake because its superiority over A^- can be concluded by transitivity, this argument is discussed and challenged in section 5.1.

3 Cf. Chang(1997, 2002).

This is illustrated by our opera example: Certainly, both tickets can be compared in terms of what it would be like to sit in the places they refer to. However, it seems possible that A and B, and B and A^-, respectively, have such different properties that they are not traditionally comparable. In this case, all of the following judgements would be false:

- A (A^-) is better than B with respect of where to sit
- A (A^-) is worse than B with respect of where to sit
- A (A^-) and B are equal with respect of where to sit

It is precisely because A and B, and A^- and B are traditionally incomparable that the problem of irrational sequences of decisions occurs without any cycle involved. The only positive traditional relation at work is betterness of A compared to A^- and, with only one positive relation, no cycle can be constructed.
Now, according to Chang, two traditionally incomparable options may still be comparable in a new sense: parity. This idea will be addressed here, although the discussion will be different from Chang's. In the following, "parity" will be taken to connote "being in the same league without being equal", a statement whose meaning must be described in detail.

Every material example of incomparable options or options on a par can be criticized.[4] However, it should be clear that, technically, it is easy to define traditionally incomplete relations,[5] and it would be surprising if in our colloquial speech there were no such relations. Yet, if there are any, then there may also be further evaluative relations to fill (part of) this space of traditional incomparability.[6]

The challenge would then be to show how the existence of parity can be revealed in everyday decision-making and what task it fulfils.
It is rather unlikely for an unorthodox evaluative relation like parity to be empirically directly verifiable. Therefore, the possibility of revealing parity in an indirect way will be discussed.

The intended context of this paper is the one set by the irrational sequences of singularly rational decisions, introduced at the beginning. The thesis is that parity constitutes a solution to that problem. Parity can be understood as a relation

4 Typically, the examples are attacked by arguing that what seem to be cases of incomparability are in fact cases of vague traditional comparability. Cf. Broome (1997), and Espinoza (2008).

5 The most prominent example is the relation of pareto betterness as used in economics: one allocation of goods to individuals is better than another if and only if it makes at least one individual better off without making any individual worse off.

6 To give a concrete example: In Brun/Hirsch Hadorn (2008) it is suggested that policy options could be on a par with respect to sustainable development. Brun/Hirsch Hadorn favour parity as introduced in Rabinowicz (2008). Neither Rabinowicz' nor Chang's parity is identical with the one presented in the following.

whose function is to credit or discredit random processes as adequate during decision making involving traditionally incomparable options. Parity will contrast a more traditional solution involving transitive closure.

2. Revealing rational preferences

Let X be a finite set of independent options for action (options) and 2^{X} the power set of X. In economics, social choice or decision theory, a choice function is a mapping $\mathrm{c}: 2^{\mathrm{X}} \supseteq \mathrm{B} \longrightarrow 2^{\mathrm{X}}, \ \emptyset \neq \mathrm{S} \mapsto \mathrm{T} \subseteq \mathrm{S}$. That is to say, it is a mapping which associates sets of options with subsets of themselves – it chooses some options out of a collection, so to speak.

In the following, we let each choice function $c(\cdot)$ be defined on all subsets of X, thus $B = 2^X \backslash \emptyset$.[7] Moreover, each choice function shall select at least one option from any set, thus $c(S) \neq \emptyset$, for all $S \in B$.

Hence, a choice function $c(\cdot)$ in general selects at least one option but not in general a single option if applied to a set of options. We consider each such choice function a solution to a *choice problem*. In contrast, a *decision problem* is a problem of how to select exactly one option out of a plurality of options.

Choice functions are typically categorized as being rational and irrational. In the comparativism's paradigm of optimization, a choice function $c(\cdot)$ is rational if and only if the choices correspond to the optimal elements of a binary relation on X. That is, $x \in S$ if and only if there is a binary relation $\succsim$ on X such that $x \in \max(S, \succsim)$. What is meant by "$\max(S, \succsim)$" depends on the version of optimization employed. Typically, the elements of $\max(S, \succsim)$ are either maxima or maximal elements of X, x being an $\succsim$-maximum if and only if $\forall y \in S[x \succsim y]$, whereas x being $\succsim$-maximal if and only if $\neg \exists y \in S[y \succ x]$.[8] The strict part $\succ$ of $\succsim$ is defined: $\forall x, y[x \succ y \ \equiv_{def} \ x \succsim y \wedge \neg(y \succsim x)]$. The symmetric part $\sim$ of $\succsim$ is defined: $\forall x, y[x \sim y \equiv_{def} \ x \succsim y \wedge y \succsim x]$.

Furthermore, rationality can be weaker or stronger. The quality is raised by asking for specific properties in the binary relation $\succsim$.

The weakest property typically asked of $\succsim$ is reflexivity. Requiring reflexivity is a conceptual necessity in order to interpret $\succsim$ as relation of *weak preference* which is exactly what $\succsim$ is to represent in these theories: "$x \succsim y$" is assumed to express that x is considered at least as attractive as y. Weak preference is distinguished from strong preference ("x is more attractive than y") and from indifference ("x is as attractive as y").

7 When giving examples, the trivial cases are mostly omitted.

8 Note that being a maximum implies being maximal.

Whatever a preference theory precisely consists of, the following seem to be conceptual truths:

(1) Weak preference is reflexive
(2) Strong preference is asymmetric
(3) Indifference is reflexive and symmetric
(4) Strong preference and indifference are disjoint

Now, if $\succsim$ is asked to be reflexive, thus if (1) is true, then this implies the truth of (2) to (4) for the strict part $\succ$ of $\succsim$, interpreted as strong preference, and the symmetric part $\sim$ of $\succsim$, interpreted as indifference.[9]
A much stronger condition is imposed on a rationalizing relation $\succsim$ by microeconomics, which deals with the choice of consumption bundles given certain prices and certain wealth. This discipline demands that $\succsim$ be reflexive, transitive and complete. In other words, $\succsim$ is assumed to be a *weak ordering*.
There are methodological reasons for demanding completeness and transitivity. Economists want to use the mighty apparatus of analysis and statistics. Therefore, preferences need to be in a shape such that for each preference relation $\succsim$ there is a function $f: X \longrightarrow \mathbb{R}$ with $x \succsim y$ if and only if $f(x) \succsim f(y)$. As long as X is finite or at most countably infinite, completeness and transitivity are sufficient for the existence of such a function $f(\cdot)$.[10] If X is uncountably infinite, $\succsim$ moreover needs to be continuous.[11]
Following is example 1 of a choice function which is so called weak-order-rational: Let $X = \{a, b, c\}$ and let $c(\{a, b, c\}) = \{a\}, c(\{a, b\}) = \{a\}$, $c(\{a, c\}) = \{a\}$, and $c(\{b, c\}) = \{b, c\}$. This choice function is weak-order-rational because the image of $c(\cdot)$ could be found by optimizing on the weak ordering

$$\succsim = \{(a,a),(b,b),(c,c),(a,c),(a,b),(b,c),(c,b)\}$$

That is, an option $x \in S$ is chosen by $c(\cdot)$ if and only if x is a maximum in that part of $\succsim$ which consists of the options in S.
The conditions for a choice behaviour expressed by a choice function to be more or less rational have been systematically axiomatized in the second half of the twentieth century. Our case, in which $c(\cdot)$ is rationalized by weak orderings and in which $B = 2^X \backslash \emptyset$, and $c(S) \neq \emptyset$ has been treated quite exhaustively in Sen (1971). In other fields, the search for new axioms continues.

9 Cf. Hansson (2001, p. 322)
10 In measurement theory such statements are called “representation theorems”. Representation theorems provide information on the axioms a given relational structure must satisfy for a homomorphism into a certain numerical relational structure to exist. Cf. Krantz et al. (1971, p. 9).
11 Cf. Mas-Colell/Whinston/Green (1995, p. 47).

The concept of formulating the axioms in terms of revealed preferences originates in the famous Samuelson (1938) paper: An option x counts as weakly preferred to an option y (xRy) if and only if x has been chosen on an occasion where y was also available. Formally,

$$\forall x, y \in X[xRy \equiv_{def} \exists S \in B[x \in c(S) \wedge y \in S]$$

The strict part of R, P, is the relation of revealed strong preference.
Here is a well-known axiom by Richter (1971, p. 33) which can directly be formulated using R:

$$\forall x \in X \forall S \in B[x \in S \wedge \forall y \in S[xRy] \supset x \in c(S)]$$

If an option must be chosen in a context and if there is an option which has been weakly preferred to all options elsewhere, then this option must be chosen in our context, too. Now, any choice function which meets this axiom is at least unqualifiedly rational.[12] That is, there is at least a binary relation $\succsim$ such that $c(S) \equiv \max(S, \succsim)$, for all contexts S — for instance the relation R.

Let us return to the fact that a choice function does not in general yield a single option in each situation $S \in B$ but a subset $T \subseteq S$ of likewise acceptable options. Assume that our choice behaviour is not rationalized by a weak order but by an even stronger *linear ordering*, a linear ordering being an antisymmetric weak order. Then, all *c(S)* are singletons. That is, in this special case the solution to a choice problem is also the solution to a decision problem. If decisions are rationalized against the background of linear orderings, the problem of which option to select among a set of likewise acceptable options disappears.

Let a decision function be a mapping $d: 2^X \supseteq B \longrightarrow X, \ \emptyset \neq S \mapsto x \in S$. Then, obviously for each choice function $c(\cdot)$ which has been rationalized against the background of a linear ordering, there is a choice function $d(\cdot)$ such that $\{d(S)\} = c(S)$, for all $S \in B$.

3. Rationalizing decision functions based on choice functions

The unpleasant example involving the opera tickets suggests the need to find criteria for the rationality not only of choice but also of decision functions. Most importantly, these criteria should exclude the possibility that a series of rational decisions renders an irrational outcome.

A first condition should govern the relation between choice and decision functions. Certainly, no rational decision can be expected which is not based on a rational choice. In fact, we act irrationally if we decide on an option $x \in S$ such that there is no rational choice function $c(\cdot)$ with $x \in c(S)$. A choice can be

12 Cf. Richter (1971, p. 33).

understood as a mechanism which restricts decisions. If a choice rationally excludes an option, we cannot expect a rational decision in favour of this option. The first condition necessary for rational decision functions $d(\cdot)$ is thus:

(I) $d(S) \in c(S)$, for a rational choice function $c(\cdot)$

A second condition must deal with sequences of decisions, as in the opera example. Assume there to be a sequence $d(S_i)$, $S_i \in B$, $1 \leq i \leq n \in \mathbb{N}$, of decisions to be taken. Assume further that we must concede each option $d(S_i)$ in favour of option $d(S_{i+1})$. The goal is to avoid selecting an option that is worse than an option we could have selected. As in the opera example, it would be irrational to adopt a decision rule that allows us to end with an option worse than we could have ended.

The second necessary condition for rational decision functions d(·) is thus:

(II) If $d(S_i) \in S_{i+1}$ and $d(S_1) = x_1 \wedge d(S_2) = x_2 \wedge \cdots \wedge d(S_n) = x_n$, then $\forall x \in S_1[x_n \in c(\{x, x_n\})]$

Thus, the final option must be at least as good as or certainly not worse than any that was available at any previous point.

It is perfectly intuitive and could easily be shown that if a decision function is being defined against the background of a choice function which is rational in terms of a linear ordering, then this decision function not only meets condition I but also condition II. As argued before, there is a choice function which quasi coincides with the decision function.

However, quite often we must deal with situations in which the choice function is not rational against the background of a linear ordering but – at best – with respect to some weaker structure. As in the case of (non-antisymmetric) weak orders, several decision functions might meet condition I.

Consulting example 1, which was introduced above, there are exactly two decision functions which meet condition I:

$$d_1(S) = \begin{cases} x, \text{ if } c(S) = \{x\} \\ b, \text{ otherwise} \end{cases}$$

$$d_2(S) = \begin{cases} x, \text{ if } c(S) = \{x\} \\ c, \text{ otherwise} \end{cases}$$

Moreover, both decision functions not only meet condition I but also condition II. This is a highly convenient property because, if all decision functions which meet condition I also meet condition II, then we can select whatever $x \in S$, whenever the upstream choice function $c(\cdot)$ does not yield a singular option if applied to S. In other words, we could flip a coin, roll a die or adopt some other random process to define $d(\cdot)$ based on $c(\cdot)$ without ever running the risk of acting irrationally when following this decision function.

The result gleaned from this example can easily be generalized: If a choice function is weak-order-rational, then each decision function which meets condition I also meets condition II. The only cases not already decided by $c(\cdot)$ are cases of equal options. These cases can be decided randomly.

4. Random processes do not generally turn rational choices into rational decisions

Problems arise from structures which are weaker than weak orderings in terms of completeness. Thus, let us look at choice functions $c(\cdot)$ which are rational against the background of a reflexive, transitive but not necessarily complete preference relation $\succsim$, a so called *preorder*.

Following is example 2 of such a choice function $c(\cdot)$: Let $X = \{a, b, c\}$ und let

$$c(S) = \begin{cases} \{a, c\}, \text{ if } S = \{a, b, c\} \\ \{a\}, \text{ if } S = \{a, b\} \\ \{a, c\}, \text{ if } S = \{a, c\} \\ \{b, c\}, \text{ if } S = \{b, c\} \end{cases}$$

Now, $c(\cdot)$ is rational, in fact, it is even preorder-rational as it could have been induced by optimizing[13] the preorder $\succsim = \{(a, a), (b, b), (c, c), (a, b)\}$.[14]

Among the decision functions which meet condition I, the following can be found:

$$d_3(S) = \begin{cases} x, \text{ if } c(S) = \{x\} \\ c, \text{ if } c(S) = \{a, c\} \\ b, \text{ otherwise} \end{cases}$$

Thus, we get $d(\{a, c\}) = c, d(\{b, c\}) = b$, but $b \notin c(\{a, b\}) = \{a\}$, and therefore $d_3(\cdot)$ does not meet condition II. That is, the choice to follow $d_3(\cdot)$ might leave us with an irrational outcome.

This example illustrates that we cannot in general opt for an arbitrary decision function which meets condition I with respect to a rational choice function. As

13 In terms of finding maximal elements.

14 $c(\cdot)$ is also complete-rational and yet not weak-order-rational. Whereas the above listed relation $\succsim$ makes it preorder-rational, it is the revealed relation $R = \{(a, a), (b, b), (c, c), (a, c), (c, a), (b, c), (c, b), (a, b)\}$ which turns it complete rational. There is no weak order, however, such that $c(\cdot)$ can be understood as induced by optimizing over it in terms of finding maxima.

soon as the choice function's rationality is based on a relational structure as weak as preorders, an arbitrary decision function cannot be trusted.

5. Two methods for generating rational decision functions

In the following, two methods are introduced to block the adoption of an irrational decision function.

The first method is based entirely on conventional evaluative relations: betterness (strict preference, respectively), and equivalence (indifference). There are no new structural elements to constrict the number of rational decision functions based on rational choice functions. What does, indeed, restrict the number of admissible decision functions is a procedural dimension: If $d(\cdot)$ is an irrational decision function, then no decision $d(S)$ is irrational per se. Whether a decision is irrational depends on the order the decisions have been made.

The second method is not based on procedural considerations. Rather, the number of admissible decision functions is restricted by an additional structural dimension: the relation of parity. In this case, the order in which the decisions have been made is extraneous to the question of whether a specific decision is rational or irrational.

5.1. A procedural restriction of the random process

In detail, the first method still allows for random processes but with restrictions. Let us assume that an arbitrary decision has been made. Then, before the random process may start again, the transitive consequences of the first decision must be taken into consideration. Technically, what is calculated is the so called transitive closure, $tr(Q)$, of a binary relation Q on a set X, defined: $tr(Q) =_{def} \{(x,y) \in X \times X$; there is a sequence $u_0, u_1, \dots, u_n$ so that $(u_i, u_{i+1}) \in Q$ for $i = 0, 1, 2, \dots, n-1$ with $u_0 = x$ and $u_n = y\}$.[15] In our case, the transitive closure of the union of two sets is relevant: First, the preorder $\succsim$ which rendered a choice function $c(\cdot)$ preorder-rational. Second, the following relation, D, which is induced by the arbitrary decision $d(\{z_1, z_2, \dots, z_i, \dots, z_n\}) = z_i$ that has already been made: $D = \{(z_i, z_1), (z_i, z_2), \dots, (z_i, z_i), \dots, (z_i, z_n)\}$. The second choice set is then preordered not simply by $\succsim$, but by $\mathrm{tr}(\succsim \cup D)$. Using transitive closures of rationalizing preorders completed by decisions that have already been made avoids selecting options that are worse than anyone could have picked before.

15 Cf. Trotter (1992, pp. 15-16)

The ticket example illustrates how this works: Assume that $d(c\{A,B\}) = d(\{A,B\})$ is supposed to be decided upon. Since it is the first decision to be made, a random process may be applied.

Let us now assume that the die has pointed to selecting A (case 1). Then, for our next decision, the transitive closure of $\{(A,A),(A^-,A^-),(B,B),(A,A^-)\} \cup \{(A,A),(A,B)\}$ is relevant. Since there are no transitive consequences in this case and since the transitive closure is thus $\{(A,A),(A^-,A^-),(B,B),(A,A^-),(A,B)\}$, we can throw the die again to decide on $d(c(\{A^-,B\})) = d(\{A^-,B\})$. However, if the die indicates B (case 2) instead of A as in case 1, then $d(\{A^-,B\})$ may not be decided by using a random process. The reason is that the transitive closure of

$$\{(A,A),(A^-,A^-),(B,B),(A,A^-)\} \cup \{(A,A),(B,A)\}$$

is

$$\{(A,A),(A^-,A^-),(B,B),(A,A^-),(B,A),(B,A^-)\}$$

Thus, we must select B in this case.

Assume that we are in a situation in which an irrational decision function has been produced. Then, if transitive closure is relevant for rationality, none of the set of produced decisions is irrational per se. Whether a decision is irrational depends on the order in which the decisions have been made.[16]
For instance, if we have $d(c(\{A,B\})) = B$ and $d(c(\{A^-,B\})) = A^-$, then each decision is irrational provided the other has been made before.
It can be formally proved that there is always a rational decision function for choice functions which have been rationalized against the background of a preorder.[17]

How convincing is transitive closure in examples such as the ticket-example, introduced at the beginning?—The problem with this conventional solution is that it cannot adequately be applied to these cases.
In order to be adequately applied, the options must be related by *one and the same* relation, preference, for instance. In our example, however, there are two relations at work. The first is the rationalizing preorder, $\succsim$, which covers the choices made. The second is the relation D generated by the random procedure.
It remains nebulous how a preference for A over A^- and a random decision for B over A can have transitive consequences for the relation between A^- and B. It is not convincing to assume that transitivity is crucial for explaining why B should be selected, when it comes to deciding on A^- or B.

16 This is a substantial claim. It has been formally proved in poset-theory, the theory of anti-symmetric preorders. Cf. Trotter (1992, p. 16). This proof can easily be generalized for preorders. Think of equivalent options as gathered in equivalence classes and represented by an arbitrary option out of any class.

17 This is implied in Szpilrajn's theorem. Cf. Trotter (1992, p. 17).

Besides this technical point, intuitively, transitivity does not seem to be what you are considering when blocking the idea that another coin should be tossed. The reason for simply deciding for B over A^- stems rather from a different thought which is discussed in the following section.

5.2. A structural restriction of the random process

The second method is based on Chang's idea that options could be comparable in a non-traditional way. There could be a value relation at work, different from betterness and equality, which does the job. Chang calls this relation parity.
Before parity, or better *revealed* parity, can be defined, some preliminary work must be done. In particular, we must demonstrate how two options can be revealed incomparable. The traditional theory of revealed preference does not account for this possibility.

As discovered above, the strict part, P, of a revealed weak preference, R, is interpreted as revealed strong preference. Accordingly, the orthodox line of the theory does interpret the symmetric part of R, named "I", as revealed indifference. Now, there are different ways to show that such an interpretation may at times be inadequate.
At an abstract level, this inadequacy can be demonstrated by pointing out that a choice function $c(\cdot)$ which is defined for all subsets of $2^X \backslash \emptyset$ and which always chooses at least one element cannot reveal incomparability among options. This results from the fact that as $c(\cdot)$ is defined for all subsets of $2^X \backslash \emptyset$ it is particularly defined for all sets $\{x, y\}$ of pairs $x, y \in X$. Consequently, for any two options $xRy \vee yRx$ applies. Thus, no two options can be incomparable.

Such an interpretation is inadequate because the circumstances of a choice situation determine whether a chooser can express incomparability between options. If the circumstances require that a choice be made for all subsets of $2^X \backslash \emptyset$, or even only for all pairs of options, then incomparability cannot be expressed. However, it should be clear that it would be more appropriate if this were possible.

On a more specific level, there are examples of options having been chosen not based on indifference but rather incomparability. The following example is adopted from Eliaz/Ok (2006, pp. 62-63):
Mrs. Watson intends to rent a movie from a video store for her two children, Alice and Tom. Since renting is costly and the kids have no time to watch more than one movie, she wants to rent exactly one movie. Among all films, $\{a, b, c\}$ is the set comprising those appropriate for children. In a quick conference call, Alice and Tom have expressed the following preferences to their mother:

- Alice: $b \succ c \succ a, b \succ a$
- Tom: $a \succ b \succ c, a \succ c$

Assuming all films are in stock, Mrs. Watson chooses $c(\{a,b,c\}) = \{a,b\}$ and plans to flip a coin in order to determine which film to carry home. This choice seems reasonable, since option c is inferior to b according to the preferences of both children, whereas such a judgement is not possible for a and b.
However, while talking to an employee, Mrs. Watson discovers that b is not available. Thus, the choice situation has changed from $\{a,b,c\}$ to $\{a,c\}$ and the relevant part of Alice' and Tom's preferences is

• Alice: $c \succ a$
• Tom: $a \succ c$

Mrs. Watson reasonably chooses $c(\{a,c\}) = \{a,c\}$ and, again, prepares to flip a coin to make the decision. Altogether, Mrs. Watson seems to have revealed the following relation of weak preference:

$$R = \{(a,a),(b,b),(a,b),(b,a),(a,c),(c,c),(b,c),(c,a)\}$$

As traditionally defined, this results in a revealed strong preference $P = \{(b,c)\}$ and a revealed indifference

$$I = \{(a,a),(b,b),(a,b),(b,a),(a,c),(c,c),(b,c),(c,a)\}$$

Since I is intransitive, according to the traditional point of view, Mrs. Watson has chosen irrationally.

Now, Mrs. Watson would probably protest against the assertion that she has intransitive preferences and that she is irrational. And indeed, there is a much more plausible way to interpret her behaviour than the one offered by the traditional approach. More plausibly, she believes that a and b, and a and c are incomparable rather than equal. Her point would be that the preferences of Tom and Alice, which are converse in a non-symmetric way, did not allow a positive judgement on the preferability of a over b (or conversly), and a and c (or conversly).

Such examples suggest that the revealed preference theory be changed such that there can be both indifference and incomparability. This has been done by Eliaz/Ok (2006, pp. 66-67). They separate incomparability from indifference in the following way:

Let "$S_{y,-x}$" be defined for all sets S that comprise x but not y and let the terminus refer to the set $(S \cup \{y\})\backslash\{x\}$. Then two options $x,y \in X$ such that $c(\{x,y\}) = \{x,y\}$, $c(\cdot)$ being a rational choice function, are revealed as incomparable $I_{\|}$ if there is a situation $S \in B$ such that $x \in S \wedge y \notin S$ and

• $x \in c(S) \wedge y \notin c(S_{y,-x})$ or

• $x \notin c(S) \wedge y \in c(S_{y,-x})$ or

• $(c(S)\backslash\{x\}) \neq (c(S_{y,-x})\backslash\{y\})$

In words, two options which so far where considered indifferent are rendered incomparable by a certain asymmetry they show in comparison with third items.
As a consequence of this partition of I, "R" must be given a different meaning. If "$xRy \wedge yRx$" no longer means revealed indifference, then "xRy" can no longer mean weak revealed preference. Eliaz and Ok offer to interpret "xRy" as "x is revealed not worse than y".

This interpretation tears a hole in the space X^2 of all pairs of options. Unlike the traditional understanding, according to which X^2 is revealed as completely ordered as soon as certain circumstances are given, certain pairs can be revealed incomparable, or rather traditionally incomparable.

They are traditionally incomparable because they are incomparable insofar as none of the traditional evaluative relations P and I can be revealed. However, according to Chang's idea, this need not mean that there could not be an additional value relation to fill that gap. She suggested that this gap might be filled by parity. In the following, we will exemplify parity as *being in the same league*.

Since equivalent options do not cause any difficulty, as argued before, they are ignored in the following. The argumentation will be based on antisymmetric relations.

In order to fill the space of traditionally incomparable options, we recursively define sets of revealed leagues. The highest revealed league of a set X that is ordered by R precisely consists of the R-maximal elements of X: $\max(X,R) = \{x \in X;\ \neg\exists y \in X[yPx]\}$. The second highest league comprises those elements which become highest after the highest league has been removed. In general: Let $X_1 = X$ and $X_{i+1} = X_i \backslash \max(X_i, R)$. Then, $\max(X_i, R)$ is the i^{th} revealed league.

Revealed parity, in turn, can be defined by membership to one and the same league. Two options $x, y \in X$ which are revealed as traditionally incomparable, $xI_{\|}y$, are revealed in the same league, $xI_{\simeq}y$, if and only if $\exists i \in \mathbb{N}[x, y \in \max(X_i, R)]$.

The relation of revealed parity, $xI_{\simeq}y$, is irreflexiv, symmetric, and distinctly transitive.[18] Harary's (1961) suggestion to call relations "parity relations" which are irreflexive, symmetric and distinctly transitive seems to have escaped the philosophical circles dealing with parity.

Parity interpreted league-wise, as presented here, differs in at least one formal and one material point from the parity Chang had in mind. The formal difference concerns the fact that the relation presented here is transitive, whereas Chang's is not. Materially, the two approaches differ in that, whereas Chang claims pari-

18 The term "distinctly transitive" is by Harary (1961). A binary relation Q on a set X is distinctly transitive if and only if
$\forall x, y, z \in X[(x \neq y \wedge y \neq z \wedge x \neq z \wedge xQy \wedge yQz \supset xQz)]$.

ty to be the only non-traditional relation, no such exclusivity is claimed here. On the contrary, further structures impose themselves.
Observe that parity does partition the set X in a way similar to that of an equivalence relation.[19] More precisely, there is a set $\{Y_1, \dots, Y_n\}$ of subsets $Y_i \subseteq X$ such that

(1) $\bigcup_i^n Y_i = X$
(2) $Y_i \cap Y_j = \emptyset$, if $i \neq j$
(3) $\forall x, y \in X[xI_{\succeq}y \equiv \exists i \in \{1, \dots, n\}[x, y \in Y_i]]$

The revealed leagues, $\max(X_i, R)$, build such a partition.

Now, we can order all options according to their league-membership. In order to do that formally, let $\mathrm{l}(x)$ be the number of the league to which x belongs. Order by league-membership then results in:

$$\forall x, y \in X[xI_p y \equiv_{def} xI_{\|}y \wedge \mathrm{l}(x) < \mathrm{l}(y)]$$

Using parity and league-wise superiority, we can make sure that condition II is met whenever condition I is met. In order to do so, the random procedure should be restricted to options among which there is parity in the highest league:

$$d(S) \in \{x \in c(S); \forall y \in c(S)[x \neq y \supset xI_{\succeq}y \vee xI_p y]\}$$

Decision functions can still be established on the basis of choice functions by applying a random procedure. However, the applicability is restricted by the additional relation of parity.

Now, how convincing is parity in examples such as the ticket-example?—According to the idea of parity, what deters you from flipping another coin is simply that A^- is an inferior option compared to an option you could also have selected. The same is not true for B, which is what makes B the better selection. This seems to be a more realistic assessment of the thought-process involved when an individual is confronted with A^- and B at the ticket counter. What seems relevant, rather than transitive closure, is that A^- is lower ranked than B with respect to those options, A^- and B, respectively, can be ranked. And this is what parity comes down to.

6. Conclusions

This paper has combined two ostensibly different problems: 1) the problem of how to deal with chains of rational decisions which lead to an irrational outcome, and 2) the problem of making sense of an additional evaluative relation of parity. As has been shown, parity can be defined as a structure which is able to

19 Cf. Harary (1961).

block an irrational outcome. This solution has been contrasted with a more orthodox solution involving transitive closure. It has been suggested that the parity solution lacks the conceptual weaknesses the orthodox solution shows and is more plausibly applied.

References

Broome, John: "Is Incommensurability Vagueness?" In: *Chang, Ruth* (ed.): Incommensurability, Incomparability, and Practical Reason. Harvard University Press, Cambridge (MA), 1997. S. 67–89

Brun, Georg/Hirsch Hadorn, Gertrude: "Ranking policy options for sustainable development". Poiesis & Praxis, Nr. 1., 5, 2008. S. 15–31

Carlson, Erik: "Incomparability and Measurement of Value". In: *McDaniel,Kris et al.* (eds.): The Good, the Right, Life and Death. Ashgate, Aldershot, 2006. S. 19–43

Chang, Ruth: "Introduction". In: *Chang, Ruth* (ed.): Incommensurability, Incomparability, and Practical Reason. Harvard University Press, Cambridge (MA), 1997. S. 1–34

Chang, Ruth: "The Possibility of Parity". Ethics, 112, 2002. S. 659–688

De Sousa, Ronald B.: "The Good and the True". Mind, 83, 197. S. 534–551

Eliaz, Kfir/Ok, Efe A.: "Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences". Game Econ Behav, 56, 2006. S. 61–86

Espinoza, Nicolas: "The small improvement argument". Synthese, 165, 2008. S. 127–139

Griffin, James: Well-Being: Its meaning, measurement, and moral importance. Clarendon Press, Oxford, 1986

Hansson, Sven Ove: "Preference Logic". In: *Gabbay, D.M./Guenthner, F.* (eds.): Handbook of Philosophical Logic Volume 4. Kluwer, 2001. S. 319–393

Harary, Frank: "A Parity Relation Partitions Its Field Distinctly". Am Math Mon, 68 Mar., Nr. 3, 1961. S. 215–217

Krantz, David H. et al.: Foundations of Measurement. Volume I: Additive and Polynomial Representations. Academic Press, New York, 1971

Mas-Colell, Andreu/Whinston, Michael D./Green, Jerry: Microeconomic Theory. Oxford University Press, New York, 1995

Parfit, Derek: Reasons and persons. Oxford University Press, New York, 1986

Rabinowicz, Wlodek: "Value Relations". Theoria, 74, 2008. S. 18–49
Richter, Marcel K.: "Rational Choice". In: *Chipman, John S. et al.* (eds.): Preferences, Utility, and Demand. Harcourt Brace Jovanovich, New York, 1971. S. 29–58

Samuelson, P.A.: "A Note on the Pure Theory of Consumer's Behaviour". Economica, 5, Nr. 17, 1938. S. 66–71
Sen, Amartya: "Choice Functions and Revealed Preferences". Rev Econ Stud, 38, Nr. 3, 1971

Verantwortung und Anfechtbarkeit. Eine Analyse der Struktur von Verantwortungszuschreibungen

Claudia Blöser und Claudia Cuadra
claudiabloeser@googlemail.com; claudia@cuadra.de
Goethe-Universität Frankfurt am Main

Abstract/Zusammenfassung

In his book *Responsibility and the Moral Sentiments* (1994) , R.J. Wallace gives an account of what it is to be a morally responsible agent. According to Wallace, holding somebody responsible implies that we are subject to emotions like resentment, indignation and guilt, or consider those emotions as appropriate („appropriateness-condition"). In this paper, we want to offer an alternative interpretation of our practice of holding each other responsible.

The first problem with Wallace' account is that Wallace himself undermines the central role of the emotions with his appropriateness-condition. The core of an ascription of responsibility seems to be not an emotion, but the *judgment* that a certain emotion would be appropriate.

The second problem concerns Wallace' emphasis of *negative* ascriptions of responsibility, i.e. blame. Wallace' assumption that negative emotions play a more important role in our practice than positive emotions and their expression, e.g. praise, can be disputed.

The third criticism points to the connection between moral and legal responsibility. The legal context constitutes an important – if not paradigmatic – practice of ascribing responsibility. An account that assigns a central role to emotions seems to exclude a similar underlying structure of ascriptions of responsibility in moral and legal contexts.

All three problems can be solved if ascriptions of responsibility (V) are understood as *defeasible judgments* in the following way:
(V) We ascribe responsibility to an adult person for her action, unless there are exempting or excusing reasons.
This general form covers two aspects or instances of ascriptions of responsibility:
(V1) A person is responsible in the sense of *accountable*, unless there is an exempting reason.
(V2) A person is responsible in the sense of *praise- or blameworthy* for action X, if the person is accountable, if X is evaluated according to a norm, and unless there is an excusing reason.

In seinem Buch *Responsibility and the Moral Sentiments* (1994) untersucht R.J. Wallace, was es heißt, eine verantwortliche Person zu sein. Unsere Praxis der Verantwortungszuschreibung sei wesentlich dadurch charakterisiert, so Wallace, dass wir verantwortlichen Personen gegenüber reaktive Emotionen wie Ärger, Entrüstung und Schuld empfinden oder diese für angemessen halten („Angemessenheitsbedingung"). Wir möchten hier für eine von Wallace abweichende Interpretation unserer Praxis der Verantwortungszuschreibung argumentieren.

Unser erster Kritikpunkt an Wallace ist, dass Wallace selbst die zentrale Stellung der Emotionen durch seine Angemessenheitsbedingung untergräbt. Was letztlich für eine Verantwortungszuschreibung zählt, sind nicht Gefühle, sondern ein *Urteil*, das die Angemessenheit der Emotionen begründet.

Der zweite Kritikpunkt betrifft Wallace' Betonung der *negativen* Verantwortungszuschreibungen, d.h. Tadel. Wallace' Annahme, dass negative Emotionen in unserer alltäglichen Praxis der Verantwortungszuschreibung eine zentralere Rolle spielen als positive und deren Ausdrücke, z.B. Lob, kann bestritten werden.

Der dritte Kritikpunkt betont den Zusammenhang zwischen moralischer und rechtlicher Verantwortung. Der gerichtliche Kontext stellt eine wichtige – wenn nicht sogar paradigmatische - Praxis der Verantwortungszuschreibung dar. Ein Ansatz, der Gefühle in den Mittelpunkt stellt, läuft Gefahr, eine ähnliche zugrundeliegende Struktur zwischen persönlichen Verantwortungszuschreibungen im Alltag und solchen, die im unpersönlichen gerichtlichen Kontext stattfinden, von vorneherein auszuschließen.

Alle drei Probleme werden gelöst, wenn Verantwortungszuschreibungen (V) als *anfechtbare Urteile* folgendermaßen verstanden werden:

(V) Wir schreiben einer erwachsenen Person Verantwortung für ihre Handlung zu, es sei denn, es liegen Entschuldigungs- oder Ausnahmegründe vor.

Diese allgemeine Form umfasst zwei Stufen der Verantwortungszuschreibung:

(V1) Eine Person P gilt als verantwortlich im Sinne von zurechnungsfähig (*accountable*), es sei denn es liegt ein Ausnahmegrund vor.

(V2) Eine Person P gilt als verantwortlich (d.h. lobens- oder tadelnswert) für Handlung X, wenn P als zurechnungsfähig gilt und wenn X in Bezug auf eine Norm bewertet wird, es sei denn es liegt ein Entschuldigungsgrund vor.

Thema unseres Vortrags ist die Analyse von moralischer und rechtlicher Verantwortung auf Basis der bestehenden Praxis der Verantwortungszuschreibung. In der Nachfolge von Peter Strawsons einflussreichem Aufsatz „Freedom and Resentment" von 1962 hat es viele Versuche gegeben, Theorien der Verantwortung in unserer Praxis zu verankern. Laut Strawson ist unsere Praxis der Verantwortungszuschreibung durch die so genannten reaktiven Gefühle wie etwa Dankbarkeit, Bedauern und Entrüstung gekennzeichnet. Einer Person Verantwortung zuzuschreiben heißt demnach, einer anderen Person gegenüber reaktive Einstellungen einzunehmen. Einer von Strawsons namhaften Nachfolgern ist R. Jay Wallace, dessen in *Responsibility and the Moral Sentiments* (1994) entwickelte Verantwortungstheorie als Ausgangspunkt unserer Überlegungen dienen wird.

Ähnlich wie bei Strawson und Wallace versteht sich unsere Untersuchung als eine Auslegung unserer Praxis der Verantwortungszuschreibung. Jedoch lehnen wir die Idee ab, dass reaktive Emotionen unserer Zuschreibungspraxis wesentlich sind. Wie wir im ersten Teil darlegen, ist Wallace' reaktive Verantwortungskonzeption drei Schwierigkeiten ausgesetzt. Unser Vorschlag ist, dass sich die drei Probleme lösen lassen, wenn man Verantwortung im Sinne einer Default-and-Challenge Struktur analysiert. Im zweiten Teil nehmen wir anhand des Ansatzes von H.L.A. Hart eine allgemeine Bestimmung anfechtbarer Begriffe vor und übertragen diese sodann auf den Begriff der Verantwortung. Abschließend legen wir dar, inwiefern unser Vorschlag eine Lösung der drei Schwierigkeiten erlaubt und darüber hinaus eine Bestimmung von Verantwor-

tung liefert, die sowohl moralische als auch rechtliche Verantwortung zu erfassen vermag.

1. Wallace' reaktive Zuschreibungstheorie moralischer Verantwortung

Wallace' zentrale These ist, dass Verantwortungszuschreibungen in reaktiven Emotionen verankert sind. Anders als Strawson fasst er darunter nicht die Gesamtheit an Emotionen, die in interpersonalen Beziehungen vorkommen können, sondern argumentiert für eine engere Auslegung. Die Einschränkung der reaktiven Emotionen auf Ärger, Entrüstung und Schuldgefühle (*resentment, indignation, guilt*) erlaubt es Wallace, das gemeinsame Element zu bestimmen, das die moralischen reaktiven Einstellungen zu einer besonderen Klasse von Emotionen macht. Dies gibt ihm die Mittel in die Hand, eine adäquate Grenzlinie zwischen moralischen reaktiven Emotionen und anderen Arten von Emotionen zu ziehen. Dieses gemeinsame Element liegt laut Wallace in der besonderen Art und Weise, wie reaktive Einstellungen mit moralischen Verpflichtungen verbunden sind. Demnach sind moralische reaktive Einstellungen durch die Überzeugung charakterisiert, dass eine moralische Verpflichtung verletzt wurde.[1]

Der Zusammenhang zwischen reaktiven Einstellungen und Verantwortungszuschreibungen lässt sich nun folgendermaßen fassen: Jemandem moralische Verantwortung zuzuschreiben impliziert, eine Person moralisch zu verpflichten, d.h. ihr gegenüber moralische Erwartungen zu haben.[2] Wenn aber Verantwortungszuschreibungen die Bindung an moralische Verpflichtungen einschließen, und die Bindung an moralische Verpflichtungen wiederum an reaktive Emotionen geknüpft ist, dann gilt, so Wallace' Folgerung, dass reaktive Emotionen ein wesentlicher Bestandteil von Verantwortung sind. Die zentrale Rolle der reaktiven Emotionen veranlasst Wallace, seine Theorie als „reaktive Verantwortungskonzeption" zu bezeichnen.[3]

Bevor wir auf die angekündigten drei Schwierigkeiten zu sprechen kommen, möchten wir einen Punkt hervorheben, der von entscheidender Bedeutung für unsere Kritik an Wallace' Konzeption sein wird. Wallace erkennt an, dass es möglich ist, einer Person für die Verletzung einer moralischen Verpflichtung Verantwortung zuzuschreiben, ohne faktisch irgendein besonderes Gefühl gegenüber dieser Person zu empfinden. Entscheidend ist nach Wallace allein, dass man die Empfindung bestimmter reaktiver Emotionen für *angemessen* hält. Die-

1 Vgl. Wallace 1994: Responsibility and the Moral Sentiments, S. 37

2 Vgl. ebd., S. 63

3 In den Worten Wallace' impliziert die grundlegende Einstellung, jemanden für moralisch verantwortlich zu halten, eine Empfänglichkeit für moralische Gefühle, wenn eine Person moralische Verpflichtungen verletzt (vgl. ebd., S. 66).

se Bedingung, die wir im Folgenden *Angemessenheitsbedingung* nennen, ist dann erfüllt, wenn gegen eine moralische Verpflichtung verstoßen wurde.[4]

Die erste Schwierigkeit knüpft direkt an diesen Punkt an. So scheint es, dass Wallace selbst mit der Einführung der Angemessenheitsbedingung die von ihm behauptete wesentliche Rolle der reaktiven Emotionen untergräbt. Er muss zugestehen, dass eine Verantwortungszuschreibung bereits allein durch das Urteil gerechtfertigt ist, welches die Angemessenheit der Zuschreibung begründet, dem Urteil also, „dass eine moralische Verpflichtung, die man akzeptiert, übertreten wurde, und dass keine entlastenden Umstände vorliegen, deren man sich bewusst wäre".[5] Im zweiten Teil dieser Untersuchung werden wir die Form eines solchen Urteils ausarbeiten.[6]

Zweitens möchten wir Wallace' Betonung der negativen Verantwortungszuschreibungen bzw. der Schuldzuschreibungen in Frage stellen. Wallace' Annahme, dass in unserer Zuschreibungspraxis eine Asymmetrie hinsichtlich der positiven und negativen Fälle besteht, insofern die negativen Fälle die zentrale Rolle einnehmen, erscheint unzureichend fundiert. Weil diese Charakterisierung kontrovers ist, sollte sie keinen konstitutiven Bestandteil bei der Analyse von Verantwortung bilden.

Drittens möchten wir die Aufmerksamkeit auf das Verhältnis zwischen moralischer und rechtlicher Verantwortung lenken. Auch wenn moralische und rechtliche Kontexte unterschiedliche Merkmale aufweisen, ist es plausibel anzunehmen, dass beide Bereiche eine gemeinsame Struktur hinsichtlich der Verantwortungszuschreibungen teilen. Der rechtliche Kontext stellt eine wichtige – wenn nicht paradigmatische – Praxis der Verantwortungs-zuschreibung dar. Es ist daher lohnend, Unterschiede und Gemeinsamkeiten beider Arten von Verantwortungszuschreibungen zu untersuchen. Nun aber spielen für die Zuschreibung rechtlicher Verantwortung Gefühle keine tragende Rolle. Angesichts dessen läuft eine Theorie, die Gefühlen eine tragende Rolle in der Analyse von moralischer Verantwortung zuspricht, Gefahr, von vornherein eine ähnliche zugrunde liegende Struktur moralischer und rechtlicher Verantwortung auszublenden.

4 Obwohl Wallace sich nicht eindeutig dazu äußert, erscheint es plausibel, die Angemessenheitsbedingung so zu verstehen, dass eine Norm faktisch verletzt sein muss. Nur dann ist die Zuschreibung fair und also angemessen.

5 Vgl. Wallace 1994, S. 77

6 Wallace selbst geht auf den Kritikpunkt ein, nennt allerdings keine befriedigende Lösung: Ohne Gefühle könne die „Tiefe" der Verantwortungszuschreibung nicht erklärt werden (Vgl. ebd., S. 76).

2. Anfechtbare Begriffe

Unser Vorschlag geht von der Annahme aus, dass Verantwortungszuschreibungen, kurz (V), eine anfechtbare Struktur besitzen, deren allgemeine Form sich folgendermaßen darstellen lässt:

> (V) *Wenn Default-Bedingungen erfüllt sind, schreiben wir einer Person für ihre Handlung Verantwortung zu, es sei denn, es liegen Entschuldigungs- oder Ausnahmegründe vor.*

Dieses Urteil folgt einem Default-and-Challenge Modell: Normalerweise können wir einer Person für ihre Handlungen Verantwortung zuschreiben, ohne explizite Gründe dafür anzuführen. Die Tatsache, dass normalerweise keine expliziten Gründe erforderlich sind, bedeutet, dass die Verantwortungs-zuschreibung standardmäßig, also *default*, gerechtfertigt ist. Dennoch müssen auch für den Default-Status gewisse minimale Kriterien erfüllt sein, die wir Default-Bedingungen nennen. Dies erkennt man sehr leicht daran, dass wir kleinen Kindern nicht ohne Weiteres Verantwortung zuschreiben dürfen – wohl aber erwachsenen Personen. Selbstverständlich ist diese Zuschreibung fallibel und kann durch bestimmte Entschuldigungs- und Ausnahmegründe angefochten werden. Solche Gründe, die dazu geeignet sind, eine Verantwortungszuschreibung anzufechten, bezeichnen wir im Folgenden als Anfechtungsgründe.

Die Rückführung der Idee anfechtbarer Begriffe auf ihre historischen Wurzeln bei H.L.A. Hart vermag zum einen die Struktur der Anfechtbarkeit näher zu charakterisieren und zum anderen unsere Verantwortungskonzeption in einen breiteren philosophischen Kontext einzubetten. In seinem Aufsatz „The Ascription of Responsibility and Rights“ (1948/49) führt Hart anfechtbare Begriffe erstmals als philosophischen Terminus ein.

In einem ersten Schritt argumentiert Hart, dass rechtliche Begriffe wie “Vertrag” oder “Mord“ anfechtbare Begriffe sind. Das entscheidende Merkmal dafür, dass diese Begriffe eine besondere Struktur besitzen, ist nach Hart die Unentbehrlichkeit des Zusatzes „es sei denn“: Unter bestimmten notwendigen und *normalerweise* hinreichenden Bedingungen kann der Begriff angewendet werden, *es sei denn* bestimmte Bedingungen liegen vor, die die Zuschreibung anfechten. Im Fall eines Vertrags beispielsweise sind die normalerweise hinreichenden Bedingungen, dass es mindestens zwei Parteien gibt, ein Angebot auf der einen Seite, Akzeptanz auf der anderen.[7] Diese Tatsachen sind es, die die Basis für bestimmte rechtliche Folgen bilden – in diesem Fall das Urteil, dass ein Vertrag besteht.

Nach Hart gibt es zwei Wege, die Anwendung eines anfechtbaren Begriffs anzufechten: Erstens durch Bestreiten der Tatsachen, die der Anwendung zu-

7 Vgl. Hart 1948/49: “The Ascription of Responsibility and Rights”, in Flew: Essays on Logic and Language, S. 148

grunde liegen (dem sog. *joinder of issue*). Und zweitens durch den Hinweis, dass in diesem speziellen Fall, obgleich alle für die Gültigkeit eines Urteils notwendigen Umstände vorliegen, besondere Umstände hinzukommen, die den Fall zu einer anerkannten Ausnahme machen. Die Konsequenz einer erfolgreichen Anfechtung ist entweder, das Urteil vollständig aufzuheben, oder es so zu revidieren, dass ein schwächeres Urteil vertreten werden kann.[8] Im Fall eines Vertrags wären solche Ausnahmen beispielsweise Falschinformation, Zwang oder Geistesstörungen.

Auch wenn Hart selbst nicht diese Terminologie verwendet, lässt sich die mit (V) eingeführte Default-and-Challenge Struktur auf Harts Erläuterung anfechtbarer Begriffe übertragen: Zum einen gibt es *Default-Bedingungen*, also Tatsachen, die normalerweise hinreichend für die Anwendung eines anfechtbaren Begriffs sind. Zum anderen ist es möglich, durch Bestreiten der Tatsachen oder Verweis auf Ausnahmen, d.h. mittels der so genannten *Anfechtungsbedingungen*, die Gültigkeit des Urteils in Frage zu stellen.

In einem zweiten Schritt argumentiert Hart, dass Begriffe, die im Strafrecht traditionellerweise ein „mentales Element" bezeichnen, so etwa mens rea, Absicht oder Freiwilligkeit, eine anfechtbare Struktur besitzen.[9] Demzufolge kann ein Begriff wie *mens rea*, d.h. dasjenige mentale Element im Strafrecht, das Schuldbewusstsein bezeichnet, nicht positiv bestimmt werden, sondern nur negativ über die Menge an Anfechtungen wie Zwang, Provokation oder Irrsinn. Am Beispiel „freiwillig" wird Hart besonders deutlich: „Das Wort ‚freiwillig' dient genau genommen dazu, eine Reihe heterogener Fälle wie physischer oder psychischer Zwang, Unfälle, Fehler etc. auszuschließen, nicht aber dazu, eine mentales Element oder einen Zustand zu bezeichnen; ebenso wenig meint ‚unfreiwillig' die Abwesenheit eines mentalen Elements oder Zustands."[10]

---

8 "a plea that although all the circumstances on which a claim could succeed are present, yet in the particular case, the claim or accusation should not succeed because other circumstances are present which brings the case under some recognized head of exception, the effect of which is either to defeat the claim or accusation altogether, or to 'reduce' it so that only a weaker claim can be sustained." (Hart 1948/49, S. 147 f.)

9 Ebd., S. 152

10 „the word ‚voluntary' in fact serves to exclude a heterogeneous range of cases such as physical compulsions, coercion by threats, accidents, mistakes, etc., and not to designate a mental element or state; nor does ‚involuntary' signify the absence of this mental element or state." (Ebd., S. 153)
Vgl. auch Austin in A Plea for Excuses (1956): "[T]o say we acted "freely" (…) is to say only that we acted not un-freely (…) Like "real", "free" is only used to rule out the suggestion of some or all of its recognized antitheses." (Ebd., S. 6) And: "For above all it will not do to assume that the "positive" word must be around to wear the trousers; commonly enough the "negative" (looking) word marks the (positive) abnormality, while the "positive" word, if it exists, merely serves to rule out the suggestion of that abnormality." (Ebd., S. 18)

Elemente wie Schuld und Freiwilligkeit werden gemeinhin als Bedingungen für Verantwortung angesehen. Indem Hart diese Bedingungen für Verantwortung als anfechtbar charakterisiert, schafft er die Grundlage für eine Analyse von Verantwortung als anfechtbarem Begriff.

Um dem Kontext gerecht zu werden, in dem Hart Anfechtbarkeit einführt, muss angefügt werden, dass sein vorrangiges Ziel eine neue philosophische Analyse des Handlungsbegriffs ist. Handlungssätze, so seine Hauptthese, sind anfechtbar und sie sind askriptiv, d.h. sie beschreiben keine Tatsachen, sondern schreiben Verantwortung zu. Wir möchten hier nicht Harts Handlungstheorie im Detail besprechen. Vielmehr folgen wir seinen Kritikern darin, dass die eigentliche Einsicht von Harts Analyse weniger den Handlungsbegriff betrifft, als vielmehr die Auslegung von Verantwortung als einem anfechtbaren Begriff.[11] Diese Idee ist unmittelbar plausibel: Wir können einer erwachsenen Person Verantwortung für ihre Handlung zuschreiben, es sei denn, Anfechtungsgründe liegen vor.[12]

3. Die Struktur von Verantwortungszuschreibungen

Harts Ansatz einer Zuschreibungskonzeption wurde dafür kritisiert, die Bedeutung des Begriffs Verantwortung im Unklaren zu lassen.[13] Wir schlagen vor, dass es zwei Bedeutungen von Verantwortung gibt, die sich im Sinne einer Default-and-Challenge Struktur auffassen lassen. Diese zwei Bedeutungen beruhen beide auf der bereits eingeführten allgemeinen Form (V): Wenn Default-Bedingungen erfüllt sind, schreiben wir einer Person Verantwortung für ihre

11 Siehe auch Georg Pitcher (1960) und Christopher Cherry, der sagt: “defeasible claims are claims to the effect that an agent is responsible for his action.” (Cherry 1974: “The Limits of Defeasibility” in: Analysis 34, S. 104)

12 Bemerkenswerterweise erfährt gegenwärtig der Gedanke der Anfechtbarkeit bzw. der einer Default-and-Challenge Struktur ein zunehmendes Interesse auch auf Seiten der Epistemologie, so bei Michael Williams (2001) und Marcus Willaschek (2007). Der terminologische Ausdruck „Default-and-Challenge“ wurde von Robert Brandom in Making it explicit (1994) eingeführt. Brandom plädiert dafür, dass unsere Praxis des Vorbringens, Kritisierens und Rechtfertigens von Behauptungen einem Default-and-Challenge Modell folgt. Die hier vorgeschlagene Konzeption teilt mindestens zwei Aspekte mit Brandoms Theorie: Erstens können Aussagen als legitim betrachtet werden, ohne explizit durch die Angabe von Gründen gerechtfertigt zu sein. Zweitens müssen bestimmte Bedingungen erfüllt sein, damit eine Aussage Default-Status genießt. Umgekehrt kann die Aussage unter Verweis auf bestimmte Gründe angefochten werden. Diese Bedingungen, die Brandom enabling und defeating conditions nennt, entsprechen den hier so bezeichneten Default- und Anfechtungsbedingungen.

13 Georg Pitcher schlägt vor, die Aussage, dass eine Person für eine Handlung verantwortlich ist, so auszulegen, dass diese Person einen Verweis oder eine Bestrafung dafür verdient (Vgl. Pitcher 1960, S. 230).

Handlung zu, es sei denn, Entschuldigungs- oder Ausnahmegründe liegen vor. Die erste Bedeutung, (V1), fasst Verantwortung als „Zurechenbarkeit“ auf:

> (V1) *Eine Person P gilt als zurechenbar bzw. als ein angemessenes Subjekt von Verantwortungszuschreibung, wenn Default-Bedingungen erfüllt sind, es sei denn es liegen Ausnahmegründe vor.*

Die Frage nach der Bewertung von Handlungen kommt erst mit der zweiten Bedeutung (V2) ins Spiel:

> (V2) *Eine Person P verdient Lob oder Tadel für eine Handlung X, wenn P als zurechenbar gilt, wenn X in Bezug auf eine Norm bewertet wird und wenn Default-Bedingungen erfüllt sind, es sei denn es liegen Entschuldigungsgründe vor.*

Wie schon angedeutet sind Default-Bedingungen minimale Bedingungen, unter denen wir ohne Angabe von Gründen Verantwortung zuschreiben können. Unserer Auffassung nach sind diese Bedingungen, dass die Person erwachsen ist und, im Fall von (V2), dass die Person die Handlung X vollzogen hat. (V1) betrifft die Frage, ob eine Person die relevanten Fähigkeiten besitzt, um für ihre Handlungen zurechenbar zu sein. Hinsichtlich der Frage nach der Bewertung einer Handlung verhält sich (V1) selbst neutral, bildet jedoch eine notwendige Bedingung für die Möglichkeit einer Bewertung, also für (V2).[14] Zuschreibungen von Verantwortung, die sich gemäß (V2) in Lob oder Tadel äußern, sind nur dann möglich, wenn eine Handlung in Bezug auf eine Norm betrachtet wird.

Dieser Vorschlag greift Aspekte der Verantwortungstheorie von Wallace auf. Indem Wallace Entschuldigungs- und Ausnahmegründe diskutiert (Kap. 5 und 6), scheint er dem Verantwortungsbegriff implizit eine anfechtbare Struktur zugrunde zu legen. Allerdings geht Wallace selbst auf diesen Aspekt nicht ex-plizit ein. Die Untersuchung von Entschuldigungen und Ausnahmen dient ihm lediglich als Mittel, um das Kriterium für die Fairness von Verantwortungszuschreibungen zu bestimmen. Dieses Ziel ist zweifellos wichtig. Wie unser Vorschlag zeigt, verkennt Wallace dabei jedoch die strukturelle Funktion, die Entschuldigungen und Ausnahmen in unseren Verantwortungsurteilen haben.

4. Anfechtbare Urteile anstatt reaktiver Einstellungen

Abschließend zeigen wir, wie die von uns vorgeschlagene Analyse von Verantwortung als einem anfechtbaren Urteil die anfangs genannten drei Schwierigkeiten von Wallace’ Theorie zu lösen vermag.

Ein erster Einwand war, dass es problematisch sei, Verantwortungszuschreibungen anhand von reaktiven Einstellungen zu analysieren. Wallace’ Angemessen-

14 In der Praxis treten V1 und V2 in der Regel gemeinsam auf, da selten die Frage nach Zurechenbarkeit per se von Interesse ist.

heitsbedingung untergräbt die zentrale Rolle von Gefühlen, da die Angemessenheit einer Zuschreibung von dem Urteil abhängt, dass eine Norm verletzt wurde und keine Entschuldigungs- und Ausnahmegründe vorliegen. Angesichts dessen gilt es in einer Verantwortungstheorie der Tatsache Rechnung zu tragen, dass Verantwortungszuschreibungen auf Urteilen basieren. Ferner gilt es, die Struktur dieser Urteile zu beschreiben. Der hier unterbreitete Vorschlag erfüllt diese Anforderungen, indem er Verantwortungszuschrei-bungen als anfechtbare Urteile der Form (V1) und (V2) auffasst.

Zweitens wurde kritisiert, dass Wallace sich auf negative Verantwortungszuschreibungen beschränkt. Auf Basis des Gesagten kann diese Kritik nun präzisiert werden. Negative Verantwortungszuschreibungen im Sinne von Schuld oder Schuldhaftigkeit besitzen die Form (V2), insofern sie sich auf die Bewertung einer Handlung nach bestimmten Normen beziehen. Angesichts der Tatsache, dass diese Bewertung prinzipiell negative, positive als auch neutrale Ergebnisse hervorbringen kann, ist es offensichtlich, dass die negative Bewertung keinen systematischen Vorrang in der Analyse des Verantwortungsbegriffs hat. Indem Verantwortungszuschreibungen gemäß (V2) analysiert werden, spielt die vermeintliche Asymmetrie zwischen positiven und negativen Zuschreibungen keine Rolle mehr.

Kommen wir zum letzten Punkt, dem Verhältnis von moralischer und rechtlicher Verantwortung. Die bisherige Analyse des Verantwortungsbegriffs gemäß der Form (V1) und (V2) lässt offen, um welchen der beiden Verantwortungstypen es sich handelt. Wir sehen nun, dass erst in (V2) und erst durch die Art der Norm, nach der die Handlung bewertet wird, sich entscheidet, welche Art von Verantwortung zur Debatte steht. Entsprechend impliziert eine Bewertung nach rechtlichen Normen rechtliche Verantwortung, während die Bewertung einer Handlung anhand moralischer Normen moralische Verantwortung impliziert. Wie plausibel unsere Analyse für rechtliche Verantwortung ist, lässt sich am deutschen Strafrecht veranschaulichen, demnach erwachsene Straftäter als verantwortlich für ihre Verbrechen angesehen werden, es sei denn, es gibt einen Grund, die Person als unzurechnungsfähig einzustufen oder bestimmte Entschuldigungen geltend zu machen.[15] Die Zurückführung von moralischer und rechtlicher Verantwortung auf eine gemeinsame Struktur erlaubt zugleich Spiel-

15 Entsprechende Ausnahmegründe werden in § 20 ("Ohne Schuld handelt, wer bei Begehung der Tat wegen einer krankhaften seelischen Störung, wegen einer tiefgreifenden Bewusstseinsstörung oder wegen Schwachsinns oder einer schweren anderen seelischen Abartigkeit unfähig ist, das Unrecht der Tat einzusehen oder nach dieser Einsicht zu handeln.") des Strafrechts aufgezählt, eine Entschuldigung wäre beispielsweise „entschuldigender Notstand" (§ 35: „Wer in einer gegenwärtigen, nicht anders abwendbaren Gefahr für Leben, Leib oder Freiheit eine rechtwidrige Tat begeht, um die Gefahr von sich, einem Angehörigen oder einer anderen ihm nahestehenden Person abzuwenden, handelt ohne Schuld").

raum für spezifische Unterschiede, sei es die höhere Gewichtung von reaktiven Gefühlen in der Zuschreibung moralischer Verantwortung, oder aber die Begrenzung auf negative Verantwortungszuschreibungen im Rechtsbereich.

Literaturverzeichnis

Austin, John L.: "A Plea for Excuses". 1956. In: Philosophical Papers. Oxford University Press, Oxford, 1961. S. 123-152

Brandom, Robert B.: Making It Explicit. Reasoning, Representing & Discursive Commitment. Harvard University Press, Cambridge, 1994

Cherry, Christopher: "The Limits of Defeasibility". Analysis, 34, Oxford University Press, Oxford, 1974. S. 101-107

Hart, Herbert L. A.: "The Ascription of Responsibility and Rights". 1948/49. In: *Flew, Antony (Hrsg.*): Essays on Logic and Language. Oxford University Press, Oxford, 1963. S. 145-166

Pitcher, George: "Hart on Action and Responsibility". The Philosophical Review, Duke University Press, Durham, 1960. S. 266

Strawson, Peter F.: "Freedom and Resentment". 1962. In: *Watson, Gary (Hrsg.):* Free Will. Oxford University Press, Oxford, 2003. S. 72-94

Wallace, R. Jay: Responsibility and the Moral Sentiments. Harvard University Press, Cambridge, 1994

Willaschek, Marcus: "Contextualism about Knowledge and Justification by Default". Grazer Philosophische Studien, 74, Rodopi, Amsterdam, 2007. S. 251-272

Williams, Michael : Problems of Knowledge. Oxford University Press, Oxford, 2001

Warum gibt es Gründe?

Mario Brandhorst
mbrandh@gwdg.de
Georg-August-Universität Göttingen

Abstract/Zusammenfassung

This paper considers the question how the existence of reasons can be explained. The question concerns reasons for action, but also reasons for belief. In the first and introductory section, I explain what I take that question to mean and what asking it presupposes.

In the section that follows, I consider a possible answer. What makes a given fact a reason in a given context? A straightforward but ultimately unsatisfactory answer would be: Reasons are simply there, exist and form part of the fabric of the universe; the facts about reasons are as they are independently of what we happen to believe about them and independently of what our attitudes towards them may be. I refer to this view as *platonism* about reasons. For the platonist, neither the fact that reasons exist, nor the fact that reasons take the particular form that they do, is susceptible of natural explanation. This appears to be arbitrary, provided that a good explanation is ready to hand.

The third and central section sketches the outline of how such an explanation may go. I suggest that we adopt a broadly *pragmatic* interpretation of the language of reasons. This interpretation is a variant of *antirealism* about the language of reasons. It therefore escapes objections not only to platonism, but also to other realist interpretations of the language of reasons. This pragmatic interpretation builds on a number of less familiar remarks from Wittgenstein's *Philosophical Investigations* that discuss the problem of induction. As I will show, Wittgenstein's interpretation directly opposes the tempting idea that reasons are simply given.

In the the fourth and final section of the paper, I discuss critical questions concerning my strategy of explanation. In conclusion, I ask what right we have to acribe to Wittgenstein such a view of the nature of reasons.

Der Beitrag geht der Frage nach, wie die Existenz von Gründen zu erklären ist. Die Frage betrifft Gründe dafür, etwas Bestimmtes zu *tun*, aber auch Gründe dafür, etwas Bestimmtes zu *glauben*. In der Einleitung werden die Frage und einige ihrer Voraussetzungen erläutert. Was macht eine Tatsache in einem gegebenen Kontext zu einem Grund?

Der zweite Abschnitt diskutiert eine einfache und letztlich unbefriedigende Antwort auf diese Frage. Sie lautet: Gründe sind als solche einfach da; sie sind Teil der Struktur der Wirklichkeit, und die Tatsachen in Bezug auf Gründe sind so, wie sie nun einmal sind, ganz gleich was wir über sie denken und sagen, oder wie wir uns zu ihnen verhalten. Diese Auffassung bezeichne ich als *Platonismus* in Bezug auf Gründe. Aus der Sicht des Platonisten gibt es weder eine natürliche Erklärung dafür, dass es Gründe gibt, noch dafür, dass sie so beschaffen sind, wie sie es sind, und nicht anders. Das erscheint willkürlich, sofern es eine gute und naheliegende Erklärung für beides gibt.

Der dritte Teil des Beitrags arbeitet die Grundzüge einer solchen Erklärung aus. Wie sich zeigt, führt sie auf eine *pragmatische* Deutung der Rede von Gründen. Diese Deutung ist zugleich eine Variante des *Antirealismus* in Bezug auf Gründe. Sie entgeht deshalb nicht nur

den Einwänden gegen die plato-nistische Deutung, sondern auch denen gegen die realistische Deutung der Rede von Gründen im Allgemeinen. Diese pragmatische Deutung der Rede von Gründen stützt sich auf einige wenig beachtete Abschnitte aus Wittgensteins *Philosophischen Untersuchungen*, die sich mit dem Induktionsproblem befassen. Wie deutlich werden wird, ist Wittgensteins Deutung der Rede von Gründen ausdrücklich gegen die Vorstellung gerichtet, Gründe seien etwas uns einfach Gegebenes.

Im vierten und letzten Teil des Beitrags diskutiere ich Einwände, die meine Erklärungstrategie betreffen, und frage, mit welchem Recht Wittgenstein selbst die entsprechende Auffassung von Gründen zugeschrieben werden kann.

1. Einleitung

Im Titel meines Beitrags steht eine Frage: „Warum gibt es Gründe?“ Die Frage zielt auf eine Erklärung, und der Vortrag geht der Frage nach, wie die Existenz von Gründen zu erklären ist. Die Frage betrifft Gründe dafür, etwas Bestimmtes zu tun, aber auch Gründe dafür, etwas Bestimmtes zu glauben.

Mir sind zwei Seiten dieser Frage wichtig. Erstens: die Frage selbst; genauer: die Tatsache, dass es hier eine *Frage* gibt. Zweitens: die *Antwort*, oder wenigstens die *Form* der Antwort, die mir überzeugend zu sein scheint. Im Folgenden werde ich eine Antwort auf die Frage, warum es Gründe gibt, diskutieren, die meines Erachtens die Rede von Gründe *nicht* überzeugend erklärt. Das wirft die Frage auf, welche Erklärung an ihre Stelle tritt, ich werde einen Vorschlag dazu machen, welche Erklärung die Rede von Gründen umfassend und insgesamt überzeugend erklärt. Die Antwort findet sich, wenn auch nur in Ansätzen, in Wittgensteins Spätwerk, insbesondere in den *Philosophischen Untersuchungen*. Die Bedeutung und die Tragweite der Antwort werden allerdings selbst in der Wittgenstein-Rezeption nicht recht deutlich.[1]

Es ist nicht ohne weiteres verständlich, was mit der Frage, warum es Gründe gibt, gemeint ist. Ein erster Grund dafür liegt in der Unklarheit der Frageform „Warum...?“ und der damit einhergehenden Unklarheit der Rede von einer „Erklärung“. Sowohl das Fragepronomen „warum“ als auch der Begriff der Erklärung sind mehrdeutig. Ein zweiter und mindestens ebenso wichtiger Grund ergibt sich aus der Unklarheit der Rede davon, dass es Gründe „gibt“. Wir wissen im täglichen Sprachgebrauch sehr genau, was ein Grund ist. Wir können sehr oft, wenn auch nicht immer, Gründe erkennen, abwägen und zu einem vernünftigen Urteil darüber gelangen, was für oder gegen eine Handlung oder Überzeugung spricht. Was dagegen ist gemeint, wenn jemand sagt, ein Grund „existiere”, und sich dann anschickt, eine „Erklärung“ dafür zu suchen?

Was, so formuliert, den Verdacht erweckt, philosophischer Unsinn zu sein, hat einen einfachen und klaren Sinn. Dass es Gründe „gibt“, oder dass Gründe

1 Interessante Perspektiven ergeben sich aus der von J. Prescott herausgegebenen Sammlung „Wittgenstein and Reason“.

„existieren“, bedeutet zunächst nichts anderes, als dass der gewöhnliche Gebrauch von Wörtern wie „Grund“ oder „Begründung“ nicht sinnlos, sondern verständlich und auch berechtigt ist, was auch immer wir aus theoretischer Sicht weiter über ihn sagen. Dass der Gebrauch solcher Wörter berechtigt ist, bedeutet nicht, dass es eine unabhängige, äußere Rechtfertigung für ihn gäbe. (Welche Rechtfertigung sollte das sein? Was würde überhaupt als eine solche Rechtfertigung zählen?) Es bedeutet, dass es gegen diesen Gebrauch keine grundsätzlichen *Einwände* gibt.

Die Sprache der Gründe hat eine Struktur, die der Struktur einer Sprache, mit der wir uns auf die Wirklichkeit beziehen und über deren Beschaffenheit Aussagen machen, zweifellos sehr eng verwandt ist. Wenn ich keine Meeresfrüchte mag, *habe* ich einen Grund, im Restaurant ein Essen ohne Meeresfrüchte zu bestellen. *Es gibt* demnach einen Grund für mich, so zu handeln, und zwar diesen. Der Grund ist, so können wir ebenfalls sagen, *vorhanden* oder *existiert*, auch wenn das im Vergleich zur Formulierung „es gibt...“ etwas künstlich klingt. Andere ebenso wie ich selbst können diesen Grund *kennen*. Wer ihn nicht kennt, aber behauptet, ich hätte in Wirklichkeit keinen Grund, im Restaurant ein Essen ohne Meeresfrüchte zu bestellen, *täuscht* sich über meine Gründe, denn die Aussage und die entsprechende Meinung sind *falsch*. Wer dagegen von mir sagt, ich hätte einen solchen Grund, sagt offensichtlich etwas *Wahres*.

Gründe sind demnach nicht nach dem Muster von *Dingen* zu deuten. Gründe sind immer die Gründe *einer Person*, und wer sagt, eine Person habe einen Grund für eine Handlung oder Überzeugung, beansprucht nicht, dass eine besondere Art von *Gegenstand* existiert, sondern dass die Aussage, dass die Person diesen Grund hat, *wahr* ist. Nun entspricht einer wahren Aussage niemals ein Gegenstand, sondern eine Tatsache, und der Begriff der Tatsachen und der Begriff der wahren Aussage sind aufeinander bezogen. Es gibt demzufolge *Tatsachen* in Bezug darauf, welche Gründe ich habe, und diese Tatsachen werden nicht durch eine gegebene Aussage über diese Gründe bestimmt. Vielmehr bestimmen die Tatsachen, welche Aussagen über meine Gründe wahr sind und welche falsch. Diese Tatsachen sind also in einem nahe liegenden Sinn *objektiv*. Wir können uns über Gründe im Irrtum befinden. Wo wir es tun, entspricht die *Wirklichkeit* nicht der Art und Weise, wie wir sie *wahrnehmen* oder *beschreiben*.

Das sind die sprachlichen Befunde. Ihre Beschreibung ließe sich weiter vertiefen und in verschiedener Weise ergänzen. Jede Analyse und Interpretation der Sprache der Gründe muss mit diesen Befunden beginnen. Sie muss aber auch mit ihnen enden, wenn die Rede von Gründen nicht instabil werden soll. Keine Analyse oder Interpretation der Sprache der Gründe kann überzeugen, wenn sie Redeweisen wie diese nicht beibehält und als verständlich und berechtigt ausweist. Andernfalls liefe die Analyse auf die sehr radikale These hinaus, die Rede von Gründen sei systematisch irreführend oder gar sinnlos. Diesen Stand-punkt

sollten wir nur im äußersten Notfall beziehen, und es gibt mehr als eine Möglichkeit, um ihn zu vermeiden.

Umgekehrt fordert die Sprache der Gründe auch dazu heraus, sich ein allgemeineres, geordneteres Bild von ihr zu machen und ihren Ort im Zusammenhang menschlicher Lebensvollzüge genauer zu bestimmen. Die sprachlichen Befunde bedürfen philosophischer Analyse und Interpretation, zumal bereits hinreichend deutlich wurde, wie sehr die Sprache der Gründe zu Existenz- und Wissensbehauptungen einlädt. Wer sich in einem gewöhnlichen Kontext dieser Sprache bedient, läuft kaum Gefahr, falsch verstanden zu werden. Es ist grundsätzlich nichts dagegen einzuwenden, wenn jemand sagt, er *habe* einen Grund, etwas zu tun oder etwas zu glauben, es *gebe* einen solchen Grund und er verfüge über entsprechendes *Wissen*. Natürlich kann er sich irren; doch wir wissen sowohl was es heißt, dass er Recht hat, als auch, dass er sich irrt. Was wir aus theoretischer, philosophischer Sicht über diese Ausdrucksweisen sagen sollen, ist dagegen alles andere als klar.

Die Frage, wie die Existenz von Gründen zu *erklären* ist, ist nun zunächst nichts weiter als die Frage, wie die Sprache der Gründe philosophisch angemessen analysiert und interpretiert werden kann. Dabei ist insbesondere zu fragen, welche theoretischen Verpflichtungen wir eingehen, wenn wir an den sprachlichen Befunden festhalten und sie als verständlich ausweisen wollen. Sind wir darauf festgelegt, an die Existenz von objektiven normativen Tatsachen zu glauben, wie die Rede von Gründen selbst nahe legt, oder gibt es Alternativen? Worauf legen wir uns überhaupt fest, wenn wir von solchen objektiven normativen Tatsachen sprechen? Welchen Ort haben Gründe in unserem Bild von der Struktur der Wirklichkeit?

Die Sprache der Gründe legt bereits eine bestimmte Antwort auf die Frage nahe. Wir sahen bereits, wie natürlich es ist, im Zusammenhang mit Gründen von Existenz, Wissen, Wahrheit, Irrtum, Tatsachen und Objektivität zu sprechen. Von hier aus ist es nur noch ein kleiner Schritt bis zu einem theoretischen Bild, in dem Gründen eine objektive Existenz zugeschrieben wird, wobei damit etwas gemeint ist, was weit über das bisher Gesagte hinausgeht. Die Struktur der Sprache der Gründe selbst verleitet dazu, Gründe als etwas Gegebenes zu betrachten. Sie verleitet darüberhinaus dazu zu meinen, Gründe seien etwas Wirkliches, das so beschaffen ist, wie es ist, unabhängig davon, was wir darüber denken und sagen, und unabhängig davon, wie wir uns dazu verhalten. Sie verleitet uns also dazu, Gründe als etwas Objektives, als Teil der Struktur der Wirklichkeit zu betrachten.

Deshalb ist es so wichtig, die Frage, warum es Gründe gibt, klar zu stellen und dann eine vernünftige Antwort auf sie zu suchen. Nicht nur, aber vor allem in der philosophischen Diskussion wird sehr oft so getan, als seien uns Gründe einfach gegeben. Sehr oft ist damit sogar die Hoffnung verbunden, die Objektivität der Moral oder des Eigeninteresses zu sichern, ohne sich dabei auf frag-

würdige psychologische oder metaphysische Annahmen stützen zu müssen. Die Rede von Gründen bietet sich dazu offenbar an, denn sie ist uns nicht nur vertraut, sondern erscheint uns auch als sehr verlässlich. Doch dabei dürfen eine entscheidende Frage nicht übersehen: Was berechtigt uns zu der Hoffnung, Gründe seien etwas in diesem Sinn Wirkliches, Objektives, Gegebenes, seien ein elementare Bestandteil der Wirklichkeit? Wie ich zeigen will, liegt eben hier der entscheidende Fehler. Oft wird das Recht, diese Annahmen machen zu dürfen, einfach vorausgesetzt. Noch häufiger wird die Frage nach diesem Recht überhaupt nicht gestellt, obwohl die Rede von Gründen im Folgenden großes Gewicht trägt.

Im nächsten Abschnitt untersuche ich die Antwort des Realisten auf die Frage, wie die Rede von Gründen angemessen zu interpretieren ist, etwas genauer. Es zeigt sich sehr bald, dass sie nicht überzeugt. In der zugespitzten Form des Platonismus ist sie noch nicht einmal eine Antwort. Die platonistische Deutung der Rede von Gründen liefert überhaupt keine Erklärung dafür, dass es Gründe gibt; sie weist die Frage nach der Erklärung einfach zurück. Wenn es eine überzeugende Erklärung für die Existenz von Gründen gibt, ist diese deshalb vorzuziehen. Im darauffolgenden dritten Abschnitt arbeite ich die Grundzüge dieser alternativen Erklärung aus. Wie sich zeigt, führt sie auf eine *pragmatische* Deutung der Rede von Gründen. Diese Deutung ist eine Variante des *Antirealismus* in Bezug auf Gründe. Sie entgeht deshalb nicht nur den Einwänden gegen die platonistische Deutung, sondern auch denen gegen die realistische Deutung der Rede von Gründen im Allgemeinen. Im vierten und letzten Teil dieses Beitrags gehe ich kurz auf die Frage ein, was aus dieser Deutung der Rede von Gründen folgt und welche philosophischen Perspektiven sich aus ihr ergeben.

2. Die platonistische Deutung der Rede von Gründen

Was genau ist mit der Rede von Gründen gemeint? Ich folge T. M. Scanlon darin, einen Handlungsgrund als etwas anzusehen, was für eine Handlung spricht.[2] Das ist, wie Scanlon sagt, weniger eine informative Definition dieses Wortes als eine Erläuterung oder Illustration. Wenn jemand fragt, wie etwas für eine Handlung sprechen kann, gibt es nur eine Antwort: indem es einen Grund dafür darstellt. Der Begriff des Grundes ist in dieser Hinsicht primitiv. Er ent-zieht sich der *Begriffsanalyse*, wenn eine Begriffsanalyse darauf abzielt, den Begriff mit einfacheren und von ihm logisch unabhängigen Mitteln zu klären. Ebenso verhält es sich mit Gründen dafür, etwas zu glauben. Hier ist ein Grund etwas, das dafür spricht, etwas für wahr zu halten. Wenn nun jemand fragt, wie etwas dafür

2 T. M. Scanlon, *What We Owe to Each Other*, S. 17.

sprechen kann, etwas anderes für wahr zu halten, gibt es wieder nur eine Antwort: indem es einen Grund dafür darstellt.

Ich merke nur im Vorbeigehen an, dass sowohl Externalisten als auch Internalisten diese Kennzeichnung von Gründen akzeptieren können. Interna-listen behaupten, dass etwas nur dann ein Handlungsgrund für jemanden sein kann, wenn es sich auf dem Weg der rationalen Überlegung aus den gegebenen Motivationen des Handelnden ergibt. Externalisten bestreiten das.[3] Nun kann es so scheinen, als sei der Internalist darauf festgelegt, die gegebenen Motivationen des Handelnden, seien es Wünsche oder Überzeugungen oder eine Verknüpfung der zwei, auch als dessen Gründe anzusehen. Tatsächlich wird Internalisten oft dieser Vorwurf gemacht. Er ist unberechtigt, denn er verwechselt zwei Begriffe des Grundes. Wünsche und Überzeugungen sind, wenn überhaupt irgend etwas, *motivierende* oder *erklärende* Gründe.[4] Internalisten und Externalisten dagegen streiten über *normative* Gründe. Sie stimmen deshalb darin überein, einen Handlungsgrund als etwas zu betrachten, was für eine Handlung *spricht*. Ihre Auseinandersetzung betrifft die ganz anders gelagerte Frage, welche Art von Bedingung dafür erfüllt sein muss, dass jemand einen solchen Grund *hat*.

Was kommt als Grund in Betracht? Gründe sind, so können wir sagen, durch die relevanten *Tatsachen* bestimmt. Handlungsgründe sind dementsprechend durch die für das Handeln relevanten Tatsachen bestimmt. Nun kann praktisch jede Tatsache in einem bestimmten Kontext ein Grund dafür sein, auf eine bestimmte Weise zu handeln. Liegt vor mir jemand hilflos auf der Straße, so ist *das* mein Grund zu helfen. Mag ich keine Meeresfrüchte, so ist *das* mein Grund, im Restaurant ein Essen ohne Meeresfrüchte zu bestellen. Tatsachen dieser Art sind in einem vertrauten und unproblematischen Sinn objektiv. Sie sind so beschaffen, wie sie es sind, unabhängig davon, was wir über sie denken und sagen.

Was aber macht eine Tatsache in einem gegebenen Kontext zu einem Grund? Eine einfache, aber wegen ihrer Einfachheit unbefriedigende Antwort lautet: Gründe sind einfach da, unabhängig davon, was wir über sie denken und sagen, und unabhängig davon, wie wir uns zu ihnen verhalten. Die Welt, so wird uns gesagt, umfasse ganz einfach noch mehr, als wir mit Aussagen über natürliche Dinge erfassen. Es gibt, so wird uns mitgeteilt, auch eine Wirklichkeit, an der sich die Wahrheit und Falschheit von normativen Aussagen bemisst. Diese Wirklichkeit ist nicht mit der natürlichen Welt identisch, und sie ist auch nicht aus Bestandteilen der natürlichen Welt konstruiert.[5] Vielmehr ist sie eine Wirk-

3 Das entspricht der Definition der Positionen bei B. Williams, auf den die Unterscheidung zurückgeht; vgl. „Internal and External Reasons“ und das spätere „Postscript“. Der Aufsatz ist zugleich die klassische Verteidigung der internalistischen Deutung.

4 Das entspricht dem Bild der kausalen Handlungstheorie; siehe D. Davidson, „Actions, Reasons, and Causes“.

5 So verstanden sind sowohl der Platonismus als auch der Konstruktivismus Spielarten des Realismus: beide lassen objektive normative Tatsachen zu. Der Konstruktivismus ist je-

lichkeit normativer, objektiver Tatsachen, die so sind, wie sie sind, unabhängig davon, was wir über sie denken und sagen, und wie wir uns zu ihnen verhalten. Diese Auffassung bezeichne ich als *Platonismus* in Bezug auf Gründe. Sie ist eine Spielart des Realismus und zeigt eine strukturelle Ähnlichkeit mit dem Platonismus in der Philosophie der Mathematik.

Es ist wichtig, die platonistische Auffassung von der offensichtlich richtigen und unbestrittenen Auffassung zu unterscheiden, dass die Tatsachen, *die Gründe sind* oder *liefern*, in der Regel unabhängig davon, was wir über sie denken und sagen und wie wir uns zu ihnen verhalten, so beschaffen sind, wie sie es sind. Das gilt sowohl für die Tatsache, dass jemand hilflos vor mir auf der Straße liegt, wie für die Tatsache, dass ich keine Meeresfrüchte mag. Die Frage, auf die es ankommt, ist was eine Tatsache dieser Art in einem gegebenen Kontext zu einem Grund *macht*. Die Tatsache, dass jemand vor mir hilflos auf der Straße liegt, *ist* für mich ein Grund, der Person zu helfen. Die Tatsache, dass ich keine Meeresfrüchte mag, *ist* ein Grund für mich, im Restaurant ein Essen ohne Meeresfrüchte zu bestellen. Die Frage ist: Was erklärt den Zusammenhang? Diese Frage ist nicht schon durch den Hinweis beantwortet, die Tatsachen, auf die wir uns hier beziehen, seien so, wie sie sind, ganz unabhängig von dem, was wir über sie denken und sagen. Das trifft zwar zu, ist aber keine Antwort auf die Frage. Die Frage lautet: Was macht diese Tatsachen *zu Gründen*? Ist die Tatsache, *dass* etwas einen Grund für eine Handlung oder eine Überzeugung darstellt, ebenfalls unabhängig davon, was wir über sie denken und sagen?

Der Platonist zeichnet sich dadurch aus, eben das zu behaupten; und mit dieser Behauptung ist der Platonismus in der Philosophie der Normativität ebenso wie der Platonismus in der Philosophie der Mathematik schwerwiegen-den Einwänden ausgesetzt, die ihn sehr unattraktiv erscheinen lassen. Abge-sehen von der Frage, wie normative Tatsachen in unser Weltbild passen und wie sie erkennbar sind, wenn die platonische Auffassung zutrifft, bleibt völlig ungeklärt, was dafür spricht, an die bloße, unerklärbare Gegebenheit von Gründen zu glauben. Es gibt ja aus der Sicht des Platonisten weder eine natürliche Erklärung dafür, dass es Gründe gibt, noch dafür, dass sie so beschaffen sind, wie sie es sind. Das erscheint willkürlich, sofern es eine naheliegende Erklärung für beides gibt.

Tatsächlich läuft die platonistische Auffassung darauf hinaus, diese Fragen nicht zu beantworten, sondern ihnen auszuweichen. Das Muster der Antworten auf Fragen nach einer Erklärung ist immer dasselbe. Warum gibt es Gründe? Wir können es nicht erklären; es gibt sie. Wie passen sie in unser Weltbild? Wir

---

doch kein Platonismus, weil er objektive normativen Tatsachen als von uns konstruiert und insofern als von uns abhängig ansieht. Der Platonist vertritt eine stärkere Unabhängigkeitsthese. Zur Problematik des Realismus im Verhältnis zum Konstruktivismus vgl. aus realistischer Sicht R. Shafer-Landau, *Moral Realism: A Defence*, insbes. Kapitel 1 und 2, und aus konstruktivistischer Sicht C. Korsgaard, „Realism and Constructivism in Twentieth-Century Moral Philosophy".

können es nicht erklären; wir müssen sie als elementaren Bestandteil der Wirklichkeit anerkennen. Wie haben wir Kenntnis von Gründen? Wir können es nicht erklären; wir haben sie. Das ist nicht glaubwürdig.

Angesichts dieser Schwierigkeiten erscheint es nun attraktiv, sich auf eine gemäßigte Position zurückziehen, derzufolge es zwar objektive Tatsachen in Bezug auf Gründe gibt, die Möglichkeit der Erklärung aber nicht ausgeschlossen wird. Entsprechend mag man vesuchen, die ontologischen und epistemischen Fragen ernster als der Platonist zu nehmen. Lassen sie sich nicht auf eine Weise beantworten, die die realistische Sicht nicht von vornherein diskreditiert?

Man kann, mit anderen Worten, versuchen, ein Realist zu sein, ohne Platonist zu werden. Dennoch bleibt eine Reihe offener Fragen bestehen: Wie können wir uns *sicher* sein, dass es objektive Tatsachen in Bezug auf Gründe gibt? Warum sollten wir es *glauben*? Was genau wird durch die Tatsache, dass es Gründe gibt, *erklärt*? Wodurch wird diese Tatsache *ihrerseits* erklärt? Was genau ist das, was in einem gegebenen Kontext *bestimmt*, dass eine Tatsache für die eine Handungsweise oder Überzeugung einen Grund darstellt, für die andere dagegen nicht? Was erklärt die Tatsache, dass Gründe abgewogen werden können und ein gegebener Grund *besser* oder *gewichtiger* sein kann als ein anderer? Was erklärt unser *Wissen* von Gründen und unser Wissen von ihrem Gewicht und ihrer Güte? Was erklärt einen Irrtum? Hinter all dem steht unverändert die Frage, die sich auch dem Platonisten stellt: Wie passen normative Tatsachen in unser Weltbild, und wie können wir, wenn es sie gibt, von ihnen wissen? Die Überzeugungskraft des Realismus hängt ganz entscheidend von seiner Antwort auf diese Fragen ab.

3. Gibt es Alternativen?

Anstatt die Fragen weiter zu vertiefen, möchte ich nun fragen, welche Alternative zum Platonismus es gibt. Ich stütze mich dabei auf einige wenig beachtete Abschnitte aus Wittgensteins *Philosophischen Untersuchungen*, die sich mit der Rede von Gründen befassen.[6] Wie ich zeigen will, führen diese Passagen auf eine *pragmatische* Deutung der Rede von Gründen. Diese Deutung ist, wie ich zugleich zeigen will, eine Spielart des *Antirealismus* in Bezug auf Gründe, und was Wittgenstein sagt, führt auf natürlichem Wege dorthin. Die pragmatische Deutung der Rede von Gründen entgeht den verschiedenen Einwänden gegen den Platonismus, der Gründe als elementaren, nicht weiter erklärbaren Bestandteil der Wirklichkeit ansieht. Sie entgeht aber auch den verschiedenen Ein-

6 Die relevanten Textabschnitte sind §§ 324-5 und §§ 472-85, im Folgenden zitiert nach der ersten Werkausgabe mit der Sigle PU.

wänden gegen den Realismus, der Gründe als objektive normative Tatsachen versteht – und, wie mir scheint, missversteht.

Wittgensteins Diskussion steht im Zusammenhang des Induktionsproblems. Hier wird die Frage, welchen Grund wir dafür haben, ein bestimmtes Geschehen mit Sicherheit zu erwarten, ja ob wir *überhaupt* Gründe für eine solche Erwartung haben, sehr dringlich. Wittgenstein fragt:

> „Warum glaubst du, daß du dich an der heißen Herdplatte verbrennen wirst?“ — Hast du Gründe für diesen Glauben; und brauchst du Gründe? (PU, § 477)

In dieser Passage zeigt sich ein naturalistischer Zug, der Wittgenstein mit Hume verbindet. Hume ist allerdings nicht nur Naturalist, sondern auch Skeptiker: er argumentiert dafür, alle Schlüsse von der Vergangenheit auf die Zukunft nicht als vernunftgeleitet, sondern als Ausdruck reiner Gewohnheit zu deuten. Hume neigt also auch in Bezug auf die Erwartung zukünftiger Ereignisse dazu, unser Verhalten nicht als begründet, sondern als praktisch notwendig, als Gegebenheit der menschlichen Natur anzusehen.[7] Wittgenstein spielt darauf an, wenn er im Nachsatz fragt, ob wir für die sichere Erwartung, uns an der Flamme zu verbrennen, überhaupt Gründe *brauchen*. Die Antwort, die er nahe legt ist: nein. Diese Antwort hat auch Hume gegeben.

Doch Wittgenstein begnügt sich nicht mit dieser Antwort. Auch wenn er Humes Einschätzung insofern teilt, als er die Bedeutung der menschlichen Natur immer wieder unterstreicht und selbst darauf verweist, dass „alle Rechtfertigung durch Erfahrung ein Ende hat“, will er nicht leugnen, dass es Gründe für eine Erwartung wie die, sich an einer Flamme zu verbrennen, gibt.[8] Er schreibt:

> Was für einen Grund habe ich, anzunehmen, daß mein Finger, wenn er den Tisch berührt, einen Widerstand spüren wird? Was für einen Grund, zu glauben, daß dieser Bleistift sich nicht schmerzlos durch meine Hand wird stecken lassen?—Wenn ich dies frage, melden sich hundert Gründe, die einander kaum zu Wort kommen lassen wollen. „Ich habe es doch selbst unzählige Male erfahren; und ebenso oft von ähnlichen Erfahrungen gehört; wenn es nicht so wäre, würde ... ; etc.“. (PU, § 478)

Wittgenstein bestreitet also nicht, dass vergangene Erfahrung uns Gründe gibt, gleichförmiges Geschehen in der Zukunft zu erwarten. Die Frage, die sich angesichts der Humeschen Skepsis stellt, ist diese: *Wie kann* bisherige Er-fahrung solche Gründe liefern? Inwiefern spricht die bisherige Erfahrung *dafür*, ein gleichförmiges Geschehen auch in der Zukunft zu erwarten?

Wittgensteins Reaktion auf diese Fragen ist radikal und verblüffend: er weist die skeptischen Fragen zurück. Wie Hume verweist er auf Gegebenheiten der natürlichen Welt, insbesondere Gegebenheiten der menschlichen Natur; doch er tut es in anderer Weise und an einem anderen Ort. Wittgenstein schreibt:

---

7 Vgl. D. Hume, *A Treatise of Human Nature*, Buch 1, Teil 3 *passim* und *An Enquiry Concerning Human Understanding*, Abschnitte 4-7.

8 PU, § 485.

> Wenn man nun fragte: Wie *kann* aber frühere Erfahrung ein Grund zur Annahme sein, es werde später das und das eintreffen? — so ist die Antwort: Welchen allgemeinen Begriff vom Grund zu solch einer Annahme haben wir denn? Diese Art Angabe über die Vergangenheit nennen wir eben Grund zur Annahme, es werde das in Zukunft geschehen. — Und wenn man sich wundert, daß wir ein solches Spiel spielen, dann berufe ich mich auf die *Wirkung* einer vergangenen Erfahrung (darauf, daß ein gebranntes Kind das Feuer fürchtet). (PU, § 480)

Die Beschaffenheit der Welt und die Beschaffenheit der menschlichen Natur sind tatsächlich *grundlegend* und deshalb wichtig: sie stellen eine Grundlage für sehr verschiedene Sprachspiele dar, die ohne dieses Fundament ihren Sinn und „Witz" verlören. Doch die Gegebenheiten der Natur, einschließlich der menschlichen Natur, sind bei Wittgenstein nicht mehr der Vernunft und dem Sprachspiel der Gründe *entgegengesetzt*. Sie sind vielmehr das *Fundament*, auf dem das Sprachspiel der Vernünftigkeit und der Gründe *aufbaut*. Die Umwelt und die Handlungen des Menschen sind der natürliche Kontext, in den dieses Sprachspiel eingebettet ist, und Handlungen sind auch das Fundament des Sprachspiels der Gründe.

Der Skeptiker, so lautet nun die Diagnose, legt den begrifflich falschen Maßstab an, wenn er aufgrund der logischen Möglichkeit einer anders als erwartet verlaufenden Zukunft an den Gründen für die Erwartung zweifelt. Wittgenstein fragt den Skeptiker, was er *meint*, wonach er *sucht*, von welcher Vorstellung von Gründen er *geleitet* ist:

> Wer sagte, er sei durch Angaben über Vergangenes nicht davon zu überzeugen, daß irgend etwas in Zukunft geschehen werde, — den würde ich nicht verstehen. Man könnte ihn fragen: was willst du denn hören? Was für Angaben nennst du Gründe dafür, das zu glauben? Was nennst du denn „überzeugen"? Welche Art des Überzeugens erwartest du dir? — Wenn *das* keine Gründe sind, was sind denn Gründe? — Wenn du sagst, das seien keine Gründe, so mußt du doch angeben können, was der Fall sein müßte, damit wir mit Recht sagen könnten, es seien Gründe für unsre Annahme vorhanden.
>
> Denn wohlgemerkt: Gründe sind hier nicht Sätze, aus denen das Geglaubte logisch folgt. (PU, § 481)

Und schließlich:

> Ein guter Grund ist einer, der *so* aussieht. (PU, § 483)

Die Einsicht Wittgensteins, die für eine pragmatische Deutung der Rede von Gründen entscheidend ist, besteht im Hinweis auf die Praxis, *etwas einen Grund zu nennen*. So verstanden führt die Frage, welche Gründe es gibt, zunächst nicht auf die Idee einer objektiven normativen Wirklichkeit, in der es Gründe einfach gäbe. Sie führt uns zurück auf *uns selbst*, auf unser *Tun*, auf die *Verwendung* der Sprache der Gründe.

Aus dieser Einsicht ergibt sich zugleich eine Form des *Antirealismus* in Bezug auf Gründe. Das ist deshalb so, weil die Frage, was ein Grund ist oder sein kann in Wittgensteins Bild nicht unabhängig davon zu beantworten ist, was wir

einen Grund *nennen* und als solchen *anerkennen*. Das wäre zwar noch damit vereinbar, Gründe als objektiv anzusehen, weil Tatsachen in Bezug auf Gründe sowohl objektiv als auch vom Sprachspiel der Gründe abhängig sein könnten. So wäre es möglich, einerseits zwar zu bestreiten, dass Gründe unabhängig vom Sprachspiel der Gründe einfach gegeben sind, andererseits aber zugleich darauf zu bestehen, dass wir dann und nur dann, wenn wir das Sprachspiel beherrschen, Gründe so sehen und so beschreiben können, wie sie *tatsächlich* sind.[9]

Es wäre jedoch unbefriedigend, bei dieser Zwischenposition stehen bleiben zu müssen. Einerseits wirft sie erneut dieselben schwierigen Fragen auf, die mit der realistischen Deutung der Sprache der Gründe unvermeidlich verbunden sind. Auch wenn die skeptischen Fragen abgewiesen werden sollen, darf die Differenz zwischen dem, was wir einen Grund *nennen* und dem, was ein Grund *ist*, keine grundsätzliche sein. Doch der wichtigste Grund, der gegen die Annahme der entsprechenden Tatsachen spricht, ist die Tatsache, dass die Annahme überflüssig und willkürlich bleibt, solange das Sprachspiel der Gründe auch ohne sie auskommt und dabei stabil, transparent und nützlich bleibt.

Das, so scheint mir, ist tatsächlich der Fall. Aus pragmatischer Sicht wird sofort verständlich, welchen Sinn und welchen Status die Rede von Gründen im Zusammenhang menschlicher Lebensvollzüge hat. Sie steht im Zusammenhang eines eigengesetzlichen Sprachspiels, das einen bestimmten Sinn und „Witz" hat. Dieser „Witz" liegt, äußerst allgemein gesprochen, darin, uns eine verlässliche Orientierung im Umgang mit anderen Menschen, mit Fragen, Gedanken, Gefühlen, Erfahrungen, und vor allem mit den Gegebenheiten einer komplizierten sozialen und natürlichen Welt zu ermöglichen. Die Notwendigkeit einer derartigen Orientierung ist absolut fundamental, was verständlich macht, weshalb deie Sprache der Gründe nicht nur fundamental ist, sondern alle menschlichen Lebensvollzüge durchdringt. Es ist dagegen nicht der Sinn des Sprachspiels, eine unabhängig von unserer sprachlichen Praxis gegebene Ordnung von Gründen oder objektiver normativer Tatsachen zu repräsentieren, ebensowenig wie es der Sinn des mathematischen Sprachspiels ist, eine gegebene Ordnung von Zahlen oder objektiver numerischer Tatsachen zu repräsentieren. Das ist eine typisch philosophische Fiktion.

4. Einwände und Perspektiven

Ich komme damit zum vierten und letzten Teil meiner Überlegungen, in dem ich einige kritische Fragen und Einwände aufgreife, die meine Konzeption von Gründen betreffen und abschließend frage, welche Perspektiven, theoretisch wie

9 Eine solche Vorstellung findet sich bei J. McDowell, auch wenn sie meines Erachtens in verschiedenen Hinsichten unklar und mehrdeutig bleibt; vgl. „Might there be external reasons?".

praktisch, sich aus ihr ergeben. Dabei werde ich fragen, mit welchem Recht Wittgenstein selbst als Kronzeuge der pragmatischen Lesart zitiert werden kann.

Man mag einwenden, die pragmatische, antirealistische Konzeption von Gründen könne letztlich *nicht* verständlich machen, warum es Gründe gibt, weil sie genau genommen gar nicht *erklärt*, sondern *leugnet*, dass es Gründe gibt. Das wäre ein Missverständnis. Mit der antirealistischen Konzeption erübrigen sich Fragen nach dem metaphysischen Status von Gründen und nach den Mitteln und Möglichkeiten, sie zu erkennen und abzuwägen. In einem bestimmten, nämlich metaphysisch verstandenen Sinn ist es deshalb ganz richtig zu sagen, dass es Gründe nicht gibt. In einem anderen und hier entscheidenden Sinn *gibt es* Gründe, und die Erklärung dafür, dass es sie gibt, liegt in der einfachen aber entscheidenden Einsicht, dass wir die Sprache der Gründe verwenden.

Diese Einsicht verdeutlicht, dass die Sprache der Gründe auch ohne metaphysische Verankerung und ohne objektive normative Tatsachen nicht nur einen klaren Sinn hat, indem sie einer Vielzahl verschiedener Zwecke dient, sondern für uns unabdingbar, praktisch notwendig und Teil der zweiten Natur ist. Das ist auch der Grund dafür, weshalb die Sprache der Gründe für uns in keinem greifbaren Sinn zur Disposition steht. Wer nun sagt, dass es für eine Handlung oder Überzeugung wirklich einen Grund *gibt*, sagt nicht zwingend etwas Falsches. Er macht einen *Zug* im Sprachspiel der Gründe. Er bezieht, anders gesagt, einen Standpunkt: er zeigt, welche Gründe er anerkennt, was er für wahr oder richtig hält, woran sein Handeln und Urteilen sich orientiert.

Wenn das so ist, zeigt sich auch, warum eine bestimme Art von skeptischen Zweifeln in Bezug auf Gründe gegenstandslos ist. Diese Skepsis zeigt sich sehr klar in der Reaktion Humes auf das Induktionsproblem. Weil Hume ein sehr anspruchsvolles Vernunftverständnis zugrunde legt und keine Gründe für die Gleichförmigkeit des Naturgeschehens angeben kann, die seinen Anforderungen entsprechen, sieht er sich gezwungen zu bestreiten, dass unsere Erwartung der Gleichförmigkeit des Geschehens überhaupt irgend etwas mit der Vernunft zu tun hat. Diese Skepsis erübrigt sich, wenn Gründe in einer sprachlichen Praxis verankert sind und der Vernunftbegriff Gründe einschließt. Ebenso erübrigt sich die Skepsis, die von der Möglichkeit lebt, wir könnten uns über unsere Gründe vollkommen täuschen. Irrtümer sind im Einzelfall möglich. Was ein Grund *ist* kann aber nie grundsätzlich von dem verschieden sein, was wir einen Grund *nennen*.

Was dürfen wir nun einen Grund nennen? Eine weitere wichtige Schlussfolgerung lautet: Gründe *sind* das, was sie zu sein *scheinen*, weil sie Tatsachen sind. Es ist ja gerade die Pointe der pragmatischen, antirealistischen Deutung der Rede von Gründen, nicht nur von dieser sprachlichen Beobachtung auszugehen, sondern auch zu ihr zurückzukehren und sie als sinnvoll auszu-weisen. Was wir einen Grund nennen und als Grund anerkennen hat in aller Regel diese grammatische Form oder lässt sich leicht in sie überführen. Der Gedanke der antirealisti-

schen Position ist deshalb uneingeschränkt mit dem Gedanken vereinbar, von dem wir ausgegangen sind: Gründe sind Tatsachen. Insbesondere sind sie *nicht* Meinungen oder Wünsche, auch wenn sie mit Meinungen und Wünschen in engem, vielleicht rationalem, Zusammenhang stehen mögen. Die pragmatische, antirealistische Deutung der Rede von Gründen untergräbt deshalb auch nicht die Unterscheidung zwischen normativen und erklärenden Gründen.

Man mag sich nun fragen, ob Wittgenstein selbst eine pragmatisch oder antirealistisch zu nennende Sicht in Bezug auf die Rede von Gründen vertritt. Viele Interpreten behaupten, er verhalte sich neutral, wenn nicht sogar kritisch gegenüber solchen Unterscheidungen. Wo sein Werk uns nicht dazu ermutige, Unterscheidungen wie die zwischen Realismus und Antirealismus für wenig hilfreich, fragwürdig, irreführend, philosophisch verwirrt oder sinnlos zu halten, entziehe es sich in jedem Fall der Zuordnung innerhalb dieser engen und künstlich geformten begrifflichen Raster.

So allgemein und kategorisch formuliert ist diese Einschätzung falsch. Wittgenstein argumentiert in der Philosophie der Mathematik, der Ästhetik, der Ethik, der Logik und der Philosophie der Religion wieder und wieder gegen die Vorstellung, die Sprache, die wir in diesen Zusammenhängen verwenden, sei in geheimnisvoller Weise auf die Wirklichkeit bezogen. Wittgenstein fordert uns auf, die tatsächliche Verwendung der Sprache zu untersuchen, um so gegen die Irrtümer und irreführenden Bilder, die uns die Struktur der Sprache aufdrängt, besser gewappnet zu sein. Zu diesen Irrtümern und irreführenden Bilder zählt auch die Vorstellung, die Sprache der Ethik, Ästhetik, Logik und Mathematik beziehe sich in geheimnisvoller Weise auf die Wirklichkeit. Aus dieser Vorstellung wird leicht der Gedanke, das, worauf wir uns mit dieser Sprache beziehen, sei eine geheimnisvolle Dimension der Wirklichkeit. Das ist die Vorstellung, die Wittgenstein unermüdlich analysiert, deren Quellen er aufspürt, und die er ganz entschieden verwirft.[10]

Nun trifft es zu, dass die wenigen Abschnitte der *Philosophischen Untersuchungen*, die sich mit der Rede von Gründen befassen, verschiedene Deutungen zulassen und insgesamt skizzenhaft bleiben. Was spricht also dafür, sie für eine bestimmte Deutung, zumal eine Deutung so kontroverser Art, in Anspruch zu nehmen?

Aus meiner Sicht sprechen dafür zunächst die sachlichen Gründe, die diese Sicht der Dinge konsequent und kohärent erscheinen lassen. Außerdem spricht dafür, dass das Problem, um das es geht, den philosophischen Problemen der Mathematik, Logik, Ethik und Ästhetik sehr eng verwandt ist und Wittgenstein hier deutlich Stellung bezieht. Es wäre überraschend und bedürfte einer guten Begründung, wenn man die antirealistische Deutung zwar für die Mathematik,

10 Vgl. dazu S. Blackburn, „Wittgenstein's Irrealism" und *Truth. A Guide for the Perplexed*, S. 129-36. Ich verteidige und diese Lesart in „Wittgenstein on Realism, Ethics and Aesthetics" im Hinblick auf das Spätwerk, insbesondere im Hinblick auf Ethik und Ästhetik.

Logik, Ethik oder Ästhetik zulassen wollte, nicht aber für die Rede von Gründen. Dementsprechend wäre es überraschend und bedürfte einer guten Begründung, wenn man Wittgenstein so lesen wollte, dass seine Reaktion auf die Frage der Gründe nicht zu seiner Reaktion auf diese anderen, aber offenkundig ähnlich gelagerten Fragen passte. Die radikale These, Wittgenstein sei im Hinblick auf diese Fragen weder eine realistische noch eine antirealistische Sicht in dem hier zur Diskussion stehenden Sinn zuzuschreiben, wird weder den Texten noch ihrem philosophischen Gewicht gerecht. Schließ-lich sprechen die Texte selbst eine recht deutliche Sprache, auch wenn es keine zwingende Lesart gibt. Der Verweis auf die sprachliche Praxis ist allgegen-wärtig. Nirgendwo wird auch nur angedeutet, dass diese sprachliche Praxis eine objektive, normative Ordnung reflektiert oder deren Existenz voraussetzt.

Diese Argumente sind meines Erachtens gewichtig, bleiben jedoch sehr abstrakt. Sie erfordern eine weit ausgreifende Analyse des Textes, die ich hier nicht ausführen kann. Doch es gibt eine wenig bekannte Passage, die in aller wünschenswerten Klarheit die Vorstellung von einer objektiven normativen Ordnung abweist. In Gesprächen mit Rush Rhees in den 40er Jahren sagt Wittgenstein:

> Wenn wir ein anderes ethisches System betrachten, mag es die starke Versuchung geben zu glauben, dass das, was *uns* die Rechtfertigung einer Handlung auszudrücken scheint, auch das sein muss, was sie dort rechtfertigt, während die wirklichen Gründe die sind, die angegeben werden. Diese *sind* die Gründe für oder gegen die Handlung. „Grund" heißt nicht immer dasselbe; und in der Ethik müssen wir uns davor hüten anzunehmen, dass Gründe in Wirklichkeit von einer anderen Art sein müssen als das, was als Grund angesehen wird.[11]

Wittgenstein wendet sich hier eindeutig und ausdrücklich gegen die Vorstellung, Aussagen über Gründe, moralische Gründe mit eingeschlossen, seien uns einfach gegeben. Zugleich macht die Passage sehr deutlich, dass diese Annahme natürlich ist und sich uns im moralischen Denken geradezu aufdrängt.

Wir neigen dazu zu meinen, die Gründe, die für eine Handlung sprechen, sprächen für diese Handlung, ganz gleich ob jemand die Gründe als Gründe anerkennt. Was diese Erwartung erfüllen kann, ist nicht mehr und nicht weniger als die Vorstellung, es gebe objektive Tatsachen in Bezug auf Gründe. *Wir*, so sagen wir uns gerne, sehen manche dieser Gründe *richtig*. *Wir wissen* etwa, so wollen wir sagen, dass es gute Gründe dafür *gibt*, Menschen mit anderer Religionsauffassung zu tolerieren, Fremde nicht auszunutzen, Frauen nicht zu benachteiligen, Kinder nicht zu manipulieren, und dergleichen mehr. Zugleich ist ganz offensichtlich, dass diese Einschätzungen nicht zu allen Zeiten von allen Menschen geteilt worden sind. Lange Zeit wurden sie nicht einmal von der Mehrheit

11 R. Rhees, „Some Developments in Wittgenstein's View of Ethics", S. 26 (meine Übersetzung).

der Menschen geteilt, und viele Menschen teilen sie heute nicht. Wenn uns nun jemand begegnet, der sagt: „Das seht *ihr* so; *wir* sehen es anders“, liegt der Schluss nahe, dass die Person wie die Gruppe, zu der sie sich zählt, eine verzerrte und falsche Vorstellung von den Gründen hat, die es tatsächlich gibt. Vielleicht ist sie für die Gründe, die wir so klar sehen, vollkommen blind; vielleicht ist sie für diese Gründe zwar nicht ganz blind, schätzt jedoch entweder ihr Gewicht oder das Gewicht der Gegengründe falsch ein und kommt deshalb zu einem falschen Urteil darüber, welche Gründe überwiegen.

In einer solchen Situation sind wir geneigt zu sagen, „dass das, was *uns* die Rechtfertigung einer Handlung auszudrücken scheint, auch das sein muss, was sie dort rechtfertigt“, auch wenn das „dort“ nicht erkannt wird. Und was Wittgenstein sagt, ist unmissverständlich: *das ist eine Illusion*. Es ist eine verführerische Illusion, aber es ist eine Illusion, weil ein falsches Bild von Gründen in diese Vorstellung eingeht und weil dieses falsche Bild die Vor-stellung trägt. Das Bild ist *falsch*, weil „die wirklichen Gründe die sind, die angegeben werden“, und das gilt „hier“ wie „dort“. Wittgenstein sagt ganz ausdrücklich: „Diese *sind* die Gründe für oder gegen die Handlung“, und die Folgerung ist nicht weniger deutlich. Wir müssen „uns davor hüten anzuneh- men, dass Gründe in Wirklichkeit von einer anderen Art sein müssen als das, was als Grund angesehen wird“.

Wenn das zutrifft, hat die These, dass was ein Grund *ist* nie ganz unabhängig von dem sein kann, was wir einen Grund *nennen* und als solchen *anerkennen*, eine viel größere Reichweite als es zunächst scheinen mag. Es ist nicht nur die vergleichsweise wenig interessante Behauptung, dass Gründe nur von einem Standpunkt innerhalb des Sprachspiels der Gründe als solche erkannt werden können; es ist auch nicht die zwar deutlich weiter reichende, aber noch immer nicht weit genug gehende These, dass was ein Grund ist nie ganz von dem verschieden sein kann, was wir einen Grund nennen, wenn damit nicht mehr gemeint ist, als eine Abhängigkeit der objektiven normativen Tatsachen von unserem Sprachspiel. Dann nämlich wäre noch immer Raum für die Annahme, es gebe objektive normative Tatsachen, und damit wären wir keinen Schritt weiter gekommen: wir hätten genau diejenige Annahme beibehalten oder wieder eingeführt, die Wittgenstein angreift. Die wichtige Einsicht, die Wittgenstein hier hervorhebt, ist, dass die sprachliche Praxis, um die es geht, eine *lokale* ist, und dass *deshalb* für eine robuste Objektivität von Gründen kein Platz bleibt.[12]

Wenn das zutrifft, werden die Hoffnungen, die eine große Zahl von ethischen Theorien leiten, unerfüllt bleiben. Doch wenn es richtig ist, dass was ein Grund *ist* nie ganz verschieden von dem sein kann, was als Grund *angesehen* wird, gilt es genauso für das, was *wir* als Grund ansehen, und das bedeutet: was wir als

12 Dieser Einwand trifft nach meiner Einschätzung insbesondere McDowell, der in der entscheidenden Hinsicht nicht Wittgenstein folgt, sondern von dessen Verständnis der sprachlichen Praxis abweicht, wenn er eine objektivistische Lesart der Rede von Grünnden vertritt.

guten Grund *ansehen*, kann nie ganz von dem verschieden sein, was ein guter Grund für uns *ist*. Niemand kann uns mit dem Hinweis auf andere Haltungen im Hinblick auf unsere eigene unsicher machen. Eine Wahrheit, die andere sehen, wir dagegen vollkommen verkennen, kann es nicht geben.

Was bedeutet das für das Verhältnis der Sprache der Gründe zur Wirklichkeit? Zweifellos ist die Sprache der Gründe auf die Wirklichkeit bezo-gen, nur nicht in geheimnisvoller Weise. Die Sprache der Gründe erfüllt einen Zweck, oder besser, eine unüberschaubare Vielfalt von Zwecken. In diesem Kontext des menschlichen Lebens hat sie ihre Berechtigung und ihren Sinn.

Wie die Verwendung der Sprache im Allgemeinen ist die Verwendung der Sprache der Gründe nicht willkürlich, sondern fest in unserer Lebensweise verankert. Sie ist veränderlich, und sie hat sich mit den Formen des menschlichen Lebens immer wieder verändert. Dessen ungeachtet gibt sie uns verlässliche Orientierung, und wir verwenden und verstehen sie als etwas, das uns wie etwas Natürliches gegeben ist. Die Sprache der Gründe entzieht sich der freien Verfügung des Individuums und gibt seiner Sprachverwendung Regeln vor. Deshalb gibt es Raum für Irrtum und Zweifel, Wahrheit und Falschheit, Erkenntnis und Ignoranz.

Diese Unterscheidungen greifen jedoch erst vor dem Hintergrund eines sprachlichen Standards, der seinerseits nicht auf Wahrheit oder Falschheit hin befragt werden kann. Was diesen Standard erzeugt, ist die Verwendung der Sprache der Gründe. Die Sprache der Gründe, die diesen Standard erzeugt, hat und braucht keine andere, angeblich tiefere oder festere, Verankerung als die Verwendung der Sprache; und diese hat und braucht keine andere, angeblich tiefere oder festere, Verankerung als das menschliche Leben.

Literaturverzeichnis

Blackburn, Simon: Truth. A Guide for the Perplexed. Allen Lane, London, 2005

Blackburn, Simon: „Wittgenstein's Irrealism". In: *Haller, R./Brandl, J. (Hrsg.):* Wittgenstein. Eine Neubewertung. Akten des 14. Internationalen Wittgenstein-Symposiums. 3 Bd., Hölder-Pichler-Tempsky, Wien, 1990, Bd. 2. S. 13-26

Brandhorst, Mario: „Wittgenstein on Realism, Ethics and Aesthetics". In: *Munz, V./Puhl K./J. Wang, J. (Hrsg.):* Language and World. Preproceedings of the 32nd International Wittgenstein Symposium. Kirchberg am Wechsel, 2009. S. 66-68

Davidson, Donald: „Actions, Reasons, and Causes". In: *Davidson, D.:* Essays on Actions and Events. Clarendon Press, Oxford, 1980. Essay 1

Hume, David: A Treatise of Human Nature. Hrsg. D. F. Norton and M. J. Norton. 2 Bd., Clarendon Press, Oxford, 2007

Hume, David: An Enquiry Concerning Human Understanding. Hrsg. T. L. Beauchamp. Clarendon Press, Oxford, 2000

Korsgaard, Christine: „Realism and Constructivism in Twentieth-Century Moral Philosophy". In: *Korsgaard, C.*: The Constitution of Agency. Essays on Practical Reason and Moral Psychology. Clarendon Press, Oxford, 2008. Essay 10

McDowell, John: „Might there be external reasons?". In*: J. McDowell:* Mind, Value, and Reality. Harvard University Press, Cambridge (MA), 1998. Essay 5

Preston, John: Wittgenstein on Reason. Blackwell, Oxford, 2008

Rhees, Rush: „Some Developments in Wittgenstein's View of Ethics". The Philosophical Review, 74, Januar 1965. S. 17-26

Scanlon, Thomas M.: What We Owe to Each Other. Harvard University Press, Cambridge (MA), 1998

Shafer-Landau, Russ: Moral Realism. A Defence. Clarendon Press, Oxford, 2003

Williams, Bernard: „Internal and external reasons". In: *Williams, B.*: Moral Luck. Philosophical Papers 1973-1980. Cambridge University Press, Cambridge, 1981. Essay 8

Williams, Bernard: „Postscript: Some Further Notes on Internal and External Reasons". In: *Millgram, E. (Hrsg.)*: Varieties of Practical Reasoning. The MIT Press, Cambridge (MA), 2001. S. 91-7

Wittgenstein, Ludwig: Philosophische Untersuchungen. In: *L. Wittgenstein*: Schriften, Band 1. Suhrkamp, Frankfurt/M., 1960

In What Sense Can an Evolutionary Meta-Ethical Sceptic Be Moral?

Gabriele De Anna
Gabriele.deanna@uniud.it
Università di Udine, Italien und Cambridge University, UK

Abstract/Zusammenfassung

Evolutionary meta-ethical scepticism is the view according to which there cannot be any justification for our ethical practices, norms, or systems, since evolutionary theory has made it clear that there is no room for moral values in the fabric of the universe. Several supporters of it have claimed that this form of scepticism leaves normative ethics untouched. I want to discuss this conclusion, and I try to argue that in fact meta-ethical scepticism has a bearing on normative ethics, and calls for a radical revision of common sense, naive normative practices. It is true that, as several supporters of this view want to claim, they may be moral, but this is only true if the word 'moral' is taken in a sense quite different from the pre-philosophical sense of common usage. My argument is that ethical conduct requires normative guidance, and that a meta-ethical sceptic about norms cannot be guided by the norms about which she is sceptic. Furthermore, I discuss how first order ethics is affected by the acceptance of evolutionary meta-ethical scepticism.

1. Introduction

Evolutionary meta-ethical scepticism (EMES) is the view according to which there cannot be any justification for our ethical practices, norms, or systems, since evolutionary theory has made it clear that there is no room for moral values in the fabric of the universe. Since it is a meta-ethical view, this form of scepticism might or might not affect first order, or normative ethics. Several supporters of it, though, have claimed that this form of scepticism leaves normative ethics untouched. In this essay I want to discuss this conclusion, and I will try to argue that in fact meta-ethical scepticism has a bearing on normative ethics, since it calls for a radical revision of common sense, naive normative practices. Ethical practices, norms, ans systems cannot have in the agency of a moral sceptic of this sort, if he is consistent, the same role they play in the agency of someone who is not sceptical on the metaethical level, or who does not have a metaethical view. Certainly, as several supporters of EMES want to claim, they may be moral, but this is only true if the word 'moral' is taken in a sense quite different from the pre-philosophical sense of common use.

In next section, I will try to define EMES, to distinguish two varieties of that view, and to present the arguments of those which deny its bearing on normative ethics. In the third section, I will suggest that ethical conduct requires normative guidance, and that a meta-ethical sceptic about norms cannot be guided by the norms about which she is sceptic. Hence, I claim, EMES affects the normative level. In the following section, I will discuss how first order ethics is affected by the acceptance of EMES. In the conclusion I will make some remarks about the general upshot of my argument for EMES, in the wider context of evolutionary ethics. In this paper, I will not question the truth of EMES, which I do not necessarily believe and which I only grant for the sake of the argument. What I want to show is that, if EMES were true, the denial of normative ethics or first order morality – in the sense in which they are taken by common sense – would follow.

2. Meta-ethical scepticism and normative ethics

Evolutionary meta-ethical scepticism arises on the backbones of a number of views, which are normally independently argued for. Here I want to present and discuss the mutual interrelation of four of these theses. For the sake of the argument, I will give for granted the truth of three of these. I will discuss the aptness of a fourth, and I will argue for improvement of it. The suggested modification will make the difference on the issue under scrutiny, i.e., the relationship between meta-ethics and normative ethics, while leaving the contribution of that these to EMES and the coherence between the two untouched.

First, evolutionary meta-ethical scepticism endorses the thesis that evolutionary theory can explain all complex phenomena of the universe, including the most sophisticated workings of the human mind (Joyce 2006, 190-199). All there is, is the result of causal interactions of proto-matter, which were determined by initial conditions and casual events, and progressively led to the formation of more and more complex entities. The complexities of ensuing entities constrain the ways in which they can interact with one another, and this progressively leads to the formation of patterns of interactions that appear to us as natural laws. The realms described by physics, chemistry, biology, psychology and behavioural sciences, ethics (and so on, if there is more), all emerged in this way. If this is the structure of the universe, any event which happens in it can be explained evolutionarily. This means that the explanation is going to be both physicalist and historical. It will be physicalist, since it will have to be compatible with the supervenience of all complex phenomena on underpinning physical events. Any event, no matter what its level of complexity, will depend on a sufficient set of physical causes. The explanation will be also historical, since it will require an account of how the physical set-up which made the physical cau-

sation of the complex event possible originated, and this will involve an account of the emergence in the past of relevant complex structures which interacted with each other in the production of that phenomena.

When I say that according to EMES evolutionary theory can explain everything, I do not intend this claim in the sense that EMES assumes that there are actual explanations for all possible complex phenomena or events. Of course, this would be plainly false, but my claim is more modest, in that it requires a twofold qualification of the possibility to which it refers. Firstly, I intend to say that EMES assumes that, on the ground of the acceptance of the evolutionary outlook of the world as the default metaphysical view, any possible event must have an evolutionary explanation, although this may not yet be available, given the current development of science. Knew science enough, we would have explanations for everything, and those explanations would be evolutionary. Secondly, the claim is not even that *one day* – when complete – science will be able to explain everything; it might well be that some facts about the past are beyond our epistemic reach. Still, evolutionary theory can explain everything in the sense that, were all the relevant facts of the past epistemically accessible, there would be an evolutionary explanation of every event. The thesis that evolution can explain everything is not epistemological, but metaphysical: evolutionism gives us the correct metaphysical account or reality, and thus all real explanations must ultimately be in or reducible to the terms of that metaphysics.

The second thesis endorsed by EMES is that moral language has to be taken at face value. When people disagree about moral issues, they use the sentences through which they make moral claims as statements about facts, not mere expressions of emotions. Furthermore, the facts in question do not involve the subjective responses of people involved, and thus moral sentences do not express facts about subjective attitudes. Rather, the facts expressed by moral language are taken to be objective matters, concerning an independent moral reality, which should be recognised by all agents involved in the circumstances. In sum, moral language is used to speak about objective facts and to persuade other speakers that some courses of action are objectively wrong, while others are objectively mandatory, and still others objectively possible (Mackie 1977, 20-25; Joyce 2006, 85-105).

The third element assumed by EMES is the existence and the explanatory priority of a human moral capacity. This is a conclusion that follows from the first two theses endorsed by EMES. If there can be evolutionary explanations for everything, and if the objectivity of moral language is a fact of human experience, there must be an evolutionary explanation of this fact. Moral language and moral systems cannot be explained evolutionarily, since they emerged in a span of time which is too short for evolution to have caused it. Hence, evolution must have shaped the human capacities which make such diversified and flexible linguistic and moral systems possible. This purports that humans must have a clus-

ter of cognitive capacities, which generates – in the presence of the right environmental conditions – moral language and moral behaviour and which must be evolutionarily explainable. This is the human moral capacity. Evolutionary explanations of ethics will have to focus on this capacity (Joyce 2006, 118-133).

Finally, the fourth thesis accepted by EMES that I need to mention is the existence of a *moral cloud*. Moral language, unlike other realms of language that also seem objective and normative at one time (aesthetics, etiquette, rules of games, etc.), is both *inescapable* and *authoritative* (Mackie 1977, 42-46; Joyce 2006, 57-64).

Ethical claims are inescapable in the sense that they can be applied to a person regardless of what her desires or wishes might be. This is common to morality and other normative systems, such as etiquette or aesthetics. If one brings the food to one's mouth with one's hands while sitting at a formal dinner, one can be reproached for that. The fact that one desired to do so, is no reason to withdraw the criticism. Similarly, if one does something morally wrong, for example steals one's neighbour's cherries from her tree, one is for that reproachable. The fact that one desires the cherries (or even the distress caused to one's neighbour by stealing them) is not a reason to suspend the disapproval. Moral statements, like other normative statements, hold good in themselves, if at all, independently from underpinning desires of the agents. Contrast these cases with someone who makes his watch in pieces for the sake of finding out how it works. This could seem strange, even if the watch was not particularly valuable. But the queerness disappears if one considers that that person's desire to learn about watch making was stronger than his desire for his inexpensive watch. Unlike these cases, moral, aesthetic and etiquette statements are inescapable. They can be applied regardless of the desires, which the involved agents may have.

Moral statements, though, are also authoritative, and this marks their difference from the normative statements of etiquette, aesthetics, etc. One should not eat with one's hands, or one should not play the violin out of tune. Not even if one desires to. But one can know that those rules apply, and decide to overlook the normative systems of etiquette or aesthetics altogether, and go on anyway. The case is different with moral rules: one cannot just decide to disregard the ethical system in which one is embedded. Morality is authoritative in a way that etiquette or aesthetics are not. One can object that deciding to overlook the norms of music, or the norms of a game might spoil the performance or the game and this might have its moral implications. This is true, but the possible moral implications do not depend on the violation of the norms of playing as such, but in the possible upshot that spoiling a performance or a game may have in certain circumstances. For example, the disappointment created in people who spent time and money to attend a concert or a match. The moral authority implicated here depends on the features of these surrounding circumstances, rather than in the violation of the norms of playing as such.

The four mentioned elements constitute the ingredients for an evolutionary explanation of ethics. If evolutionary theory can explain everything (first thesis), it must account also for the fact that moral language purports to make objective claims (second thesis), and that it is inescapable and authoritative (fourth thesis). The existence of an evolved human moral capacity (third thesis) is the evolutionary explanation of the character of moral experience (Joyce 2006). There is a rich literature about evolutionary explanations of ethics.[1] Several aspects of human behaviour (sympathetic and altruistic tendencies, competitiveness, kin preferences, etc.) are combined with known facts about the environments of our species' historical past to construct models which might explain why those behavioural traits were selected. Eventually, this explains the emergence of our moral capacities, which ground all possible ethical systems.

Although all this can explain ethics ("can" in the sense qualified above), it fails to justify it. Why should people stick to the rules that ethical systems and traditions furnish them with? Why should they abide by those rules even in case in which their desires would lead them to other directions? Evolutionary explanations seem capable to explain why we follow rules, and may contribute to explain why we follow the rules we follow rather than others. But has it anything to say about the reasons why we should abide by them? This is the meta-ethical question.

According to EMES, evolutionary explanations can answer the question about justification, and the answer is sceptical. The supporters of EMES claim that moral language purports to be objective and that we are always in the grip of the moral cloud. Thus, ethical discourse is cognitive, i.e. of a sort that could be justified. But evolutionary theory shows that *in principle* no justification can be given for our ethical practices and principles. Of course, other evolutionary ethicists claim that the question about justifications is misplaced, since evolutionary explanations of ethics have nothing to do with its justification (Kitcher famously held this view at a point: cf. Kitcher 1994; see also Boniolo 2006). But supporters of EMES have countered this claim by means of theses two and three above. If moral language is objective, it can be evaluated as true or false (Joyce 2006, 51-57). Although I tend to be convinced by the supporter of EMES on this point, I cannot discuss it here, and I ask to accept it at least for the sake of the argument.

Other evolutionary ethicists contend that the question about justification can indeed be answered, but the answer need not be sceptic: they propose naturalistic, evolutionary accounts of justification (Campbell 1996, Casebeer 2003, Dennett 1995, Richards 1986). Supporters of EMES have argued against this view at length as well, both by criticising attempts to naturalise ethical justification (Joyce 2006, Ch. 5), and by arguing in favour of scepticism (Joyce 2006, Ch. 6,

1 For a survey, see Joyce 2006, Ch. 4.

Sober 2006). I will not be able to discuss the criticisms to the naturalisation of ethical justification, and I will only mention, in what follows, the main arguments for scepticism. Again, I accept only for the sake of the argument that the supporter of EMES is right in arguing against these attempts of justification. I mention these debates, here, in order avoid possible misunderstandings and make it clear that my arguments do not engage with those alternative possible views about evolutionary ethics.

Before mentioning the argument for EMES, I must note that there are two main forms of evolutionary metaethical scepticism, two versions or EMES. According to the *stronger version* of EMES, our moral language purports to be objective, but it is systematically erroneous (Mackie 1977, 15 and 35; Ruse 2006). It speaks as if there were objective moral values, or truth-makers of moral claims, but in fact there is nothing of that sort in the fabric of the world. All moral statements are thus false. The *weaker version* of EMES, claims that we cannot have any justification for the truth of moral claims. They could be true, but there is no way to find out whether they are true or false (Joyce 2006, 223). We can now turn to the argument in favour of each version of EMES.

Famously, strong EMES is grounded on two arguments put forward by Mackie, the *argument from relativity* and the *argument from queerness*. The argument from relativity contends that there is a great variety of diverse and incompatible moral systems, both across different cultures and within a particular culture. This would not be the case, if there were objective moral facts to which moral statement referred. Therefore, there are no such facts. The argument from queerness has a metaphysical version and an epistemological version. The metaphysical version claims that moral facts, if existed, were unlike anything existing in the natural world. The epistemological version claims that moral facts, if existed, could not be known through any of the natural cognitive faculties we have. The conclusion of both arguments (i.e., from relativity and from queerness) is that there cannot be any moral facts. If there cannot be any moral facts, statements presupposing the existence of those facts must be false. Hence, strong EMES must be true.

Weak EMES is based on the consideration that the human moral capacity and ensuing ethical systems have been selected since they increased fitness. (It does not matter, for our purposes, whether this is fitness of individuals or groups). Fitness, however, is in no way dependent on the truth of the moral beliefs which might have served it and were hence selected. Things are different in the case of doxastic beliefs: the fact that 2+2 really equals 4, for example, was crucial for someone's survival in an environment in which calculating the actual number of predators running after him was essential for escaping. Unlike our epistemic capacities, our moral capacities are not a reliable process for the formation of true moral beliefs. Therefore, we have no way to know whether any and, in case, which of our moral beliefs are true: we have to suspend our judgement about

each of them. This is scepticism in the old, classical sense: "no moral judgements are epistemically justified" (Joyce 2006, 224).

We can already note that the distinction between the two forms of EMES is not trivial for our purposes, since the two views might have different upshots on the normative level. Indeed, the weak sceptic is uncertain about the epistemical status of his moral judgements, and this opens the possibility that he might *accept* them, even if he is not in the position to *believe* them. Accepting that *p* is an epistemic stand that can be hold in cases in which there are not enough grounds for believing that *p*, but there are other non-epistemic reasons in its favour. For example, if I lend my favourite book to my best friend, and it is stolen from him in dubious circumstances, I might accept his awkward explanation even if it is hard to believe, just for the sake of safeguarding our friendship. Accepting is more subject to the will than believing. On the other hand, if the epistemic status of *p* is not uncertain, and there are reasons to believe that *p* is false, accepting *p* would be an irrational act. *Prima facie*, we can grant the weak sceptic the possibility to accept – on the normative level – moral judgements he has no reason to believe, whereas the strong sceptic should not accept moral judgements, since he believes that they are false. This distinction will have to be considered in the discussion to follow.

Both weak and strong sceptics have claimed the same point about their attitude toward normative ethics: meta-ethical scepticism does not entail the rejection or abandonment of normative ethics. This follows from the conjunction of the above-mentioned theses. Normative ethics concerns first order ethical discourse, i.e. the discourse concerning the application of a ethical systems to action. We all act within an ethical system of moral beliefs and associated dispositions, habits, and attitudes, since we all are subject to the moral cloud (fourth thesis) and we all use moral discourse objectively (second thesis). That we do this is the necessary consequence of us having the human moral capacity (third thesis), and we all have this capacity because of the way in which we evolved (first thesis). Even the supporter of EMES cannot help being in the grip of his biology, ad thus he objectifies and keeps reasoning within her moral system, no matter what her second order, meta-ethical beliefs might be.

These are some examples of famous statements of this view. John Mackie wrote: "what I am discussing is a second order view, a view about the status of moral values and the nature of moral valuing, about where and how they fit in the world. These first and second order views are not merely distinct but completely independent: one cold be a second order moral sceptic without being a first order one, or again the other way around" (Mackie 1977, 16). Another example can be taken from Michael Ruse: "once we recognize [that there is no justification for our moral norms], we see the sentiments as illusory – although, because we objectify, it is very difficult to recognize this fact. That is why I am fairly confident that my having told you of this fact will not now mean that you

will go off and rape and pillage, because you now know that there is no objective morality (Ruse 2006, 23).

Richard Joyce has a more articulated view. He recognises that EMES amounts to a debunking of morality, and he is initially ready to bite the bullet: "if your thinking on some matter presents itself as a faithful representation of the world but in fact there are no grounds for supposing that it is, then, by epistemic standards, its being undermined is a *good* thing" (Joyce 2006, 222). EMES should lead to scepticism on the normative level. But then he feels the pressure of someone who might find such a conclusion appalling, and he surprisingly concedes, in the relatively short space of the conclusion of his book, that EMES does not need to lead to normative scepticism, because we can still *decide* to keep our moral beliefs, or possibly we even *should* keep them, in order to keep our motivational mechanism serving the function for which it was selected. "We can go further and say not merely that people *could* carry on allowing moral thoughts ad moral emotions to have some motivational influence in their lives, but that many individuals *should* do so". He can maintain this view since he holds weak EMES, and, as we have seen, this is compatible with the possibility of accepting epistemically unjustified statements, if they are not known to be false.

3. Two ways of following a rule

In what fallows, I want to challenge the claim that EMES leaves normative ethics untouched. My complaint has mainly to do with the fourth thesis supported by the EMES, i.e. what Joyce called the moral cloud. I do not want to protest against the facts about ethical behaviour which are grouped under that label. To that extant, I believe that I will leave untouched the theoretical contribution of the fourth thesis to the characteristics of EMES. What I would like to suggest is that a supporter of EMES should consider other, related facts about moral behaviour, and that these have their consequences on the relationship between meta-ethical scepticism and normative ethics.

In the light of those new facts, it may be questioned whether there are any set of features of our agency which jointly constitute what may be correctly described as "the moral cloud". That expression is indicative. It aims at conceptualising together a number of facts concerning our experience of moral behaviour and moral language, i.e. the fact that moral judgements are inescapable and authoritative. At the same time, though, it does that but suggesting that moral judgements exercise a sort of force or constriction on us, as if they mesmerized us with their contents and we where forced to follow them by the moral capacities which were wired in us throughout the process of evolution.

A supporter of strong EMES, can hold the thesis of the sharp separation between normative ethics and meta-ethics, only if he also holds that our moral capacities work in us a little like our perceptual capacities. We cannot help seeing what we see, even if we know that our senses are stimulated by deviant causes. For example, when I experience a perceptual illusion I cannot stop being in the grip of it, even if I know that it is an illusion. Moral capacities need to be akin to this, if I can be in the grip of moral judgements, even if I know that they are false. And this is precisely what the thesis of the sharp separation between normative ethics and meta-ethics presupposes.

Things are more complicated, but ultimately identical with weak EMES. In this case, the separation thesis can be maintained even if some connection between our ethical response mechanism and our belief formation system is granted. The supporter of weak EMES grants that, were we to believe that moral judgements are false, we would not be in the grip of them. But it is allowed that for some non-epistemic reasons we may accept a judgement even if we have no reason to believe it, and we can be motivated by it just as if we believed it. Our "moral mechanism" is seen as dependent on our epistemic mechanism, but in ways, which give us the freedom to play with it and allow us to direct it to whatever aims we might pick. (I say "pick" rather than "choose", since – by hypothesis – the selection procedure we are dealing with at this point is not epistemically guided). Thus, we can be sceptical about the truth of the statements included in a normative ethical system, and still be motivated by them, if we decide to embrace them for some non epistemic virtue they might have.

I think that this way of conceiving of the inescapability and authority of moral judgements as constituting the moral cloud is mistaken, since it considers some of the facts to be explained (the inescapability and authority of ethical discourse), but leaves others out. We do not follow moral judgements in the mechanical manner envisaged by strong EMES, nor in the more flexible and partially epistemically determined manner suggested by weak EMES. The thesis that I want to support in this section is that it is a fact of our moral experience that we act morally to the extent that we let ourselves to be guided by moral judgements *on the ground that we believe in them*. If we consider this fact together with the inescapability and the authority of moral judgements, we should not accept the existence the moral cloud, i.e. the view that moral judgement impose themselves in us *blindly*, i.e. *regardless of our awareness of their lack of justification,* but we should endorse a more complex account of human agency, and that – I will claim in next section – does not allow for a sharp distinction between first and second order ethics.

The fact of our moral experience that, I claim, EMES fails to consider is best discovered if we look at our agency from within, i.e. from the point of view of the agent who evaluates moral judgements – together with other relevant beliefs – in the process of making a decision. This could be thought of as an unaccept-

able move. Considering agency from within involves a certain dose of introspection, and whatever results this analysis could give, they are certainly not empirically observable facts. This is problematic, since EMES presupposes, in the first thesis, an empirical approach to reality, and here I am not engaging in a discussion of that presupposition, the truth of which I have granted for the sake of the argument. I am trying to show that EMES affects the level of normative ethics *on its own grounds*, and thus I cannot challenge its presuppositions.[2] Since I am discussing the implication of EMES, not its assumptions, I cannot assume views which are inconsistent with its empirical perspective.

However, I do not think that turning to introspection might necessarily entail the denial of an empiricist perspective. The empiricist assumption beyond EMES is committed to the evolutionary explanation of complex phenomena, which involves the possibility of offering scientifically acceptable accounts of the relevant facts, and an account about the possibility – at least in principle – of reducing them to evolutionarily significant unities of selection. This will involve also a scientific account and evolutionary explanation of mental and psychological phenomena. There are two main ways in which psychological facts can be treated scientifically: behavioural analyses, and personality accounts (cf. Shrout and Fiske 1995). I cannot get involved in the discussion about the merits of each, but I cannot see why the supporter of EMES should not be satisfied with either, unless he had other reasons, quite independent from the assumptions of EMES. And I believe that the facts I am going to point to through the introspective analysis to follow could be empirically investigated by means of both methodologies. The cases I present could be object of empirical investigation, and thus they could be empirically supported or disproved. The use of introspection in this case, thus, could be seen just as a means to focus our attention to some aspects of our empirical external experience, which could then be investigated, by other, empirical means. To this extent, I do not think that my appeal to introspection is inconsistent with the empiricist assumptions of EMES.

Let us then start the analysis from the point of view of the agent. When supporters of EMES describe our moral practices and our deployment of moral language, they refer to "moral judgements". This is ambiguous: it can refer to judgements about a particular situations ("I cannot steal my neighbour's cherries"), to judgements about a situation types ("it is wrong to steal things from neighbours"), or to more general moral principles ("stealing is wrong"). It must be recognised that this ambiguity is not harmful to EMES: all those examples are sentences, which, in our linguistic practice, can be observed in a realm of

2 I am indeed questioning the existence of a "moral cloud", the fourth thesis; but, as I said above, I grant all those fact about moral behaviour that Joyce called by that name, and that contribute to EMES. My criticism questions the aptness of that label in the light of the other facts about our agency, precisely those under discussion, and thus does not deny the grounds of EMES, but rather call for an addition to them.

discourse that we would call "moral". There is nothing wrong in putting them all under the same label, "moral judgements". However, from the point of view of the agent, the ambiguity of the expression "moral judgement" is more significant. The realm of discourse that we would call moral embraces a number of *immediately* visible practices, uses, and institutions. Linguistic practices – including "moral judgements" – are among these. But underneath the *immediately* visible practices, uses, and institutions, there is an array of mental activities, which make those practices, uses, and institutions work in the way they do.[3] If we pay attention to this, we can note that different kinds of moral judgements play very different roles in moral agency.

A moral action does not necessarily involve the entertainment of a moral judgement in the process of thinking of the agent which leads to deliberation. Sometimes one just does the right thing unreflectively, i.e. without entertaining a thought embedding a judgement about that action. I see my neighbour's juicy cherries and this immediately generates a desire for cherries in me, maybe even for *those* cherries. However, I do not even think about stealing them, and I do not steal them without forming the thought that this would be morally wrong. One could protest that this is just a case of doing the right thing accidentally, i.e. not because it is the right thing. Had my desires been different (e.g., were my desires for those cherries stronger), I would have picked the cherries. Not stealing was not something I did intentionally, and thus there is no moral worth. Surely, this is a possible scenario, but it is not the only possible one. The possibility I am thinking of is that in which I did not even think about doing the wrong think since I am not that kind of agent: I notice niche cherries, but if they do not belong to me they do not move my desires. That I am that sort of agent can only be clear by seeing how I react in similar conditions, which I slightly different in some relevant respects. Sometimes, when I am really hungry and the cherries look really nice, I might feel a stronger desire, and end up entertaining the thought of taking them. But – given the sort of agent I am – the very thought gives me a sense of distress and guilt. The possibility of taking the cherries is to me a *temptation*, something which I want to do, but goes against "another part" of me. A process of reasoning, which will also involve moral judgements, eventually begins. At the end of the process, I overcome my desire and restrain from stealing the cherries. The fact that I an this sort of agent suggests that when I restrain from stealing without even thinking about it I do not restrain accidentally.

3 I claimed that the relevant mental activities lay underneath *immediately* visible practices, uses and institutions, since I leave it open that some of those mental activities might be *indirectly* empirically accessible, by structuring the visible aspects of moral life. As I said above, it is this possibility that makes my turn to introspection acceptable to an empirically oriented thinker.

I just do – without thinking about it – what I would have wanted to do,[4] had I thought about it (under the pressure of temptation, but other reasons could also be envisaged).

All this might seem suspect. I am trying to persuade the reader that an agent might do a moral action without entertaining a moral judgement, and I support my claim by appealing to how that agent would reason about a similar action in a case in which his desires are slightly different. This seems just a dispositional account of moral action, and dispositional accounts of moral agency point to the theory of virtue. However, the theory of virtue is dubious since the empirical results of social psychology and ensuing situationalism seem to suggest that there are no strong character traits, as virtue theory supposes (Doris 2002). Although I am not completely convinced that empirical results make a case for a version of situationalism that is incompatible with any plausible version of virtue theory, I think I do not need to address this debate here, since my appeal to counterfactual thinking about the behaviour of an agent does not imply that he has a disposition which resists across a diversity of eliciting conditions which is countered by the empirical evidence (Doris 2002, 22-3). I believe that the appeals to counterfactual situations that I have made so far – and those that I will make in what fallows – do not presuppose the existence of strong character traits, and can be acceptable for both a virtue theorist and a situationalist.

Let us then continue the analysis from the point of view of the agent. As mentioned, sometimes I might have a desire to do something, but feel a sense of distress about it, as in the case of temptation. I desire to overcome the distress and I start conceptualising the object of my desire, by looking for principles regulating the situation and trying to square my desires with my judgements. Thus, I entertain the thought "Taking one's neighbour's cherries is wrong". This might just be enough to stop my desire to steal the cherries,[5] but it might not. My desire for the cherries leads me to think that after all "My neighbour is a jerk and often he does not even pick his cherries anyway". My conflict of desires turns into a conflict of moral judgements: "would this be an action of stealing something from my neighbour or an action of taking something which my neighbour do not really care about anyway"? I try to evaluate the options open ahead of me by deploying more and more general kinds of moral judgements. Eventually, I

4 What really matters here is what I want to do when I reflect about the case, not what I could actually do, when failing to do what I want (*akrasia*).

5 Even if it is, this case can still be empirically distinguishable from the case in which I do the right thing without even thinking about it. A personality test could spot the difference between the two mental activities, and behavioural observation can detect signs of hesitation, sweating, and facial movements. As mentioned above, this introspective analysis aims at highlighting facts which could give rise to empirically testable hypotheses. The same point should hold also for subsequent steps of the analysis, and thus I will not repeat it.

come to subsume my case under few general moral principles that I believe to hold true. 'Stealing is wrong, and this would just be stealing'. I finally determine. And I do not steal.

This description of a plausible deliberative process suggests that, even if one does not always turn to norms, one's moral conduct is ultimately shaped by norms.[6] In our conduct, we have an inclination or a tendency to follow norms and to justify what we do by deploying norms. Even actions in which we do not consciously engage in a deliberative process involving relevant norms, are our actions (rather than things which we do accidentally) if – given the sorts of agents we are – we should and could have justified them by turning to norms. But norms can play the role of shaping our processes of deliberation, the formulation of our more particularised moral judgements, and ultimately our actions only if they have enough hold on us to constitute a possible hinder for our desires. When we act morally – i.e. out of duty, not out of mere desire – the moral norms we deploy in our reasoning must be non-negotiable for us, we need to trust them. In other words, they must guide our actions.

This is the new fact of our moral agency to which I wanted to point to, and which combines with the other two claims about moral judgements which the supporters of EMES embrace, i.e. the inescapably and the authority of moral judgements. This combinations leads to the *requirement of normative guidance*: the moral action of an agent must be guided by moral norms, which cannot be avoided (inescapability), and are not disposable for the sake of other reasons, such as desires, utilities, aesthetical considerations, etc. (authority); furthermore, no matter how complex the deliberative process might be, those moral norms need to be deeply trusted by and non-negotiable for the agent. To appreciate the implications of this requirement for the relations between EMES and normative ethics, we have to pay further attention to the notion of normative guidance.

Normative guidance has been discussed by Peter Railton (2006), and his account can be taken in here, even if some aspects of his analysis will have to be discussed and modified below. Railton notes that, intuitively, conduct *C* is guided by norm *N* only if *C* is in accord with *N*, but this is unsatisfying in a number of ways and needs refinement. Firstly, one could aim at following a norm, but fail. Whatever he does, in this case, is still guided by that norm. What counts is

6 A particularist would certainly object to this claim, but I have reasons to discontent against particularism: the particularist grants that the same (kind of?) reason might bring different weights in different situations (CITA), but must allow that something in each situation makes the difference, and I cannot see how this can be stopped from lading to the possibility of specifying how differences among situations affects the weights that that reason would have across the different cases. However, this would lead us back to norms, o, at least, universal characterisations of the weights of reasons. I do not have the space to spell out my complaint and its implications here, but I can at least ask the reader to grant me credit against the particularist for the sake of the argument.

that the conduct of the agent is informed by that norm over a certain span of time, even if most of the times he fails to abide by it. Second, conformity of *C* to *N* must be one of the purposes of the action, that is *N* must be a reason to *C*. We would not say that an action is guided by a certain norm if that norm plays a purely instrumental role for the achievement of another end. That end would then be guiding the action. Third, very often people abide by norms out of habit or education, without consciously entertaining a thought having the norm as its content. Normative guidance can be explicit or implicit in this sense. Under this respect, having a regulative role is sufficient for normative guidance: actions are guided by a norm if there is a mechanism that keeps those actions conform to the norm, no matter whether the agent is conscious or unconscious of it. Fourth: norms are different from plans, in that, even if they have no consequences and do not lead to sanctions (nor even to internal sanctions, such as a sense of discomfort), the agent has a tendency to make up for failures. All these considerations lead Railton to the following definition of normative guidance:

> (NG) Agent *A*'s conduct *C* is guided by norm *N* only if *C* is a manifestation of *A*'s disposition to act in a way conductive to compliance with *N*, such that *N* plays a regulative role in *A*'s C-ing, where this involves some disposition on *A*'s part to notice failures to comply with *N*, to feel discomfort when this occurs, and to exert effort to establish conformity with *N*, even when the departure from *N* is unsanctioned and non-consequential.

For our purposes, the crucial point of this is that when *N* is required to play e regulative role, in Railton's analysis, *N* is not an instrumental reason for some further purpose or end, but it is (one of) the purpose(s) of acting in certain ways. A norm can be followed as a means to gain something else, but then it has no e regulative role, it plays an instrumental role, and it is aimed at a further end *E*. Were the end *E* to be attainable also by other means, let us say by following another norm *N'*, *A* could follow *N'*, just as well as *N*. This means that *A*'s conduct *C* is not regulated by *N*, or by *N'*, but by *E*. Since moral judgements are authoritative, they cannot be given up for the sake of other reasons, and thus, *from the point of view of the agent*, they should guide his actions. For an action to be a moral action, it does not suffice that the agent follows moral norms when he performs it: it is also required that the agent is guided by it. More generally, an agent endorses a moral normative system only if he is normatively guided by the norms belonging to that system. This is what I called the requirement of normative guidance.

In order to be guided by a norm an agent needs to believe that that norm is justified, or at least she does not have to believe that it is false or that lacks justifications. If she believes that it is false or that it is not justified, she would not endorse it for its own sake, but for some other reason. In that case, though, the norm would have a merely instrumental role for her, and thus she could not be regulated by it. Hence, we accept moral norms and are guided by them, only if

they are norms that are inescapable and authoritative *to us*, and thus we follow them for their intrinsic worth, i.e. because we believe that they are justified, or at least we trust them. (All the "at least" qualifications of this paragraph are meant to address the case of agents which are not meta-ethical sceptics because they do not have a meta-ethical view at all, and they never worried about justifying the moral norms they trust).

I would like to argue that a supporter of EMES cannot but fail to meet the requirement of (NG), for any moral norm *N*. And since meeting (NG), for each norm belonging to an ethical system, is necessary for an agent to endorse that ethical system as such, EMES has important consequences on the level of normative ethics, for the supporter of it. Let us consider why the supporter of EMES fails to meet (NG). We must start from the supporter of strong EMES.

The problem with strong EMES is that it assumes a too simplified view of our moral agency. The inescapability and authority of moral judgement is not just a mater of moral cloud: it is the result of the role that certain norms play in the architecture of our agency, of the part they play in shaping our dispositions and our habits, in the face of our past history and of current rational concerns. When this is taken into account, the sharp distinction between first and second order ethics cannot be maintained. The supporter of EMES must also be sceptical on the normative level.

The supporter of strong EMES could insist that he can indeed be normatively guided by norms he believes to be false. (NG) suggests that normative guidance can be unconscious, if there is a mechanism which makes the agent keep conformity to the norm through a process of feelings of discomfort. Doesn't this imply that one can be normatively guided regardless a lack of beliefs in the relevant norms? The answer is the negative. Such mechanisms cannot but be implemented in our habits, and humans habits are plastic. For this reason a mechanism can normatively guide action – as NG suggests – only in contexts in which the relevant norm keeps playing a rational constraint on the agent.

I will try to make this point by considering the case of Luca, who, at some point in his life, realises that there are no objective values and that all moral judgements he had so far endorsed, but also all those he had not endorsed, are in fact false. At this point, he will keep having the dispositions he always had toward the norms he formerly endorsed, for example stealing cherries. These norms are the result of his upbringing and of acquired habits, and they cannot be easily given up. The argument of the supporter of strong E<ES could then be that Luca still meets (NG) for each of the norms that he previously believed in. This seems to show that a conduct may be regulated by a norm even if the agent is not aware of it, or does not believe in it.

I think that this conclusion does not follow. Let us pay further attention to the example. At some point, Luca feels a strong desire for cherries. He knows that the only way to get hold of some cherry is to steal them from his neighbour, and

he knows that no one will notice him (it is the day of a big game, and everyone in the neighbourhood is watching television). He also knows that he will feel discomfort for a while. But he knows that this discomfort is just the result of his upbringing, or maybe a hardly wired upshot of evolution: there is no justification for the moral judgement "do not steal these cherries". Before endorsing scepticism, in cases of temptation, he used to turn to a moral judgment like this, and eventually to moral norms such as "do not steal", in order to determine the courses of actions to take. Now he cannot do that: he knows that there is no point in those rules. He can just follow whatever desire is strongest in him. When he is hungry enough and the cherry juicy enough, e would give up, and go for the cherries. And after repeating actions of this sort he would even end up changing his habits, and stop feeling discomfort. Of course, in the old times he could have failed to conform to the norm "do not steal cherries", for *akrasia*. But such failures were still manifestations of a disposition to conform to that norm, i.e. the norm still guided him. After he endorsed scepticism, Luca could conform his conduct to the rule "do not steal cherries", but this would be for the desire of avoiding a sanction, for the desire to obtain the advantages deriving from the trust of others, etc. His following the rule would be instrumental and his conduct would not be regulated by that rule, i.e. it would not be normatively guided by that rule. The point is that there might be a mechanism inducing an agent to conform to a rule, but the rule needs to keep playing a supporting and reinforcing role for that mechanism if the actions elicited by that mechanism have to be cases of actions which are regulated by that norm, rather than cases in which conformity to the norm plays a purely instrumental role.

Things are more complicated for the weak sceptic. As we have seen above, the weak sceptic can accept moral judgments in face of their lack of epistemic justification, for other properties which they might have. If Luca were a weak sceptic, he could still hold that stealing cherries is wrong, just for the sake of being part of the moral community, and nourish all the dispositions and the habits, which can make that easier. Would he then be guided by the norm that stealing cherries is wrong, i.e. would he meet (NG)?

Railton faces the problem in his analysis, and his answer seems to be the positive. This is the part in which I wish to modify his proposal, as mentioned above. He notes that the epistemic distinction between accepting p and believing that p – where p is a proposition – crosses over onto the normative realm, in the distinction between accepting N and endorsing N – where N is a norm. He considers the example of a person who had a normal, well-balanced, moral upbringing, who eventually converts to a morally strict religion, which prescribes mutual scrutiny among the faithful and public accusation of transgressors. Given his conversion, he endorses those strict norms, but, given his upbringing, he find it hard to abide by them. He is not hard to him-self, and he accepts what he is and what he was, he accepts also his upbringing; thus, sometimes he does not do

what the norms he now endorses require. These are occasions of *akrasia*, but also manifestations of the fact that his conduct is still regulated by the norms he was brought up with and which he accepts, even if they clash with those which he now endorses. Eventually, he may come to realise that the new views he had endorsed after his conversions were wrong, and go back to his previous outlook. This proves, according to Railton, that that person was guided by the norms received during his upbringing also in the period when he merely accepted them, and endorsed others. Hence, normative guidance is compatible with both acceptance and endorsement of norms. Since the supporter of weak EMES claims that moral beliefs are epistemically unjustified, but can be accepted on other, non-epistemic grounds, he can accept some moral judgments and thus be guided by them, even if he does not believe in them. If this is right, normative guidance of moral norms is possible for him. He can embrace a first order ethical outlook, while being a metaethical sceptic.

As I mentioned, I do not think that Railton's analysis of normative guidance should be followed under this respect. My complaint is with the way in which he describes his example and with the theoretical consequences he wants to draw from it. In his example, Railton describes the person in question as having *endorsed* new norms while going on merely *accepting* his old ones. This is however contentious. If the person in question, while embracing the new faith, remains true to his old self, keeps accepting what he is by upbringing, i.e. he does not really die to be reborn into a new life, then he is not merely accepting his old norms. He keeps seeing some truth in his old outlook of the world. Hence, he does not merely accept the old norms, but goes on endorsing them, at least some of them. Were his conversion complete, we would expect him to reject all which is dependent on his old self. In the example discussed by Railton, we seem to be presented with an incomplete or partial conversion, and the character of the story seems to be trapped in a contradictory situation in which he holds on to two partial views of the world which belong to two general, incompatible outlooks. He embraces the new faith, but with reservations. He renounced to his old self, but not to whole of it. In his contradictory situation, he uncomfortably feels the pressure of conflicting norms. Indeed, in Railton's example, he resolves the conflict and restores consistency by giving up the new faith he had temporarily and partially embraced. Thus, this example does not show the possibility of normative guidance of both endorsed and accepted norms, since it involves only cases of endorsement.

I want now to argue that my complaint does not depend on specific features of Railton's examples, and can be generalised to all possible examples or situations. This will lead me to my theoretical complaint to the endorsing *N*-accepting *N* distinction. What I contest is that a case can be presented in which an agent accepts one or more norms non-instrumentally and without endorsing them. Indeed, one could accept a norm one does not endorse, for some other rea-

son. For example, for the sake of being accepted in a community, or for the desire to be trusted by others, even if one thinks that there is no justification for the norm as such. But Railton would not consider following rules in this way as cases of normative guidance. In all these cases, the norms in question are accepted only instrumentally: the are accepted for whatever reason makes them appealing. However, (NC) excludes all cases of following rules instrumentally. On the other hand, if one accepts a norm for no other reason than the norm itself, one accepts it for its own worth, i.e. one endorses it. I am questioning that there might be a middle way between endorsing a norm and accepting it merely instrumentally: either the reason for accepting the norm is the norm itself, and the agent can be normatively guided by it, or it is different from the norm, and the norm can guide then agent only instrumentally. There is no middle way.

This explains, I hope my claim that the person of Railton's example is said to be a convert, but he remains non-instrumentally attached to norms which he previously accepted and which do not fit in his new world view. I would like to suggest that, even if we grant the plausibility of this example, we cannot infer from it the existence of a state between endorsing and accepting instrumentally. The fact is that, from the standpoint of the new religion endorsed by the convert, the norms he grew up with turn out to be wrong, but he still hold on to them non-instrumentally. How is this possible? Since he does not accept those norms for some other reasons, i.e. instrumentally, he must accept them for their intrinsic worth. But he cannot endorse them on the ground of his new religion. Still, he must have the conceptual resources to appreciate some good in them. There seems to be a contradiction, but this needs not be a sign that the example is logically impossible, nor that our (Railton's and my) conceptual framework is inconsistent. The inconsistency can simply be in the beliefs framework of the convert: he is not completely converted to the new religion, and still maintains some aspects of his previous world view. He is hesitating between two competitive and inconsistent world-views. He values aspect of each, but cannot embrace unconditionally either of them in its totality. Thus he endorses some of the norms which can be grounded in each framework, but are inconsistent with the other framework. Eventually, he resolves the inconsistency by going back to a full endorsement of his old world-view. But we do need to grant him any state between endorsing and accepting instrumentally, in order to make sense of his behaviour.

If there is no middle way between endorsing and accepting instrumentally, the supporter of weak EMES cannot be normatively guided by moral norms. When someone accepts a norm he has no reason to believe, he does not accept it in virtue of its value, but in virtue of some other value to which it might conduce. This means that he accepts it instrumentally. Were the norm to stop conducing to that value, he would lose any interested in it. Let us again consider Luca desiring his neighbour's cherries while everybody is watching the game. He might accept the rule that stealing cherries is wrong. That rule fosters good

neighbour relationships, allows one to go to work without fear of finding no cherries on his tree since the whole neighbourhood protects them by abiding to this rule. But now he knows that violating the rule will have no consequence: no one will see him, his neighbour will not notice that few cherries are missing, he is certain that stealing them will not make him feel distressed than he already is for the desires of cherries, he knows that he will not take on any uncontrollable vice. Why should he stop himself? He should not. Actually, he should get them, given his actual and future-projected desires. He accepted the norm 'do not steal' for some of its non-epistemic virtues; let us say its social bearings. Now those social results would not be endangered by a violation of that norm, and so there is no reason to follow it. This means that accepting the rule is purely instrumental: Luca was not normatively guided by it. The supporter of weak EMES cannot be guided by moral norms he knows to be unjustified.

A meta-ethical sceptic, no matter whether weak or strong, could now question the requirement of normative guidance entirely, on two grounds, at least. First, he could complain that it is not clear whether that requirement is psychological or normative. Second, he could insist that an agent could hold onto a moral normative system for purely instrumental reasons, and acquire a number of corresponding habits, to the point that the systems becomes a second nature to him. He would thus be in the grip of that moral normative system, regardless of his second order scepticism. I will address the first issue here, and leave the second for the next section.

The first complaint is that it is not clear whether the requirement of normative guidance is psychological or normative, i.e. whether it is a description of how agents deploy norms or a requirement concerning the way in which they *should* deploy them. I introduced it through a psychological analysis from the point of view of the agent, but I then seem to claim it as a norm for agency, since I call it a requirement. Either way, the meta-ethical sceptic can reject it. It is psychologically implausible, since many people seem not to conform to the rules they claim to endorse or even to any rules at all. As a norm, the supporter of EMES can reject it, since there is no natural fact which could constitute the relevant normative fact.

I would answer that, in the above discussion, I have only supported a psychological role for the requirement of normative guidance: the analysis from the first personal perspective which I have proposed above tries to show that humans are inclined to be guided by norms which they believe to have intrinsic worth, i.e. to be justified. I also believe that the requirement has a normative role, but I have not argued for this view, which – however – does not play any role in my argument here.

When I refer to the *requirement* of normative guidance, I am not referring to a requirement for an agent, i.e. I am not claiming that normative guidance is something that the agent should endorse. I only claim that it is a requirement

that an action needs to meet in order to be e moral action, and it follows from the proposed analysis of moral facts, including the inescapability and authority of moral claim and our tendency to justify our moral actions through norms.

The fact that many people seem not to conform to the rules they claim to endorse is no objection to the requirement of normative guidance, since (NC) claims that an agent's conduct may be guided by a norm also in cases in which that agent fails to abide by it. On the other hand, I do not think that there are people who to not conform to rules at all, i.e. who are not normatively guided. I tried to deploy a first-person analysis to show that we have an inclination to follow norms. I think that this inclination cannot be resisted. Imagine someone acting in a way which might resist (NC). Either he does it for a reason (even the idle, hopeless reason to show that I am wrong), or for no reason. In the latter case, his going is not even an action (Anscombe 1957, section 5). In the former, his reasons must depend on norms to which he is ultimately conforming in order to describe the possible course of events to be determined by his doing as worth pursuing (in the example, the norms to pursue truth and show wrong people – me – that they are wrong).[7] Then he is normatively guided.

I conclude this section with a short summery and a brief note about its content. The summery: I have suggested that the supporter of EMES rightly recognises that moral judgements are inescapable and authoritative, but fails to recognise a further fact about our moral agency, i.e. the fact that we tend to follow norms which we believe to be justified, or at least – if we do not engage in meta-ethics – that we do not think to be unjustified. In other words, he overlooks the fact that if there is morality, this is because we follow the requirement of normative guidance in our moral conduct. The note: pointing to the requirement of normative guidance, I am not simply saying that we tend to follow moral norms. This would be trivial, since the supporter of EMES recognises it by himself, when he describes the moral capacity, and the inescapability of moral illusions. What the analysis of normative guidance points to is the fact that we tend to be *guided* by norms which we trust. Indeed, there are two ways of following a moral norm. We can abide by it just instrumentally, while we are aiming at something else, or we may be guided by it, for no further reason than the worth of the norm itself.

4. Meta-ethical scepticism and normative ethics

In what sense can a meta-ethical sceptic be moral? A supporter of EMES, no matter whether weak or strong, can be moral in the sense that she might follow a

7 A particularist about reasons could object to this claim and block my line of argument here. As I mentioned above, I have reasons to discontent about particularism, but I cannot spell them out here. To this extent, mu argument has to be taken conditionally.

normative system of ethics. But she cannot be moral in the sense that she is guided by the norms embedded in that system. Since normative guidance is a requirement for morality, if she is consistent, she cannot be moral in the strict sense, but only in the loose sense that she keeps following a certain moral system in so far as it is conducive (or even, the best means) to whatever ends are regulating her actions. To the sceptic, moral systems are not justified in themselves, and can only be accepted for the sake of other goods that they might bring about. They can however be dispensed as soon as they are not any longer conducive to those ends (since, for example, situations have changed), or a better means for those ends has been found, or the choice of favoured ends has changed.

This allows me to answer the second complaint which I have mentioned at the end of the last section. Recall that an objector could claim that an agent could hold onto a moral normative system for purely instrumental reasons, and acquire a number of corresponding habits, to the point that the systems becomes a second nature to him. He would thus be in the grip of that moral normative system, regardless of his second order scepticism, and even if the system loses its instrumental value. The point is now that a moral system, which was accepted without endorsement, acquires – by habit – such a motivational force that it cannot be changed at will, or by a change in the agent's beliefs framework. This point is similar, but different from two objections which I have considered in the previous section: that concerning the example of an agent who is guided – through a reinforcement mechanism – by a norm he does not trust, and that of an agent who accepts a view without endorsing it. What makes the difference here is the notion second nature: the agent does not merely accept the normative system, but he enforces it on himself, for some external value it has, to the extent that his agency loses its plasticity.[8]

Maybe this is a psychologically possible scenario. Of course, it would be unreasonable to let a set of habits known to correspond to unjustified rules to assume such a strong power on one's motivational set-up. The acquired system of norms gains a hold on the agent only if he resists his inclination to normative guidance. The system is not followed because of its reasonableness (whether instrumental or substantial does not matter), and its stability in face of varying circumstances, far from being a sign of moral strength, is a sign of the weakness of the agent: not only he is not guided by the norms of the system, but he even fails to take advantage of the instrumental role that the system could have, by becoming unable to change it as the varying circumstances require. The norms of the system have no authority on the agent (i.e. do not have a guiding role in his choices), but limit his agency by constraining the range of possibilities that he

8 I am grateful to Melissa Lane for this objection, in which she extended to morality a case Peter Lipton had made for religion (Lipton 2007).

can face with his action. In sum, this case does not disprove my denial of a middle way between accepting instrumentally and endorsing (systems of) norms, nor the role for normative guidance which I have recognised in moral action. To the extent that it is realistic, it shows that one can "destroy" one's agency.

This remark leads us into the final issue. So far I have being claiming that evolutionary meta-ethics has bearings on the normative level, contrary to what many supporters of EMES claim. Now, we have to consider in what those bearings are. In what way is a supporter of EMES different, on the normative level, from someone who is not sceptic. The case of a sceptic who inforces a system on norms on himself to the point of making it a second nature is a first extreme case, and this is way I dealt with that case in this section. Let us consider other possible scenarios.

On the surface, from the outside, the difference between a moral person in the strict sense, who is normatively guided by a moral system, and a moral person in the loose sense, who follows a moral system only instrumentally or regardless of its lack of justification, might not be noticed in normal circumstances. Consider the instrumental adoption of a moral system. The system originated in circumstances which are normal for the society in which one was brought up. The sceptic knows that all the norms he learned from birth are not justified at all. But he also knows that most people in his society – the profane, those who do not know the sceptical truth – believe that the norms of their moral system are justified, usually follow them, and expect everyone to do the same. Thus, like a Nietzschean overman, the sceptic walks around the world being as nice as possible to everyone. He pretends he also believes in those norms, he follows them, but he is not guided by them. He is ready to give any of them up, when he is sure that this may help him reach his goals better and has no disadvantages. This might only happen in rare circumstance, and then his diversity from normal, "profane" people shows up to the surface. An akin example can be made for the case of someone following moral system out of habit, regardless of its worth.

This is another possible scenario in which the moral difference between a sceptic and a moral realist may show up. Let us imagine someone moving to a different culture, which is quite different from that of his upbringing. Many sorts of actions that are mandatory in his original culture are here forbidden and vice versa. Were he a normal, naïve believer in the justification of moral norms, he might not even survive this radical change. He would be crashed by psychological distress in the impossibility of changing the norms guiding his conduct, since he believes that those norms are objective and justified. Or he should undergo serious punishment in order to abide by his rules. He would be a martyr. Nothing of this would happen if he were a meta-ethical sceptic. Of course, he could have some psychological distress in the span of time needed to acquire the habits, the dispositions and the expertise needed to follow the new norms. But he

would have no problem in taking on new norms, to the extent that they allow him to reach his goals.

These examples suggest that the moral difference between a sceptic and a non-sceptic about the justification of moral norms is not only a matter of different psychological attitudes, of what goes on in their heads in the process of deliberation, and it might show up at the level of behaviour in certain circumstances.

The moral difference between a supporter of EMES and a moral realist can become apparent also if we consider the counterfactual situation in which meta-ethical scepticism were a generalised position. If the counterfactual situation is not too different from our, in that we assume that no one is aware that EMES is so widespread, things would not be too different from the actual world. If most people believe that moral judgements are unjustified, but do not know that everyone believes it too, they keep utilising the current moral system, in the belief that this would be instrumentally useful for their life in society. Apart for the fact that metaethical scepticism is much more common, that world would not be too different from our on the surface. (Actually, as far as we know we could well live in that world). But if we take a possible world which is still further away from ours in that (almost) everyone is a meta-ethical sceptic *and* (almost) everyone is aware that (almost) everyone else is also a meta-ethical sceptics, then mistrust and suspicion would spread around. Probably most human institutions would not hold, and human communities would be in danger. Again, meta-ethical scepticism makes a difference on the normative level, and this difference could be empirically detected, even if even it might become apparent only in extreme circumstances.

5. Conclusion

This essay has a conditional form. It claims that if some four assumptions are granted, EMES follows, but if EMES is true, contrary to what many of its supporters suggest, the normative level is also affected. In fact, the meta-ethical and the normative levels cannot be sharply separated, since an agent can be guided by a norm only if he beliefs that that norm is justified (or at least he does not think that it is false or unjustified), and normative guidance is a necessary condition of moral behaviour as it is experienced by the agent.

I do not deny that the meta-ethical sceptic might still follow the norms of moral systems. But I deny that he follows those norms because he is guided by them. He follows them instrumentally, if they can serve the attainment of whatever ends he might have, or just out of a deeply rooted habit which counts against the best current judgements of the agent and hence constrains his agency.

Of course, nothing of what I have said counts against EMES, its coherence or its truth. If I am right, though, and if EMES wins over rival versions of evolutionary ethics as I granted above, the upshot of my considerations is that evolutionary thinking should be much more revisionary on ethical matters than it is often assumed, for example in the quotations at the end of section 2 above.

Acknowledgments

I started work for this paper while holding a Visiting Fellowship at the Center for the Philosophy of Science at the University of Pittsburgh, and concluded it while holding a Marie Curie Fellowship (funded by the European Commission) at the Centre for Research in the Arts, Social Sciences, and Humanities, at the University of Cambridge. Drafts of the paper were presented at conference and seminars at the Universities of Cambridge, Rjieka, and Bremen. I am grateful to many people who gave me comments in that occasion. I am particularly grateful for comments to Kevin Brosnan, Craig Delancey, Christian Illies, and Melissa Lane.

References

Anscombe, G. Elisabeth: Intention. Basil Blackwell, Oxford, 1957

Boniolo, Giovanni: "The Descent Of Instinct And The Ascent Of Ethics". In: *Boniolo, Giovanni/De Anna, Gabriele (Hrsg)*: Evolutionary Ethics and Contemporary Biology. Cambridge University Press, Cambridge and New York, 2006. S. 27-42

Campbell, Richmond: "Can Biology Make ethics Objective?". Biology and Philosophy, 11, 1996. S. 21-31.

Casebeer, William: Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition. MIT Press, Cambridge (MA), 2003

Dennett, Daniel: Darwin's Dangerous Idea. Simon and Schuster, New York, 1995

Doris, John M: Lack of Character. Personality and Moral Behavior. Cambridge University Press, Cambridge, 2002

Joyce, Richard: The Evolution of Morality. MIT Press, Cambridge (MA), 2006

Kitcher, Philip: “Four ways of ‘Biologizing’ Ethics”. In: *Sober, Elliot (Hrsg.)*: Conceptual Issues in Evolutionary Ethics. MIT Press, Cambridge (MA), 1994

Lipton, Peter: “Science and Religion: The Immersion Solution”. In: *Moore A./ Scott M. (Hrsg)*: Realism and Religion: Philosophical and Theological Perspectives. Ashgate, Aldershot, 2007

Mackie, John: Ethics. Inventing Right and Wrong. Penguin Books, Harmondsworth, 1977

Railton, Peter: “Normative Guidance”. In: *Shafter-Landau R. (Hrsg)*: Oxford Studies in Metaethics, Vol. I. Oxford University Press, Oxford, 2006. S. 3-33

Richards, Robert: “A Defence of Evolutionary Ethics”. Biology and Philosophy 1, 1986. S. 265-293

*Ruse, Michael: “*Is Darwinian Metaethics Possible (And If It Is, Is It Well Taken)?”. In: *Boniolo, Giovanni/De Anna, Gabriele (Hrsg)*: Evolutionary Ethics and Contemporary Biology. Cambridge University Press, Cambridge and New York, 2006. S. 13-26

Shrout Patrick E./Fiske Susn T. (Hrsg.): Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske. Lawrence Erlbaum Associates, Hillsdale (N.J.), 1995.

Modals versus Morals: Supervenience and Conceptual Relativity

Daniel Dohrn
Daniel_Dohrn@yahoo.com
Universität Konstanz, Konstanz

Abstract/Zusammenfassung

In order to critically scrutinize the well-known modal argument Simon Blackburn forwards against moral realism, I propose to distinguish weak and strong conceptual relativism: In contrast to the former, the latter takes into account all natural facts. Given the former, I want to show how moral realism can solve Blackburn's problem while projectivism as endorsed by Blackburn cannot. Given the latter, I firstly attempt a consistent formulation of Blackburn's challenge. Secondly, I argue that Blackburn's version of projectivism cannot adequately deal with it. Thirdly, I uncover general difficulties for moral realism resulting from moral relativism.

Um Blackburns modales Argument gegen den modalen Realismus zu rekonstruieren, schlage ich eine Unterscheidung zwischen schwachem und starkem begrifflichen Relativismus vor. Im Gegensatz zum ersten bezieht der letztere alle natürlichen Fakten ein. Gegeben den ersteren, möchte ich zeigen, wie der moralische Realismus Blackburns Problem löst, während der Projektivismus, wie ihn Blackburn unterstellt, es nicht löst. Gegeben den letzeren, versuche ich erstens eine konsistente Formulierung von Blackburns Herausforderung. Zweitens argumentiere ich, dass Blackburns Version des Projektivismus nicht angemessen damit umgehen kann. Drittens lege ich allgemeine Schwierigkeiten des moralischen Realismus dar, die aus dem begrifflichen Relativismus resultieren.

Blackburn's Supervenience Argument

Blackburn considers the following claims:

(S) $N_c((\exists x)(Fx \,\&\, G^*x \,\&\, (G^*x \,U\, Fx)) \supset (y)(G^*y \supset Fy))$
(?) $N_c((\exists x)(Fx \,\&\, G^*x \,\&\, (G^*x \,U\, Fx)) \supset N_c(y)(G^*y \supset Fy))$
(P) $P_c(\exists x)(G^*x \,\&\, \neg Fx)$

"N_c" stands for conceptual necessity, "P_c" for conceptual possibility, "F" for some suitably chosen moral property, "G*" for some suitably chosen natural property, "U" for the metaphysical relation of *underlying*.

Roughly, these claims can be paraphrased thus:

(S) By conceptual necessity, if there is a case of G* underlying F, all cases of G* *in the same world* must underlie cases of F.

(?) By conceptual necessity, if there is a case of G* underlying F, *by conceptual necessity* all cases of G* must underlie cases of F.

(P) It is conceptually possible that there be some case of G* without being a case of F.

For Blackburn's argument to start, we must add:[1]

(PS) P_c ((∃x)(Fx & G*x & (G*x U Fx))

Roughly, this claim can be paraphrased thus:

(PS) It is conceptually possible that there be some case of G* underlying F.

According to Blackburn, the conceptual structure of morality contains (P) and (S). (S) seems simply to spell out what it is for a property to supervene on another one.[2] There are two arguments for (P) available: One basic argument is that (S) leaves open the possibility of a thing´s being G* without being F. What (S) requires is only that in worlds in which G* underlies F, it does so for all cases of G*. If there are no additional conceptual constraints to as to eschew (P), one seems to be committed to (P).

Blackburn provides another argument for (P):

> ...it is not plausible to maintain that the adoption of some particular stand is 'constitutive of competence' as a moralist. People can moralise in obedience to the conceptual constraints that govern all moralising, although they adopt different standards, and come to different verdicts in the light of a complete set of natural facts. (Blackburn 1985, 56)

Conceptual constraints of being competent in using moral vocabulary do not constrain moral outlooks to uniqueness, even in light of "a complete set of natural facts", the consequence being that there is no uniquely true moral outlook given all non-normative facts available. I propose to distinguish between weak and strong conceptual relativity. The former is required to maintain both (S) and (P); the latter to maintain Blackburn's statement as just quoted.

---

1 Elliot suggests "(∃x)(Fx & G*x & (G*x U Fx))"(Elliot 1987, 133). But the modally weaker (PS) works as well. (PS) captures what Blackburn intends by claiming: "I am going to build it into the sense of `G*´, that at least in one possible world, the set of properties it denotes is sufficient to underlie F."(Blackburn 1985, 51) Yet this is really a weird claim. G* was supposed to belong to a naturalistic vocabulary. Why should it be built into this vocabulary that there is a possible world in which G* underlies F? Rather we would expect it to be built into the sense of `F´ that G* underlies F. In any case, (PS) accounts for such meaning relationships.

2 "Supervene" in a sloppy sense according to which F supervenes on G* if G* being exemplified is in a modally qualified manner sufficient for F being exemplified.

Blackburn concludes:

> If we put (?) into abeyance we should be left with a possible form of doctrine which accepts both (S) and (P): the (S)/(P) combination. It is this which I originally claim to make a mystery for realism ... there are relations between F and G vocabularies which are properly characterized by (S) and (P). In the moral case I think that this is best explained by a projective theory of the F predicates.(Blackburn 1985, 52-3)

The mystery is the following: Since there are possible worlds in which G* underlies F such that all x that are G* also are F, and there are possible worlds in which there is some y that is G* but not F, a principle of plenitude about possibilities has it that there must be mixed worlds which contain some x for which the *underlying* relation holds and some y which is G* but not F.

According to moral realism, value judgements track moral properties. According to projectivism, in contrast, value judgements determine but do not track moral properties. Blackburn deems projectivism more suited to deal with the above conceptual structure than moral realism:

> The difficulty is that once we have imagined a G*/F world, and a G*/O [no F], it is as if we have done enough to imagine a G*/FvO world, and have implicitly denied ourselves a right to forbid its existence. At least, if we are to forbid its existence, we need some explanation of why we can do so… in the moral case, projectivists can do this better than realists.(Blackburn 1985, 54)

Weak Relativity

It is weak relativity that is usually assumed in discussing Blackburn's challenge. Weak relativity requires that it be conceptually open whether G* underlies F or some case of G* is not F. In contrast, strong relativity requires that it be open given the constraints of being a moral outlook and a complete set of natural facts.

Mixed Worlds without Underlying

Blackburn´s argument has been challenged by several critics. Anthony Brueckner argues that all that follows is that mixed worlds do not allow for the *underlying* relation to hold.

> What actually follows
> from the (S)/(P) combination is that there is no possible world w meeting the following three conditions:
> (1) Something in w is G* and F.
> (2) Something in w is G* and -F.
> (3) F supervenes upon G* in w (G* underlies F in w).

> That there is no world meeting conditions (1)-(3), however, is obviously consistent with there being mixed worlds in which something is G* and F while something else is G* and -F. In such mixed worlds, F simply does
> not supervene upon G* (G* simply does not underlie F).(Brueckner 2008, 70)

Mixed worlds can be reconciled with the possibility of the underlying relation. For this possibility allows for worlds in which some x which are G* are F and some y which are G* are not F, the consequence being that there is no *underlying* relation in such worlds. And that is all plenitude requires.

Now Blackburn suggests that one must also be able to conceive the following world: At the beginning, all x which are G* are F. Condition (S) of the underlying relation is fulfilled. In this world, G* underlies F. Then the world changes so that one of the G* cases ceases to be F (Blackburn 1985, 53). The underlying relation is not any longer fulfilled. Or one might imagine some G*-case which is not F being added to this world. Why should thereby the metaphysical status of the other G*-cases be changed so that the relation of underlying does not hold any longer? Hence there seem to be mixed worlds which contain cases of G* underlying F. I want to suggest that these considerations count against Brueckner's argument (provided they do) only because they show that Blackburn's account of underlying enshrined in (S) is unconvincing. One needs a stronger conceptual necessity (?).

The Modal Strength of the *Underlying* Relation

Once a metaphysical notion of underlying is accepted which allows to sustain (S), it cannot be prevented from spilling over to a stronger necessity as expressed in (?). Robert Elliot proposes against Blackburn to strengthen (S) so as to yield (?). Assume F to be the property of being good and G* the natural property which underlies being good:

> I notice that Sw is characterized by the set of natural properties G* and I judge that it is good. I am then asked to imagine a possibility which is not actual. This possibility contains a state of affairs, Sp, which is likewise characterized by G*. I am asked whether Sp is good and I judge that it is. Whatever moves me to judge Sw as good in virtue of exemplifying G* should move me to judge S as good since it similarly exemplifies G*. Moreover, the force of 'should' here is conceptual. States of affairs characterized by G* are not found only in the actual world; they are scattered across the whole structure of possible worlds. Judging that Sw is good but not that Sp is good would just as much breach the conceptual constraint on moralizing as would failure to abide by the ban on mixtures within the actual world. It is 'constitutive of competence' with the moral vocabulary that mixtures be eschewed across the structure of possible worlds.(Elliot 1987, 135f.)[3]

3 Essentially the same point is already made by I.G. McFetridge (1985). J. Klagge replies that although a person taking G* to underlie F should do so for all possible exemplifications of G*, this does not hold for another person who may endorse an outlook which yields not-F for some case of G* (Klagge 1987, 314). However, this argument either does

There are two difficulties with this argument.

(i) Elliot does not show the second necessity operator in (?) to denote conceptual necessity. Even if an underlying relation between G* and F obtains, it does not make is *conceptually* necessary that every G* is F. Thus it seems as if Blackburn had an answer to Elliot avant la lèttre:

> For (?) to help, we would need it to be read so that, if something is F and G*, and G*-ness underlies its being F, then it is analytically necessary that anything G* is F. And this we will not have in the moral case, for we want to say that there are things with natural properties underlying moral ones, but we also deny necessities of the form (Na)[conceptual or analytic necessity]. (?) would not help if the necessity of the consequent were interpreted in any weaker sense. For instance, we may want to accept (?) in the form
>
> $(?_{MN})\ MN((\exists x)(Fx \,\&\, G^*x \,\&\, (G^*x\ U\ Fx)) \supset MN(y)(G^*y \supset Fy))$
>
> and then there will be metaphysical necessities of the form of the consequent, … but they will not help to resolve the original mystery, since that is now proceeding at the level of analytical necessity.(Blackburn 1985, 57)

Blackburn grants that (?) may be devised so as to contain only *metaphysical* necessities. However, this does not impose a constraint on *conceptual* possibilities which are at issue. Of course, Elliot denies that the first necessity operator can be read as metaphysical necessity, and rightly so. Still one may question G* underlying F to be conceptually necessary, even given the antecedent of (?).[4]

I deem the main step towards a solution to lie in Nick Zangwill proposing *mixed* modalities:

> ($S_{\mathbb{cm}}$) Necessarily { (∃F …)(∃x)(Fx → (∃G …)[Gx & necessarily (∀x)(Gx → Fx)])}(Zangwill 1995, 257)

The indices $_c$ and $_m$ specify the occurrences of "necessary" in due order. Let "F" stand for a moral property and "G" for a natural property: It is *conceptually necessary* that if there is a case of F, there is a G such that the former case is also a case of G and it is *metaphysically necessary* that all cases of G also are cases of F.[5]

---

not help against everyone being committed to (?), or it just resumes strong relativism according to which one may endorse incompatible outlooks which do equally well in terms of truth.

4 This problem is not solved either by Shafer-Landau claiming: "Lack of conceptual entailment doesn't make supervenience queer at all. If nonmoral base properties metaphysically necessitate the presence of specified moral properties, then the conceptual possibility that they fail to do so reveals only a limitation of our appreciation of the relevant metaphysical relations."(Shafer-Landau 1994, 149)

5 Zangwill describes the problem and his solution as follows:
"To satisfy Blackburn, we had to explain how it could be that although each of the conjuncts of [(∃x)(∃w)(G*xw & Fxw) & (∃y)(∃w')(G*yw' & -Fyw')] is individually conceptually possible, the conjunction is conceptually impossible …Thus [given $S_{\mathbb{cm}}$], it is

However, in my opinion Zangwill incurs the second difficulty of Elliot's argument:

(ii) Zangwill provides a reasoning quite similar to Elliot's in favour of supervenience being conceptually necessary:

> …striving for consistency is part of what it is to moralize. But this quest for consistency rests on a commitment to moral supervenience. So moral supervenience is a conceptual truth.(Zangwill 1995, 243)

But even if consistency requires to hold any possible case of G* to be F upon holding one of them to be, it does *not* require the following conception of strong supervenience: By conceptual necessity, once a property F is devised, there must be some property G* *in the focus of the valuer* on which it supervenes.

Let me explain. The very idea of projectivism is that we project our attitudes onto the world, usually without explicitly noticing this. We attend to whether kicking dogs is a sufficient condition of an action being wrong and not to whether kicking dogs + our attitude towards it are. In a quote to be more closely scrutinized below, Blackburn considers

> …the moral view that the feature which makes it wrong to kick dogs is our reaction. But this is an absurd moral view, and not one to which a projectivist has the least inclination. Like anyone else he thinks that what makes it wrong to kick dogs is that it causes them pain.(Blackburn 1981, 179)

To me the lesson the projectivist should draw is to distinguish two kinds of supervenience claims. One is the claim that moral properties supervene on a narrower basis, properties *in the valuer's focus*, the other is the claim that they supervene on a broader basis, namely facts the valuer considers + *other facts about the valuer as such*, namely her dispositions to value something. The projectivist reserves the explanatory "making F (e.g. wrong)"-relation to the relationship between natural facts belonging to the *narrower* basis and moral facts. Making wrong and supervenience come apart. But do we not take into consideration facts about our attitudes and desires to flourish? Of course, sometimes we do. Yet the main point is that some of them might exert their influence on our evaluations without being in focus. Elliot's and Zangwill's arguments also relate to supervenience on the narrower basis.[6] For they require consistence of attitude towards properties considered and not with regard to supervenience on facts in general including facts about the valuer as such which are not taken into account.

---

conceptually necessary that either in all metaphysically possible worlds G* things are F or in all metaphysically possible worlds G* things are -F,even though neither disjunct is conceptually necessary taken by itself."(Zangwill 1995, 258)

6 Supervenience on the narrower basis is expressly endorsed by Jackson (1998, 120).

The main difficulty Elliot's and Zangwill's accounts share is that they rule out from the outset a certain variant of the projectivist account as endorsed by Blackburn. Below I will discuss in how far Blackburn is committed to this variant. Of course, it might be that the projectivist account can be ruled out or at least restricted by a strong supervenience claim. But since according to Elliot and Zangwill this claim is so obviously crucial to conceptual competence, there should not be a shadow of doubt whether to exclude on conceptual grounds the following projectivist eventuality:

> If (as Blackburn believes) any outlook deserving the name `moral´ requires consistency of attitude, supervenience regarding morality is a conceptual truth. But it is a conceptual truth that concerns our attitudes and desires to flourish, rather than any realistic relations obtaining between moral and other properties."(Shafer-Landau 1994, 146)[7]

If moral truths supervene at all, they do not supervene on the natural facts *in focus*, but on a complex whole of these facts and facts about the valuing subjects.[8] According to Blackburn, the latter may be given a naturalistic explanation (Blackburn 1993, 72). To him this implies their being contingent. Let G* be a natural fact in focus which is deemed sufficient for moral property F. If F supervenes, it will supervene on G* considered *and on some fact about the valuer*. But this is not the supervenience basis which Elliot and Zangwill take to follow from their argument. They require supervenience on facts considered (if any), not on such facts and facts about the valuer which are not considered. Projectivism cannot claim to ensure supervenience on the narrower but only on the broader basis.

Once one deems some property G* to be sufficient for some moral property F, one should do so consequently such that any case of G* is by metaphysical necessity a case of F. But on pain of begging the question against the projectivist one should not, as Zangwill does, build supervenience on natural properties (in the narrow sense not taking into account facts about the valuer) into the conceptual constraints any moral outlook must fulfil. Nor should one, as Elliot does, read the second modal term in (?) as *conceptual* necessity. Hence I propose:

(?*) $N_c((\exists x)(Fx \,\&\, G^*x \,\&\, (G^*x \,U\, Fx)) \supset N_m(y)(G^*y \supset Fy))$

---

7 In the quote to be discussed more thoroughly below, Blackburn's tells the projectivist to beware of „...the mistake of thinking that after all there is a state of affairs making the projected commitment true, only one about us. He must not think this, nor is there any reason for him to do so, provided he has a proper appreciation of the theory of meaning which must be attached to his metaphysic."(Blackburn 1981, 179) This statement could be read so as to exclude that a moral claim supervenes on any facts. But this reading is not reconcilable with the conceptual structure of moral theory Blackburn builds his argument on.

8 I doubt that this contention squares with Blackburn insisting that the feature making it wrong to kick dogs is not our attitude but the fact that it causes them pain. However, the projectivist seems committed to a certain attitude and certain natural facts being sufficient for being wrong.

It is conceptually necessary ($\mathbf{N_c}$) that if there is some case of G* underlying F, it is metaphysically necessary ($\mathbf{N_m}$) that any case of G* is a case of F.

Why does this solve Blackburn's original problem? The following statements are conceptually possible: The relation of G* underlying F obtains in some world. There is some x which is G* but not F. The following is a conceptual necessity: Either the relation of G* underlying F holds in some world. Then there are no metaphysically possible worlds in which G* does not underlie F. Or there is no world at all in which the relation of G* underlying F holds. Neither alternative admits of mixed worlds in which the relation of underlying holds together with some instantiation of G* not being an instantiation of F. Since both alternatives are conceptually possible, the second necessity operator in (?*) must be interpreted as metaphysical necessity. In light of the conceptual possibility of G* not underlying F, it cannot be conceptually but only metaphysically necessary that all cases of G* are F. If these considerations are correct, moral realism does not have to deal with the problem of mixed worlds.

However, the above solutions require that knowledge of moral supervenience be empirical or at least non-conceptual. The gap between conceptual and metaphysical modalities is usually associated with the epistemic gap between a priori and empirical knowledge. Even assuming that, in the present context, there is no interesting boundary between a priori and conceptual knowledge, the issue is intricate. If there are metaphysical necessities which can only be detected a posteriori, such necessities cannot be conceptual, or so it seems.[9] Now there are reasons why moral theory is a priori. One may deny that moral truths are to be empirically detected. In contrast, it seems as if moral truths could be detected by considering possible cases in an armchair. Consider for instance the numbers problem: When we may save firstly A,B,C or secondly D, are we bound to save the greater number? Scholarly discussion of this problem proceeds by a priori considering possible cases and not by empirical scrutiny.

Yet there also are reasons to be doubtful about the aprioricity of morals: Enabling conditions like suitable education may prove indispensable for coming to appreciate moral truths. It is a familiar claim that moral concepts cannot be learnt in an armchair but only by being confronted with, say, social paradigm scenarios. Furthermore, affective-emotional attitudes acquired by encountering such scenarios may prove indispensable in evaluating moral issues.[10]

---

9 Yablo 2002 supposes otherwise.

10 To meet this challenge, we may recur to a concession which Yablo shares with most approaches to a priori knowledge. Conditions under which a concept is learnt are to be distinguished from prerequisites of justifying certain claims which presuppose mastery of this concept: "If experience cannot be appealed to at all, then shouldn´t be enough to stop S from being a priori if it is through experience that we *understand* S? The answer to this is that we are interested in how S is *justified*, our understanding taken for granted."(Yablo 2006, 255, cf. Kitcher 1980, 5-6, Peacocke and Boghossian 2000, 2)

In spite of all reservations, let us consider what happens if moral statements including supervenience claims are claims to a priori knowledge, a priori knowledge being knowledge for which mastering the concepts at stake is sufficient. If so, we should be in a position to a priori figure out whether G* necessarily underlies F. In this case, the second necessity operator in (?) must denote conceptual necessity. Consequently, either (?) or (P) has to go. If both are intuitively plausible, we have a puzzle which will be tackled in a moment:

> (P1) Given that knowledge of moral supervenience is a priori, how can the conceptual structure of morals requires to reconcile (?) and (P)?

Projectivism cannot deal with mixed worlds either

I now want to criticize Blackburn's claim that projectivism is better suited to deal with the conceptual structure of morals than realism.

Firstly projectivism cannot incur the commitments enshrined in (S) given (P), (PS). Given the Elliot-Zangwill argument, there is no mechanism which ensures that one's contingent evaluative dispositions mark an eligible boundary carving a possible world at the joints such that just those x which are G* in this world also are F. There is no reason why the valuers' dispositional structure should incline her to focus on all x which are G* within one world.

Secondly, projectivism cannot deal with (?*) given (P), (PS) either. The above discussion leads to two variants of projectivism. According to the first, F

However, Timothy Williamson questions the distinction between enabling conditions of knowing that p, e.g. conditions of understanding p, and evidential or justificatory conditions. Williamson considers the counterfactual: "(25) If two marks had been nine inches apart, they would have been at least nineteen centimeters apart."
Mastering the concepts "inch" and "centimeter" may well involve perceptually evaluating certain distances. Williamson asks: "Do I know it a priori or a posteriori? Sense experience plays no direct inferential role in my judgment. I do not consciously or unconsciously recall memories of distances encountered in the past, nor do I deduce (25) from general premises I have inductively or abductively gathered from experience… Nevertheless, the causal role of past sense experience in (25) far exceeds enabling me to grasp the concepts relevant to (25); … My possession of the appropriate skills depends constitutively, not just causally, on past experience for the calibration of my judgments of lengths in those units."(Williamson 2007, 166)
Carrie Jenkins suggests "…that what experience does is (not merely supply) but *epistemically ground* the relevant concepts."(Jenkins 2008, 700) Jenkins concludes for knowledge exclusively based on competences required to grasp the concepts involved "…that this way of knowing is both empirical in the standard sense and also *a priori* in a wide range of accepted senses."(Jenkins 2005, 743) Yet she insists that upon understanding a priori justification as "independent of all experience", such knowledge is not a priori justified. In light of such considerations and the devastating criticism of the concept of a priori knowledge which philosophers like Hawthorne (forthcoming) advanced, one may doubt the a priori-a posteriori distinction to be useful at all.

necessarily supervenes on facts in the valuer's focus and the valuer's dispositions which allow for naturalist explanation. According to the second, they do not. I deem the second variant prima facie implausible. Projection would be really queer if it did not somehow reside in properties of the projector or the world into which features are projected. Given the first variant, surely one may commit oneself to G* underlying F. But given the dependence of one's valuations on one's dispositions of valuing which are contingent, it is metaphysically possible that one come to diverging valuations upon suitably changed valuing dispositions. In a possible situation, one has different valuations. Or one may change one's actual valuations. In the latter case, one's original commitments and one's valuations will conflict. Why is this a problem? Say one considers a given situation in which dogs are kicked. Following one's dispositions, one deems kicking dogs wrong. Now one adds a further constraint to that situation: One's dispositions are changed such that one accepts kicking dogs as right. When one evaluates this situation, which dispositions are the ones to follow? Probably the actual ones. After all, one has committed oneself to kicking dogs being sufficient for being wrong. But as an enlighted projectivist, one knows that evaluation follows one's dispositions. How can the actual dispositions override one's possible dispositions simply because they are actual? So it seems that one has no rationale of being actuality-chauvinistic. If moral properties supervene on dispositions and other facts, possible dispositions are as good candidates as actual ones. The same holds for changing one's dispositions. One cannot commit oneself to keeping one's dispositions. After all, value judgements follow dispositions and do not determine them. What about the "consistency of attitude" Shafer-Landau requires? As already argued, this consistency does not relate to the content judged, but to a broader range of facts which includes value dispositions themselves upon which valuations are alleged to be contingent. Suppose that one judges to be wrong what displeases one in a qualified way, a case of causing pain displeases one in this way, and one by projection judges that causing someone to be in pain is sufficient for an action to be wrong. There is a possible case in which it does not displease one in a qualified way to cause another's being in pain, say because one has sadistic inclinations in this situation. In this case, one refrains from judging it to be wrong. This is not what one commits oneself to if one judges some natural property G* such as causing pain which does not include one's evaluative disposition to underlie some moral property F such as "is wrong". Commitments and dispositions grounding values are at odds.[11] Since a projectivist analysis allows for this scenario, projectivism cannot deal with mod-

11 Of course, one may subject one's commitments to the condition that one's value dispositions are such and such. But as the distinction between facts in the valuer's focus and facts about the valuer not in focus shows, this is not how moral discourse functions.

al requirements of the underlying relation. The only solution available to projectivism is that one cannot consistently commit oneself to an *underlying* relation.[12]

Strong Conceptual Relativity

I now want to come to strong relativity. Strong relativity is implied by Blackburn maintaining:

> People can moralise in obedience to the conceptual constraints that govern all moralising, although they adopt different standards, and come to different verdicts in the light of a complete set of natural facts.(Blackburn 1985, 56)

Another expression of strong relativity is this:

> Hare's logical necessity is not a constraint on the thing judged, but on the judge. Two things, even in the same possible world, could be judged descriptively indiscernible and yet receive different moral valuations if they were judged by different people employing different moral principles. (Klagge 1988, 463)

If I am right, these statements cannot be interpreted such as to exclude the "different verdicts" and the "different moral valuations" to be all true.[13] In contrast, depending on the diverging moral standards, incompatible moral valuations are true to the facts.[14]

Formulating Strong Relativity

Blackburn does not provide an adequate formulation of strong relativity. The following model is an attempt to do so: Let the intension of a concept be a function which determines for any possible world what the concept applies to in this world. Let difference of intension imply difference of concepts. In order for A's

12 To be sure, one can incur commitments one cannot hold. But then projectivism turns out not to be reflection-proof. If projectivism is conceptually true, the enlighted projectivist who endorses the projectivist view cannot incur commitments to an underlying relation and hence cannot accept (PS). If projectivism is not conceptually true (but must be figured out empirically), the following is: If there is an *underlying* relation, projectivism cannot account for it and hence cannot be true. If projectivism is true, there cannot be an *underlying* relation. Projectivism and the underlying relation are conceptually incompatible.

13 Does the distinction between a broad and a narrow basis play a role here? No, for the basis is always the broadest possible, the complete set of natural facts.

14 A more cautious view is expressed by Williamson. There might be irresolvable moral disputes:
„Jackson cannot exclude the possibility that on some significant issues folk morality lacks the resources ever to reach reflective equilibrium. Some moral disputes may be irresolvable-which of course does not imply that there is no right answer, just that we cannot agree on what it is."(Williamson 2001, 630)

verdict "kicking dogs is always wrong" and B's verdict "there is a case of kicking dogs not being wrong" to be both true, they must deploy different concepts. Let [F] be the concept which represents being wrong and [G*] be the concept which represents kicking dogs. Then we cannot simply state conceptual relativity in this way: A truly holds that every case of G* is a case of F and B truly holds that there is a case of G* which is not a case of F. Instead, we must state conceptual relativity as follows: In B's conceptual scheme there is a concept [F´] which competes with A's concept [F] for fulfilling the same role in meeting the conceptual requirements of a moral outlook.[15] [F´] stands for the property F´. Two concepts compete for the same role only if the conceptual structure of morals allows to replace one by the other and one cannot conform to the role of both simultaneously. Let [F] and [F´] compete for the role of what we express by "being wrong". Arguably the role of the concept of being wrong is to guide action. One should refrain from doing what is wrong. Let G* be the property of kicking dogs, F be underlain by G*, and F´ be such that some case of G* is not F´. Then the conflict between [F] and [F´] becomes manifest when I should refrain from kicking a dog as kicking dogs is F and I may kick a dog as kicking dogs is not F´. In such a case there are two systems of morals M and M´ such that if I conform to M, I thereby violate the prescriptions of M´.

Blackburn's strong relativity can be captured by the following claims. Let D_w stand for the complete description of the natural facts in some canonical vocabulary including a self-locating description of the moral subject. Let D stand for the set of complete descriptions measuring the space of metaphysical possibility. Strong conceptual relativity has it that given knowledge of $D + D_w$, the conceptual competence required to develop a moral theory is reconcilable with endorsing alternative moral theories M and M´ such that M yields G* underlying F for all possible instantiations of G* while M´ yields some instantiation of G* not underlying an instantiation of F´. The result is that the following claims both come true:

(DF) $P_c((D + D_w) \supset N_c(x)(G^*x \supset Fx))$
(DF´) $P_c((D + D_w) \supset P_c\exists y(G^*x \,\&\, \neg F´x))$

These claims can be paraphrased as the following conceptual possibilities:

(DF) The full canonical descriptions of the world and metaphysical possibilities entail that all cases of G* by conceptual necessity are F.
(DF´) The full canonical descriptions of the world and metaphysical possibilities entail that there is a possible case of G* which is not a case of F.

15 For a discussion of the notion of conceptual counterparts [F] and [F´] cf. Bealer (2002).

Blackburn's Challenge Relativisited

There are two issues to be tackled in light of strong conceptual relativism: (P1) introduced above and a new problem

(P2) How can M and M´ be both true to the facts and in tune with conceptual requirements of being a moral outlook?

I want to distinguish two stages (A, B) of conceptual requirements:

At stage A, there are requirements of being a conceptual system capturing a certain region of discourse, for instance moral discourse. Blackburn's claim to strong relativity depends on there being several alternative moral outlooks. All these outlooks fulfil the criteria of being a moral outlook at all. The metascheme of a moral theory can be filled in several ways.

At stage B, there are conceptual truths given a specific system of moral concepts. Conceptual possibility as it is usually understood is located at this second stage.

With this distinction in hand, we can solve (P1): At stage A, the metascheme leaves open whether a moral theory yields G* underlying F or not underlying F´. It leaves open whether to endorse a scheme which yields G* underlying F or a scheme which yields G*x and not F´x. However, once at stage B a moral outlook involving a certain conceptual scheme is adopted which yields F underlying G*, some x being G* but not F´ is not any longer an option.

We can also solve (P2): At stage A, D + D_w leave open whether to endorse a moral system M which contains [F] or a moral system M´ which contains [F´]. But once one system is chosen, at stage B the other system is not any longer an option. The combination of (DF) and (DF´) can be rejected. It would have to be placed at stage B. But the intuition on which it rests relates to viable conceptual alternatives at stage A. The combination of (DF) and (DF´) rest on confusing viability at stage A with conceptual possibility at stage B.

Projectivism and Relativism

In order to show that in spite of this consistent formulation strong relativism raises difficulties as well for Blackburn's projectivism as for moral realism, I want to start from the following passages:

> …suppose a projective theory must involve us in believing things like [C] `If we had different attitudes it would not be wrong to kick dogs´… Then clearly it is refuted, because these things are absurd. Fortunately, however, the projective account of indirect contexts shows quite clearly how to avoid them. The counterfactual `If we had different attitudes it would not be wrong to kick dogs´ expresses the moral view that the feature which makes it wrong to kick dogs is our reaction. But this is an absurd moral view, and not one to which a projectivist has the least inclination. Like anyone else he thinks that what makes it wrong to kick dogs is that it causes them pain. … A projectivist is only tangled

> in these unlovely counterfactuals if he makes the mistake of thinking that after all there is a state of affairs making the projected commitment true, only one about us. He must not think this, nor is there any reason for him to do so, provided he has a proper appreciation of the theory of meaning which must be attached to his metaphysic.(Blackburn 1981, 179)

I begin with considering whether Blackburn's projectivist analysis can be upheld. Blackburn insists that "People can … adopt different standards, and come to different verdicts in the light of a complete set of natural facts." In the non-projectivist reading of counterfactuals, this contention seems to involve at least a variant of C: If people had adopted a different moral standard, they would have come to different verdicts. Why then is Blackburn so concerned about C?

In order to appreciate his argument, some conceptual clarifications are indispensable. Even if we were to use the word "being wrong" differently so as to capture a different conceptual possibility (M´) which yields that one may sometimes kick dogs, kicking dogs would still be wrong as the counterfactual is formulated from our actual viewpoint in our language. Blackburn's counterfactual must be reformulated using words branded to cover the newly realized conceptual possibility, say "schrong" instead of "wrong". There is the role the concept of being wrong has in our moral outlook. The concept associated with "schrong" must be what competes for this role in the newly coined moral outlook as [F´] competes with [F]. On the whole, it must fulfil the conceptual constraints imposed on any moral outlook as well as the concept of being wrong does. Then we get

> C´ If we had different attitudes, kicking dogs would not be schrong.

Blackburn offers a somewhat revisionary reading of conditionals, be they subjunctive or indicative, in anti-realist discourse. The consequent expresses a commitment incurred when committing oneself to the antecedent: If kicking dogs were not wrong, kicking cats would not be either (cf. Blackburn 1981, 178). However, Blackburn's anti-realism is not global. There are discourses which are realist, for instance physics. If one does endorse a non-truth-conditional semantics in such discourses, there must be mixed conditionals such as: "If there were dogs which did not feel anything, kicking them would not be wrong". The antecedent of this conditional expresses a natural metaphysical possibility without any glimpse of moral discourse, the consequent according to Blackburn a moral commitment. Consider the indicative conditional without modal commitments: "If there are dogs which do not feel anything, kicking them is not wrong". Again the projectivist must hold that the antecedent sentence is truth-conditional, the consequent is not. Analogously, the best way of understanding C´ is to regard it as a mixed conditional which is true. Blackburn insists that if C´ were true, we would have to accept that the "feature that makes it wrong to kick dogs is our reaction". And that's what to him sounds absurd be-

cause "what makes it wrong to kick dogs is that it causes them pain".[16] But the counterfactual C´ states only a counterfactual dependence. It does not replace the feature that underlies making wrong by another feature, our attitude. In the same vein, we may say that I would not muse about anti-realism if Blackburn's arguments for it had not been forwarded. But granting there is a physical brain-state underlying my mental activity, the same holds for this brain-state obtaining: I would not muse about anti-realism if the brain-state did not obtain. Both counterfactuals seem perfectly in order.

Now Blackburn is a modal anti-realist, too. Modal statements do not track an independent modal reality but reflect what we find conceivable (Blackburn 1993). Hence he might endorse a non-truth-conditional semantics for counterfactuals in general. Then we would not have mixed counterfactuals. But all the better for C´: Surely we can conceive of people adopting different standards and consequently coming to different verdicts, for instance that kicking dogs is not schrong. At least Blackburn cannot exclude the latter possibility by the reasons he provides. To him the difference between moral and modal discourse is that the former but not the latter can be naturalistically explained as a naturalistic explanation implies the metaphysical contingency of what is explained.[17] Projectivism explains valuations like "kicking dogs is wrong" by value dispositions. If these dispositions can be given a naturalistic explanation, it is conceivable that

16 In a similar vein, Jackson argues:"We want rightness to be what makes an action right, not in the causal sense but in the sense of being what ought to be aimed at. Now what we should aim at is not doing what is right qua what is right. I should rescue someone from a fire because if I don't they will die, not because that is the right thing to do."(Jackson 1998, 141)
But a Kantian would insist that I should rescue people because it is the right thing to do. Of course to explain why one should rescue people by their being in danger to die is perfectly in order in a certain context. But so is – albeit somewhat pleonastic- explaining it by its being the right thing to do. Both explanations are incomplete. I suggest that a complete realist explanation why to do something cannot be confined to the range of natural facts underlying moral ones but must specify suitable relationships to normative vocabulary.

17 At least this is the only way I can make sense of Blackburn noting a fundamental asymmetry between morals and modals: "the moralist can be quite completely aware of the genesis and justification of his activity, whereas … the modalist cannot be."(Blackburn 1993, 72) I cannot help being suspicious that this amounts to a defeater. For how can one grant that value dispositions can be given a naturalistic explanation but other judgemental dispositions cannot? Pust maintains that knowledge of necessities must be maintained to be necessary itself and that what is necessary cannot be explained. For such an explanation would have to depend on counterfactual claims how modal judgement would change if necessary truth changed (Pust 2004, 72). But it is questionable in how far this explanation is open to the enlighted projectivist. How could she hold that her dispositions of modal judgement are necessary? For what should exclude her to imagine a being with diverging dispositions whithout the projectivist being in a position to dismiss these dispositions as deviant?

they be different, namely if necessary conditions specified in the explanation do not obtain. Of course, some moral statements like "kicking dogs is wrong" might be necessary provided they are at the same time modal statements, for instance when the *underlying* relation holds. But Blackburn wants to generally eschew counterfactuals like (C´), be their consequent necessary or not.

Moral Realism and Strong Relativism

Notwithstanding my criticism of Blackburn rejecting (C´), I join Blackburn in deeming (C´) awkward. Since Blackburn's explanation of this feeling proves insufficient, I suggest that moral realism provides a better one. An inclination to reject C´ reveals an uneasiness about strong relativism giving rise to a certain sort of moral scepticism. Provided M and M´ fare equally well in terms of being true to the facts, why should I behave according to, say, M and not M´ when they conflict? This concern reveals a crucial asymmetry between terms like "water" and moral ones. While no one would bother about the eventuality of reshaping the concept of water, say changing from a Kripke-Putnam style to a more functional analysis according to which we call `water´ whatever fulfils the role of a watery stuff, a license to transcend from M to M´ in a conflict case is irritating. And this indicates that one cannot live with moral relativism as it threatens the action-guiding role of morals. However, the question is how moral realism can avoid relativism when considering the parallel case of "water". As canonical world descriptions do not seem to constrain our reshaping the term "water", how can they be reckoned to constrain moral outlooks to uniqueness?[18]

When we are looking for further constraints of an eligible moral outlook besides $D + D_w$, the following dilemma arises: On the one hand, if moral realism is true, the right kind of reason to decide between M and M´, for instance in order to know what to do in a certain case, seems to be whether M or M´ are true to the facts. But M and M´ are supposed to be equally true to the facts. On the other hand, there may be further differences between M and M´ one might hope to provide a decision. But these differences do not seem to amount to the right kind of reason.

To make the point more vivid, consider a variant of Jackson's famous Mary-example (Jackson 1986). Assume Mary to be moral- or value-blind but to know the complete canonical description D_w of the actual world including de se-beliefs which locate Mary in the actual world + the set D of canonical descriptions of metaphysical possibilities. She might have a system of non-moral preferences or desires which guide her action. But when she asks herself which moral outlook to embrace, nothing rationally commits her to prefer one to the other, say

18 Of course, one may reject the very conception of such a description.

one yielding that there can be no x which is G* but not F or one yielding that there can be some y which is G* but not F´.

To put the point otherwise: Suppose someone is a perfectly moral person by virtue of endorsing a moral outlook according to which it is wrong to kick dogs. Given Blackburn's account, preserving all her moral integrity she might as well come to the result that in the very same situation, kicking dogs is not wrong (respectively shrong), provided she endorses a diverging moral outlook. Of course, some things would have to change in order for her to come to the diverging verdict. But what is crucial is that all these things seem somehow morally contingent compared to what would be preserved while radically changing her moral outlook: the integrity of her character, her moral standing, her moral competence, metaethics, the characteristics of the situation to be evaluated.

While proposals such as a motivational set restricting admissible moral outlooks fall prey to the "wrong kinds of reasons"-reproach, I want to finish mentioning three approaches which do not:

(i) Theoretical virtues: Although several incompatible conceptual schemes are possible which fare equally well in terms of truth given D + D_w, some scheme fares better in terms of theoretical virtues such as simplicity and naturalness (cf. Lewis 1984). However, to moral realists, their lacking such virtues must be the reason why moral outlooks entailed by the canonical world description nevertheless do not figure in it. Judging from theoretical virtues, one should give up M and M´ altogether. But if further constraints of another kind are added, the "wrong kind of reasons"-problem looms. The most appealing story is that a canonical world description being infinitely complex, one needs some shortcut to moral truth. One may say that a moral outlook is distinguished by the virtue of making infinitely complex truths available to us:

> The disjunctive descriptive story D that is equivalent to the ethical story E may be an infinite disjunction we need ethical terms to handle… there is nothing more `there´ other than the relevant similarities among those descriptive ways.(Jackson 1998, 124)

There are certain similarities or perhaps some "gestalt that we cannot break down exhaustively into its components"(Jackson 1998, 66) guiding use of moral terms. Hence the best moral outlook is what suits best our capacities. But it seems doubtful that this virtue of being cognitively more comfortable is sufficient to privilege a moral outlook compared to another one which is only a bit less comfortable to handle. Solution (i) implicitly presupposes that the best moral outlook is distinguished from otherwise equal rivals only by a metatheoretical advantage. It is the best way to represent facts of the matter. Given this presupposition, the point of a moral truth should be discernible from the theoretical standpoint of someone knowing all moral facts about the world. Moral truth should not be distinguished merely by our finite cognitive capacities. For instance, they should not cease to be binding upon increasing one's capacity of representing the world.

(ii) Moral Mary: The conceptual restrictions of being a moral outlook could be strengthened so as to constrain moral theory to uniqueness given $D + D_w$. But such constraints cannot consist in facts to be appreciated. The general strategy to add moral constraints can be derived from Jackson's Mary: Find some kind of fact about the valuer which does not interfere with moral realism. As there is something new about the world Mary learns albeit her world description has been incomplete, the moral subject must appreciate something new about the world which is entailed by her world description without succumbing to projectivism. Eligible candidates of such constraints can be found in functional theories of moral concepts. According to such theories, moral concepts must play a certain role in triggering motivations to act. One problem of such theories is how understanding moral concepts and acting on a certain outlook are related. If on the one hand one can have a purely theoretical understanding of such concepts, it seems as if one could appreciate a moral theory and hence conform to the conceptual constraints of such a theory without conforming to the functional role of moral concepts, for instance being motivated by the theory. One could know what a moral outlook requires without being motivated. But then again the problem is to choose between M and M´. If on the other hand one cannot have a purely theoretical understanding of moral concepts, the additional practical role of such concepts will be quite difficult to reconcile with the function defining the essence of a concept: to represent. To be sure, one might need to be acquainted with how it is to feel red or sad in order to fully appreciate the meaning of "red" and "sad", but what can an actual motivation to act contribute to the representational function of a concept that cannot be contributed by *simulating* motivation?[19] Furthermore, Blackburn's relativism might require an ability to fully understand diverging moral conceptions.

19 Jackson and Pettit are well aware of this problem (Jackson and Pettit 1995). Nevertheless they suggest that the understanding-motivation link holds. They illustrate their claim by distinguishing two modes of understanding, an *in abstracto*, general mode and a case by case or *in sensu diviso* understanding. The differences in understanding can be illustrated by Pettit using them to evade a circularity in understanding rules. In order to apply a rule, one must use modus ponens. But modus ponens is a rule itself. Jackson and Pettit suggest that there is another way of knowing rules. I do not have to know them as general rules in order to competently apply them. Unfortunately, the analogy fails. In the case of modus ponens, one sometimes cannot come to know what the correct application of a rule is unless one knows it in sensu diviso. Otherwise one would lack certain pieces of propositional knowledge. But the same does not hold in the case of being motivated. One may know well and even articulate what applying a moral term in tune with its rules requires without conforming to these requirements. There is nothing cognitive one lacks. Could it be argued that the case of refusing to conform to a certain moral rule resembles the case of refusing to follow modus ponens? No, one cannot refuse to follow modus ponens in the way one can refuse to follow a moral rule. Even if one tries to withhold assent to some reasons involving modus ponens, one still knows this reasoning to be valid. In the sense of "as-

(iii) Irreducible moral properties: The most radical solution is to build irreducible moral terms into the canonical world descriptions $D + D_w$. They contain moral terms as they contain physical ones. This alternative raises the well-known queerness problems (Mackie 1977, 38–42).

In spite of these concerns, (iii) might prove the most elegant way of dealing with the following considerations: As argued against (i), there must be some reason for using moral vocabulary beyond its achievement as a theory of the world which explains why it is used over and above the canonical world description. One insufficient answer was the otherwise plausible hypothesis that moral theory abbreviates infinitely complex natural truths. Now what enables us to trace the relevant standards? There must be some unifying feature. For instance, in case of colour predicates, this unifying feature may be given by the way our perceptual apparatus is built, or perhaps the irreducible quality of red-experiences. One may be inclined to call this unifying feature the real ethical property. Now this unifying feature distinguishes a moral outlook even from the viewpoint of a canonical description. Furthermore, unity is relative to to the role of a feature in a moral outlook. It must be explainable why this feature qualifies for the role it plays in a moral outlook. I cannot see what else can fulfil the task than an irreducible moral property. Apart from questions of language competence, there is a further consideration leading in the same direction. One may require that if moral realism is true, it should allow to understand the point of an ethical predicate which purports to guide one in life, even or better: especially from the vantage point of the canonical world descriptions. An infinitely complex disjunction without a unifying point does not live up to this task.

References

Benacerraf, Paul: "Mathematical Truth". The Journal of Philosophy, 70, 1973. S. 661–679

Blackburn, Simon: "Rule-Following and Moral Realism". In: *Holtzmann, S./ Leich, C. (ed.)*: Wittgenstein: To Follow a Rule. London, 1981. S. 163–190

Blackburn, Simon: Morals and Modals. Essays in Quasi-Realism. Oxford University Press, Oxford, 1993. S. 52–74

---

sent" at stake, this knowledge is all that assent requires.

Jackson (1998) suggests that moral judgements à la "it is right to ϕ" always involve a component of ϕ being what one would desire under ideal circumstances. Since one is motivated by what one judges to desire under ideal circumstances, one cannot endorse moral judgement without being motivated. But the problem is to cash out "under ideal circumstances". If Jackson must understand them as "if one were a morally ideal agent", the problem is that moral judgement is not any longer be made true by natural facts but by natural facts plus a normative fact: what a morally ideal agent would do.

Blackburn, Simon: "Supervenience Revisited". In: *Hacking, Ian (ed.):* Exercises in Analysis. Cambridge University Press, Cambridge, 1985
Brueckner, Anthony: "Blackburn's Modal Argument against Moral Realism". Theoria, 68, 2008. S. 67–70

Elliot, Robert: "Moral Realism and the Modal Argument". Analysis, 47, 1987. S. 133–137

Hawthorne, John: "A Priority and Externalism". In: *Goldberg, Sanford (ed.)*: Internalism and Externalism in Semantics and Epistemology. i.V.

Jackson, Frank: "What Mary didn't know". Journal of Philosophy, 83, 1986. S. 291–295
Jackson, Frank: From Metaphysics to Ethics. Clarendon Press, Oxford, 1998.
Jackson, Frank/Pettit, Philip: "Moral Functionalism and Moral Motivation". The Philosophical Quarterly, 45, 1995. S. 20–40
Jenkins, Carrie S.: "Knowledge of Arithmetic". British Journal for the Philosophy of Science, 56, 2005. S. 727–747
Jenkins, Carrie S.: "Modal Knowledge, Counterfactual Knowledge and the Role of Experience". The Philosophical Quarterly 58, 2008. S. 693–701

Kitcher, P.: "A Priori Knowledge". The Philosophical Review, 89, 1980. S. 3–23
*Klagge, James C.: "*Supervenience: Perspectives v. Possible Worlds". The Philosophical Quarterly, 37, 1987. S. 312–314

Lewis, David: "Putnam´s Paradox". Australasian Journal of Philosophy, 62, 3, 1984. S. 221–236

McFetridge, Ian: "Supervenience, Realism, Necessity". The Philosophical Quarterly, 35, 1985. S. 245–258
Mackie, John L.: Ethics. Inventing Right and Wrong. Penguin, London, 1977

Peacocke, Christopher/Boghossian, Paul: "Introduction". In: *Peacocke, Christopher/Boghossian, Paul (ed.)*: New Essays on the a Priori. Clarendon Press, Oxford, 2000. S. 1–11
Pust, Joel: "On Explaining Knowledge of Necessity". Dialectica, 58, 2004. S. 71–87

Shafer-Landau, Russ: "Supervenience and Moral Realism". Ratio, 7, 1994. S. 145–152

Sonderholm, Jorn: “A logical Response to Blackburn’s Supervenience Argument”. Sats- Nordic Journal of Philosophy, 8, 2007. S. 178–187

Williamson, Timothy: The Philosophy of Philosophy. Blackwell, Oxford, 2007

Yablo, Stephen: “Coulda, Woulda, Shoulda”. In: *Gendler, T.S./ J. Hawthorne, John (ed.)*: Conceivability and Possibility. Clarendon Press, Oxford, 2002. S. 441–492

Zangwill, Nick: “Moral Supervenience”. Midwest Studies in Philosophy, 20, 1995. S. 240–262

Desire and Action-Directed Tendencies

Daniel Friedrich
daniel.friedrich@philosophie.hu-berlin.de
Berlin School of Mind and Brain, Humboldt Universität zu Berlin

Abstract/Zusammenfassung

According to the Standard View desiring that *p* entails being disposed to act in ways one believes to promote the desired end. The Standard View is clarified and two arguments against the Standard View are considered: Strawson's argument based on passive creatures and an argument based on desires for states of affairs to obtain without the agent's intervention. While it is argued that the first argument is unsuccessful, the second argument seems to show that not all desires entail a disposition to act; some desires seem to have a content that makes it impossible for them to play that role. This conclusion is defended against three objections.

Einer weit verbreiteten Auffassung zufolge impliziert der Wunsch das *p* die Disposition, Handlungen auszuführen, von denen man glaubt, sie trügen zur Wunscherfüllung bei. Diese These wird erläutert, und zwei potentielle Gegenargumente werden diskutiert. Zum einen ein Argument von Strawson, das zu zeigen versucht, dass passive Wesen Wünsche haben können, ohne dass sie eine Handlungsdisposition haben. Zum anderen ein Argument, das sich auf eine spezielle Art von Wunsch konzentriert: den Wunsch, etwas möge der Fall sein, ohne dass die Realisierung durch den Wünschenden vorangetrieben wird. Das erste Argument wird sich als wenig überzeugend herausstellen. Das zweite Argument scheint allerdings zu zeigen, dass nicht alle Wünsche eine Handlungsdisposition implizieren: Einige Wünsche scheinen einen Inhalt zu haben, der dieses unmöglich macht. Dieses Ergebnis wird gegen drei Einwände verteidigt.

According to Michael Smith, desire is a disposition to act. More precisely, Smith holds that desiring that *p* is identical to being disposed to act in ways one believes to further that *p*.[1] However, this view faces a serious objection. Desire is a pro-attitude. To desire a state of affairs is in some sense to view it in a positive light. This is not to say that the desired state cannot also be seen to have its flaws. But it appears to be impossible to desire something and to be *wholly* neutral or *wholly* negative towards it. Desiring, it seems, requires at least *some* kind of endorsement.[i] Moreover, being motivated to X can be rationalized by desiring *p* and believing that X-ing promotes that *p*. Yet, how could desire play this rationalizing role unless it were a pro-attitude? The problem with identifying desire with a disposition to act is that this leaves the pro-attitudinal character of de-

1 Smith (1994): 113. See also Stalnaker (1987).

sire unaccounted for. After all, merely being in a functional state that moves one in various practical directions in accordance with one's instrumental beliefs does not entail that one has a pro-attitude towards being so moved.[2]

For this reason, few people follow Smith in identifying desire with a disposition to act. However, most people think there is more than a small grain of truth in Smith's contention. Indeed, it is generally thought that desiring entails a disposition to act. More precisely, it is generally thought that desiring that *p* entails being disposed to act in ways one believes to further that *p*. Call this the Standard View. In the first section I will sharpen the Standard View. Thereafter, I will consider two arguments designed to show that the Standard View is false. First, an argument by Strawson focusing on passive creatures, and second, an argument focusing on desires for state of affairs to obtain without the agent's intervention. While I shall argue that the first argument is unsuccessful, the second argument should convince us that not all desires entail a disposition to act; some desires have a content that makes it impossible for them to play that role.

Standard View

Saying that desiring that *p* entails being disposed to act in ways one believes to promote that *p* can be understood in two ways. Firstly, it can be understood to mean that desiring something entails some, however weak, actual motivation to act in certain ways one does believe to promote the desired end. Secondly, it can be understood to mean that desiring something entails being disposed to be motivated to act were one to believe that acting thus-and-so promotes the desired end.

Of these two interpretations, the latter is to be preferred for the simple reason that one can desire something without having any beliefs as to how one could act to promote the desired end. I may, for example, desire that the sun shines tomorrow or that Australia wins the World Cup without having any beliefs whatsoever as to what I could do to promote these state of affairs. Yet, without some such belief I cannot be motivated to do anything with the aim of promoting these state of affairs. Consequently, desiring something cannot entail some, however weak, actual motivation to act in certain ways one does believe to promote the desired end.[3]

Better then to understand the Standard View in terms of the claim that desiring that *p* entails being disposed to be motivated to act in ways one believes to promote the desired end. The Standard View, therefore, holds that desire entails

2 Quinn (1993)

3 Might it be said that since desire aims at the attainable (Velleman (2000)) one cannot desire that *p* while also lacking any belief as to how to promote that *p*? No, for this would conflate what is attainable with what is attainable by the agent's own actions.

a certain *disposition*. Dispositions in turn support counterfactuals. To say that an sugar is disposed to dissolve in water may not be *analysable* in terms of counterfactuals, but if a piece of sugar is so disposed then it must be true that it would dissolve in water – at least absent extraneous factors such as the disposition being masked, finked, etc. To further explicate the Standard View we should ask what counterfactuals are supported by the disposition it takes to be entailed by desire. One possible answer would be that the Standard View holds the following counterfactual to be supported by desires: an agent who desires that *p* and believes that she can promote that *p* by X-ing would have at least some minimal motivation to X – or at least, would be so motivated in the absence of any extraneous factors such as the disposition being masked, finked, etc. However, this isn't quite right. Suppose Brian wants to have a drink and believes he could do so by getting a beer out of the fridge. Despite this, Brian might not be motivated to go to the fridge because he might fail to *rationally combine* his desire and belief. Consider an analogy. Suppose Brian believes that *p* and believes that if *p* then *q*. It doesn't follow that he must also believe that *q*. For while *q* must be true if *p* and if *p* then *q* are true, Brian may simply fail to put two and two together. Sometimes such mistakes are due to inattention, but this need not be so. Brian may attend to both of his beliefs yet nonetheless fail to draw the required conclusion. As most of us have plenty of experience to attest, a solution can stare us in the face without us recognising that it is there. Moreover, as Lewis Carroll forcefully argued, the failure to draw a conclusion is not just a matter of a missing piece of knowledge – such as not knowing that if *p* is true and that if it is true that if *p* then *q*, then *q* must be true. Add as many premises as you like, the possibility always remains that one fails to draw the implied conclusion.4 What Brian needs is not more attention or knowledge, but more *rationality*. The same applies in the case of desire. In the absence of sufficient rationality, an agent can always fail to rationally combine her desires and beliefs to come to be motivated. Per contra, it would seem that an *ideally rational agent* who desires that *p* and believes that she can promote that *p* by X-ing must be motivated to X – any failure to be motivated would *ipso facto* be a failure of rationality.[5]

4 Carroll (1895)

5 Perhaps this is not *quite* right. Suppose that Brian also believes that, all things considered, he ought not to go and get a beer out of the fridge. In that case it seems clear that Brian could be fully rational yet lack the *intention* to get a beer out of the fridge. Indeed, it might be said that Brian intending to get a beer out of the fridge would, under these circumstances, constitute a breach of rationality. But then it might be argued that lacking any *motivation* to get a beer out of the fridge can't constitute a failure of rationality. For it might be said that it is odd to suppose it is ever a condition on full rationality that one is both motivated to perform an act and lack the intention to do so. Even if we should accept this – and I'm not sure we should – it doesn't impact upon the argument in any substantive way. For we could deal with this and similar complications by restricting the claim to cases in which the agent has a desire and a suitable belief and does not have any *additional* mental

Pulling these threads together, we arrive at the following explication of the Standard View: to desire that *p* entails being disposed to be motivated to act in ways one believes to promote the desired end. Being so disposed, in turn, makes it true that if one desires that *p* one would be motivated to X were one to believe that X-ing promotes that *p* and were one ideally rational.

The Weather Watchers: A Counterexample?

One potential objection to the Standard View has been put forward by Galen Strawson in his book *Mental Reality*.[6] Strawson asks us to consider the case of the Weather Watchers. These are creatures that are in many respects just like us. They have sensations, thoughts, emotions, beliefs, desires. But there is one fundamental difference: the Weather Watchers are *constitutionally unable to act*. They are just not the kind of creatures that can *do* anything. Their lives are completely passive. They watch and observe (particularly the weather). They have feeling and beliefs, desires and hopes. But all of this occurs in a passive mode. The possibility of the Weather Watchers, Strawson claims, should convince us that desires do not entail a disposition to act.

It is not obvious how this argument is supposed to work. We can't change the weather. Still, it is plausible that if we desire a rainy day, we thereby have a disposition to be motivated to act in ways we believe to promote rainy days. Obviously, since we know that we can't change the weather, we won't be motivated to do anything. Nonetheless, having that desire does seem to dispose us towards action: if we had some suitable belief about how to promote the desired end and were fully rational, then our desire would motivate us to take action. Why should matters be any different in the case of the Weather Watchers?

One suggestion would be that since the Weather Watchers can't act, they could not possibly form a suitable belief. However, this suggestion is rather implausible. Even though we can't change the weather, we might still *falsely believe* we can do so (think of rain-dancers). There seems to be no reason to suppose that the Weather Watchers could not likewise falsely believe that they can do something to influence the weather. A different suggestion would be that just as a super-vase that can't break isn't even disposed to break, so a Weather Watcher that can't act isn't even disposed to act or disposed to be motivated to act. However, once again this does not seem right. We can't act to change the weather, yet this does not undercut us being disposed to be so motivated. True, we can perform other actions, whereas the Weather Watchers can't perform any. Thus, it might be said, we can act in ways we falsely believe to change the

states that would conflict with the rationality of coming to be motivated to perform the action the agent believes will promote the desired end.

6 Strawson (1994): Ch.9.

weather, but the Weather Watchers can't even act on the basis of false belief. However, suppose that unbeknownst to us we have been paralysed. In that case, we can't change the weather nor perform a bodily action that we falsely believe will change the weather. Desiring a rainy day, we could nonetheless be motivated to do a rain dance if we believed that this would make it rain. Again, why should matters be different in the case of the Weather Watchers?

Strawson's reasoning seems to be that one is disposed to be motivated to act only if there are circumstances under which one would act. Yet, there are no circumstances under which the Weather Watchers would act on their desires. This is because *ex hypothesi* the Weather Watchers are *essentially* passive. Thus, any putative circumstances in which they would act on their desires would first have to involve changing their very *nature* and would, therefore, not genuinely be circumstances in which the *Weather Watchers* would act on their desires. Saying that the Weather Watcher who desires a rainy day is disposed to act would, therefore, be much like saying that "a lump of plastic is now disposed to conduct electricity because its constituent parts could be reorganized to constitute gold" (Strawson (1994): 275).

The trouble with this argument is that it depends upon a very strong interpretation of the idea that the Weather Watchers are constitutionally unable to act. A weak interpretation is that the Weather Watchers are unable to act and would only acquire the capacity to act under circumstances that are *remote.* Yet for the argument to go through Strawson requires the strong interpretation according to which it would be strictly *metaphysically impossible* for the Weather Watchers to act. Once this is made explicit, we lose our grip on the example. It is not clear that we can make sense of creatures that are metaphysically unable to act. It is even harder to say whether such creatures could have desires. Arguing on such contentious grounds against a well-established thesis is unlikely to be successful and will invite the response that one person's *modus ponens* is another's *modus tollens*. Nonetheless, I think the Standard View can be shown to be false and in the next section I will try to mount a more convincing argument to that effect.

Non-Actable Desires

According to the Standard View to desire that *p* entails being disposed to be motivated to act in ways one believes to promote the desired end. Being so disposed, in turn, makes it true that if one desire that *p* one would be motivated to X were one to believe that X-ing promotes that *p* and were one ideally rational.

But what if being motivated on the basis of a desire would be *irrational*? In that case, the condition of full rationality would work to *prevent* the agent from coming to be motivated. Consequently, the agent could have the desire, a suitable belief and be fully rational yet not be motivated. Moreover, this lack of mo-

tivation would not be due to any extraneous interference, but be based upon the same structural conditions that are typically responsible for ensuring that the agent is motivated. Thus, we would have to conclude that, at least in these cases, desire does not involve a disposition to be motivated. Consequently, desire would not entail such a disposition and the Standard View would be mistaken.

Is it ever irrational to be motivated in accordance with one's desire? One interesting case concerns *irrational desires.* Which desires, if any, are to count as irrational is controversial.[7] But suppose a desire D is irrational. Suppose further, that the agent believes she could satisfy D by X-ing. Would it be irrational under these circumstances to be motivated to X? This is a difficult question, but on balance I think it wouldn't. This, in any case, is suggested by reflection upon an analogy with theoretical rationality. Suppose a person irrationally believes that *p*. It still seems that, if the person also believes that *p* implies *q*, there is something rational in the agent coming to believe that *q*. We are inclined to say that her starting point is flawed, but that the way she moves forward can't be faulted. Likewise, one could argue that even though D provides a flawed starting point for practical reasoning, the way in which the agent comes to be motivated to X on that basis is flawless.

A second type of desire provides a stronger challenge. Suppose Harry fancies Paula. Suppose, further, that there are a number of things Harry could do that would make Paula fall for him, and that he is well aware of this. But Harry abhors the thought of doing anything in order to make Paula fall for him. Acting with such an end in mind, Harry feels, is contrary to the ideal of romantic love, and in any case, simply not what he desires. Rather, what he wants is for Paula to come to love him without his engaging in any intentional effort designed to make her fall in love with him. In short, Harry has what I call a *non-actable desire*. What is distinctive about non-actable desires is their content. They are desires for a state of affairs to obtain without the intervention of the agent. The fully specified content of a non-actable desire is "*p* without me doing anything to promote that *p*". With non-actable desires there is of necessity *nothing* that the agent can do to promote the desired state of affairs. *Any* attempt designed to further the obtaining of the desired state of affairs is self-defeating.

Non-actable desires are not irrational. But being motivated to act so as to bring the non-actably desired state about is. Or so I shall argue. First, however, let me set aside a potential misinterpretation. Desires motivate agents in light of their beliefs about how to promote the desired end. In the case of a non-actable desire, such motivation would require the belief that one can X so as to promote that *p* without promoting that *p* by X-ing. This is obviously absurd.8 Indeed, it

7 See Parfit (2001) for an interesting discussion.

8 Since such a belief could not possibly be true, this also highlights the difficulties non-actable desires pose for a Stalnaker-style account of the content of desires, that is, an account that holds "to desire that *p* is to be disposed to act in ways that would tend to bring it

isn't even clear whether the belief is so much as *possible*.[9] If it isn't, non-actable desires refute the Standard View simply because the disposition would have impossible manifestation conditions. In that case, the corresponding counterfactual may be true, but this truth would be too weak to support a disposition ascription, since by the same reasoning it could be shown that non-actable desires involve *every* possible disposition, including a disposition *not* to be motivated.

Even if the required belief is possible, it would obviously be a highly irrational belief. And this, it might be thought, is why I claim that motivation on the basis of the non-actable desire would be irrational. Put another way, I could be understood as claiming that it is simply because the motivation would be based on an irrational state that it can't be rational. So understood, however, the case of non-actable desires would only give rise to the same kinds of difficulties present in the case of irrational desires, and an argument based on the latter was already found to be inconclusive. Needless to say, this would not quite capture the argument I want to make. I should also caution against a second misunderstanding. It might also be said that in the case of a non-actable desire, a fully rational agent could not even have the required belief about how to act as to promote the desired end. This, it might be said, shows that the antecedent of the relevant counterfactual in the case of non-actable desires must be impossible: no-one could have the required belief and be fully rational. And in that case, it might be said that a disposition ascription would be unjustified. But again, this is not the basis upon which I object to the Standard View. For it seems that this objection fails to refute the spirit of the Standard View. More precisely, it seems only to show that we have been somewhat imprecise in formulating the exact conditions under which the said disposition manifests itself. Instead of requiring full rationality, we ought to have required merely that the agent is *fully rational in the way she combines her desires and beliefs*.

The real reason why motivation by non-actable desires is irrational is best illustrated by a comparison with the case of theoretical rationality. Typically, it is rational to believe the believed logical consequences of one's beliefs. But not necessarily. Suppose you believe that *p* and also believe not *p*. What rationality requires of you in that case is to resolve that inconsistency and not to believe the conjunction. This is because rationality is an *a priori* guide to true belief, and it is an *a priori* truth that *p* and not *p* can't both be true. So, rationality can't dictate that you believe that *p* and not *p*. *Mutatis mutandis* for desire. It is typically rational to be motivated to do what one regards as promoting one's desires. But not necessarily. Not if you believe you can satisfy your non-actable desire by X-ing. It isn't rational for you to be motivated to X because rationality is an *a pri-*

about that *p* in a world in which one's beliefs, whatever they are, were true" (Stalnaker (1987): 15.)

9 See Stalnaker (1987); Lewis (1986) for proponents of the view that beliefs with contradictory content are impossible. See Crane (1988); Mellor (1988) for the opposing view.

ori guide to the satisfaction of your desires, but it is an *a priori* truth that you can't possibly satisfy your non-actable desire by X-ing. So, rationality can't dictate that you are motivated to X if you non-actably desire that *p* and believe you can satisfy your non-actable desire by X-ing, any more than rationality can dictate that you believe a contradiction just because you hold contradictory beliefs.

We might sum up the argument as follows: Suppose an agent is *not* fully rational in the way she combines her desires and beliefs in her motivations. Then this agent could desire that *p*, hold a suitable belief and not be motivated. To maintain that this agent is nonetheless disposed to be motivated in virtue of her desire, we must make the further claim that she would be so motivated if she held the desire, suitable belief and were *also* fully rational in her motivations. But while this would go to show that desires other than non-actable desires entail a disposition to be motivated, it would not show the same about non-actable desires. For in the case of non-actable desires it would simply not be rational to be motivated to act on the basis of the desire and belief. It would not be rational because rationality is an *a priori* guide to the satisfaction of desire, but in the case of non-actable desires it is an *a priori* truth that every action designed to satisfy the desire would have to fail.

I shall finish this section by briefly considering three objections. First, it might be argued that cases of 'non-actable desires' are, strictly speaking, cases of *wishing* or *hoping*. They are, therefore, *desire-like* states but not desires. Clearly, this objection will go through only if desire-like states do not imply desires – a controversial claim at best. Let us, however, grant that they don't. Yet, why should we assume that cases of 'non-actable desires' *must* be cases of wishing or hoping?

Desire differs from wishing, for example, in so far as desire aims at the *attainable*, whereas wishing aims at what is taken to be *unattainable*.[10] Since in the case of 'non-actable desires' the agent cannot do anything to attain the desired end, it might be said, we are really dealing with wishes and not desires. In reply to this objection we should distinguish between a state of affairs being attainable *tout court* and a state of affairs being attainable through the agent's own actions. It's plausible that desire presupposes believing the desired end to be attainable *tout court*; that is, believing that it is a live possibility. But desire does not also presuppose that one believe that it is attainable by one's own deeds. Once that distinction is in place, it is clear that the objection fails because someone with a non-actable desire can believe that the desired state of affairs is attainable, even though it is not attainable through her own actions.

It's even less clear on what basis one could argue that cases of non-actable desires must be cases of hoping, if only because it's unclear in virtue of what a

10 Velleman (2000): 116-117. Note, Velleman is quite clear that attainable does not mean "attainable by the agent" but rather "being a possible future outcome" (Ibid.).

state is one of hoping rather than desiring. One possibility is that hoping that *p* is set apart from desiring that *p* by being a state that involves being *less confident* about the prospect of *p*. But this would not do to support the present challenge, because someone who non-actably desires that *p* can be quite confident that *p* will come to obtain; it's just that she also thinks this won't be so in virtue of her own doing. To illustrate, it is compatible with Harry knowing that he can't get what he wants by acting so as to make Paula fall in love with him, that he should also be brimming with confidence that she will sooner or later have to recognise what a fantastic catch he is and fall in love with him. Another possibility is that hoping that *p* is distinguished from desiring that *p* because it involves being *anxious* about the possibility of not *p*. But there is no reason to think people like Harry need be troubled by anxiety. Finally, it might be said that hope is desire minus any disposition to act. While this would immunise the Standard View from any challenge, it is quite unpersuasive. Surely, the athlete who hopes to win the race is still disposed to run as fast as she can. In short, there is little reason to think that non-actable desires can't be desires.

Second, it might be argued that I have mischaracterised the nature of non-actable desires. To illustrate, consider another example. Abe wants his son to win the chess championship on his own. He realises that he can't further the desired state of affairs by bribing or drugging the opponent. Abe, it might be claimed, has a non-actable desire. But close attention reveals that there are nonetheless some actions that Abe can take to try and satisfy his non-actable desire. His desire may motivate him to practise with his son or drive him to the tournament if he believes this will promote his son winning on his own. So, it seems that, contrary to what I claimed, the required belief in the case of non-actable desires would not have to be incoherent.

No doubt there are people like Abe. For the objection to succeed, however, it must be shown that there can be no people like Harry – people who desire the obtaining of a state of affairs without them taking *any* action whatsoever. In short, it must be shown that *genuine non-actable* desires are impossible and that any appearance of a non-actable desire only demonstrates a lack of imagination as to what action would, after all, be compatible with the desired state of affairs. I do not see how this claim could be established. Suppose Bert also wants his son to win the chess tournament on his own. But when Bert says 'on his own', he really means it. Bert will shun any action that could improve the chances of his son winning the tournament. He will not practise with his son, he will not drive him to the tournament, and he will not even offer encouraging words. Bert may be a tough guy and a terrible father, but his desire seems possible.

Third, in his book *Mental Reality* Galen Strawson hints at the argument from non-actable desires when he writes

> When I want Wimple rather than Ivanov to win the World Chess Championship, I do not wish – let alone necessarily wish – that I could affect the outcome. I want something to happen. I do not wish that I could do anything about it. Desire does not necessarily involve the will. (Strawson (1994): 287)[11]

In reply, Michael Smith has suggested that the proponent of the Standard View can avoid the challenge by introducing gambles. He says

> a subject desires that Wimple beats Ivanov fairly and squarely only if, in the closest possible world in which she has a desire to gamble, and is offered a choice between a gamble in which the pay-off is that Wimble beats Ivanov fairly and squarely and a gamble in which the pay-off is something that she wants less but assigns only a somewhat higher probability, she chooses the gamble in which the pay-off is that Wimple beats Ivanov fairly and squarely. (Smith (1998): 450)

However, introducing gambles will not save the standard view either. There are two ways to understand Smith's reply, neither of which is successful. According to the first interpretation the desired state of affairs is the *object* of a gamble. Gustav desires England to win. Bet A promises $10 if England wins. Bet B offers $10 if England loses. Gustav thinks it somewhat more likely that England loses. In that case he should take bet B because it has the higher expected pay-off. This is *all* that matters. The fact that Gustav desires that England wins is completely irrelevant. Accepting bet A because of his desire for England to win would be irrational. The same, of course, would be true if the object of a gamble were a non-actably desired state of affairs. So understood, therefore, Smith's reply fails because, contrary to what he claims, the desire is irrelevant to the choice of gambles. According to the second interpretation the desired state of affairs is the *pay-off* of a gamble. Gustav desires England to win and Bet A promises that England wins if X happens. Bet B offers $10 if Y happens. Gustav thinks that Y is somewhat more likely to happen than X. In that case, desiring that England wins will be relevant to the rationality of accepting the bet, as it affects the expected utility of the bet. However, for an agent with a genuinely non-actable desire, there can be no bet that offers the desired state of affairs as a pay-off. For wanting Wimple to beat Ivanov without any intervention of the agent rules out, amongst other things, Wimple beating Ivanov as a result of the subject having made and won a bet with this pay-off.

11 However, Strawson does not elaborate upon this idea, and at another place seems to deny it: "Any desire has the following property: it is necessarily true that there are beliefs with which the desire can combine in such a way as to give rise to, or constitute, a disposition to act or behave in some way" (Strawson (1994): 276). Schroeder (2004): 17-18 may have a similar argument in mind, though his discussion is very brief and the argument remains underdeveloped.

Conclusion

The Standard View has considerable appeal as it is hard to deny that there are intimate connections between desire and action. Nonetheless, I have argued that we should ultimately reject the Standard View because not all desires dispose us towards action. Some desires have a content that makes it impossible for them to play that role.

References

Carroll, Lewis: "What the Tortoise said to Achilles". Mind, 4 (14), 1895. S. 278–80

Crane, Tim: "The Waterfall Illusion". Analysis, 48, 1988. S. 142-147

Lewis, David: On the Plurality of Worlds. Basil Blackwell, Oxford, 1986

Mellor, D. H.: "Crane's Waterfall Illusion". Analysis, 48, 1988. S. 147-150

Parfit, Derek: "Rationality and Reasons". In: *Egonsson, D./Josefsson, J./ Petterson, B./Rønnow-Rasmussen, T.*: Exploring practical philosophy: from action to values. Ashgate, Aldershot, 2001

Quinn, Warren: "Putting Rationality in Its Place". In: Morality and Action. Cambridge University Press, Cambridge, 1993

Schroeder, Timothy: Three faces of Desire. Oxford University Press, Oxford, 2004

Smith, Michael: The Moral Problem. Blackwell Publishing, Oxford, 1994

Smith, Michael: "Galen Strawson and the Weather Watchers". Philosophy and Phenomenological Research, 58 (2), 1998. S. 449-454

Stalnaker, Robert: Inquiry. MIT Press, Cambridge (MA), 1987

Strawson, Galen: Mental Reality. MIT Press, London/Cambridge (MA), 1994

Velleman, J. David: "The Guise of the Good". In: The Possibility of Pratical Reason. Clarendon Press, Oxford, 2000

Drei Arten des Naturalismus

Thomas Hoffmann
Thomas.Hoffmann@ovgu.de
Institut für Philosophie, Otto-von-Guericke-Universität Magdeburg

Abstract/Zusammenfassung

According to John McDowell, the ›aristotelian naturalism‹ in Philippa Foot's *Natural Goodness* is ambiguous. Because Foot omits to make clear, what kind of naturalism her kind of ›aristotelian naturalism‹ is. Following McDowell, there are two fundamental different sorts of naturalism, namely the reductive kind of naturalism, that is typical for modern scientism, and McDowell's own approach of a ›naturalism of second nature‹. In McDowell's opinion, Foot has to decide between these two kinds of naturalism. But if she doesn't want to draw a distorted picture of the relation between ethics, morals, and reason on the one hand and human nature on the other, she has to choose McDowell's naturalism of second nature. In the current article I will question this alternative, suggested by McDowell. I will emphasize that McDowell's criticism goes astray because of a problematic assumption about the so called ›first nature‹, that is characteristical for McDowell's own approach. Then I will outline, how a productive and favourable interpretation of Foot's aristotelian naturalism will lead us to a *third sort of naturalism* that is neither identical with the reductive kind of naturalism nor identical with McDowell's naturalism of second nature. This third kind of naturalism is more suitable to help us to circuit the two horns of the notorious dilemma that appear whenever we think about the role that ethics and morals play in human life: the horn of a reductive naturalism and the horn of a relativistic culturalism.

Der von Philippa Foot in *Natural Goodness* vertretene ›aristotelische Naturalismus‹ ist nach Ansicht von John McDowell missverständlich. Denn Foot versäume es, klar zu machen, welche Art von Naturalismus ihre Art von ›aristotelischem Naturalismus‹ sei. Nach McDowell gibt es zwei grundsätzlich unterschiedliche Arten von Naturalismus, nämlich den reduktiven Naturalismus, der typisch für den modernen Szientismus ist, und McDowells eigenen Ansatz eines ›Naturalismus der Zweiten Natur‹. Laut McDowell steht Foot vor der Wahl zwischen diesen zwei Arten des Naturalismus. Will sie jedoch kein verzerrtes Bild des Verhältnisses zwischen Ethik, Moral und Vernunft auf der einen Seite und der menschlichen Natur auf der anderen Seite zeichnen, so muss sie sich für McDowells Naturalismus der Zweiten Natur entscheiden. Im vorliegenden Artikel werde ich die von McDowell behauptete Alternative in Frage stellen. Ich werde dafür argumentieren, dass die McDowellsche Kritik fehlgeht, da ihr eine problematische Annahme hinsichtlich der so genannten ›Ersten Natur‹ zu Grunde liegt, die kennzeichnend ist für McDowells Position. Dann werde ich zu zeigen versuchen, wie eine produktive und wohlwollende Lesart von Foots aristotelischem Naturalismus uns zu einer *dritten Art von Naturalismus* führt, die weder identisch ist mit dem reduktiven Naturalismus noch mit McDowells Naturalismus der Zweiten Natur. Diese dritte Art des Naturalismus kann uns helfen, die zwei Hörner des notorischen Dilemmas zu umsegeln, das regelmäßig auftaucht, wenn wir darüber nachdenken, welche Rolle Ethik und Moral im menschlichen Leben spielen: das Horn des reduktiven Naturalismus und das Horn eines relativistischen Kulturalismus.

1. Wie schon der Titel *Natural Goodness*[1] verrät, verfolgt Philippa Foot in ihrer 2001 veröffentlichten Abhandlung das Ziel, den Begriff des ethisch (und moralisch) Guten dadurch zu erhellen, dass er in Zusammenhang gebracht wird mit dem Begriff des für den Menschen natürlich Guten. Dies ist die grundsätzliche Stossrichtung ihres Ansatzes, den wir als ›aristotelischen Naturalismus‹ bezeichnen könnten.

Foots Ansatz wurde in unterschiedlicher Weise verstanden, und es hat Verwirrung darüber gegeben, welche Funktion der Verweis auf die Natur und auf das natürlich Gute letztlich genau haben soll. Ein prominenter Ausdruck dieser Verwirrung ist m. E. die als Warnung verpackte Kritik John McDowells an Foots Ansatz, die er in seinem Aufsatz *Two Sorts of Naturalism*[2] geäußert hat. In diesem Aufsatz behauptet McDowell, dass es zwei grundsätzlich verschiedene Arten von Naturalismus gäbe, nämlich den reduktiven szientistischen Naturalismus, der typisch für die Moderne sei, und McDowells eigenen Ansatz eines *Naturalismus der Zweiten Natur*[3], in dem zwischen einer Ersten und einer Zweiten Natur unterschieden wird. Laut McDowell stünde Foot vor der Wahl zwischen diesen zwei Arten des Naturalismus. Will sie jedoch kein falsches Bild des Verhältnisses zwischen Ethik, Moral und Vernunft auf der einen Seite und der menschlichen Natur auf der anderen Seite zeichnen, so sollte sie sich für seinen Naturalismus der Zweiten Natur entscheiden.

Ich möchte nachfolgend die von McDowell behauptete Alternative zwischen dem szientistischen Naturalismus und seinem Naturalismus der Zweiten Natur in Frage stellen, indem ich folgendes tue. Ich werde *erstens* sehr kurz skizzieren, wovor McDowell Foot warnt oder, weniger freundlich ausgedrückt, was er an ihrem Ansatz kritisiert. Ich werde *zweitens* das m. E. zutreffende Verständnis des Footschen Grundgedankens skizzieren. Ich werde *drittens* deutlich machen, inwiefern die McDowellsche Warnung fehlgeht, da ihr eine problematische Annahme hinsichtlich der Ersten Natur zu Grunde liegt, die kennzeichnend ist für McDowells eigene Position. Und ich werde schließlich *viertens* andeuten, inwiefern eine produktive und wohlwollende Lesart von Foots aristotelischem Naturalismus zu einer *dritten Art von Naturalismus* führt, die weder identisch ist mit dem szientistischen Naturalismus noch mit McDowells Naturalismus der Zweiten Natur.

Diese dritte Art des Naturalismus fügt sich nicht nur besser als McDowells eigene Ausführungen in seine m. E. zutreffende metaphilosophische Position ein, die die Unmöglichkeit eines ›Seitenblick-Verständnisses‹ (»*sideways-on view of understanding*«[4]) derjenigen umfassenden Praxis unseres In-der-Welt-Seins betont, in die wir initiiert sind. Vielmehr besteht der Vorteil dieser dritten

1 Vgl. Foot 2001.
2 Vgl. McDowell 1996.
3 Vgl. v. a. McDowell 1994, 4., 5. und 6. Vorl.
4 Vgl. McDowell 1994, 34ff.

Art des Naturalismus auch darin, die zwei notorischen Hörner des Dilemmas umsegeln zu können, die regelmäßig auftauchen, wenn darüber nachgedacht wird, welchen Status Ethik und Moral im menschlichen Leben haben: das Horn des szientistischen Naturalismus und das Horn eines relativistischen Kulturalismus. Mit McDowell gelingt diese Umsegelung m. E. nicht vollends. Mit Foot kann sie dagegen gelingen.

2. Ich beginne mit der sehr kurzen Skizze von McDowells Warnung bzw. Kritik, auf die Philippa Foot selbst leider nie mit der nötigen Klarheit reagiert hat.

McDowell warnt in *Two Sorts of Naturalism* vor der Gefahr, moralphilosophische Überlegungen auf einen szientistischen Naturalismus zu gründen, der Tugend, Ethik und Moral fundieren möchte in den Tatsachen einer naturwissenschaftlich beschriebenen Natur, die McDowell ›Erste Natur‹ nennt.[5] Die Gefahr, die vom szientistische Naturalismus ausgeht, besteht laut McDowell darin, dass er den Raum der Gründe nicht als Raum *sui generis* anerkennt und das begriffliche Vermögen der Vernunft nicht als die natürliche Fähigkeit darstellt, sich im Raum der Gründe zu bewegen.[6] Vielmehr muss, dem szientistische Naturalismus gemäß, alles, was überhaupt als natürlich gelten kann, als Teil der Ersten Natur dargestellt werden, die im kausalen Reich der Naturgesetze beheimatet ist und nicht im rationalen ›Raum der Gründe‹.[7]

Würde man nun – ganz im Sinne des szientistischen Naturalismus – versuchen, Ethik, Moral und Vernunft durch den Verweis auf Notwendigkeiten zu fundieren, die in der ›bloßen‹ Ersten Natur verbürgt seien, so würde man, nach Ansicht McDowells, zu einer fehlgehenden Begründung der Wichtigkeit von Tugenden im Leben von Menschen gelangen. Diese fehlgehende Begründung würde besagen, dass die Orientierung an den Tugenden für jedes Exemplar der menschlichen Spezies etwas natürlich Gutes sei, weil tugendhaft zu sein, einen biologischen Vorteil für Menschen darstellt, welcher durch die empirische Untersuchung der Ersten Natur des Menschen erkannt werden kann. Jeder Mensch, der nicht tugendhaft ist, ist daher irrational oder unvernünftig. Denn er verweigert sich dem, was gemäß seiner Ersten Natur gut für ihn ist. (Dies versucht McDowell in *Two Sorts of Naturalism* an seinem bekannten Beispiel eines Wolfes, der über rationale Fähigkeiten verfügt, zu verdeutlichen.[8])

In einem solchen Bild der menschlichen Praxis wird Ethik, Moral und auch Vernunft auf etwas Non-Normatives und Arationales – auf etwas *biologisch Vorteilhaftes* reduziert. Und dies ist in der Tat eine unbefriedigende Spielart von Naturalismus, die von McDowell zu Recht zurückgewiesen wird. Nur ist es kei-

5 Vgl. McDowell 1996, 167f.
6 Vgl. v. a. McDowell 1994, 4. Vorl. und McDowell 2004.
7 Vgl. hierzu Sellars 1956, 76.
8 Vgl. McDowell 1996, 169-173.

neswegs Foots Position in *Natural Goodness*. Denn die Argumentation verläuft dort merklich anders, nämlich in etwa wie folgt.

3. Laut Foot muss ›... ist gut‹ von seiner logischen Rolle her als attributives Adjektiv begriffen werden.[9] Demnach ist das Urteil ›Dies ist gut‹ ganz analog zu Urteilen wie z. B. ›Dies ist groß‹ oder ›Dies ist gesund‹ zu analysieren, deren Wahrheitswerte nicht unabhängig davon sind, zu welcher Art von Gegenständen derjenige Gegenstand gehört, auf den mit ›dies‹ referiert wird.[10]

Diese logische Grammatik des attributiven Gebrauchs von ›... ist gut‹ bleibt unverändert bestehen – ganz gleich, ob wir uns nun auf Pflanzen, Tiere, Menschen oder auch auf unbelebte Gegenstände beziehen. Während die Qualität von Unbelebtem ausschließlich daran bemessen wird, welchen Nutzen der jeweilige Gegenstand für uns oder für andere Lebewesen hat, kann die Qualität von Lebendigem aber über den Fremdnutzen hinaus beurteilt werden. Lebewesen weisen nämlich auch Eigenschaften und Vollzüge auf, deren Besitz und Ausführung für die jeweiligen Lebewesen selbst gut sind. Diese Form der Gutheit kann als intrinsisches Gutsein oder als natürliche Güte dieser Lebewesen bezeichnet werden.[11]

Der Maßstab der natürlichen Güte besteht in den Eigenschaften der jeweiligen Spezies bzw. Lebensform, von der das einzelne Lebewesen ein Exemplar ist. Diese Eigenschaften werden der Lebensform in den von Michael Thompson so genannten *natural-historical judgements* zugesprochen, welche in Form nonquantifizierter generischer Aussagen (*aristotelian categoricals*) artikuliert werden.[12] Wahre *natural-historical judgements* sagen Tatsachen über eine Spezies aus und implizieren, laut Foot, in Verbindung mit einer entsprechenden ›Teleologie der Spezies‹ ein ›Muster natürlicher Normativität‹, das die speziesimmanenten Kriterien zur Beurteilung der natürlichen Qualität eines Einzelexemplars der jeweiligen Spezies bereitstellt.[13]

Diese formale Struktur des Musters natürlicher Normativität bleibt – so Foots zweite zentrale These (neben derjenigen des attributiven Gebrauchs von ›... ist gut‹) – unverändert bestehen, solange es um die Bewertung der natürlichen Qualität von Lebewesen geht. Was sich jedoch selbstverständlich ändert, sind die Kriterien, aufgrund derer beurteilt werden kann, welcher Wahrheitswert einem

9 Vgl. Geach 1956.

10 Diese Auffassung der logischen Rolle von ›... ist gut‹ wendet sich gegen Moores Analyse; vgl. Moore 1903, Kap. 1, Abs. 4 und 9.

11 Vgl. Foot 2001, Kap. 2.

12 Vgl. Thompson 1995, Thompson 2008, v. a. 65-73, 201-206; vgl. auch Thompson 2004 und Foot 2001, Kap 2. Zur logischen Struktur generischer Aussagen vgl. auch Stekeler-Weithofer 2004, Rödl 2005, Kap. VI, Heuer 2008, 171-187.

13 Foot 2001, 38.

jeweiligen Urteil über die Qualität eines bestimmten Lebewesens zukommt. Denn die Kriterien unterscheiden sich je nachdem, welcher Spezies oder Lebensform dasjenige Lebewesen angehört, auf das sich das Urteil bezieht.[14]

So entspricht es etwa einer natürlichen Norm der Spezies ›Wolf‹, wenn Exemplare dieser Spezies im Rudel jagen. Und daher ist ein Wolf, der im Rudel jagt, ein (natürlich) guter Wolf. Zu den natürlichen Normen der Lebensform Biene gehört es dagegen, dass Exemplare dieser Lebensform einen Stachel besitzen. Und eine Biene, die keinen Stachel besitzt, ist daher ein (natürlich) defektives Exemplar der Lebensform Biene.[15]

4. Ebenso, wie das Jagen im Rudel oder das Verfügen über Stachel zu den signifikanten Vollzügen und Eigenschaften dieser oder jener Spezies gehören kann, kann es auch Lebensformen geben, deren Exemplare als natürliche Eigenschaft das Vermögen einer speziestypischen Form praktischer Rationalität aufweisen. Und in der Tat gibt es Exemplare zumindest einer Lebensform, von denen wir nicht lediglich aufgrund von Beobachtung wissen,[16] dass sie über dieses natürliche Vermögen verfügen: nämlich uns.[17]

Gesunde Exemplare der menschlichen Lebensform verfügen über die Fähigkeiten der praktischer Rationalität und können einen vernünftigen Charakter erwerben sowie einen entsprechenden Willen entwickeln. Und folglich ist es nach Maßgabe des speziesimmanenten Musters natürlicher Normativität auch natürlich gut für den einzelnen Menschen, wenn er dieses Vermögen möglichst vollständig ausbildet, während ein Mensch, der dieses Vermögen nicht besitzt oder nicht bzw. nicht vollständig ausbildet, natürlich defektiv ist.

Die praktische Rationalität des Menschen reduziert sich, so Foot, allerdings nicht auf eine rein formal darstellbare Form von instrumenteller Rationalität oder eine material reichhaltigere Form des rationalen Egoismus. Vielmehr muss zudem die nicht nur prudentielle Qualität der Zwecke bzw. der angestrebten Ziele berücksichtigt werden, wenn es darum geht, zu erläutern, was es heißt, dass ein Mensch praktisch rational ist.

Mit Blick auf die Qualität der Ziele ist es sicherlich richtig, dass es zur praktischen Überlegung eines gesunden Menschen gehört, über seine eigenen Bedürfnisse und Belange zu reflektieren und sich zu fragen, was diesen praktisch am dienlichsten ist. Aber zu behaupten, praktisch rational sei nur derjenige Mensch, dessen praktische Deliberation lediglich am individuellen Eigennutz orientiert ist, wäre eine Einengung des Rationalitätsbegriffs, die gänzlich davon

14 Vgl. Foot 2001, Kap. 2, Kap. 3; vgl. hierzu auch Thompson 2003.

15 Foot 2001, 35; vgl. hierzu auch Geach 1977, 17: »Men need virtues as bees need stings.«

16 Vgl. zum Begriff des Wissens, das nicht auf Beobachtung beruht, d. h. *praktisches Wissen* ist, Anscombe 1957 und Rödl 2007, v. a. Kap.1 u. Kap.2.; vgl. hierzu auch Haase 2010.

17 Vgl. Foot 2001, Kap. 4.

absieht, dass Menschen auch uneigennützige Ziele verfolgen.[18] Entgegen einer reduktiven Metaphysik des rationalen Egoismus, muss eine treffende Darstellung der praktischen Rationalität des Menschen, laut Foot, sowohl die prudentielle Sorge um sich selbst erfassen als auch diejenigen Momente praktischer Deliberation, die nicht auf lediglich eigennützige Ziele ausgerichtet sind.[19] Aber ganz gleich, ob es sich nun um eigennützige oder uneigennützige Ziele handelt, sind ethisch relevante Ziele, dann gute Ziele, wenn sie – so Foot – an dem orientiert sind, was in den mehr oder minder traditionellen Tugenden zum Ausdruck kommt.[20]

Ist der Charakter und der Wille eines Menschen entsprechend orientiert, so ist dieser Mensch nicht nur praktisch rational, was die Wahl der Mittel zur Erlangung beliebiger Ziele betrifft, sondern sein praktisches Denken und sein Handeln sind rational im Sinne der praktischen Weisheit (*phronēsis*). Denn er zeigt nicht nur darin Geschick, die angemessenen Mittel auszuwählen, sondern auch darin, die für Exemplare der menschlichen Lebensform guten Ziele anzustreben. Das Vermögen der praktischen Rationalität manifestiert sich bei diesem Menschen in der gleichsam material bestmöglichen Ausformung praktischer Rationalität. Er verfügt, so könnte man auch sagen, über die vollständig entfaltete Form praktischer Rationalität.[21]

5. Zusammenfassend können wir nun zunächst also festhalten: Praktische Weisheit und die mit ihr verbundene Orientierung praktischer Deliberation an den Tugenden stellt sich bei Foot als die perfekte Form der praktischen Rationalität des Menschen dar.

Dabei ist praktische Weisheit das Resultat einer erfolgreichen Initiation in eine begrifflich strukturierte Praxis, einer gelungenen Erziehung, einer gediegenen Bildung (was auch die kritische Reflexion auf Teile der etablierten Praxis umfasst). Kurzum: praktische Weisheit ist das Resultat von Vorgängen, die wir dem Gebiet der Kultur und nicht demjenigen der Natur zurechnen würden. Nichtsdestotrotz ist die auf kulturellem Wege ausprägbare praktische Weisheit als perfekte Form praktischer Rationalität die perfekte Form von etwas, das – laut Foot – zur Natur des Menschen gehört. Und zwar zu seiner Natur *punktum*! Wollte man dies in McDowells Terminologie ausdrücken, so müsste man wohl sagen: Praktische Weisheit ist die perfekte Form von etwas, das zur *Ersten* Natur des Menschen gehört.

18 Vgl. Foot 2001, Kap. 5.

19 Vgl. Foot 2001, 9f., 16f.

20 Vgl. Foot 2001, Kap. 6.

21 Foot setzt jedoch den Begriff des phronetischen und tugendhaften Lebens nicht *eo ipso* gleich mit dem Begriff des glücklichen oder gelingenden Lebens, wie z. B. McDowell dies tut. Vgl. hierzu Foot 2001, 97f., Foot 2004, 129f., McDowell 1995, McDowell 1998a.

Wenn das die Footsche Position ist, gerät Foot dann aber nicht in Gefahr, einem reduktiven und irreführenden szientistischen Naturalismus das Wort zu reden? McDowell legt dies in *Two Sorts of Naturalism* nahe. Aber warum sieht er diese Gefahr? Der Grund besteht m. E. nicht in einem Manko der Footschen Position, sondern in McDowells eigenem Ansatz eines Naturalismus der Zweiten Natur – genauer gesagt: in der Art und Weise, wie in diesem Ansatz zwischen einer Ersten und einer Zweiten Natur unterschieden wird.

Die Zweite Natur wird von McDowell, so können wir näherungsweise und abkürzend sagen, als eine sehr umfassende Praxis in einem sehr holistischen Sinne dargestellt. Sie ist – wie man auch mit Heidegger sagen könnte – die Gesamtheit einer ›vorgängigen Welterschließung‹.[22] Die Fähigkeit des Menschen, sich in dieser Praxis zu bewegen, beruht vor allem auf seinem Vermögen der Spontaneität, d. h. auf dem begrifflichen Vermögen, das im Denken, im Sprechen, im Handeln und auch in der Erfahrung zum Einsatz kommt. Dieses menschliche Vermögen könnten wir abkürzend auch als das Vermögen bezeichnen, sich in dem von Sellars so genannten ›Raum der Gründe‹ zu bewegen.

Diese Zweite Natur ist aber nicht nur die Zweite Natur des Menschen selbst, sondern damit zugleich auch die Zweite Natur der Welt. Denn mit seiner Initiation in die gleichsam ›geistvolle Praxis seiner Ahnen‹ stellt sich dem Menschen immer auch schon die Welt in Form begrifflich erschlossener Sinn- und Verweisungszusammenhänge dar. Denn der Mensch erfährt qua Initiation in die Praxis die Dinge der Welt zuvorderst *als* diese-und-jene Dinge, und er kann wahrnehmen, *dass* das-und-das der Fall ist. Und bei gelungener Initiation – und vor allem gediegener Bildung – erwirbt er auch die nötige rationale Offenheit gegenüber den jeweiligen situativen empirischen, moralischen u. a. Gründen, die die Welt ihm liefern kann, eben *weil* er eine Zweite Natur erworben hat.

Dieser gleichsam doppelte Begriff der Zweiten Natur wird bei McDowell nun aber vom Begriff der Ersten Natur insofern unterschieden, als die Erste Natur im Gegensatz zur Zweiten Natur das bloße körperliche Wesen des Menschen meint sowie die ›bloße‹ Natur der Außenwelt. Und diesen Begriff von Erster Natur setzt McDowell ohne Weiteres gleich mit dem kausalen ›Reich der Naturgesetze‹, das in den Beschreibungen der Naturwissenschaften zum Ausdruck kommt.

Aber genau diese Gleichsetzung ist m. E. das eigentlich Problematische bei McDowell. Denn aufgrund dieser Gleichsetzung erscheint es nun so, als sei der Bezug auf die bloße Natur ausschließlich den Kausalbeschreibungen der Naturwissenschaften vorbehalten, während wir uns mit dem nicht-naturwissenschaftlichen Vokabular unserer alltäglichen Praxis, das im Raum der Gründe beheimatet ist, offenbar nicht auf die bloße Natur beziehen können. Warum aber, sollte dem so sein? Warum sollten wir annehmen, dass z. B. eine teleologische

22 Vgl. hierzu Heidegger 1927, z. B. 66-88, 95-130, 148-160.

Beschreibung, die Beziehungen formuliert, die im Raum der Gründe daheim sind, nicht mindestens ebenso sehr eine Beschreibung der bloßen Natur sein kann, wie eine kausale Beschreibung, die das Reich der Naturgesetze betrifft?

Meines Erachtens ist es gerade dann, wenn man McDowells Ansatz für eine im Grunde richtige Reaktion auf den szientistischen Naturalismus der Moderne hält, sehr viel plausibler, zu sagen, dass es sich einfach um zwei unterschiedliche Betrachtungsweisen der bloßen Natur handelt, die vor dem begrifflich-normativen Hintergrund derjenigen Praxis, in die wir initiiert sind, vollzogen werden können. Allerdings läuft die Art und Weise, wie McDowell z. B. in *Two Sorts of Naturalism* die Erste Natur als Reich der Naturgesetze abgrenzt gegenüber der Zweiten Natur, eigentümlicher Weise Gefahr, die bloße Natur loszulösen vom Raum der Gründe – und den Naturwissenschaften nicht nur einen privilegierten, sondern den *einzigen* Zugang zur bloßen Natur (bzw. Ersten Natur) einzuräumen.

Eben diesen Zug könnte man durchaus als eine szientistische Tendenz in McDowells eigenem Ansatz betrachten, die jedoch nicht in McDowells Absicht liegen kann, will er seinen programmatischen Versuch der non-szientistischen »Wiederverzauberung«[23] der Natur nicht völlig desavouieren und seine ausdauernde und resolute Kritik am Szientismus der Moderne nicht selbst unterwandern.

6. Wollen wir diese szientistische Tendenz vermeiden und Philippa Foot von der Alternative befreien, entweder einen szientistischen oder einen McDowellschen Naturalismus vertreten zu müssen, so besteht m. E. ein wirksamer Weg darin, einfach McDowells Gleichsetzung von ›Erste Natur‹ und ›Reich der Naturgesetze‹ zu verabschieden, um unter der bloßen Natur schlicht dasjenige zu verstehen, was wir innerhalb unserer Praxis nicht als Kultur und Artefakte auffassen.

Auf eine derart verstandene bloße Natur können wir sowohl mit nicht-naturwissenschaftlichem als auch mit naturwissenschaftlichem Vokabular Bezug nehmen. Und je nachdem welcher Art von Vokabular wir uns bedienen, können die Dinge und Vorkommnisse der bloßen Natur dann entweder als Gegenstand nomologischer Kausalbeschreibungen oder als Gegenstand teleologischer Beschreibungen hervortreten, die im Raum der Gründe beheimatet sind. Aber daran, dass wir uns auf die Natur beziehen, ändert unsere Wahl teleologischer Beschreibungen nichts, sofern wir gewillt sind, den Begriff der Natur aus den naturwissenschaftlichen Klammern zu befreien, die McDowell ihm mit der Gleichsetzung von ›Erste Natur‹ und ›Reich der Naturgesetze‹ immer noch umlegt.

23 Vgl. McDowell 1994, 70.

Ein derartig verstandener Naturbegriff läuft auf eine *dritte Art des Naturalismus* hinaus, die jenseits der Alternative von szientistischem Naturalismus und McDowells Naturalismus der Zweiten Natur liegt. Wir könnten diese dritte Art als *›hermeneutischen Naturalismus‹* bezeichnen – oder eben als ›aristotelischen Naturalismus im Stile Foots‹. Denn es ist m. E. genau die Art von Naturalismus, die Foots Ausführungen zu Grunde liegt – oder zumindest: zu Grunde liegen müsste. Das Bild, das der hermeneutische Naturalismus zeichnet, könnte man kurz und knapp wie folgt skizzieren:

Unser In-der-Welt-Sein ist dadurch gekennzeichnet, dass uns Menschen die Welt sowohl mit ihren natürlichen Elementen als auch mit ihren kulturellen Elementen vorgängig erschlossen ist durch den alltäglich besorgenden Umgang innerhalb derjenigen umfassenden holistischen Praxis impliziter und expliziter begrifflicher Verweisungszusammenhänge, in die wir initiiert sind – und in der sich der Raum der Gründe entfaltet.[24] Naturwissenschaftliche Beschreibungen der Welt, stellen dagegen eine ganz spezielle Teilpraktik dieser umfassenden Praxis des In-der-Welt-Seins von Menschen dar. Und zwar eine bestimmte Praktik des instrumentellen Umgangs mit den Dingen, Vorgängen und Geschehnissen in der uns vorgängig erschlossenen Welt.[25] Diese Praktik des instrumentellen Umgangs ist eine ›Privation‹ der umfassenderen Praxis des In-der-Welt-Seins, da das naturwissenschaftliche Vorgehen die Welt gleichsam um all jene teleologischen, normativen, rationalen, mentalen, semantischen, sozialen und geschichtlichen Aspekte ›beraubt‹, die die Welt mit-konstituieren.

Die menschliche Praktik der naturwissenschaftlichen Beschreibung kann dabei vor dem Hintergrund naturwissenschaftlicher Fragestellungen durchaus zu geeigneten Darstellungen dessen führen, was in der Welt vor sich geht. Nichtsdestotrotz sollte man diese spezielle Praktik mit den Dingen, Vorgängen und Geschehnissen in der Welt nicht hypostasieren. Denn obwohl naturwissenschaftliche Beschreibungen der Welt auf naturwissenschaftliche Fragestellungen geeignete Antworten geben können, die besagen, was in der Welt vor sich geht, ist es allein deshalb nicht schon sinnvoll, auf *alles*, was in der Welt vor sich geht, naturwissenschaftliche Fragestellungen anzuwenden. Weder liefern naturwissenschaftliche Beschreibungen eine umfassende Beschreibung dessen, was Welt ist, noch sind sie die fundamentale Praxis des menschlichen In-der-Welt-Seins. Und ebenso wenig sind sie die grundlegenden, geschweige denn die einzig möglichen Beschreibungen der Natur des Menschen.

7. Fasst man den zuvor betrachteten Gedankengang Foots aus *Natural Goodness* so auf, wie ich ihn zu skizzieren versucht habe, und setzt man dabei das jetzt

24 Vgl. Hoffmann 2007, 293-394.

25 Vgl. hierzu auch Wellmer 2008.

skizzierte Verständnis der Natur des Menschen voraus, so wird klar, dass man sich bei der Erläuterung des für den Menschen Guten durchaus, wie Foot das tut, auf die ›bloße‹ Natur beziehen kann, ohne dabei in die von McDowell vermutete Gefahr zu geraten, szientistische Seitenblicke auf die praktische Vernunft, die Ethik und die Moral zu werfen.

Die Footsche Sicht entgeht nicht nur der vermuteten Gefahr, sondern hat nicht zuletzt auch den Vorteil, dass sie das von McDowell mit den Begriffen ›Erste Natur‹ und ›Zweite Natur‹ nur angedeutete *Kontinuum* zwischen diesen beiden Sphären unseres In-der-Welt-Seins deutlich macht, indem Kulturelles als *praktische Ausgestaltung* und *Vervollkommnung* unserer Natur erscheint.

Dieses Kontinuum kann man m. E. exemplarisch kaum besser verdeutlichen als Philippa Foot mit ihrer Darstellung der praktischen Weisheit als vollständige und beste Manifestation des natürlichen Vermögens der praktischen Rationalität, das uns als Exemplare der menschlichen Lebensform eigen ist.[26]

Literaturverzeichnis

Anscombe, G. E. M.: Intention. Harvard University Press, Cambridge (MA), Paperback-Ausgabe nach 2. Auflage 2000, 1957

Foot, P. (Hrsg.): Theories of Ethics. Oxford University Press, Oxford, 1976/2002

Foot, P.: Natural Goodness. Clarendon Press, Oxford, Paperback-Ausgabe 2003, 2001 [dt.: Die Natur des Guten, übers. v. M. Reuter, Frankfurt a. M.: Suhrkamp 2004]

Geach, P.: "Good and Evil". 1956. In: *Foot, P. (Hrsg.):* Theories of Ethics. Oxford University Press, Oxford, 1976. S. 64-73

Geach, P.: "The Virtues". Cambridge University Press, Cambridge, 1977

Haase, M.: „Drei Formen des Wissens vom Menschen". In: *Hoffmann, T./Reuter, M. (Hrsg.):* Natürlich gut. Aufsätze zur Philosophie von Philippa Foot. Ontos, Heusenstamm b. Frankfurt, 2010. S. 25-74

Halbig, C./Quante, M./Siep, L. (Hrsg.): Hegels Erbe, Suhrkamp, Frankfurt/M., 2004

Heidegger, M.: Sein und Zeit. Niemeyer, Tübingen, 17., durchges. Aufl. 1993, 1927.

26 Eine ausführlichere Argumentation des hier dargelegten Standpunkts findet sich in Hoffmann 2010.

Heuer, P.: Art, Gattung, System. Eine logisch-systematische Analyse biologischer Grundbegriffe. Karl Alber, Freiburg, 2008
Hoffmann, T.: Welt in Sicht. Wahrheit – Rechtfertigung – Lebensform. Velbrück Wissenschaft, Weilerswist, 2007
Hoffmann, T.: „Erste Natur, Zweite Natur und das Gute für den Menschen". In: *Hoffmann, T./Reuter, M. (Hrsg.):* Natürlich gut. Aufsätze zur Philosophie von Philippa Foot. Ontos, Heusenstamm b. Frankfurt, 2010. S. 75-104
Hoffmann, T./Reuter, M. (Hrsg.): Natürlich gut. Aufsätze zur Philosophie von Philippa Foot. Ontos, Heusenstamm b. Frankfurt, 2010
Hursthouse, R./G. Lawrence/W. Quinn (Hrsg.): Virtues and Reasons. Clarendon Press, Oxford, 1995

McDowell, J.: Mind and World. Harvard University Press, Cambridge (MA), 2. Auflage, Paperback-Ausgabe 1996, 1994
McDowell, J.: 1995, "Eudaimonism and Realism in Aristotle's Ethics". In: *McDowell, J.:* The Engaged Intellect. Philosophical Essays. Harvard University Press, Cambridge (MA), 2009. S. 23-40
McDowell, J.: "Two Sorts of Naturalism". 1996. In: *McDowell, J.:* Mind, Value, and Reality. Harvard University Press, Cambridge (MA), 1998. S. 167-197
McDowell, J.: Mind, Value, and Reality. Harvard University Press, Cambridge (MA), 1998
McDowell, J.: "The Role of Eudaimonia in Aristotle's Ethics". In: *McDowell, J.:* Mind, Value, and Reality. Harvard University Press, Cambridge (MA), 1998a. S. 3-22
McDowell, J.: 2004, "Naturalism in the Philosophy of Mind". In: *McDowell, J.:* The Engaged Intellect. Philosophical Essays. Harvard University Press, Cambridge (MA), 2009. S. 257-275
McDowell, J.: The Engaged Intellect. Philosophical Essays. Harvard University Press, Cambridge (MA), 2009
Moore, G. E.: Principia Ethica. Cambridge University Press, Cambridge, Ausgabe von 1993, 1903

O'Hear, A. (Hrsg.): Modern Moral Philosophy. Cambridge University Press, Cambridge, 2004

Rödl, S.: Kategorien des Zeitlichen. Eine Untersuchung der Formen des endlichen Verstandes. Suhrkamp, Frankfurt/M., 2005
Rödl, S.: Self-Consciousness. Harvard University Press, Cambridge (MA), 2007

Sellars, W.: Empiricism and the Philosophy of Mind. Harvard University Press, Cambridge (MA), Ausgabe von 1997, 1956

Stekeler-Weithofer, P.: „Formen, Normen und Begriffe“. In: *Halbig, C./Quante, M./Siep, L. (Hrsg.): Hegels Erbe*, Suhrkamp, Frankfurt/M., 2004. S. 368-400

Thompson, M.: “The Representation of Life”. In: *Hursthouse, R./G. Lawrence/W. Quinn (Hrsg.):* Virtues and Reasons. Clarendon Press, Oxford, 1995. S. 247-297

Thompson, M.: “Tre gradi di bontà naturale”. Iride 38, 2003. S. 191-197; deutsch: “Drei Stufen natürlicher Güte”. In: *Hoffmann, T./Reuter, M. (Hrsg.):* Natürlich gut. Aufsätze zur Philosophie von Philippa Foot. Ontos, Heusenstamm b. Frankfurt, 2010. S. 253-263

Thompson, M.: “Apprehending Human Form”. In: *O’Hear, A. (Hrsg.):* Modern Moral Philosophy. Cambridge University Press, Cambridge, 2004. S. 47-74

Thompson, M.: Life and Action. Elementary Structures of Practice and Practical Thought. Harvard University Press, Cambridge (MA), 2008

Wellmer, Albrecht: “Bald frei, bald unfrei: Reflexionen über die Natur im Geist“. WestEnd. Neue Zeitschrift für Sozialforschung, 5, 2, 2008. S. 3-21

Ist Liebe als Vereinigung eine Bedrohung für die Autonomie der Liebenden? Zum Zusammenhang zwischen Liebe, Identität und Autonomie

Michael Kühler
michael.kuehler@uni-muenster.de
Westfälische Wilhelms-Universität Münster

Einleitung

Liebe im Sinne einer Vereinigung der Liebenden ist mindestens seit Aristophanes' Mythos in Platons *Symposium*[1] eine prominente und zentrale Vorstellung in der Explikation personaler Liebe. Bekanntlich teilte in dem Mythos Zeus die einstmals kugelgestaltigen „Doppelmenschen" in zwei Hälften, also unsere heutige Gestalt, um sie davon abzuhalten, den Olymp zu übernehmen. Jede der Hälften sehnte sich nun verzweifelt nach der Vereinigung mit ihrer ursprünglich zugehörigen anderen Hälfte. Liebe ist also nichts anderes als das Streben nach Vereinigung bzw. als erfüllte Liebe die Vereinigung selbst.[2]

Auch in der aktuellen Diskussion spielt die Vorstellung von Liebe als Vereinigung noch immer eine prominente Rolle. In jüngerer Zeit haben etwa Robert Solomon, Robert Nozick und Mark Fisher Varianten dieser Vorstellung vertreten.[3] Natürlich spricht niemand (mehr) von einem wörtlichen Verschmelzen zweier Menschen zu einem einzigen. Die Liebenden bleiben als körperliche Wesen weiterhin getrennt. Die Metapher der Vereinigung bezieht sich vielmehr auf die Vorstellung, dass die Liebenden eine *Wir*-Identität bzw. ein gemeinsames „*Selbst*" teilen. Es geht also auf psychologischer Ebene um die *qualitative* Identität der Liebenden, nicht um deren numerische Identität. Eine sich im Sinne der Vereinigungsidee verstehende liebende Person wird die Frage, *wer* sie ist, demnach mehr oder weniger stark durch einen Bezug auf die geteilte *Wir*-Identität beantworten. Schwächere Varianten der Vereinigungsidee rekurrieren dann lediglich auf ein bestimmtes Verhältnis zwischen weiterhin vorhandenen indivi-

1 Vgl. Platon: *Gastmahl*, 189c-194e.

2 Das Verhältnis zwischen dem Streben nach Vereinigung und der Vereinigung selbst ist in der Explikation personaler Liebe – im Sinne besonders von *eros* – notorisch umstritten. Für den von mir hier verfolgten Zweck spielt natürlich die Vorstellung einer erfolgreichen Vereinigung die Hauptrolle.

3 Siehe Solomon 1988, bes. 194-199, Nozick 1989, bes. 70-74, sowie Fisher 1990, bes. 26-35. Siehe außerdem, partiell kritisch, Delaney 1996, Friedman 1998, Merino 2004 sowie Kühler 2009.

duellen Identitäten der Liebenden und einer zudem geteilten *Wir*-Identität.[4] Eine starke Variante enthält hingegen die Behauptung, dass sich die Frage *nurmehr* durch einen Verweis auf die gemeinsame *Wir*-Identität beantworten lässt. Alles, was die Liebenden – auch zunächst je einzeln – in relevanter Weise betrifft, betrifft sie demnach *stets* gemeinsam. Die individuellen Identitäten der Liebenden lösen sich somit *vollständig* in der *Wir*-Identität auf.[5]

Es ist vor allem diese starke Variante, die sich dem Einwand ausgesetzt sieht, dass sie die personale Autonomie der Liebenden untergrabe.[6] Wenn Autonomie nicht ohne eine eigenständige Identität zu denken sei, diese aber gemäß der Vereinigungsidee in einer gemeinsamen *Wir*-Identität aufgehe, so könne Liebe als Vereinigung nicht mit personaler Autonomie vereinbar sein. Der Einwand ist damit vor dem Hintergrund des systematischen Zusammenhangs zwischen Liebe, Identität und Autonomie zu sehen.

Im Folgenden werde ich diesen Einwand kritisch auf seine Überzeugungskraft prüfen. Dazu werde ich zunächst noch ein paar erläuternde Worte zur starken Variante der Vereinigungsidee sowie zu den Details des Einwandes sagen. Anschließend werde ich den Einwand im Rahmen zweier prominenter und diametral entgegengesetzter Positionen zu personaler Autonomie diskutieren. Die erste Position ist eine existenzialistische und geht davon aus, dass wir unsere Identität und das, was bzw. wen wir lieben, frei wählen können. Liebe ist damit ein *Resultat* von personaler Autonomie und Identität. Die zweite Position folgt Harry Frankfurts Vorstellung, nach der Liebe im Sinne „volitionaler Notwendigkeit“ die notwendige *Grundlage* der eigenen Identität und Autonomie darstellt.[7] Meine These ist, dass sich der Einwand in beiden Fällen als nicht stichhaltig erweist und die Vorstellung von Liebe als Vereinigung damit keineswegs eine Bedrohung für die personale Autonomie der Liebenden darstellt.

Liebe als Vereinigung und Identität

Zunächst also zur starken Variante der Vereinigungsidee. Von einem – ja durchaus naheliegenden – grundsätzlichen Zweifel daran, ob und wie die Vereinigungsidee überhaupt plausibel expliziert werden kann oder doch nur eine irre-

4 Für eine pointierte Übersicht über die drei wesentlichen Varianten der Vereinigungsidee (kompletter Ersatz der individuellen Identitäten durch eine geteilte *Wir*-Identität; Ergänzung der individuellen Identitäten durch eine zusätzliche *Wir*-Identität; partielle Veränderung der individuellen Identitäten durch wechselseitigen Bezug aufeinander) siehe Merino 2004.

5 So mit Einschränkungen etwa Solomon 1988, bes. 194-208, und vor allem Fisher 1990, 26-35.

6 Vgl. etwa Soble 1997, Friedman 1998 und Helm 2009, Abschnitt 2.

7 Siehe v.a. Frankfurt 1999a, 1999b und 2004, Kap.2.

führende Idealisierung darstellt, möchte ich hier absehen. Ich gehe schlicht davon aus, dass selbst die starke Variante sinnvoll ausbuchstabiert werden kann, so dass sie sich dann dem Einwand zu stellen hat, sie untergrabe die personale Autonomie der Liebenden.

So vertritt etwa Mark Fisher eine relativ starke Variante der Vereinigungsidee.[8] Ihm zufolge entwickelt sich bei gegenseitiger Liebe ein „verschmolzenes Selbst" (*a fused self*). Zum einen „absorbieren" und teilen die Liebenden ihre sämtlichen Vorstellungen und Einstellungen und zum anderen entwickelt sich daraus eine – allerdings niemals vollständig abgeschlossene – „personale Fusion", also eben ein *einziges* „verschmolzenes Selbst". Wahrnehmung, Gefühle, Entscheidungen, Handlungen, an allem haben die Liebenden gemeinsam Anteil und sind insofern als eine einzige Person aufzufassen.[9]

Als körperliche Wesen bleiben die Liebenden natürlich dennoch weiterhin getrennt. Mit der Rede von einem „verschmolzenen Selbst" ist demnach keineswegs notwendig eine ontologische These hinsichtlich einer *neuen Entität* verknüpft.[10] Viel eher handelt es sich um eine These darüber, wie sich die *jeweiligen Identitäten* der Liebenden hin zu einem Bezug auf ein „*geteiltes* Selbst" entwickeln, wie Robert Solomon und Robert Nozick es formulieren.[11] Keiner der Liebenden versteht und sieht sich *selbst* demnach unabhängig von der geliebten Person. Und selbst Außenstehende trennen häufig nicht mehr zwischen den Liebenden, sondern sehen sie als „zusammengehörend". Wer etwa auf die Frage zu antworten versucht, wer Romeo ist, der wird ihn üblicherweise in seiner Beziehung zu Julia beschreiben – und umgekehrt. Beide gehören insofern zusammen, es geht stets um Romeo *und* Julia, wie Solomon weiter erwähnt.[12]

In der Vereinigungsidee geht es realistischerweise also nicht um die Schaffung einer neuen Entität, sondern schlicht darum, dass die individuellen Identitäten der Liebenden durch einen wechselseitigen Bezug aufeinander *redefiniert* werden.[13] Daraus wiederum resultiert die geteilte *Wir*-Identität.[14] Die starke Va-

8 Für das Folgende vgl. Fisher 1990, bes. 26-35.

9 Fisher gibt dabei ausdrücklich zu, dass die individuelle Autonomie der Liebenden dadurch bedroht ist. Das „verschmolzene Selbst" wiederum hat jedoch seine eigene Autonomie. Vgl. Fisher 1990, 27f. Siehe sehr deutlich in diesem Sinne bspw. auch bereits Ortega 1917, 95f., und 1925, 215, in seiner frühen Position zur Liebe.

10 Auch wenn sich etwa Nozick 1989, 70-73, stellenweise in diesem Sinne verstehen lässt.

11 „Love is shared identity" (Solomon 1988, 193). „In a *we*, the people *share* an identity and do not simply each have identities that are enlarged" (Nozick 1989, 82).

12 Vgl. Solomon 1988, 192f.

13 „That is what shared identity means – not a loss of individual identity but a redefinition of personal identity in terms of the other person" (Solomon 1988, 193).

14 In diesem Sinne auch Nozick 1989, 71 und 74, der dabei betont, dass die Autonomie der Liebenden ebenfalls vereinigt wie auch jeweils erhalten bleibt. „People who form a *we* pool not only their well-being but also their autonomy. They limit or curtail their own de-

riante der Vereinigungsidee besteht demnach darin, dass die individuellen Identitäten der Liebenden durch den wechselseitigen Bezug aufeinander *umfassend* und *vollständig* im Sinne der geteilten *Wir*-Identität redefiniert werden.

Eine derart umfassende Redefinition legt natürlich in der Tat den Einwand nahe, dass sie die Gefahr in sich birgt, die jeweilige personale Autonomie der Liebenden zu untergraben. Wie lautet dieser kritische Einwand nun im Detail?

Liebe als Vereinigung und personale Autonomie: die Kritik

Dezidiert formuliert hat den Einwand insbesondere Alan Soble.[15] In einer Vereinigung, in der die Liebenden ihre je eigenständige Identität vollständig zugunsten einer geteilten *Wir*-Identität aufgeben, geben sie zugleich ihre Unabhängigkeit bzw. eben ihre Autonomie auf. Ausgangspunkt für Soble ist Fishers These, dass in einem „verschmolzenen Selbst" selbst bei trivialen Dingen des alltäglichen Lebens kein Unterschied mehr gemacht werden kann, *wer* nun z.B. auf die Idee gekommen ist, erst Pizza zu essen und dann ins Konzert zu gehen.[16] Die Liebenden verlieren demnach ihre jeweils unabhängige Perspektive auf die Welt.[17] Diese aber bräuchten sie, um *selbst* bzw. eben *autonom* Entscheidungen treffen zu können. Anders formuliert: Personale Autonomie ist auf eine *unabhängige* und *eigenständige* Identität angewiesen, die in der starken Variante der Vereinigungsidee durch das „geteilte Selbst" gerade explizit ausgeschlossen wird.[18]

Sobles Beispiele für personale Autonomie beziehen sich auf Entscheidungen des alltäglichen Lebens, z.B. ob man italienisch oder chinesisch essen gehen möchte. Kann man dies nicht mehr *alleine* bzw. *eigenständig* entscheiden, so fasst er dies als Verlust personaler Autonomie auf.

Nun kann man allerdings die wenigsten Entscheidungen des täglichen Lebens alleine bzw. eigenständig treffen. Sobald man in einen sozialen Zusammenhang eingebettet ist, hängen die Entscheidungen und das darauf folgende Handeln nicht mehr nur von einem selbst ab. So bin ich in diesem Sinne keineswegs völlig autonom darin, wann, wo und ob überhaupt ich beispielsweise ein bestimmtes philosophisches Seminar anbiete. So verstanden betrifft der Einwand nicht nur die Liebe als Vereinigung, sondern sämtliche Entscheidungen im Rahmen

---

cision-making power and rights; some decisions can no longer be made alone" (Nozick 1989, 71).

15 Siehe Soble 1997, bes. 70-77.

16 Vgl. Soble 1997, 71

17 „[I]t cancels cognitive autonomy", wie Soble 1997, 72, hervorhebt.

18 Vgl. Soble 1997, 74f.

eines sozialen Eingebundenseins. In dieser Trivialität könnte der Einwand denn auch problemlos zugegeben werden. Er wäre kaum der Rede wert.[19]

Sobles Argument setzt jedoch natürlich an einem interessanteren Punkt an. Denn in unserem „normalen" sozialen Eingebundensein wissen wir sehr wohl, welche Entscheidungen wir selbst treffen würden, d.h. was wir wollten und täten, wenn es nur an uns läge. Im Rahmen der starken Variante der Vereinigungsidee ist genau dies hingegen nicht mehr möglich. Die Liebenden wissen gerade nicht (mehr), wie *sie selbst*, d.h. je einzeln und unabhängig voneinander, entscheiden würden, d.h. was sie je einzeln wollten. Entscheidungen werden von beiden stets auf der Basis der geteilten *Wir*-Identität getroffen und auch das eigene Wollen bildet sich stets sozusagen in einem „*Wir*-Modus".

Die Art personaler Autonomie, die zur Debatte steht, ist also diejenige, die sich auf die Konstitution der eigenen Identität und auf die Bildung des eigenen Wollens bezieht. Es geht somit um die Fähigkeit der Liebenden, den eigenen Willen und die eigene Identität autonom, d.h. frei und eigenständig bzw. authentisch, auszubilden. An diesem Punkt gilt es, den Einwand gegenüber der Vereinigungsidee ernst zu nehmen und kritisch zu prüfen.

Hervorzuheben bei Sobles Einwand ist insofern nochmals, dass er *personale Autonomie* mit *Unabhängigkeit* gleichsetzt und *Unabhängigkeit* wiederum im Rahmen einer *eigenständigen Identität* expliziert. Ist diese Ineinssetzung von eigenständiger Identität und personaler Autonomie, von der die Durchschlagskraft seines Einwandes maßgeblich abhängt, nun aber wirklich überzeugend?

Entscheidend für die Frage, ob Liebe als Vereinigung eine Bedrohung für die personale Autonomie der Liebenden darstellt, scheint mir deshalb zu sein, wie die Vereinigung zustande kommt. In welchem Verhältnis also steht personale Autonomie zur Veränderung der zunächst je individuellen Identitäten der Liebenden hin zu einem „verschmolzenen" bzw. „geteilten Selbst"?

Um diese Frage und damit den Einwand Sobles kritisch zu prüfen, werde ich nun auf die beiden angekündigten und maßgeblichen *individualistischen* Autonomiekonzeptionen zurückgreifen:[20] erstens auf eine *existenzialistische* Position, die davon ausgeht, dass wir unsere Identität und das, was bzw. wen wir lieben, frei wählen können, und zweitens auf eine vor allem von Harry Frankfurt vertretene Position, derzufolge Liebe im Sinne „volitionaler Notwendigkeit" unsere jeweilige Identität festlegt und die so festgelegte Identität wiederum notwendige Grundlage unserer Fähigkeit zu personaler Autonomie ist.

---

19 Für eine gute Übersicht über die verschiedenen Dimensionen der Einschränkung einer so verstandenen personalen Autonomie siehe Friedman 1998, 169-172.

20 Von einem Einbezug *sozial-relationaler* Thesen hinsichtlich des Autonomiebegriffs sehe ich ab, da ihnen zufolge Autonomie von vornherein auf einen sozialen Bezug angewiesen ist (und bleibt) und der Einwand somit ebenso von vornherein ein gewisses Maß an Plausibilität verliert. Ähnliches lässt sich mit Blick auf die Konstitution des Selbst behaupten, eine These, die etwa Solomon 1988, bes. 204-208, denn auch explizit vertritt.

Zurückweisung des kritischen Einwandes 1: Liebe als Vereinigung und Identität als Resultat existenzialistischer personaler Autonomie

Zunächst zur existenzialistischen Autonomiekonzeption. Ihr zufolge haben wir die Fähigkeit zur *Selbstwahl*, d.h. wir können unsere Identität – nach Sartre in radikaler Weise – frei wählen. Zwar findet diese Wahl immer im Rahmen bestimmter biographischer Vorgegebenheiten sowie bestehender Umstände in der Welt statt. Dennoch ist die Selbstwahl keine Willkürentscheidung aufgrund (heteronom) gegebener Motive. Vielmehr betrifft sie unsere grundsätzliche Fähigkeit zur Entscheidung darüber, diese Vorgegebenheiten und Umstände zu reflektieren, so dass wir letztlich unabhängig von diesen Bedingungen entscheiden können, wer wir sein wollen. In dieser beständigen Fähigkeit zur Selbstwahl also besteht unsere personale Autonomie hinsichtlich der Konstitution der eigenen Identität und der Bildung unseres Wollens.

Entscheidend für meine Diskussion hier ist, dass die existenzialistische Autonomiekonzeption folglich mit einer beständigen, grundsätzlichen *Reflexivität* der eigenen Identität einhergeht. Dies wiederum bedeutet zudem, dass die eigene Identität stets *reversibel* bleibt, wir uns hinsichtlich unserer Selbstwahl also stets umentscheiden können.

Geht man nun davon aus, dass die Liebe generell zu den wesentlichen Konstituenten des eigenen Selbst bzw. der eigenen Identität gehört, so umfasst die existenzialistische Autonomie offenbar auch eine „Liebeswahl". Liebe ist demzufolge als ein *Resultat* existenzialistischer Autonomie zu verstehen. Was heißt dies nun wiederum für die Vorstellung von Liebe als Vereinigung?

Existenzialistische Autonomie schließt damit offenbar eine vornehmlich *aktivische* Deutung der Vereinigung ein. Das „geteilte Selbst" muss insofern durch die Liebenden aktiv gewollt und hervorgebracht werden. Auch wenn Solomon, Nozick und Fisher in dieser Hinsicht nicht ausdrücklich Stellung nehmen, so scheinen sie doch zumindest implizit auch die aktive Beteiligung der Liebenden in der Hervorbringung des „geteilten Selbst" im Blick zu haben.[21] Die angestrebte *Wir*-Identität drückt infolgedessen aus, was bzw. wer die Liebenden sein wollen. Diese Selbstwahl müssen sie offenbar zunächst je eigenständig treffen und zudem beständig aufrecht erhalten. Denn gemäß der existenzialistischen Autonomiekonzeption bleibt jegliche Selbstwahl jederzeit reversibel, somit also auch

21 Siehe Solomon 1988, 125 und 199-208, der immerhin ausdrücklich auf Sartre verweist, eine rein individualistische Konzeption der Selbstwahl jedoch ablehnt und stattdessen, wie erwähnt, die sozialen Bedingungen der Konstitution des Selbst hervorhebt. Deutlicher ist Solomon 1988a, 513, wenn er Liebe ausdrücklich als eine Tugend versteht. Nozick 1989, 70-74 und 82f., und Fisher 1990, 22-35, bleiben ebenfalls unklar in dieser Hinsicht, enthalten jedoch ebenfalls zumindest implizit aktivische Elemente. Für eine ausdrücklich primär aktivische Deutung der Vereinigung siehe Kühler 2009.

die Entscheidung für ein „geteiltes Selbst“. Auch die *Wir*-Identität bleibt für beide Liebenden damit jederzeit reversibel. Zwar kann gemäß der starken Variante der Vereinigungsidee bei einem „Rückzug“ von der *Wir*-Identität nicht einfach auf eine parallel noch vorhandene individuelle Identität zurückgegriffen werden. Wie bei individuellen Identitäten, so ist es jedoch auch hier „lediglich“ nötig, eine neue (radikale) Selbstwahl zu treffen.

Solange gemäß der existenzialistischen Autonomiekonzeption also Reflexivität und Reversibilität der eigenen Identität gewährleistet sind, betrifft die Vereinigungsidee stets lediglich die Identität der Liebenden, niemals aber deren Autonomie. Auf diese bleibt Liebe als Vereinigung vielmehr grundsätzlich angewiesen. Liebe als Vereinigung erweist sich hier demnach als *Resultat* einer *existenzialistischen autonomen Entscheidung* beider Liebenden, d.h. als jeweils eine beständige, bewusste und aktive Festlegung auf eine geteilte *Wir*-Identität. Von einer Bedrohung der personalen Autonomie der Liebenden kann hier folglich keine Rede sein.

Wie sieht es nun im Rahmen der zweiten Position zu personaler Autonomie aus?

Zurückweisung des kritischen Einwandes 2: Liebe als Vereinigung als Grundlage von Identität und personaler Autonomie

Das Herzstück der existenzialistischen Autonomiekonzeption ist die Konzeption einer radikalen Selbstwahl, die wiederum zugleich die dort ausschlaggebende personale Autonomie markiert. Diese Konzeption sieht sich allerdings der Kritik ausgesetzt, dass sie keineswegs eine *freie Wahl* sei, sondern bestenfalls eine willkürliche und zufällige und damit letztlich weder eine Wahl sei, noch personale Autonomie ermögliche.[22] Stattdessen hat in jüngerer Zeit prominent vor allem Harry Frankfurt dafür argumentiert, dass personale Autonomie auf *nicht frei gewählte* Aspekte der eigenen Identität angewiesen ist.[23] Denn erst durch diese für das eigene Selbst *wesentlichen* Aspekte sind einem die Kriterien an die Hand gegeben, um die einem bei einer Entscheidung offen stehenden Optionen allererst in autonomer Weise gegeneinander abwägen zu können.

Frankfurt expliziert diese für das eigene Selbst *wesentlichen* Aspekte wiederum als „volitionale Notwendigkeiten“. Sie formen unseren Willen und konstituieren damit unsere Identität. Eine besondere Form volitionaler Notwendigkeit sieht Frankfurt in der Liebe. Liebe ist für Frankfurt dabei ausdrücklich *keine* Sache der Wahl. In der Liebe wird der eigene Wille vielmehr vollständig *genötigt*.

22 Siehe hierfür etwa Taylor 1977, 28-38, und Frankfurt 1999a, 109f.

23 Siehe für das Thema hier vor allem Frankfurt 1999a, 1999b, 1999c und 2004.

Insofern handelt es sich hier um eine passivische Vorstellung von Liebe.[24] Dennoch handelt es sich bei der Liebe für Frankfurt nicht um einen externen Zwang, sondern im Gegenteil um einen *authentischen* Ausdruck des eigenen Willens und somit der eigenen Identität. Deshalb versteht Frankfurt den durch Liebe genötigten Willen denn auch als einen autonomen.[25]

Zwar vertritt Frankfurt nicht explizit eine Vorstellung von Liebe als Vereinigung. Er hebt jedoch die volitionale *Identifikation* mit der geliebten Person und deren Wollen sehr wohl hervor.[26] Die Interessen der geliebten Person werden zu den eigenen. Systematisch gesehen ist es von hier aus denn auch nur noch ein kleiner Schritt hin zur Vereinigungsidee und der Vorstellung eines „geteilten Selbst". Liebe als Vereinigung im Sinne volitionaler Notwendigkeit würde demnach zu einer *Wir*-Identität führen, die sich durch von den Liebenden *geteilte volitionale Notwendigkeiten* auszeichnet. Die Liebenden wären insofern durch ihren je eigenen Willen in authentischer Weise zu einem „geteilten Selbst" „genötigt". Die dabei geteilten volitionalen Notwendigkeiten würden ihnen wiederum die *gemeinsamen* Kriterien an die Hand geben, um – jeweils wie auch gemeinsam – autonome Entscheidungen treffen zu können.

Liebe als Konstituens der eigenen, geteilten Identität bildet im Anschluss an Frankfurts Position also die *authentische* und *notwendige Grundlage* für die personale Autonomie der Liebenden – ganz analog zu individuellen volitionalen Notwendigkeiten. Die Vereinigungsidee erweist sich somit auch an dieser Stelle keineswegs als Bedrohung personaler Autonomie.

Fazit

Folgt man meiner Argumentation, soweit ich sie hier in aller Kürze skizzieren konnte, so müssen sich Vertreter der Vorstellung von Liebe als Vereinigung von dem kritischen Einwand der Bedrohung personaler Autonomie also nicht allzu sehr beeindrucken lassen. Liebe als Vereinigung hat zwar natürlich einen immensen Einfluss auf die Redefinition der Identität der Liebenden – keine Frage. Eine Bedrohung für deren personale Autonomie aber ist sie in keinem der beiden Fälle.

Eine noch verbleibende Möglichkeit, dem Einwand eine gewisse Kraft zuzusprechen, könnte allenfalls darin bestehen, Liebe als Vereinigung im Sinne eines *direkten* Einflusses auf die Autonomie der Liebenden zu verstehen, ohne dabei

24 Frankfurt selbst bezeichnet „seine" Liebe zwar im Gegenteil gerade als aktiv, zielt gegenüber meiner Einordnung hier damit jedoch auf einen anderen Aspekt der Liebe, nämlich letztlich auf eine Gegenüberstellung zwischen egoistischer (passiver) und altruistischer (aktiver) Ausrichtung der Liebe. Vgl. Frankfurt 1999b, 135.

25 Vgl. Frankfurt 1999b, 135ff., sowie 2004, 44ff., 50

26 Vgl. Frankfurt 1999c, 168f., und 2004, 61f.

den Umweg über deren Identität zu gehen. In der *Wir*-Identität zeige sich demzufolge lediglich dieser negative Einfluss. So verstanden müsste Liebe allerdings in Analogie zu negativen Einflussfaktoren, wie z.B. Zwang, Hypnose o.Ä., verstanden werden, da diese allesamt unstrittig als Bedrohung von Autonomie gesehen werden. Insofern wäre Liebe denn auch weder eine authentische Grundlage der eigenen Identität noch Ergebnis existenzialistischer autonomer Entscheidung, sondern vielmehr ein externes, die „gesunde" Beziehung zwischen Autonomie und Identität manipulierendes Element, dem wir mehr oder weniger lange unterworfen wären. Die resultierende *Wir*-Identität der Liebenden wäre demnach letztlich eine pathologische, die treffender vielleicht als (wechselseitige) psychische Abhängigkeit zu beschreiben wäre denn als Liebe.

Der Einwand setzt damit jedoch schlicht eine entsprechend negative Interpretation der Vorstellung von Liebe als Vereinigung voraus – die man ohnehin nicht teilen muss –, anstatt die Gefahr für die Autonomie der Liebenden von einem neutralen Standpunkt aus plausibel aufzuzeigen. Auch in dieser Form erweist sich der Einwand somit als nicht durchschlagend.

Denkt man hingegen an das Phänomen der *Verliebtheit*, so ließe sich dem so verstandenen Einwand wohl eher eine gewisse Plausibilität zusprechen. Denn die Verliebtheit gilt schließlich häufig genug als ein solch „störender" Einflussfaktor. Wenn beispielsweise Ortega von einer „Übereignung durch Bezauberung" spricht, in der man „durch den anderen lebt",[27] und Fisher eine „Absorption der Perspektive der geliebten Person" betont,[28] so scheint diese einseitige Übernahme der Identität der anderen Person die Identität der verliebten Person in einer Weise zu „überschreiben", die in der Tat auch deren Autonomie gefährdet. Da es somit aber um das Phänomen der *Verliebtheit* ginge, erweist sich der Einwand gegenüber der Vorstellung von *Liebe* als Vereinigung schließlich auch in dieser Form als verfehlt.[29]

Literaturverzeichnis

Delaney, Neil: „Romantic Love and Loving Commitment: Articulating a Modern Ideal". American Philosophical Quarterly, 33, 1996. S. 339-356

Fisher, Mark: Personal Love. Duckworth, London, 1990

27 Vgl. Ortega 1925, 215ff.

28 Vgl. Fisher 1990, 26f.

29 Für eine andere Strategie, den (vermeintlichen) Gegensatz zwischen Liebe und Autonomie aufzulösen, siehe Keith Lehrers (Lehrer 1994 und 1997) Konzeption autonomer Liebe. Zwar bezieht sich Lehrer dort nicht auf die Vorstellung von Liebe als Vereinigung. Sein Ansatz lässt sich – analog demjenigen Frankfurts – jedoch durchaus in diese Richtung erweitern.

Frankfurt, Harry G.: Necessity, Volition, and Love. Cambridge University Press, New York, 1999

Frankfurt, Harry G. (1999a): „On the Necessity of Ideals". In: *Frankfurt, Harry G.:* Necessity, Volition, and Love. Cambridge University Press, New York, 1999. S. 108-116

Frankfurt, Harry G. (1999b): „Autonomy, Necessity, and Love". In: *Frankfurt, Harry G.:* Necessity, Volition, and Love. Cambridge University Press, New York, 1999. S. 129-141

Frankfurt, Harry G. (1999c): „On Caring". In: *Frankfurt, Harry G.:* Necessity, Volition, and Love. Cambridge University Press, New York, 1999. S. 155-180

Frankfurt, Harry G.: The Reasons of Love. Princeton University Press, Princeton, 2004

Friedman, Marilyn: „Romantic Love and Personal Autonomy". Midwest Studies in Philosophy, 22, 1998. S. 162-181

Helm, Bennett: „Love". In: *Zalta, Edward N.*, The Stanford Encyclopedia of Philosophy, Fall 2009 Edition. http://plato.stanford.edu/archives/fall2009/entries/love/ [21.05.2010]

Kühler, Michael: „Liebe als Vereinigung im Anschluss an Adam Smith". Allgemeine Zeitschrift für Philosophie, 34, 2009. S. 197-220

Lamb, Roger E. (Hrsg.): Love Analyzed. Boulder, Westview Press, Colorado, 1997

Lehrer, Keith: „Liebe und Autonomie: Ein Vortrag". Conceptus: Zeitschrift für Philosophie, 27, Nr. 70, 1994. S.3-20

Lehrer, Keith: „Love and Autonomy". In: *Lamb, Roger E. (Hrsg.):* Love Analyzed. Boulder, Westview Press, Colorado, 1997. S. 107-121

Merino, Noël: „The Problem with ‚We': Rethinking Joint Identity in Romantic Love". Journal of Social Philosophy, 35, 2004. S. 123-132

Nozick, Robert: „Love's Bond". In: *Nozick, Robert:* The Examined Life. Philosophical Meditations. Simon & Schuster, New York, 1989. S. 68-86

Ortega y Gasset, José: Für eine Kultur der Liebe. In: *Ortega y Gasset, José*: Gesammelte Werke, Bd. 1. Stuttgart, 1978, 1917. S. 91-97

Ortega y Gasset, José: Zur Psychologie des interessanten Mannes. In: *Ortega y Gasset, José*: Gesammelte Werke, Bd. 2. Stuttgart, 1978, 1925. S. 210-228

Platon: Gastmahl. In: Platon. Sämtliche Dialoge, Band III, herausgegeben und übersetzt von Otto Apelt. Meiner, Leipzig, 1993/1926

Soble, Alan: „Union, Autonomy, and Concern". In: *Lamb, Roger E. (Hrsg.):* Love Analyzed. Boulder, Westview Press, Colorado, 1997. S. 65-92

Solomon, Robert C.: About Love: Reinventing Romance for Our Times. Simon & Schuster, New York, 1988

Solomon, Robert C. (1988a): „The Virtue of (Erotic) Love". Midwest Studies in Philosophy 13, 1988, 12-31 (Zitiert nach Wiederabdruck in: *Solomon, Robert C./Higgins, Kathleen M. (Hrsg.):* The Philosophy of Erotic Love, mit einem Vorwort von Arthur C. Danto. The University Press of Kansas, Lawrence, 1991. S. 492-518

Taylor, Charles: „What is Human Agency?". 1977. In: *Taylor, Charles*: Human Agency and Language. Philosophical Papers, 1, Cambridge University Press, Cambridge, 1985 S.15-44 (Zitiert nach der deutschen Übersetzung: „Was ist menschliches Handeln?" In: Taylor, Charles: Negative Freiheit? Zur Kritik des neuzeitlichen Individualismus, 2. Auflage, Suhrkamp, Frankfurt/M., 1995. S. 9-51

Attributive Verantwortung – eine Theorieskizze

Christoph Lumer
lumer@unisi.it
Università di Siena

Abstract/Zusammenfassung

The aim of this contribution is to sketch a general theory of attributive responsibility, which in particular clarifies the semantical and the practical sense of the concept of responsibility.
First, the various meanings of the German expression for responsibility ("Verantwortung") are systematized. Three theoretical fields can be distinguished to which the various concepts of responsility belong: 1. attributive responsibility, i.e. responsibility as ascription of actions, omissions and other events; blame / culpability; liability; 2. responsibility as a not precisely defined obligation, in particular duty of care and accountability; 3. autonomous responsibility: to be aware of and recognize one's responsibility. In the rest of the contribution only attributive responsibility is dealt with. (Sect. 2.)
Clarifying the practical sense of the concept of attributive responsibility is the key to its precise definition and to establishing its exact conditions. The basic idea of attributive responsibility is to identify points of vantage for socially influencing socially relevant events. Certain kinds of actions and omissions are such vantage points, however only under certain conditions. Society can impinge on them by rewards and punishments. Hence, the three main conditions of an ideal attributive responsibility are: the faculty to act intentionally; social controllability of the action; efficient allocation of responsibility. Responsibility for events has to fulfill all of these three conditions; 'responsbility for events' is defined accordingly. (Sect. 3)
Responsbility for actions and for omissions, however, have only to fulfil the first two of these conditions. On the basis of determining the practical sense of attributive responsibility, the conditions for attributive responsbility for actions and omissions are specified. The most important conditions are: acting intentionally or knowingly or culpable omission; the faculty to act differently (in a somewhat weaker version); sanity; lack of shielding the responsibility by interposed actions; no unacceptability. The shielding condition constitues social liberty of the person. (Sect. 4)
The theory elaborated so far, finally, is used to give substantiated answers to some topical questions in the philosophical debate about responsibility: the principle of alternative possibilities and the counterfactual intervener; responsbility and duress; responsibility despite metaphysical lack of alternatives and determinism. (Sect. 5.)

Ziel des Beitrags ist, eine allgemeine Theorie der attributiven Verantwortung zu skizzieren, die insbesondere den semantischen und praktischen Sinn des Verantwortungskonzepts klärt.
Zunächst werden die verschiedenen Bedeutungen des deutschen Ausdrucks "Verantwortung" systematisiert. Es können drei Theoriefelder unterschieden werden, zu denen die diversen Verantwortungsbegriffe gehören: 1. attributive Verantwortung: Verantwortung als Zurechnung von Handlungen, Unterlassungen und sonstigen Ereignissen; Schuld und Haftungspflicht; 2. Verantwortung als nicht genau geregelte Pflicht, insbesondere Fürsorge- und Rechenschaftspflicht; 3. autonome Verantwortung: Verantwortungsbewusstsein und -wahrneh-

mung. Im Rest des Beitrags wird nur die attributive Verantwortung weiter thematisiert. (Abschn. 2.)
Die Klärung des praktischen Sinns der Konzeption der attributiven Verantwortung ist der Schlüssel für deren präzisere Definition sowie für die Festlegung ihrer genauen Bedingungen. Die Grundidee der attributiven Verantwortung ist, günstige Angriffspunkte für eine gesellschaftliche Beeinflussung von sozial relevanten Ereignissen zu identifizieren. Diese günstigen Angriffspunkte sind bestimmte Arten von Handlungen und Unterlassungen, aber auch nur unter bestimmten Bedingungen. Die Gesellschaft kann auf sie durch Belohnungen und Strafen Einfluss nehmen. Die drei Hauptbedingungen idealer attributiver Verantwortung sind dann: Handlungsfähigkeit, soziale Steuerbarkeit des Handelns und effiziente Verantwortungsallokation. Ereignisverantwortung muss alle drei Bedingungen erfüllten; dieser Begriff wird definiert. (Abschn. 3.)
Handlungs- und Unterlassungsverantwortung hingegen müssen nur die ersten beiden Hauptbedingungen erfüllen. Auf der Basis der Bestimmung des praktischen Sinns von attributiver Verantwortung werden dann die Bedingungen für das Vorliegen attributiver Handlungs- und Unterlassungsverantwortung präzisiert. Die wichtigsten Bedingungen sind: absichtlichtes oder wissentliches Tun oder schuldhafte Unterlassung; die Fähigkeit, anders zu handeln (in einer leicht abgeschwächten Version); Zurechnungsfähigkeit; Fehlen einer Abschirmung der Verantwortung durch zwischengeschaltete Handlungen; keine Unzumutbarkeit. Die Abschirmungsbedingung konstituiert soziale Freiheit des Subjekts. (Abschn. 4.)
Die bis hierhin ausgearbeitete Theorie wird schließlich verwendet, um eine begründete Antwort auf einige aktuelle Fragen in der philosophischen Debatte um Verantwortung zu geben: Prinzip der alternativen Möglichkeiten und kontrafaktischer Intervenierer; Verantwortung und Nötigung; Verantwortung trotz metaphysischer Alternativenlosigkeit und Determinismus. (Abschn. 5.)

1. Einleitung und Überblick

In diesem Beitrag wird eine allgemeine Theorie der Verantwortung skizziert, die insbesondere den semantischen und praktischen Sinn des Verantwortungskonzepts klärt und dann auch zu aktuellen Fragen der philosophischen Diskussion um Verantwortung auf einer verbesserten Grundlage Stellung nimmt. Der Beitrag beginnt mit einer Differenzierung der verschiedenen Verantwortungsbegriffe (2), entwickelt dann eine Konzeption des praktischen Sinns einer dieser Arten von Verantwortung, nämlich der attributiven Verantwortung (3). Anschließend werden auf der Basis dieser Bestimmung genauere Bedingungen der attributiven Verantwortung vorgestellt und begründet (4). Schließlich wird angerissen, wie einige der Probleme der aktuellen Diskussion um die Verantwortung mit dieser Konzeption gelöst werden können (5).

2. Überblick über die verschiedenen Verantwortungsbegriffe

Die Ausdrücke "Verantwortung" und "verantwortlich" sind nicht nur reichlich vieldeutig, die so bezeichneten Dinge sind auch miteinander verwandt, stehen aber in einem schwer durchschaubaren Verhältnis zueinander, so dass sie zunächst einer Systematisierung bedürfen. Es können drei Begriffsfelder von 'Verantwortung' unterschieden werden, die jeweils in unterschiedlichen Theorien thematisiert werden. Innerhalb dieser drei Begriffsfelder gibt es dann z.T. mehrere untereinander zusammenhängende Verantwortungsbegriffe. Die folgende Systematik erfasst die wichtigsten Verantwortungsbegriffe. Die Erläuterungen stellen keine Definitionen dar, sondern sollen nur helfen, die diversen Begriffe zu differenzieren.[1]

1. Attributive Verantwortung, häufig auch "retrospektive Verantwortung" genannt: Bei der attributiven Verantwortung geht es direkt um die Frage, wo die aus gesellschaftlicher Sicht entscheidende Ursache für ein Ereignis liegt, aber mit Blick auf die weitere Frage, wen nach einem bestimmten Geschehen gegebenenfalls Sanktionen oder Belohnungen oder Wiedergutmachungspflichten zu treffen haben.

1.1. Zurechnung: Wendungen: "Der Minister übernimmt die Verantwortung." 'Zurechnung' ist der allgemeinste attributionstheoretische Verantwortungsbegriff: Das Handeln oder Unterlassen einer Person werden als die zentrale Ursache bzw. als zentrales Versäumnis bei der Verhinderung von bestimmten Ereignissen ausgemacht. – Zurechnen kann man Handlungen, Unterlassungen und sonstige Ereignisse; die entsprechenden Begriffe sind jeweils leicht unterschiedlich definiert.

---

1 Die folgende Differenzierung verwendet Material von Björn Burkhardts [2000] sehr reicher und verdienstvoller Klassifikation der juristischen Verantwortungsbegriffe. Sie unterscheidet sich allerdings von der Burkhardts in vielerlei Hinsicht. Abgesehen von den hier z.T. anderen Bezeichnungen, kommt bei Burkhardt beispielsweise das, was hier "autonome Verantwortung" genannt wird, nicht vor (dies ist wohl eine Folge des juristischen Fokusses bei Burkhardt); die Rechenschaftsverantwortung ist bei ihm eine eigenständige Form der Verantwortung neben der retrospektiven und der prospektiven, während sie hier als Unterform der prospektiven Verantwortung eingeordnet wird. Der wichtigste Unterschied ist aber, dass Burkhardt die retrospektive Verantwortung als *Verpflichtung* für etwas Geschehenes einzustehen definiert [Burkhardt 2000, 672], während sie hier als einen Schritt vorher angesiedelte soziale Zuschreibung der zentralen Stelle in der Erklärung von Ereignissen konzipiert wird; aus dieser Zuschreibung können dann diverse Konsequenzen gezogen werden: die Haftungspflicht, aber z.B. auch die Bestrafung, die für den Bestraften meist gerade nicht die Form einer Pflicht hat. – Z.T. mögen diese und andere Unterschiede darauf beruhen, dass Burkhardt juristische Verantwortungsbegriffe analysiert, während hier moralische und allgemein zwischenmenschliche Verantwortungsbegriffe thematisiert werden. In der Regel sind diese Verantwortungsbegriffe aber weitgehend gleich.

1.1.1. Handlungszurechnung: Wendungen: "Er ist (selbst) dafür verantwortlich, sein Vermögen verschleudert zu haben / dass er den Präsidenten im Zorn beleidigt hat." Handlungszurechnung bezieht sich darauf, ob ein Verhalten dem Verhaltenssubjekt als von ihm kontrollierte Handlung zugerechnet werden kann.
1.1.2. Unterlassungszurechnung: Wendungen: "Sie ist dafür verantwortlich, dass sie die Bremsen nicht hat kontrollieren lassen." Bei Unterlassungszurechnung wird dem Subjekt eine (ontisch) negative Handlung zugerechnet, nämlich dass es eine Handlung unterlassen, nicht ausgeführt hat (insbesondere eine, die es hätte ausführen sollen), die es hätte ausführen können. (Unterlassungszurechnung macht praktisch meist nur Sinn in Verbindung mit einer Ereigniszurechnung, dass dem Subjekt ein (schlechtes) Ereignis zugeschrieben wird, das nur wegen der Unterlassung eintreten konnte.)
1.1.3. Positive Ereigniszurechnung: Wendungen: "Er ist für den Tod dreier Menschen verantwortlich"; "(weil er die Sicherheitsvorschriften nicht beachtet hat,) ist er dafür verantwortlich, dass sich das Feuer so schnell ausbreiten konnte". Positive Ereigniszurechnung bezieht sich darauf, ob irgendein ontisch positives, stattgefundenes Ereignis einem Subjekt 1. als (Mit-)Verursacher dieses Ereignisses zugerechnet werden kann oder 2. im Zuge der Unterlassungszurechnung zugerechnet werden kann in dem Sinne, dass nicht das Handeln des Subjekts die Ursache des Ereignisses war, dass das Subjekt aber die Möglichkeit gehabt hätte, das Ereignis durch die unterlassene Handlung zu verhindern, und diese Möglichkeit nicht wahrgenommen hat. Ein Sonderfall der positiven Ereigniszurechnung ist, dass das Ereignis wiederum eine andere Handlung (oder ein Charakterzug, Motivation) desselben Subjekts selbst ist; Handlungen, für die keine Handlungszurechnung in Frage kommt (wegen Vollrausch, Zwang oder Nötigung), können dem Handelnden auf diese Weise doch zugerechnet werden [vgl. Aristoteles, NE 1114a-b].
1.1.4. Negative Ereigniszurechnung: Wendungen: "Der Chef des Katastrophenschutzes / die lokale Mafia ist dafür verantwortlich, dass die Katastrophenopfer immer noch keine Notunterkünfte haben." Bei negativer Ereigniszurechnung wird das Ausbleiben eines (erwünschten) Ereignisses – das Ereignis ist also ontisch negativ, nicht existent – einem Subjekt zugerechnet: Dieses Subjekt hat es unterlassen, das erwünschte Ereignis herbeizuführen (insbesondere durch eigene Handlungen oder Beauftragte) oder wesentliche Voraussetzungen für sein Eintreten zu schaffen; oder es hat (aktiv) verhindert, dass der sonstige Gang der Dinge zum Eintreten des Ereignisses geführt hätte (die Mafia hat die Notunterkünfte "abgezweigt"). Die negative Ereigniszurechnung geht also wie die positive immer auf Handlungs- oder Unterlassungszurechnung zurück.

1.2. Schuld: Wendungen: "Die Verantwortung für den Schaden / für das Fiasko liegt ganz bei ihm." Schuld ist zu einem Teil ein Sonderfall von Zurechnung, nämlich Zurechnung von unglücklichen, schädlichen Ereignissen, sie geht aber darüber hinaus dadurch, dass sie Konsequenzen hinsichtlich Sanktionen oder Haftung hat.

1.3. Haftung: Wendungen: "Dafür ziehe ich Sie zur Verantwortung / mache ich Sie verantwortlich." Haftung ist oft, aber nicht zwingend eine rechtliche Folge der Schuld. Haftung ist aber trotz Schuld ausgeschlossen, wenn der Schaden so groß ist, dass der Schuldige gar nicht haften kann. (Umgekehrt setzt Haftung im Zivilrecht nicht immer die Schuld und Zurechnung voraus. Aus moralischer Sicht wird dies oft als fragwürdig angesehen. [Hart 1968, 132])

2. Prospektive Verantwortung, Aufgabenverantwortung, Pflicht i.w.S.:

2.1. (Nicht genau geregelte) Pflicht: Wendungen: "Der letzte Benutzer trägt die Verantwortung dafür (ist dafür verantwortlich), das Haus abzuschließen." 'Verantwortung' im Sinne von Pflicht ist ein Grundbegriff unter den Verantwortungsbegriffen, der nicht auf die attributive Verantwortung zurückgeführt werden kann. Verantwortung in diesem Sinne unterscheidet sich von Pflicht dadurch, dass im Verantwortungsbegriff eine Komponente der Entscheidungsbefugnis und -freiheit enthalten ist.[2]

2.1.1. Fürsorgepflicht: Wendungen: "Verantwortung gegenüber künftigen Generationen"; "Verantwortung für die Kinder"; "unverantwortlich gegenüber allen Betroffenen". Die Fürsorgepflicht ist eine Pflicht mit besonderem Inhalt, eben der Fürsorge für bestimmte Wesen. Weil diese Fürsorgepflicht unbestimmt und nicht klar regelbar ist, wird sie oft als "Verantwortung für" bezeichnet.

2.2. Zuständigkeit: Wendungen: "Die Verantwortung für die Heizung liegt beim Hausmeister"; "ein verantwortungsvoller Posten". Zuständigkeit ist ein spezieller Skopus von Pflicht, nämlich die Pflicht, in einem bestimmten (Zuständigkeits-)Bereich einen gewissen (vage bestimmten) Zielzustand herbeizuführen oder aufrechtzuerhalten.

2.3. Rechenschaftsverantwortung: Wendungen: "Verantwortung vor Gott / gegenüber dem Parlament." Rechenschaftsverantwortung bedeutet, zur Rechenschaft verpflichtet sein, Rechenschaft schulden. Die Rechenschaftspflicht ist eine sekundäre *Pflicht*, nämlich zu belegen, dass man seine primäre Pflicht eingehalten hat. Das Ergebnis dieser Rechenschaft kann dann wie-

2 Die gesetzestechnische Verwendung des Verantwortungsbegriffs resultiert daraus, daß der Gesetzgeber präzise Begrenzungen (der kompetentiellen und sanktionsrechtlichen) Verantwortung vermeiden und das Ausmaß der Bindung des Verantwortlichen verbergen will. Positiv resultiert diese Unbestimmtheit daraus, dass es Aufgaben gibt, deren Bewältigung nicht mittels eines inhaltlich bestimmten Pflichtenkatalogs gesteuert werden kann. [Burkhardt 2000, 673]

derum relevant sein für die attributive Verantwortung und eventuelle Sanktionen – etwa wenn das Subjekt nicht zeigen kann, seine Pflicht erfüllt zu haben, oder wenn es beweist, dass es bestimmte Ereignisse nicht verhindern konnte.

3. Autonome Verantwortung, Aufgabensuche und -wahrnehmung, verantwortungsvoller Charakter: Wendungen: "Er ist ein verantwortlicher / verantwortungsvoller Mensch"; "Verantwortungsethik"; "Verantwortungsbewusstsein". "Verantwortung" in diesem Sinne bedeutet: die gezielte und umsichtige Wahrnehmung moralischer Aufgaben; Möglichkeiten der Einflussnahme erkennen, sich ihnen nicht entziehen, sie in moralisch guter Weise wahrnehmen; (glauben, eine bestimmte Pflicht zu haben).

Häufig wird noch zwischen juristischer, moralischer und allgemeinmenschlicher Verantwortung unterschieden. In der Tat sind beide zu unterscheiden, innerhalb des juristischen Bereichs sogar noch zwischen zivilrechtlicher und strafrechtlicher Verantwortung. Aber dies sind hauptsächlich Unterschiede hinsichtlich der bei der Verantwortungszuordnung anzuwendenden Normen, keine Unterschiede im grundlegenden Verantwortungsbegriff.

Im Folgenden geht es nur um die attributive Verantwortung.

3. Der praktische Sinn der attributiven Verantwortung

Die erste, retrospektive Art der Verantwortung wurde oben "attributive Verantwortung" genannt; damit ist schon eine bestimmte, nun zu entwickelnde Theorie dieser Art von Verantwortung angedeutet. Der Inhalt der attributiven Verantwortung ist, dass bestimmte Ereignisse bestimmten Subjekten (oder, im uneigentlichen Sinne, auch Situationen) attribuiert, zugerechnet werden. Attribution ist im einfachsten Fall die Identifizierung von Ursachen des attribuierten Ereignisses, also eine simple Form der Erklärung dieses Ereignisses. In komplizierteren Fällen geht Attribution darüber hinaus: Menschen werden auch für Unterlassungen für verantwortlich erklärt oder für Gemische aus Handlungen und Unterlassungen sehr indirekter Ursachen – man denke etwa an die Verantwortung von Eltern für die von ihren Kindern angerichteten Schäden. Es geht also nicht einfach um eine wissenschaftliche Erklärung des attribuierten Ereignisses, sondern, grob gesagt: um die Identifikation von gesellschaftlich steuerbaren zentralen Angriffspunkten zur Herbeiführung oder Verhinderung von sozial relevanten Ereignissen. Genauer gesagt, geht es bei der Unterlassungsattribution und bei der Attribution unerwünschter Ereignisse darum, ob man überhaupt und wo man mit zumutbarem Aufwand oder günstig hätte ansetzen können, um ein unerwünschtes Ereignis zu verhindern oder ein erwünschtes herbeizuführen. Übergeordnetes Ziel dieser Attributionen ist es, dazu beizutragen, unerwünschte Er-

eignisse dieser Art in Zukunft auf eine gemessen an ihrer Wünschbarkeit zumutbare und ökonomische Weise zu verhindern. Bei der Attribution erwünschter Ereignisse geht es entsprechend darum, günstige Ausgangspunkte für die Herbeiführung dieser Ereignisse ausfindig zu machen. Insgesamt sucht man also nach Hebeln, an denen man besonders leicht ansetzen kann, um das unerwünschte Ereignis zu verhindern oder ein erwünschtes herbeizuführen. Der Hebel, der gesellschaftlich unmittelbar beeinflussbar ist, ist immer das Handeln der Subjekte; entsprechend bezieht sich attributive Verantwortung letztlich immer auf das Handeln oder Unterlassen von Subjekten: Das Subjekt ist verantwortlich für das ihm zugerechnete Handeln oder Unterlassen selbst und für die ihm zugerechneten Ereignisse. Gibt es keinen Verantwortlichen, ist die Attribution i.e.S. gescheitert. Die fraglichen Ereignisse können dann anderen Entitäten oder Ereignissen i.w.S. attribuiert werden. Man sagt dann – im uneigentlichen Sinn – "die schlechten Wetterverhältnisse waren verantwortlich für den Unfall", um anzudeuten, dass man bei diesen Wetterverhältnissen kaum Sinnvolles oder Ökonomisches hätte tun können, um den Unfall zu verhindern. Ein anderes Beispiel für das Scheitern einer Attribution i.e.S. ist *pathologisches Zwangshandeln*, das die attributive Verantwortlichkeit des Subjekts aufhebt: die Emotionen waren pathologisch unkontrollierbar stark; in diesem Fall werden gewissermaßen die Emotionen für "verantwortlich" erklärt; die ökonomischste Gegenmaßnahme zur Verhinderung dieser Handlungen ist, die Emotionen durch psychotherapeutische o.ä. Behandlung auf ein normales Maß zu reduzieren. Hätte umgekehrt beispielsweise ein Fehlverhalten durch ein zumutbares Maß an Selbstkontrolle verhindert werden können, wird das Fehlverhalten der (tugendtheoretischen) Willensschwäche und damit dem Handelnden zugeschrieben; dem Handelnden können Vorwürfe gemacht werden, um ihn und andere dazu zu bringen, in Zukunft mehr Selbstkontrolle auszuüben.

Dass das fragliche Subjekt die Geschehnisse in der gewünschten Weise hätte handelnd beeinflussen können (im konditionalen Sinn[3]), ist die eine Hauptbedingung für attributive Verantwortung (Handlungsfähigkeit). Die zweite ist, dass die Gesellschaft durch entsprechende Anweisungen und Anreize den potentiell Handelnden so steuern kann, dass er seine Einflussmöglichkeit auch wahrnimmt (soziale Steuerbarkeit). Bei der Ereignisverantwortung kommt als dritte Hauptbedingung noch hinzu, dass das Handeln bzw. Unterlassen des Subjekts ein kausal günstiger Ansatzpunkt für die Beeinflussung des fraglichen Ereignisses ist (effiziente Allokation).[4]

---

3 Ein Handelnder kann im konditionalen Sinn eine Handlung *A* ausführen, gdw. wenn der Handelnde sich entscheiden würde, *A* zu tun, dann würde er auch *A* tun [Moore <1912>/1975, 127].

4 Peter Strawsons [1962] Konzeption der Verantwortungszuschreibung als natürliche und unabänderliche zuschreibende emotionale Reaktion ist von der hier entwickelten ziemlich radikal verschieden. Abgesehen davon, daß Verantwortungskonzepte sich – entgegen

Liegt die attributive Verantwortung für ein (tatsächliches oder mögliches Ereignis) bei einem Subjekt, dann kann dieses Subjekt also das Ereignis herbeiführen bzw. verhindern; und die Gesellschaft kann dieses Ereignis durch Einflussnahme auf das Subjekt steuern. Die klassischen Mittel dieser Steuerung sind Anweisungen verbunden mit Sanktionsandrohungen oder Belohnungsversprechen. Auf diese Weise wird die Konzeption der attributiven Verantwortung auch zu einem Bestandteil der Straftheorie (und natürlich auch der Theorie positiver Anreize). Attributive Verantwortung ist die Voraussetzung für eine sozial sinnvolle Bestrafung. Und die sozial steuernde Form der Strafe, die die günstigen Ansatzpunkte für soziale Einflussnahme auf den Weltverlauf ausnutzt, ist natürlich die generalpräventive Form der Strafe: Sie bedroht jeden, der den günstigen Ansatzpunkt für die sozial erwünschte Beeinflussung des Weltverlaufs nicht wahrnimmt, mit Strafe; etwas genauer bedroht sie jeden mit Strafe, der diesen günstigen Ansatzpunkt hat und verstehen kann, was von der Gesellschaft gefordert wird, und der versteht, dass die Nichteinhaltung dieser Forderungen Strafen zur Folge hat.

Innerhalb der attributiven Verantwortung wurde oben zwischen Handlungs-, Unterlassungs- und positiver und negativer Ereigniszurechnung unterschieden. Von diesen vier Begriffen enthalten die Handlungs- und Unterlassungszurechnung nur die ersten beiden Hauptbedingungen der attributiven Verantwortung: Handlungsfähigkeit und soziale Steuerbarkeit. Bei Ereignisverantwortung muss hingegen auch die dritte Hauptbedingung der attributiven Verantwortung erfüllt sein, die effiziente Verantwortungsallokation, dass das für das Ereignis verantwortlich gemachte Subjekt auch ein guter Ansatzpunkt für die Beeinflussung dieses Ereignisses ist. Aus diversen Gründen ist es aber oft alles andere als offensichtlich, wessen Handeln ein solcher günstiger Ansatzpunkt wäre: Viele könnten mit ungefähr gleichem Erfolg Einfluss nehmen; die Fähigkeiten und Ressourcen zur erwünschten Einflussnahme sind aber unterschiedlich und nicht allseits bekannt; ein in gleich effizienter oder gar effizienterer Weise einflussreiches Subjekt mag an anderer Stelle noch viel Besseres bewirken können; viele erwünschte Effekte sind nur durch gemeinschaftliches Handeln oder Unterlassen erreichbar; den Subjekten ist nur ein gewisses Maß an Engagement zumutbar usw. Ein extremes, aber leider äußerst virulentes Beispiel dafür ist: Wahrscheinlich könnten jeweils Milliarden Menschen hundert Millionen andere Menschen je einzeln vor dem Hungertod retten – ohne dies zu tun. Ist ein beliebiges von jenen Milliarden Subjekten dann dafür verantwortlich, wenn einer der hundert Millionen Menschen hungers stirbt? In einer arbeitsteiligen Welt mit gleichzei-

---

Strawsons Annahme – historisch gewandelt haben, also nicht natürlich und unabänderlich sind, was Strawsons Theorie das anthropologische Fundament entzieht, ist ein großer Nachteil dieser Konzeption, daß sie den praktischen Sinn der Verantwortungszuschreibung nicht erklären kann. Dieser praktische Sinn wird natürlich wichtiger zur Orientierung, wenn Verantwortungskonzepte historisch wandelbar sind.

tig weltweiten Möglichkeiten, aber unterschiedlichen Fähigkeiten und Ressourcen zur Einflussnahme ist eine Regelung von Zuständigkeiten zur koordinierten Verbesserung der Welt erforderlich. Solche Zuständigkeiten werden durch juristische und moralische sozial verankerte Normen geregelt; wegen der Komplexität der Koordinationsaufgabe und der unglaublichen Vielfalt der Möglichkeiten bei gegebenen sozialen Verhältnissen können solche Festlegungen nicht effizient durch (individuell ermittelbare) ideale Regeln vorgenommen werden, sondern größtenteils nur durch sozial geltende Normen.[5] Eine genauere Analyse der Bedingungen der effizienten Verantwortungsallokation und damit eine detaillierte Definition der 'Ereignisverantwortung' kann hier nicht geleistet werden. Stattdessen wird in der folgenden Begriffsbestimmung nur auf die sozial geltenden moralischen Normen verwiesen werden – wohl wissend, dass diese die angerissenen Probleme nicht zufriedenstellend lösen. Genauere Bedingungen werden im Folgenden also nur für die Handlungs- und Unterlassungsverantwortung entwickelt werden.

Bevor die Begriffe der attributiven Verantwortung genauer gefasst werden können, ist noch eine weitere Differenzierung erforderlich. Die eben gegebene Erläuterung des Sinns der attributiven Verantwortung ist *konsiliativ*; sie erläutert, wie Verantwortung nach einem bestimmten, von mir unterstützten Vorschlag *idealiter* organisiert sein sollte. Sie greift jedoch Elemente auf, die in empirisch realisierten Konzeptionen der Verantwortung zu finden sind. Und m.W. orientieren sich alle empirisch realisierten Verantwortungskonzeptionen ein gutes Stück an diesen Ideen. Die konsiliative Konzeption greift also auf empirisch realisierte Konzeptionen der Verantwortung oder Elemente solcher Konzeptionen und deren – bei entsprechender Interpretation – erkennbare Begründungen zurück, kombiniert sie aber auf der Basis einer praktischen Begründung zu einem Ideal, das in dieser Reinheit wahrscheinlich bisher nirgends umgesetzt ist. Deshalb nenne ich die Methode, auf dem dieser Vorschlag beruht: idealisierend-hermeneutisch und praktisch-konstruktiv: Sie rekonstruiert verstehend bisher realisierte Konzeptionen, kombiniert aber deren Elemente und ergänzt sie auf der Basis einer praktischen Begründung zu einer idealen Konzeption.[6]

Man muss also einen allgemeinen normativen Begriff der attributiven Verantwortung, der spezielle Ausprägungen in unterschiedlichen Systemen sozialer Normen zulässt, und die konsiliative, ideale Konzeption der attributiven Verantwortung unterscheiden.

Der allgemeine, sozial-relative Begriff der 'attributiven (positiven Ereignis-)Verantwortung' kann etwa so definiert werden:

---

5 Ob man wegen unterlassener Hilfeleistung z.B. für den Tod eines Menschen juristisch verantwortlich ist, hängt davon ab, ob man zu dieser Hilfeleistung verpflichtet ist. [Duff 1990, 85.]

6 Allgemeine Darlegung dieser philosophischen Methode: Lumer 1989; 1990, 10-19.

Eine Person *s* ist in einem System sozialer Normen *n attributiv verantwortlich* für ein Ereignis *p* := 1. Handlungen von *s* sind gemäß den Regelungen und kausalen Konzeptionen von *n* zentrale Ursachen von *p*, ohne die *p* nicht passiert wäre, oder mögliche Handlungen von *s* hätten gemäß *n p* verhindern können, und 2. *s* ist gemäß *n* jemand, auf den die Normen von *n* Einfluss nehmen oder hätten Einfluss nehmen sollen, um ihn zur Herbeiführung oder Verhinderung von *p* zu bewegen.

Einige Erläuterungen: 1. Im Spezialfall kann das Ereignis *p* selbst eine Handlung von *s* sein (wenn man aus technischen Gründen einmal annimmt, dass die geforderte Verursachung auch Selbstverursachung einschließt.) 2. Nach dieser Definition *machen* die Normen der sozialen Normensystems *n* den *s* verantwortlich; die Verantwortung wird attribuiert, zugesprochen. Dies geschieht allerdings u.U. ohne gute Begründung, in einem fast dezisionistischen Sinn. Nach dieser Definition ist nicht gesichert, dass die Person *s* wirklich *p* hätte herbeiführen oder verhindern können, und schon gar nicht, dass *s* das ideale Subjekt zur Beeinflussung von *p* ist; und es ist auch nicht gesichert, dass die sozialen Instrumente des Normensystems *n* das Subjekt *s* wirklich in der von diesem System gewünschten Weise beeinflussen können.

Nach der Definition der '*idealen* attributiven (positiven Ereignis-)Verantwortung' hingegen wird Verantwortung ohne diese Mängel zugeschrieben:

Eine Person *s* ist *idealiter attributiv verantwortlich* für ein Ereignis *p* := 1. Handlungen von *s* sind zentrale Ursachen von *p*, ohne die *p* nicht passiert wäre; oder mögliche Handlungen von *s* hätten *p* verhindern können; und 2. diese Handlungen von *s* sind / wären ein günstiger Punkt zur Beeinflussung von *p* gewesen, und 3. Belehrungen, Informationen, Belohnungen, Sanktionen sind / wären ein günstiger Anatzpunkt zur sozialen Beeinflussung dieser Handlungen von *s* gewesen; und 4.1. wenn *p* moralisch wünschenswert ist: *s* ist moralisch verpflichtet, oder es ist *s* freigestellt, *p* herbeizuführen; und 4.2. wenn *p* moralisch unerwünscht ist: *s* ist moralisch verpflichtet, *p* zu verhindern.

Erläuterungen: Die Bedingungen 4.1 und 4.2 sehen vor, dass man sowohl für moralisch unerwünschte als auch für moralisch erwünschte Ereignisse verantwortlich sein kann. Und im Fall der moralisch erwünschten Ereignisse ist es nicht erforderlich, dass man zu deren Herbeiführung verpflichtet ist; man kann also nach dieser Konzeption auch für die Folgen supererogatorischer Handlungen verantwortlich sein.

Auf der Basis der eben gelieferten Erläuterung der attributiven Verantwortung kann der Zusammenhang zwischen den wichtigsten Verantwortungsbegriffen erklärt werden.

1. Ausgangspunkt der Konzeption von sozialer Steuerung ist eine basale Form der attributiven Verantwortung: Es gibt eine personale Ursache für (sozial wichtige) Ereignisse, durch deren Veränderung diese Ereignisse besonders leicht zu verhindern oder herbeizuführen sind: Die Person kann diese Ereignisse handelnd beeinflussen; und diese Person ist offen oder sensibel für eine soziale Steuerung durch Strafandrohungen oder Belohnungsversprechen.
2. Die Tatsache, dass jemand etwas tun könnte, das ihm dann attribuiert wird, also die basale attributive Verantwortung, wird dann zur Grundlage einer Regelung gemacht; der Person wird vorgeschrieben, dass sie etwas tun muss. Die Aufgabenverantwortung entsteht.
3. Die Aufgabenverantwortung wird schließlich unmittelbar mit einer stärkeren Form der attributiven Form der Verantwortung verknüpft: der Rechenschaftsverantwortung, der Zurechnung und gegebenenfalls mit der Schuld und Haftung. Man wird für die Nichteinhaltung der Pflicht zur Verantwortung gezogen.

Einige Kommentare zu dieser Ordnung der Verantwortungsbegriffe mögen zum Verständnis beitragen und naheliegende Fragen beantworten:

1. Viele, aber nicht alle Pflichten konstituieren eine attributive Verantwortung, insbesondere dann nicht, wenn das Subjekt zwar die Pflicht erfüllen kann, aber noch nicht voll verantwortlich ist: Kinder, anderweitig Unzurechnungsfähige.
2. Nicht jede retrospektive Verantwortung führt auch zur Verantwortung im Sinne von Haftung, insbesondere dann nicht, wenn der Betreffende zu einer solchen Haftung gar nicht in der Lage ist – die berühmte Übernahme von Verantwortung durch hohe Entscheidungsträger für Schäden, die jedes Maß an individueller Wiedergutmachung überschreiten.
3. Aufgabenverantwortung kann auch durch Festlegung von attributiver, insbesondere retributiver Verantwortung geschaffen werden. Dies gilt z.B. bei Erziehern, denen die retributive Verantwortung für ihre Zöglinge übertragen wird, überantwortet wird. Daraus ergibt sich dann eine bestimmte prospektive Verantwortung, auf die Kinder Einfluss zu nehmen.

4. Spezifizierte Bedingungen attributiver (Handlungs- und Unterlassungs-)Verantwortung

Die Definition der 'idealen attributiven Verantwortung' gibt nur allgemeine Bedingungen an; sie identifiziert keine speziellen Regelungen, durch deren Einhaltung die ideale attributive Verantwortung garantiert ist. Ich werde im folgenden solche spezielleren Bedingungen etwas mehr spezifizieren, ohne aber Detailregelungen anzuführen, sondern nur ein *Schema* angeben, die Richtung der Spezifikationen beschreiben. Außerdem werde ich mich aus den genannten Gründen auf die Begriffe der attributiven Handlungs- und Unterlassungsverantwortung beschränken, also die Ereignisverantwortung vernachlässigen.

Viele Handlungsbeschreibungen haben analytisch die Form: '*s* tut etwas, das verursacht, dass *p*', z.B. 'Sebastian kränkt Hans', was bedeutet: 'Sebastian tut etwas (*x*), was verursacht, dass Hans gekränkt ist.' Durch diesen Bezug auf das

durch die Handlung hervorgebrachte Ereignis entsteht der Anschein, als handele es sich bei der Zuschreibung solcher komplexer Handlungen bereits um die Zuschreibung von Ereignisverantwortung. Dies ist aber schon aus formalen Gründen falsch: Eine Beschreibung der Art '*s* tut etwas, das *p* verursacht' referiert allein auf das Verhalten von *s*; dieses wird lediglich über seine Folgen identifiziert ('dasjenige Verhalten, das verursacht hat, dass *p*';[7] Ereignisverantwortung hingegen schreibt dem Subjekt *s* die Folge *p* zu, macht *s* auch für die Folge verantwortlich, nicht nur für das Verhalten. Der inhaltliche Unterschied ist, dass bei der Ereignisverantwortung zusätzlich die Bedingung der effizienten Verantwortungsallokation erfüllt sein muss. Nun scheint es merkwürdig, dass Sebastian für das Kränken von Hans, nicht aber für Hans' Gekränktsein verantwortlich sein soll. In der Tat fallen die Handlungsverantwortung für über ihre beabsichtigten Folgen identifizierte Handlungen und Ereignisverantwortung für diese Folgen meist zusammen; es gibt aber auch Ausnahmen (s.u.). Bei der Unterlassungsverantwortung vs. Ereignisverantwortung für die nicht herbeigeführten Folgen ist die Diskrepanz aber wegen der vielen nicht wahrgenommenen Möglichkeiten, Gutes zu tun, ziemlich groß. Sollte man dann nicht die Extension der Unterlassungsverantwortung (und auch der Handlungsverantwortung) entsprechend beschränken? Nein. Bei der Handlungs- und Unterlassungsverantwortung geht eben nur um die ersten beiden Hauptbedingungen der attributiven Verantwortung (Handlungsfähigkeit und soziale Steuerbarkeit), insbesondere ob ein Verhalten zunächst individuell und dann auch sozial steuerbar ist. Eine Daumenregel dabei ist, dass wir meist für unsere absichtlichen Handlungen und Unterlassungen verantwortlich sind.

Bedingungen für ideale attributive (Handlungs-)Verantwortung: Ein Subjekt *s* ist (ideal attributiv) verantwortlich für eine Handlung *a*, dass *s* zu *t A* tut, gdw.

H1. *Objektiver Tatanteil: Handlung:* Dass *s A* tut, ist eine Handlung.

H2. *Abgeschwächtes Prinzip der alternativen Möglichkeiten:* Grob: Das Subjekt *s* hätte (konditional) anders handeln können; oder *s* hätte (konditional) anders handeln können, wenn *s* nicht durch Überdeterminierung daran gehindert worden wäre.

H3. *Subjektiver Tatanteil:* absichtliches oder wissentliches Tun: *s* tut *A* zielabsichtlich, mittelabsichtlich oder nur wissentlich;[8] wenn die Tat sozial sehr

7 Dies ist Davidsons 'Ziehharmonikaeffekt' [Davidson <1971>/1985, 81-84; 87-98].

8 '*s* tut *A* zielabsichtlich' bedeutet ungefähr: '*s* tut *A*' impliziert, dass mit dem Handeln eine bestimmte Folge herbeigeführt wird; diese Folge war in *s*' Deliberation das mit der Handlung angestrebte Ziel; *s* führt die Handlung aufgrund dieser Deliberation gesteuert aus; und die angezielte Folge tritt ungefähr auf dem von *s* angedachten Weg ein. '*s* tut *A* mittelabsichtlich' ist analog definiert. 'Dass *s A* wissentlich tut' schließlich ist ebenfalls analog definiert; es gilt aber insbesondere: *s* nimmt in seiner Deliberation an, dass sein Handeln

schädlich ist, tut *s A* mindestens wissentlich; wenn die Tat weniger wichtig, aber noch keine Bagatelle ist, tut *s A* mindestens mittelabsichtlich; wenn die Tat eine Bagatelle ist, tut *s A* zielabsichtlich (d.h. *s* muss *A* schon zielabsichtlich tun, um verantwortlich zu sein). (Kurz: je gravierender die Tat ist, desto geringer kann der subjektive Tatanteil sein, ab dem man schon verantwortlich ist.) Wenn die Tat *a* hingegen sozial erwünscht ist, hat *s* mindestens mittelabsichtlich *A* getan. Verringerte subjektive Wahrscheinlichkeit der relevanten Handlungsfolge führt auch zu einer verringerten Verantwortung für diese Folge; denn mit der verringerten subjektiven Wahrscheinlichkeit sinkt auch die Kontrolle und Steuerung dieser Folge [Roughley 2007]. – Durch diese Bedingungen (insbesondere die Absichtlichkeit) werden auch abwegige Absichtsrealisierungen als Objekte der Verantwortung ausgeschlossen, also beabsichtigte, aber nur zufällig eingetretene Taten und Handlungsfolgen.[9]

H4. Zurechnungsfähigkeit:

H4.1. Generelle Zurechnungsfähigkeit: Das Subjekt *s* hat ein gewisses Alter erreicht, ist nicht völlig schizophren, kann deliberieren und nach den Deliberationsergebnissen handeln etc.

H4.2. Aktuelle Zurechnungsfähigkeit: Die Handlung *a* muss aus Überlegungen von *s* entstammen, darf z.B. nicht durch Hypnose verursacht sein; und die Deliberationsfähigkeit von *s* darf zum Entscheidungszeitpunkt nicht substantiell eingeschränkt sein. (Zwanghaftigkeit beispielsweise schließt Zurechnungsfähigkeit im Sinne von Beeinflussbarkeit durch gute Gründe aus. Zwanghaftigkeit liegt vor, wenn gilt: Selbst wenn das Subjekt *s a* als suboptimal bewertet hätte und *s* zur Verfügung stehende Selbststeuerungstechniken eingesetzt hätte, hätte *s* immer noch *a* ausgeführt [vgl. Kennett 2001, 157]. Willensschwäche, die mit synchronen, also im Moment der Entscheidung anwendbaren, Selbststeuerungsmechanismen überwunden werden könnte, ist noch keine Unzurechnungsfähigkeit. Denn der in diesem Sinn Willensschwache hätte die Möglichkeit zur Selbststeuerung, übt sie aber nicht aus [ibid. 155].)

H5. Keine Abschirmung der Verantwortung durch andere Verantwortliche:

H5.1. Ausschluss der Fremdsteuerung: Die Entscheidung von *s* für *a* war nicht durch eine Strukturierung der Entscheidungssituation durch andere subjektiv nötigend; es lag also keine echte Nötigung, zwingende Anstiftung, Handlung ex ufficio, kein Befehl o.ä. vor [Duff 1990, 83]. (Die Verantwortung liegt in solchen Fällen bei dem, der die Situation strukturiert hat. Beispiel: Wenn der Kultusminister die Schulen schließt (indem er entsprechende Anweisungen gibt), schließt zwar am Ende der einzelne

---

die in der Beschreibung '*s* tut *A*' implizierte Folge mit mindestens mittlerer Wahrscheinlichkeit herbeiführen wird; aber diese Folge ist kein Handlungsziel von *s*.

9 Ausführlich zur abwegigen Absichtsrealisierung: Lumer 2008.

Hausmeister die Schule ab; aber er ist dafür nicht handlungsverantwortlich – ebenso wenig die ganze zwischengeschaltete Befehlshierarchie –, wohl aber der Kultusminister für das Schließen der Schulen.)

H5.2. Keine zwischengeschaltete Handlung: Der Handlungserfolg ist nicht vermittelt über das mit Blick auf diesen Handlungserfolg wissentliches Handeln einer anderen zurechnungsfähigen Person eingetreten. (In diesem Fall ist der in der Folgenkette früher Handelnde nicht verantwortlich. Beispiel: Sebastian kränkt Hans, indem er Dora vertraulich erzählt, wie einfältig Hans doch sei, wohl wissend, dass die klatschsüchtige und gehässige Dora dies Hans kolportieren werde. In diesem Fall ist Sebastian nicht dafür handlungsverantwortlich, Hans gekränkt zu haben, wohl aber Dora.[10]) Ausnahmen von dieser Regel sind: Die früher agierende Person hat die Entscheidungssituation des später Handelnden strukturiert durch Nötigung, Befehl, Auftrag, Anstiftung o.ä.; oder der später Handelnde erfüllt andere der hier genannten Verantwortungsbedingungen nicht, er ist z.B. nicht (voll) zurechnungsfähig (und die früher agierende Person ist zuständig für das Handeln der später agierenden Person); oder früher Handelnde schafft wissentlich notwendige oder hinreichende Voraussetzungen für eine vom später Handelnden geplante kriminelle Handlung. (Beispiel: *s* verkauft wissentlich einem Killer eine Waffe.) [Duff 1990, 85 f.]

H6. Zumutbarkeit: A (oder ein substantiell ähnliche Handlung A^*) nicht zu tun war dem Subjekt *s* zumutbar – die Unterlassung war z.B. nicht zu kostspielig für *s* im Verhältnis etwa zum Wert des durch ein Verbot von *a* zu schützenden Guts.

10 Man kann sich aber auch auf den Standpunkt stellen, diese Bedingung sei zu juristisch; mindestens moralisch sei Sebastian durchaus mitverantwortlich, weil er die verwerfliche Handlungsweise Doras vorausgesehen und gezielt ausgenutzt hat. In diesem Fall wird die unten noch angeführte Aufhebung der Abschirmung bei Mithilfe zu einer kriminellen Handlung weiter ausgelegt. – Handlungsutilitaristische oder, allgemeiner, handlungswelfaristische Ethiken erkennen die Abschirmbedingung überhaupt nicht an; aber für sie macht die ganze attributive Verantwortung als Teil einer stark auf Normen aufbauenden Ethik ohnehin keinen Sinn. Normenutilitaristische oder normenwelfaristische und deontologische Ethiken hingegen sehen unterschiedlich starke Abschirmbedingungen vor. Kants berüchtigtes Verdikt über das vermeinte Recht, aus Menschenliebe zu lügen, setzt eine äußerst starke Verantwortungsabschirmung voraus. Über den Sinn der Verantwortungsabschirmung in normwelfaristischen Ethiken wird unten noch einiges ausgeführt.

Bedingungen für ideale attributive (Unterlassungs-)Verantwortung: Ein Subjekt *s* ist (ideal attributiv) verantwortlich dafür, eine Handlung *a*, dass *s* zu *t A* tut, unterlassen zu haben, gdw.

U1. *Objektiver Tatanteil:* Unterlassung: *s* tut nicht absichtlich *A*.

U2. *Ausführungsmöglichkeit: s* könnte (konditional) *A* tun; d.h. wenn *s* sich entscheiden würde, *A* zu tun, würde *s* auch *A* tun.

U3. *Subjektiver Tatanteil:* absichtliche oder schuldhafte Unterlassung: *s* unterlässt absichtlich *A* (d.h. hat an die Handlungsmöglichkeit *a* gedacht, sich aber für eine andere Alternative entschieden); oder (wenn *s* nicht an *a* gedacht hat) es ist *s*' Pflicht, sich über die obligatorischen Ausführungsbedingungen zum A-Tun und ihre eventuelle aktuelle Erfüllung auf dem laufenden zu halten.

U4. *Zurechnungsfähigkeit:*

U4.1. *Generelle Zurechnungsfähigkeit:* Wie H4.1: Das Subjekt *s* hat ein gewisses Alter erreicht, ist nicht völlig schizophren, kann deliberieren und nach den Deliberationsergebnissen handeln etc.

U4.2. *Aktuelle Zurechnungsfähigkeit:* Die Deliberationsfähigkeit von *s* darf zum Entscheidungszeitpunkt nicht substantiell eingeschränkt sein.

U5. *Keine Abschirmung der Verantwortung durch andere Verantwortliche:*

U5.1. *Ausschluss der Fremdsteuerung:* Die Entscheidung von *s* gegen *a* war nicht durch eine Strukturierung der Entscheidungssituation durch andere subjektiv nötigend; es lag also keine echte Nötigung, zwingende Anstiftung, Handlung ex ufficio, kein Befehl o.ä. vor. (Die Verantwortung liegt in solchen Fällen wieder bei dem, der die Situation strukturiert hat.)

U5.2. *Keine vorgelagerte Unterlassung:* Die unterlassene Handlung *a* erforderte nicht die Autorisierung, Anordnung o.ä. durch ein anderes Subjekt, die ebenfalls unterlassen worden ist. (In diesem Fall ist das andere Subjekt verantwortlich. Beispiel: Der Schulleiter unterlässt es, die Lehrer über das – ihm bekannte – Ministerialdekret zu unterrichten; das Schulamt hat das Dekret nicht weitergeleitet.)

U6. *Zumutbarkeit: A* (oder ein substantiell ähnliche Handlung A^*) zu tun war dem Subjekt *s* zumutbar.

Die meisten dieser Bedingungen sind aus der Literatur einigermaßen bekannt; ich habe sie hier nur systematisiert. Auch sind aus der Literatur in der Regel Begründungen für diese Bedingungen bekannt, die zum oben skizzierten attributionstheoretischen Ansatz passen. Generelle und aktuelle Zurechnungsfähigkeit (H4, U4) sind beispielsweise erforderlich, damit der rationale Einflussversuch der Gesellschaft über die Strafandrohungen und Belohnungsversprechen überhaupt funktionieren kann; wer nicht zurechnungsfähig ist, lässt sich auch durch solche Drohungen und Versprechen nicht beeindrucken und in die richtige Richtung steuern. Und ohne den subjektiven Tatanteil (H3, U3) sind das Verhalten

des Subjekts und seine Folgen wiederum vom Subjekt aus nicht gesteuert, also auch der angestrebten gesellschaftlichen Steuerung entzogen. Auch das abgeschwächte Prinzip der alternativen Möglichkeiten (H2) soll eine soziale Steuerung ermöglichen: Wenn das Subjekt nicht anders handeln kann, ist der Einfluss auf seine Entscheidungen kein effektiver Weg zur Steuerung der Welt. Die Überdeterminierungsklausel (in H2) ist eine Antwort auf Fälle à la Frankfurt (1969): Auch wenn Jones aufgrund von Blacks Maßnahmen nicht anders handeln konnte, ist eine soziale Steuerung von Jones' Handeln *nicht* ausgeschlossen; Jones' Handeln ist durch Blacks mögliches Eingreifen überdeterminiert; und die soziale Steuerung muss sowohl Jones als auch Black beeinflussen; beide sind also verantwortlich. Die Zumutbarkeitsbedingung (H6, U6) geht aus von der Annahme, dass es unmoralisch wäre, von den Subjekten den vollständigen Einsatz ihres Lebens zu fordern; Moral hat Grenzen, jenseits derer man nicht mehr moralisch verpflichtet und verantwortlich ist.

Es ist aber noch einiges zu der üblicherweise ziemlich vernachlässigten Abschirmungsbedingung (H5, U5) zu erläutern, weil sie für das Verständnis des Werts der hier konzipierten Form von Verantwortung sehr aufschlussreich ist.

Jede zurechnende Verantwortung muss festlegen, *wer* verantwortlich ist. Natürlich sollten nur diejenigen verantwortlich sein, die das fragliche Geschehen beeinflussen können, sonst kann man mit dem System der Verantwortung dieses Geschehen auch nicht steuern. Da meist aber sehr viele Menschen das Geschehen beeinflussen können, bleiben immer noch große Spielräume, wie die Verantwortung unter ihnen verteilt werden soll. Es gibt insbesondere kollektivistische Systeme der attributiven Verantwortung, bei denen die Verantwortung auf mehr oder weniger große Gruppen verteilt ist; und es gibt individualistische Systeme. Die soeben spezifizierte attributive Verantwortung ist – vermittelt über die genauen Regelungen zur Verantwortungsabschirmung – streng individualistisch. (Dies schließt nicht aus, dass auf der Ebene der Ereignisverantwortung auch kollektive Verantwortungen konzipiert werden können – aber mit genau geregelten individuellen Anteilen.) Im Normalfall ist nur der zuletzt Handelnde verantwortlich und niemand, der sein Handeln beeinflusst hat oder der sein Handeln hätte beeinflussen können. Eine wesentliche Begründung dafür ist – naheliegenderweise –, dass die soziale Steuerung des Geschehens viel präziser und effizienter sein kann, wenn sie den zuletzt Handelnden zu beeinflussen versucht. Andere Handelnde können den zuletzt Handelnden ja auch nicht völlig kontrollieren, sein Agieren nicht sicher voraussehen, ihn nicht 100%ig überwachen; und den anderen fehlt auch oft die Macht, ihn zu beeinflussen. Aus diesem Grunde verletzen kollektivistische Systeme der attributiven Verantwortung häufig auch die Minimalbedingung für attributive Verantwortung: Es werden Angehörige des Kollektivs mitverantwortlich gemacht und mitbestraft, die das Geschehen gar nicht hätten beeinflussen können.

Die individualistische Verantwortungsabschirmung hat aber noch ganz andere Vorteile, die im Vergleich zum kollektivistischen System deutlich werden. Was passierte in einem kollektivistischen System? Ohne die Abschirmung würden andere, die Sorgen hätten, für unsere Handlungen zur Rechenschaft gezogen zu werden, laufend versuchen, Einfluss auf unsere Entscheidungen zu nehmen. Wenn sie befürchten, dass wir sozial Unerwünschtes tun, ja schon bei Anzeichen dazu, etwa wenn wir nonkonformistische Ansichten und Einstellungen entwickelten, würden sie laufend in unser Entscheiden und Handeln eingreifen: Sie würden uns bedrängen, unsere Planungen preiszugeben, uns bei unseren Entscheidungen beschwören, mit Konsequenzen drohen, nach bestimmten Entscheidungen solche Konsequenzen wahrmachen, die von uns gewünschten Handlungsfolgen wieder beseitigen usw.; es wäre eine dauernde Einmischung. Das Ausmaß dieses Einflusses kann man sich auch an einer Situation verdeutlichen, in der die Abschirmung aufgehoben ist, nämlich bei Kindern, für deren Handlungen die Eltern verantwortlich sind. Individualistische Verantwortungsabschirmung gewährt uns demgegenüber garantierte *soziale Handlungsfreiheit*; sie schirmt uns auch vor der Einmischung und Kontrolle durch die anderen ab. Dies ist für die Individuen eine Befreiung und ein Fortschritt. Außerdem konstituiert die individualistische Verantwortungsabschirmung klare Zentren der Verantwortlichkeit. Das Individuum ist nicht nur Träger seiner Entscheidungen – dies ist ein natürliches Fakt, kein soziales Konstrukt –, sondern wird auch für die sozialen Institutionen und für andere Individuen zum vollen Ansprechpartner, Rechtssubjekt, mündig, jemand, an den sie sich wenden können, mit dem sie verbindliche Verabredungen treffen, an den sie sich *wieder* wenden können, ohne dass er dann auf andere verweist, weil er nämlich haftbar ist. Aus diffuser Verantwortung wird konzentrierte individuelle Verantwortung. Das Subjekt erhält soziale Autonomie. Der Preis dieser sozialen Autonomie und Mündigkeit ist aber gleichzeitig, dass wir voll verantwortlich sind für unser Handeln, für es geradestehen müssen. Mündigkeit geht einher mit Strafmündigkeit und Haftung. Die Folgen werden uns zugerechnet, und wir müssen die Konsequenzen für sie tragen. Zudem erzeugt die Abschirmung auch Vorteile für diejenigen, denen die Verantwortung abgenommen wird, also für alle nicht unmittelbaren Täter; sie werden in zweifacher Hinsicht entlastet: Es ist eine Arbeitserleichterung für sie; sie müssen nicht dauernd beobachten, im Vorfeld zu beeinflussen versuchen und eingreifen, schon wenn jemand nonkonformistische Ansichten hat. Die Straftatverhinderung wird arbeitsteilig und viel effizienter dem Polizeiapparat überlassen, der allerdings auch nur genau beschränkte Möglichkeiten zum Eingriff in Überlegungen, Entscheidungen und Handlungen von potentiellen Tätern hat. Allerdings ist auch der Polizeiapparat auf die Mitwirkung des Rests der Bevölkerung angewiesen; und dieser ist – zumindest in Deutschland – zu einer relativ genau definierten Mitwirkung verpflichtet. Die Entlastung der Nichttäter von der Straftatverhinderung bei anderen geht also nur bis zu einem gewissen Punkt. Ein

anderer Aspekt der Entlastung ist, dass man nicht mehr für das Versagen der anderen zur Rechenschaft gezogen wird, es ist also eine Exkulpierung.

Ein weiterer Vorteil der individualistisch konzipierten Verantwortung und der dadurch vorgenommenen Konstituierung des *Individuums* als soziales Subjekt ist die Rationalisierung des sozialen Lebens: Die kreativen und energetischen Ressourcen der Individuen werden voll freigesetzt. Die Individuen müssen bei solchen individuellen Einsichten, die weit über die Einsichten der Mehrheit der Kollektivs hinausgehen können, nicht immer den Gruppenkonsens einholen. Entscheidungen werden dadurch, bei intelligenten Individuen, intelligenter und flexibler. Durch die Individualisierung der Verantwortung entsteht auch erst eine echte soziale Arbeitsteilung, bei der jeder seinen eigenen individuellen Bereich optimieren kann. Auch für andere, die mit sozialen Subjekten kooperieren wollen, ist die Individualisierung ein Vorteil: Sie haben klare Partner und klare Verantwortlichkeiten; die Subjekte sind kalkulierbarer.

5. Anwendungen der Idee und der Bedingungen der attributiven Verantwortung – Konsequenzen für einige offene Fragen der aktuellen Verantwortungsdiskussion

Diese Grundideen und die Bedingungen der attributiven Verantwortung können angewendet werden zur Beantwortung einiger offener Fragen der Debatte um Verantwortung.

Zum Prinzip der alternativen Möglichkeiten und zum Problem kontrafaktischer Intervenierer liefert die Bedingung H2 einen Lösungsvorschlag, der oben schon erläutert wurde.

Das Verhältnis von Verantwortung und Nötigung wird in den einschränkenden Bedingungen der Verantwortungsabschirmung thematisiert (H5, U5): Die Entscheidungssituation des Genötigten ist für eine bestimmte Entscheidung ziemlich stabil vorstrukturiert; bei einer effektiven Nötigung ist der Nutzenunterschied zwischen den Alternativen aufgrund der (glaubwürdigen) Drohungen des Nötigers so groß, dass der Genötigte rationaliter nur die vom Nötiger gewünschte Handlung wählen kann. Aufgrund dieser Strukturierung der Entscheidungssituation durch den Nötiger setzt die soziale Steuerung des Geschehens also, anders als sonst, besser beim Nötiger an als beim Genötigten; der Nötiger muss also verantwortlich sein, der Genötigte hingegen nicht.

Hier ist nicht der Ort, eine systematische Antwort zum Thema 'Verantwortung trotz metaphysischer Alternativenlosigkeit und Determinismus' zu geben. Nur so viel: Selbstverständlich ist die vorgelegte Verantwortungskonzeption kompatibilistisch. Dem Kompatibilismus wird u.a. vorgehalten, er sei intuitiv unfair; Menschen würden zur Verantwortung gezogen, die metaphysisch nicht anders handeln konnten. Nach der hier vorgelegten Konzeption konnten sie aber

immerhin konditional anders handeln (H2, U2). Und die skizzierten Vorteile der individualistischen Verantwortung und Mündigkeit besagen, dass es für geistig halbwegs normale Erwachsene viel besser ist, dieses individualistische Paket aus sozialer Handlungsfreiheit, voller Geschäftsfähigkeit, Mündigkeit, aber auch Strafmündigkeit anzunehmen, als ein Paket aus permanenter Einmischung, Bevormundung und dann eventuell stark eingeschränkter Strafmündigkeit. Immerhin drängen sich Jugendliche normalerweise nach dem ersten Paket und bereuen, wenn sie mündig geworden sind, nicht, es erhalten zu haben – trotz seiner straf- und haftungsrechtlichen Brisanz. Wenn, wie ich annehme, die Jugendlichen diese Vor- und Nachteile einigermaßen realistisch sehen, dann kann dieses Paket – einschließlich seiner strafrechtlichen und Haftungskomponente – nicht so unfair sein.

Literaturverzeichnis

Burkhardt, Björn: „Verantwortung, rechtlich“. In: *Korff, Wilhelm/Beck, Ludwin Mikat, Paul (Hrsg.)*: Lexikon der Bioethik. Bd. 3. Gütersloher Verlagshaus, Gütersloh, 2000. S. 671-673

Davidson, Donald: „Handeln“ („Agency“. 1971.). In: *Davidson, Donald*: Handlung und Ereignis. Übers. v. Joachim Schulte. Suhrkamp, Frankfurt/M., 1985. S. 73-98

Duff, R[obin] A[ntony]: Intention, Agency and Criminal Liability. Philosophy of Action and the Criminal Law. Blackwell, Oxford/Cambridge (MA), xiv, 1990. 219 S.

Frankfurt, Harry G: “Alternate Possibilities and Moral Responsibility”. Journal of Philosophy, 66, 1969. S. 829-839

Hart, Herbert L[ionel] A[dolphus]: “Intention and Punishment”. In: *Hart, Herbert L[ionel] A[dolphus]*: Punishment and Responsibility. Essays in the Philosophy of Law. Clarendon, Oxford, 1968. S. 113-135

Kennett, Jeanette: Agency and Responsibility. A Common-sense Moral Psychology. Clarendon, Oxford, viii, 2001. 229 S.

Lumer, Christoph: „Ziele und Methoden der Philosophie“. In: Aufgaben der Philosophie heute. Arbeitstagung des Fachbereichs Kultur- und Geowissenschaften (Universität Osnabrück) in Verbindung mit dem Istituto di Filosofia (Università degli Studi di Urbino), 24. - 26. Oktober 1988. Osnabrücker Philosophische Schriften, Osnabrück, 1989. S. 108-132

Lumer, Christoph: Praktische Argumentationstheorie. Theoretische Grundlagen, praktische Begründung und Regeln wichtiger Argumentationsarten. Vieweg, Braunschweig, xi, 1990. 474 S.

Lumer, Christoph: „Abwegige Absichtsrealisierung und Handlungssteuerung. Eine intentional-kausalistische Erklärung". Internationale Zeitschrift für Philosophie, 1, 2008. S. 9-37

Moore, G[eorge] E[dward]: Grundprobleme der Ethik. (Ethics. 1912.) Vorwort v. Norbert Hoerster. Aus d. Englischen v. Annemarie Pieper. Beck, München, 1975. 155 S.

Roughley, Neil: "The Double Failure of 'Double Effect'". In: *Lumer, Christoph/Nannini, Sandro (Hrsg.)*: Intentionality, Deliberation and Autonomy. The Action Theoretic Foundation of Practical Philosophy. Ashgate, Aldershot, 2007. S. 91-116

Strawson, Peter F.: "Freedom and Resentment". Proceedings of the British Academy, 48, 1962. S. 1-25

Betrachtung zweier libertaristischer, aber nicht akteurskausalistischer Konzeptionen von Willensfreiheit und Verantwortung

Jacob Rosenthal
jacob.rosenthal@uni-bonn.de
Universität Bonn, Deutschland

Abstract/Zusammenfassung

According to libertarians, freedom and moral responsibility are incompatible with determinism and we have good reason to believe that under normal circumstances we are indeed free and responsible agents. Thus, they need to develop a conception of freedom that distinguishes free and responsible actions from random or otherwise unexplainable events. So-called theories of agent causation introduce heavy metaphysical machinery to provide such a conception, but even if one grants that, it remains doubtful whether they can achieve their goal. There are, however, also libertarians that claim to get by with thoroughly down-to-earth measures, namely Robert Kane and Geert Keil. Both stick to normal causal relations between events, and according to both an action is only free and responsible if it is selected from genuinely open alternatives in a way that markedly distinguishes it from a random event. Now, it is easy to imagine processes of consideration and decision in which indeterministic elements are involved. But the difficulty is to explain how these very elements can make the action free and responsible instead of just weakening the control of the agent over what she does or turning the action into something that is partially inexplicable. The accounts of Kane and Keil are introduced and discussed with regard to the latter problem. I argue that they do not achieve a satisfying solution to this problem of explanatory gaps. Given their assumptions, it is impossible to answer contrastive questions of the type "Why does a certain person in a certain situation act in a certain way instead of acting in another way that is also genuinely open to her in her situation?" This defect holds no matter whether the question is meant to ask for reasons or for causes of action.

Libertarier meinen, dass sich eine realistische indeterministische Konzeption der Willensfreiheit und Verantwortlichkeit ausarbeiten lässt, die freie Handlungen von unerklärlichen Ereignissen unterscheidet. Die meisten derartigen Konzeptionen führen eine besondere Form der Verursachung ein, die sog. Akteurskausalität, bei der Handelnde als Initiatoren jeweils neuer Kausalketten aufgefasst werden. Die metaphysischen Kosten und sonstigen Probleme solcher Ansätze sind beträchtlich; zwei Libertarier, die es anders machen, sind Robert Kane und Geert Keil. Beide beschränken sich auf die gewöhnliche Ereigniskausalität, und für beide kann von freiem Handeln nur die Rede sein, wenn objektive Alternativen vorhanden sind, von denen in nicht-zufälliger, aber indeterminierter Weise eine ergriffen wird. Nun sind selbstverständlich Überlegungs- und Entscheidungsprozesse denkbar, in die genuin indeterministische Elemente einfließen, die Schwierigkeit besteht aber darin zu erklären, wie gerade diese den Akteur frei und verantwortlich machen können, anstatt einfach seine Kontrolle über die Handlung zu vermindern oder diese als etwas Zufälliges oder in anderer Weise Unerklärliches erscheinen

zu lassen. Die Ansätze von Kane und Keil werden jeweils vorgestellt und im Hinblick auf das letztgenannte Problem diskutiert. Ich komme zu dem Schluss, dass beide keine befriedigende Lösung dieses sog. Erklärungslückenproblems erreichen. In ihrem Rahmen kann keine Antwort auf die kontrastive Frage gegeben werden, warum eine bestimmte Person in einer bestimmten Situation etwas Bestimmtes tut und nicht etwas anderes, das sie in der nämlichen Situation auch tun könnte. Dies gilt unabhängig davon, ob die Warum-Frage als eine nach Ursachen oder als eine nach Gründen des Handelns verstanden wird.

Wer erstens die Vereinbarkeit von freiem Handeln und Determinismus bestreitet, zweitens meint, dass sich eine positive Freiheitskonzeption ausarbeiten lässt, die freie Handlungen von zufälligen Ereignissen befriedigend unterscheidet, und drittens der Auffassung ist, dass die Voraussetzungen dieser Konzeption in unserem Fall entweder erfüllt sind oder wir zumindest von ihrem Erfülltsein ausgehen dürfen, wird als Libertarist oder Libertarier bezeichnet. Die meisten Libertarier wollen ihr Ziel dadurch erreichen, dass sie eine besondere Form der Verursachung annehmen, die sog. Akteurskausalität. Dieser zufolge verursacht das „Selbst“ einer Person, ihr substantieller Kern, oder auch die Person als Ganze, die Handlungen, ohne hierin durch etwas anderes determiniert zu sein. Die Akteurskausalität unterscheidet sich dadurch von der gewöhnlichen Spielart, dass ihre Relata nicht Ereignisse sind, sondern es sich bei der Ursache um einen Gegenstand handelt, eben um ein „Selbst“. Dieses erscheint als ein unbewegter Beweger – allerdings nicht im aristotelischen Sinne als Finalursache, sondern als Wirkursache – der aus dem Nichts heraus oder von selbst neue Kausalketten in Gang setzt. Die metaphysischen Kosten akteurskausalistischer Ansätze sind grundsätzlich hoch, und auch wenn man sie zu tragen bereit ist, bleibt fraglich, ob die Idee überzeugend durchführbar ist. Immerhin sind „Selbste“ persistierende Entitäten, und wenn man sie selber, und nicht etwa Veränderungen an ihnen oder Geschehnisse, in die sie involviert sind, als Ursachen anführt, ist nicht zu sehen, warum diese Ursache gerade zu dem jeweiligen Zeitpunkt und nicht früher oder später auf die jeweilige Weise wirkt (Datiertheitseinwand).

Zwei Libertarier, die es anders machen und ohne solche „metaphysischen Zumutungen“ auskommen möchten, sind Robert Kane und Geert Keil. Beide beschränken sich auf die gewöhnliche Ereigniskausalität und verwerfen die Idee der Ersturheberschaft, die den Akteur als den Ausgangspunkt einer jeweils neuen Ursachenkette auffasst. Für beide kann von freiem Handeln aber auch nur die Rede sein, wenn für den Akteur genuine Alternativen objektiv offen stehen. Selbstverständlich sind Überlegungs- und Entscheidungsprozesse denkbar, in die indeterministische Elemente involviert sind, die Schwierigkeit besteht aber darin zu erklären, wie gerade diese den Akteur frei und verantwortlich machen können, anstatt einfach seine Kontrolle über die Handlung zu vermindern, indem sie sie teilweise zu einer Angelegenheit des Zufalls machen, oder zumindest zu etwas Unerklärlichem und auch für die handelnde Person selbst Rätselhaftem.

Dieses sog. Zufallsproblem oder vorsichtiger: Problem der Erklärungslücke ist sicherlich eines der Haupthindernisse für die Akzeptanz libertarischer Freiheitskonzeptionen. Im Folgenden möchte ich Robert Kanes und Geert Keils Reaktionen auf dieses Problem untersuchen.

Zunächst eine Vorbemerkung. Die Erklärungslücke im Fall des Indeterminismus besteht primär bei *kontrastiven* Erklärungen. Wir können annehmen, dass, wenn eine Person eine bestimmte Handlungsoption ergreift, es im Normalfall auch *Gründe* gibt, die aus der Perspektive der Person für diese Option sprechen. Und wir dürfen, wenn wir einen indeterministischen Verursachungsbegriff zulassen, auch annehmen, dass es irgendwelche *Ursachen* für das Tun der Person, oder vorsichtiger: für die damit einher gehenden physischen Ereignisse gibt. Diese Ursachen machen die Handlung bei einer indeterministischen Kausalitätsauffassung ja nicht notwendig. Wir können also für das, was eine Person tut, auch im Indeterminismus sowohl Erklärungen durch Ursachen wie auch Erklärungen durch Gründe finden, allerdings keine kontrastiven Erklärungen. Wenn die Person nämlich *stattdessen* etwas anderes täte, so würde es auch dafür sowohl Ursachen als auch Gründe geben. Warum tut eine Person in einer bestimmten Situation aus bestimmten Gründen *H anstatt* aus anderen Gründen *H** zu tun? Es ist diese Warum-Frage, die weder in einer kausalen noch einer rationalen Lesart beantwortbar zu sein scheint. Insofern und nur insofern ist der Ausgang des Entscheidungsprozesses rätselhaft, und nur insofern der Frage „Warum diese Handlung?“ ein kontrastiver Sinn beigelegt wird, macht sie einem Libertarier Schwierigkeiten. Insbesondere scheint die handelnde Person selber nichts Informatives darüber sagen zu können, warum sie im Falle eines Indeterminismus und genuinen Anders-Handeln-Könnens in der konkreten Situation unter dem Eindruck der Gründe für *H H* tut *anstatt* unter dem Eindruck der Gründe für *H* H** zu tun oder noch weiter zu überlegen mit offenem Ausgang. Sie tut eben das eine und nicht das andere; mehr scheint an dieser Stelle nicht zu sagen zu sein. (Der Einfachheit halber tue ich so, als ginge es bei Entscheidungen immer um zwei Handlungsoptionen.)

Die Ansätze von Kane und Keil sollen nun jeweils kurz vorgestellt und problematisiert werden. Ich beschränke mich jeweils auf das für unsere Zwecke Wesentliche. Robert Kane hat seine Position in dem Buch *The Significance of Free Will* (Oxford University Press 1998) und zahlreichen Aufsätzen entfaltet. Ihm zufolge darf keine relevante Warum-Frage in Bezug auf eine Handlung unbeantwortet bleiben, wenn ihre Zurechenbarkeit gewährleistet sein soll. Es muss *erklärt* werden können, warum der Akteur in einer bestimmten konkreten Situation diese und nicht jene Handlungsoption ergriff, *obwohl* er in der nämlichen Situation auch anders hätte handeln können. Kanes Lösung sieht so aus, dass die Frage, warum der Akteur den Willensbildungsprozess in dieser und nicht in je-

ner Weise, mit dieser und nicht mit jener Entscheidung beendete, dadurch zu beantworten ist, dass die entsprechenden Gründe dem Akteur als die gewichtigeren erschienen. Wenn man nun fragt, warum *dies* so war, warum also der Akteur zu der Auffassung kam, die Gründe für *H* seien im Vergleich zu denen für *H** die gewichtigeren, dann soll die Erklärung dafür nun gerade umgekehrt darin bestehen, dass der Akteur den Willensbildungsprozess eben mit *H* und nicht mit *H** beendete. Damit soll keine relevante Warum-Frage offen bleiben. Die Zirkularität dieser Erklärungsform wird von Kane explizit vermerkt, aber nicht als entscheidendes Problem angesehen.

Zur Kritik. Zunächst ist zwar das so oder so Terminieren der Willensbildung gut erklärbar durch Berufung auf das relative Gewicht der Gründe, aber die umgekehrte Erklärung wirkt weit weniger plausibel. Es klingt im Allgemeinen merkwürdig zu sagen, es seien einem die Gründe für *H* gewichtiger erschienen als die für *H**, eben weil man sich für *H* und gegen *H** entschieden habe. Für bestimmte Situationen freilich ist das, was Kane sagt, phänomenologisch nicht unplausibel. Wenn sich auch bei gründlichem Nachdenken nicht sagen lässt, ob die Gründe für *H* oder die Gründe für *H** die stärkeren sind, aber eine Entscheidung getroffen werden muss, könnte man die Idee haben, dass das Subjekt die Pattsituation dadurch auflöst, dass es sich zu einer der Optionen durchringt und eben dadurch und im selben Atemzug die entsprechenden Gründe für sich zu den stärkeren macht. Es ist allerdings fraglich, ob hier nicht eine Selbsttäuschung des Subjekts im Spiel wäre. Ohne eine solche Selbsttäuschung, so könnte man vermuten, müsste die zutreffende Beschreibung lauten, dass sich das Subjekt in einer unklaren Situation oder angesichts gleich starker Gründe willkürlich für eine Option entscheidet, ohne dass sich dadurch an seiner Einschätzung des relativen Gewichtes der Gründe etwas verändert. Aber auch wenn wir die Kanesche Beschreibung für Fälle gleich gewichtiger oder, was wahrscheinlich der wichtigere Fall ist, inkommensurabler Gründe akzeptieren, ist doch, so denke ich, klar, dass sie auf alle anderen Fälle jedenfalls nicht passt. Es kommt auch häufig vor, dass sich beim Nachdenken herausstellt, dass die einen Gründe ganz klar die gewichtigeren sind und es daher irrational, im Extremfall geradezu verrückt wäre, die entsprechende Option nicht zu ergreifen. In solchen Fällen kann man gewiss nicht sagen, die jeweiligen Gründe würden durch die Entscheidung zu den gewichtigeren gemacht. Angesichts dessen beißt Kane in den sauren Apfel und konzediert, dass dies eben keine Fälle freien Entscheidens seien. Solche sind für ihn selten und ereignen sich nur bei bestimmten grundlegenden Weichenstellungen im Leben eines Menschen. Dies ist eine für Libertarier gewiss unerwünschte Wendung, wenn auch verständlich ist, wie Kane dazu kommt.

Lassen wir diese erste Schwierigkeit nun beiseite und konzentrieren uns auf Situationen, in welchen jede der beiden Handlungsoptionen vom Subjekt in gut nachvollziehbarer Weise ergriffen werden kann. Damit sowohl die Entscheidung

durch die relative Gewichtung der Gründe als auch umgekehrt diese Gewichtung durch die Entscheidung erklärt werden kann, darf keiner dieser Vorgänge zeitlich vor dem anderen liegen. Eine zeitliche Asymmetrie wäre nämlich ein guter Grund, auch eine entsprechende Erklärungs-Asymmetrie anzunehmen. Und Kane sagt auch ganz konsequent, dass es sich nur um zwei verschiedene Aspekte oder Beschreibungsweisen desselben Vorgangs handelt: eine volitive und eine kognitive. Damit geraten aber beide Erklärungsansprüche unter erheblichen Druck, denn was ist das für eine Art von Erklärung, bei der sich zwei Aspekte derselben Sache wechselseitig erklären, oder bei der ein Phänomen unter einer Beschreibung (einer kognitiven) sich selbst unter einer anderen Beschreibung (einer volitiven) erklärt, und umgekehrt? Wenn man hier überhaupt von Erklärung sprechen möchte, ist es jedenfalls kein gewöhnlicher Typ, insbesondere keine Kausalerklärung. Dieser Befund bedeutet nichts Gutes für Kanes Anspruch, es besser als die Anhänger einer Akteurskausalität zu machen, die eine besondere Art der Verursachung postulieren müssen, während er selber mit der normalen Ereigniskausalität auskommt. Denn Kane muss nun zwar keinen besonderen Verursachungstypus, aber doch eine besondere Erklärungsweise in Anspruch nehmen.

Schließlich ist, auch wenn man diese Art wechselseitiger Erklärung zweier Aspekte des Willensbildungsprozesses akzeptiert, doch klar, dass dieser Prozess, der uns hier mit einem volitiven und einem kognitiven Aspekt entgegen tritt, nun *als Ganzes* gerade keine Erklärung besitzt. Man muss ja nur fragen, wie es kommt, dass der Willensbildungsprozess so und nicht anders abgeschlossen wurde, und, damit einher gehend, dem Akteur diese und nicht jene Gründe als die gewichtigeren erschienen, um zu sehen, dass nun *diese* Frage, die sich auf das Gesamtphänomen mit allen seinen Aspekten bezieht, keine Antwort mehr findet, die Handlung also in dem besagten kontrastiven Sinn rätselhaft und unerklärlich bleibt.

Geert Keil bestimmt in seinem Buch *Willensfreiheit* (De Gruyter 2007) Freiheit als die Fähigkeit zur überlegten hindernisüberwindenden Willensbildung. Der Naturverlauf ist indeterministisch. Die Naturgesetze, aufgefasst als Dispositionen oder essentielle Eigenschaften der natürlichen Dinge, legen nicht fest, was geschieht, sondern schränken nur den Möglichkeitsspielraum ein. In der Regel bleiben zahlreiche Möglichkeiten objektiv offen. So besitzt eine Person im Normalfall die Fähigkeit zu überlegen, was sie tun soll, und ebenso die Fähigkeit, eine Überlegung an diesem oder jenem Punkt abzubrechen und dann im Lichte derjenigen Gründe, die ihr zu dem jeweiligen Zeitpunkt als die besten erscheinen, zu handeln. Auch wenn eine Person faktisch nicht überlegt, so hätte sie es im Normalfall doch tun können, und zwar in genau der nämlichen Situation und in einem unbedingten Sinne, und ist deshalb verantwortlich auch für

Versäumnisse und Unterlassungen, bei denen sie gar nicht überlegt und gar keine bewusste Entscheidung trifft. Und wenn sie überlegt und dann aus bestimmten Gründen etwas Bestimmtes tut, so hätte sie doch in genau der gegebenen Situation auch weiter überlegen oder die Überlegung vorher abbrechen und dann ggf. aus anderen Gründen etwas anderes tun können. All dies steht im Normalfall bei ihr, sie hat normalerweise die entsprechenden Fähigkeiten in einem unbedingten Sinne.

Zur Kritik. Zunächst habe ich den Eindruck, dass Keil mit Situationen extrem unbalancierter Gründe dieselben Schwierigkeiten hat wie Kane. Inwiefern hat der Akteur die Fähigkeit des Anders-Handelns, wenn offensichtlich ist, dass für die eine Option deutlich stärkere Gründe sprechen? Das Ergreifen der anderen Option wäre dann eigentlich nur durch einen Lapsus erklärbar, der ein Zeichen von Unfreiheit wäre, etwas, das einem zustößt, aber nicht zurechenbar ist: Rätselhafterweise handelt der Akteur in einer für ihn wichtigen Angelegenheit völlig ohne Überlegung, oder: rätselhafterweise erkennt er beim Überlegen die stärkeren Gründe nicht, obwohl sie offen zu Tage liegen, oder: merkwürdigerweise ergreift er im vollen Bewusstsein der Asymmetrie die klarerweise weniger vernünftige Option. Generell hat die libertarische Intuition Schwierigkeiten, sobald leicht einsehbare und relativ gesehen starke oder gar zwingende Gründe für eine Handlungsweise vorliegen. Das behauptete „anders können“ wird dann zweifelhaft: „Ich müsste schier verrückt sein, um in dieser Situation *H** und nicht *H* zu tun, aber ich bin nicht verrückt“, so könnte das Subjekt sagen. Dieser Situationstyp ist keineswegs einer, in welchem gar nichts zu überlegen und zu entscheiden ist. Vielmehr kann man sich vorstellen, dass das Subjekt mit zwei Handlungsoptionen konfrontiert ist, und sich erst nach kurzem Überlegen erweist, dass die Gründe auf der einen Seite wesentlich stärker sind. Man könnte *H** dann auch als eine prima-facie-Option ansprechen, die schnell ausgeschieden werden kann.

Konzentrieren wir uns nun, wie bei Kane, auf Situationen, in welchen die Idee der „zwei-Wege-Rationalität“ plausibel ist, in welchen also jede der beiden Optionen in gut nachvollziehbarer Weise ergriffen werden kann. Zentral ist bei Keil der Begriff der Fähigkeit. Was unterscheidet Fähigkeiten oder, Kantisch gesprochen, Vermögen von bloßen Möglichkeiten, wie sie auch bei Zufällen bestehen? Die Ausübung einer Fähigkeit ist etwas *Aktives*: „Eine Fähigkeit, die sich in bestimmten Bedingungen gleichsam automatisch aktualisiert, wäre von einer passiven Disposition nicht zu unterscheiden. […] Menschen, die eine Fähigkeit ausüben, müssen […] stets etwas hinzutun, damit das Fragliche geschieht.“ (Keil 2009, Absatz 29) Was müssen sie denn hinzutun? Nun, sie müssen unter den gegebenen Umständen die Fähigkeit ausüben *wollen*. Keil sagt das möglicherweise deshalb nicht besonders deutlich, weil man dann sofort auf einen Regress stößt. Die Ausübung einer Fähigkeit erscheint selbst wieder als eine Handlung, zu der sich das Subjekt entscheiden muss. Sie verweist wiederum auf

ein Wollen: Was eine aktive Fähigkeit von einer passiven Disposition unterscheidet, ist, dass sich unter den einschlägigen Umständen eine Disposition von selbst aktualisiert, ggf. mit einer bestimmten Wahrscheinlichkeit, die Fähigkeit hingegen nur dann ausgeübt wird, *wenn die Person das will*.

Aus diesem Grunde kann die für Keil zentrale Rede von „Fähigkeiten" für die Erläuterung der praktischen Freiheit nichts leisten. Dasselbe gilt für die Kantische Rede vom „Vermögen, den eigenen Willen zu bestimmen". Diese Willensbestimmung, sofern sie als die Ausübung eines Vermögens oder einer Fähigkeit dargestellt wird, erscheint selber als eine (innere, mentale) Handlung, der ein weiteres Wollen entspricht. Sieht man davon ab, dann bleibt von der Rede von Fähigkeiten und Vermögen doch nichts weiter übrig als die objektive *Möglichkeit* des Anders-Überlegens und Anders-Wollens in derselben Situation, wie sie auch bei bloßer Zufälligkeit gegeben wäre. Keil zeigt also in Wahrheit kein die beiden Situationstypen – Zufälligkeit versus Freiheit bei der Willensbildung – unterscheidendes Merkmal auf. Da die Einführung von Fähigkeiten oder Vermögen wiederum auf einen Entscheidungs- und Handlungskontext verweist, kann sie nicht benutzt werden, um zu erläutern, was „Freiheit" in solchen Kontexten bedeutet, und wie sie sich von Zufälligkeit unterscheidet. Auch Keils Konzeption ist daher nicht geeignet, den Zufallsverdacht gegen libertarische Freiheitsauffassungen zu zerstreuen.

Die Unerklärlichkeit kontrastiver Fakten der genannten Art, sowohl im kausalen wie im rationalen Sinne, impliziert nicht von selbst, dass es sich hierbei tatsächlich um Zufälle handelt oder die Person keine Kontrolle über das Geschehen hat. Der Schluss vom Fehlen des (rationalen oder kausalen) Determinismus auf den Zufall ist nicht ohne weiteres gültig. Es könnte eine dritte Möglichkeit geben, um deren positive Ausgestaltung sich Kane und Keil allerdings meines Erachtens vergeblich bemühen. Diese dritte Möglichkeit scheint ein bloßes Postulat zu bleiben. Es läuft darauf hinaus, dass es in einem Entscheidungsprozess eben bestimmte nicht weiter zu erhellende Elemente der reinen Willkür gibt, die dem Subjekt zurechenbar sind. Ich bin mir letztlich nicht sicher, ob derartige Stipulationen legitim sind oder nicht. Nehmen wir zum Vergleich den Begriff des Naturgesetzes, der als einen Aspekt die natürliche Notwendigkeit enthält. Was lässt sich denn über notwendige Verknüpfungen in der Natur Erhellendes sagen? Wie sieht die Verbindung aus, vermöge derer ein Ereignis ein anderes herbeizwingt oder auch objektiv so-und-so wahrscheinlich macht? Wir können dazu nichts weiter sagen, aber daraus folgt nicht, dass es solche Verknüpfungen nicht geben könnte. Dass das Unbehagen in diesem Fall wesentlich geringer ist als bei unserem Thema, liegt meines Erachtens daran, dass man es mit den Fundamenten der Natur und nicht mit so hochstufigen Phänomenen wie menschlichem Handeln, Zurechenbarkeit und moralischer Verantwortung zu tun hat. Bei solchen scheint

mir der Rückzug auf die bloße Auskunft „Es ist weder Zufall noch Determinismus, sondern ein Drittes, das den Akteur verantwortlich macht“ nicht statthaft zu sein, solange diese dritte Möglichkeit nicht positiv ausgestaltet werden kann.

Eine Erklärungslücke besteht bei libertarischen Freiheitskonzeptionen immer, aber dies wäre kein Einwand, wenn es gelänge, Lücken dieses Typs von solchen zu unterscheiden, die bei rein zufälligen Vorgängen bestehen. Und zwar nicht bloß verbal zu unterscheiden, indem man sie sprachlich anders etikettiert, sondern *so* zu unterscheiden, dass man einsehen kann, wie Freiheit und Verantwortlichkeit der handelnden Person in dieser Lücke zustande kommen. Hier wäre eine alternative Struktur sichtbar zu machen. Es ist nicht befriedigend, einen relevanten Unterschied einfach zu postulieren – unter Hinweis auf die verwendeten Begriffe – und den Opponenten die Beweislast zuzuschieben. Diese sollten nicht behaupten, dass a priori klar ist, dass es außer Determinismus und Zufall nichts Drittes geben kann, sondern sich lediglich die Freiheit nehmen, diese dritte, sehr anspruchsvolle Möglichkeit wenigstens ansatzweise *sehen* und nicht bloß glauben zu wollen, bis dahin aber von ihrer Existenz als Möglichkeit, erst recht von ihrer Realität in unserer Welt, nicht überzeugt zu sein.

Literaturverzeichnis

Kane, Robert: The Significance of Free Will. Oxford UP, New York, 1998
Keil, Geert: Willensfreiheit. De Gruyter, Berlin, 2007
Keil, Geert: „Wir können auch anders. Skizze einer libertarischen Konzeption der Willensfreiheit“. In: Erwägen-Wissen-Ethik, 20, Heft 1, 2009

Moral Principles despite Particularism

Julius Schönherr
Juliusschoenherr@gmx.de
Humboldt-Universität Berlin

Abstract/Zusammenfassung

There has long been much effort to give an informative and counterfactually robust account of deontological moral principles. In this paper I propound the metaethical view that there are two irreducible, relatively independent levels of moral discourse: the realm of *categorical and principled value application* and that of *value justification*. This distinction justifies the assumption that moral principles do exist, even though, ultimately, they might not be counterfactually robust or justifiable.

My strategy is to start from a weak conception of moral principles – the Rossian account of prima facie values – and then work my way down to even weaker conceptions. I argue alongside Jonathan Dancy that prima-facie values fail to do the job, since they are, as Dancy points out, susceptible to counter examples. I, then, turn my attention to what Dancy calls "default reasons", "a consideration which is reason-giving unless something prevents it from being so." (Dancy 2004, p. 112) This type of reason, however, is likewise insufficient. In the course of my argument I present a new and even weaker picture of the form of deontological principles that does away with the idea that valid moral principles have to be counterfactually robust in a strong sense. I consider a pragmatic account for these principles and argue that a single justification for this categorical value application on the pragmatic level can, ideally, be asked for *in any one case*, but *not in all cases*. This move confirms that these principles are governed by moral reasons and that they are therefore not mere contingent social norms.

Es hat viele Versuche gegeben, moralische Prinzipien so zu formulieren, dass sie einerseits informativ, und andererseits kontrafaktisch Robust, d.h. gegenbeispiel-resistent sind. In diesem Paper verteidige ich die metaethische These, dass es zwei irreduzible und relativ unabhängige Ebenen gibt, um moralische Phänomene zu verstehen. Dies ist zum einen die Ebene der *kategorischen Anwendung*, zum anderen die der *partikularistischen* Rechtfertigung moralischer Prinzipien. So lässt sich die Annahme untermauern, dass moralische Prinzipien zwar existieren, obwohl sie, letztendlich, nicht zu rechtfertigen sind.

Ich beginne mit einer schwachen Konzeption moralischer Prinzipien, der Ross'schen Konzeption der prima facie Pflichten, und wende mich in der Folge noch schwächeren Positionen zu. Mit Jonathan Dancy vertrete ich die These, dass prima-facie Pflichten fehlgehen, da sie, wie Dancy zeigt, ebenfalls anfällig für Gegenbeispiele sind. Sodann richte ich mein Augenmerk auf sogenannte "default reasons", "a consideration which is reason-giving unless something prevents it from being so" (Dancy 2004, p. 112). Leider ist auch diese abgeschwächte Form der moralischen Prinzipien unzureichend. Im Gange meiner Argumentation lenke ich die Aufmerksamkeit auf ein noch schwächeres Bild moralischer Prinzipien, das Abstand von der Idee nimmt, moralische Prinzipien müssten in einem starken Sinne kontrafaktisch robust sein. Ich schlage ein pragmatisches Bild moralischer Prinzipien vor und argumetniere, dass eine singuläre Rechtfertigung dieser *in jedem einzelnen*, aber *nicht in allen Fällen*

verlangt werden kann. Diese punktuelle Verknüpfung pragmatischer Regeln mit moralischen Gründen garantiert, dass diese Prinzipien als *Moral*prinzipien und nicht als kontingente, gesellschaftliche Normen angesehen werden können.

Ross' prima facie duties.

The doctrine of prima facie values as stated by David Ross in *The Right and the Good* can be interpreted as follows:

PFD There is a set of prima facie reason-types[1] each of which functions as an incontrovertible moral reason to do an action X if properly applied within a certain situation Y.

On this account, the very fact that an act fulfills a promise is always a reason to perform that act. I consider this interpretation of prima facie duties to be the somewhat uncontroversial core idea of the doctrine. It stays neutral with respect to the questions of how these reasons relate to what one actually ought to do, as opposed to what there is reason to do, how one can know about these principles, and how these principles intertwine. All of which are aspects that Ross addresses but I am not. All my description captures is the generality of these moral reasons, the fact that each prima facie reason retains its normative force in each situation it appears.

I deem Ross' theory to be a particularly weak conception of normative principles, since it (i) does not provide for any *one* principle from which the rest should, in some sense, follow and (ii) does not evaluate any particular action-type but rather makes claims about the normative function of reasons that have "the tendency" (Ross 1930, p. 28) to make actions right.

Jonathan Dancy has raised a number of objections against this line of thought throughout his career. I will discuss only one of his central objections, which simply asserts that prima facie duties are susceptible to counterexamples. His objection to the value of beneficence (say) is that "there are many acts that would benefit others that I have no particular duty or reason to do, and in many such cases the act would not even be for the better. If someone does not deserve a benefit, giving her that benefit is not something one has a prima facie duty to do." (Dancy 2004, p. 120) Furthermore, in the case of a promise given with a depraved content, having given that promise is not always a reason for keeping it. There are other well known counterexamples such as: Suppose a person who wishes to murder another asks you to tell him where the other is hiding so that only a lie on your part will save the person in hiding. Here, there is no moral

1 According to Ross values of fidelity, reparation, gratitude, justice, beneficence, self-improvement, non-maleficience. (See Ross 1930, p. 21)

reason to tell the truth at all. There is no reason to feel any moral regret (Shope 1965, p. 280).

Default Reasons

As an alternative to Ross' prima facie reasons, Dancy admits of what he calls "default reasons" which are, at first blush, "pro tanto reasons unless undermined" (Cullity 2002, p. 188). He also suggests that this minimal account of moral principles is too weak to pose any threat to moral particularism the doctrine that "moral thought and judgment does not depend on the provision of a suitable supply of moral principles" (Dancy 2004, p. 73).

The term "pro tanto reason" is just another expression for the rather unfriendly formulation "prima facie reason". At the cost of some generality, the "unless undermined" amendment is designed to cope with the few counterexamples that pro tanto reasons cannot account for. The default reason account is weaker than the Rossian version because any pro tanto reason is a default reason – one that is never undermined – but not vice versa. As such, the default reason account would seem promising. However, my claim for this section is that there is no non-trivial version of default reasons that Dancy could build into the framework of moral particularism. Unfortunately his take on this issue is scarce. In showing that default reasons can't threaten moral particularism, he is more occupied showing what they cannot do than what they can do. In what follows, I will, one by one, consider and criticize the relevant interpretations and practical implications concerning default reasons homed in on by Dancy.

Statistical generalizations

SG Default moral principles are statistical generalizations. They are about the proportional applicability of a moral property X to a certain class of non-moral cases Y. This class contains actual rather than possible cases.

A statistical generalization about the moral wrongness of lying might say that the property of being wrongful is applicable to, say, 80% of all actual lies. I think it is uncontroversial that the Rossian reason-types are perfect candidates for such statistical generalizations. Nobody would deny that an action's being a lie would, most of the time, give reason not to do it. This interpretation applies to actions as well as reasons. Since statistical inquiries are based on a set of actual rather than possible cases they are largely contingent. They fit the default reason pattern for two reasons. (i) They are generalizations of some kind and (ii), most importantly, they are defeasible. However, they fail for the following reason.

Moral statistical generalizations are not explanatory (Dancy 2004, p. 113). They merely correlate a moral value and a class of non-moral cases. The fact that the percentage of car accidents with red cars is higher than with brown ones could hardly be explained with respect to the colour of the car but much rather with respect to the behaviour of the drivers. The fact that most lies are morally wrong does not yet imply that this is *because* they are lies. A certain action is not right *because* it is of a certain non-moral type that usually has the property of being right; rather it is right because moral reasons endow it with this property. The explanatory primacy, therefore, is with the moral reasons and not with the statistical generalizations.

Moral relief

MR Default moral reasons are necessary because "rationality needs something rule-like to work from" (Dancy 2004, p. 185). Morality demands too much from moral agents and without rules moral judgement is lost in the entanglement of moral reasons.

One explanation of why rationality could make use of such principles would be that the contributory reasons in each morally relevant situation are too complex and too numerous to grasp. Principles could, therefore, provide some relief. This type of argument is stock-in-trade for act-consequentialists. They typically claim that rightness is determined by the actual consequences of individual acts. Nevertheless, for the less than ideally rational agents that we are, these consequences are unforeseeable. Therefore, we have to fall back on moral principles. Rules, according to this view, are of heuristic value to rationality. This argument might have some intuitive appeal when talking about consequentialism; however, I don't think that such an argument lies within the scope of Dancy's particularist programme. For one thing, in discussing Hookers consequentialism, Dancy himself rejects the idea of principles serving as an aid to rationality. "We can perfectly well rely on people by and large to do what is right in the circumstances. We don't need principles to tell them what to do, or determine what is right [...]." (Dancy 2004, p. 133) For another, and more importantly, the counterexamples to pro tanto reasons that I propounded were perfectly well discernible cases. No intricate philosophical reflection was necessary to make us see that there is no particular reason to keep a promise if its content is somewhat depraved. And no hard thinking was needed to realize that we have reason to lie when lying would save someone's life. Quite normal sensitivity to the shape of specific situations is enough to determine the moral rightness or wrongness of a particular act. I conclude that there is no need for the particularist to employ moral principles to give rationality "something rule-like to work from".

Reason giving unless undermined

RU A default moral principle always has the same right-making force, other things being equal. "That contribution can be reversed or annulled by untoward circumstances." (Dancy 2004, p. 103)

The most obvious sense in which this interpretation is true is, again, the statistical sense. Without considering any circumstances we can be sure, on statistical grounds, that the presence of a certain non-moral property will, *most likely*, make a certain moral contribution. For all we know, lying here would, *most likely*, not be the thing to do. But, as I have shown above, more is needed. The stronger claim would be that, before any circumstances are considered, lying (say) is, in fact, always a reason against doing it.

It is a platitude that moral properties supervene on descriptive properties (Smith 1994, p. 23), paradigmatically on actions in particular circumstances (Dancy 1993, p. 78). These actions, and along with them their moral properties, *necessarily*, not merely typically, occur in circumstances. Under these considerations, it seems misguided to think of certain moral action types as generally exhibiting the same moral value before any specific circumstances are considered.

The particularist, I conclude, cannot make good sense of default reasons. Statistical generalizations are no explanatory principles, no principles are needed to provide for relief from the overly demanding moral reasons, and the "unless undermined"-version proves uninformative because moral reasons never occur independently of circumstances. I will now put forward my pragmatic version of default reasons to show how sense can be made of this concept.

Two levels of moral discourse

I suggest that there are two irreducible, relatively independent, levels of moral discourse: The social realm of *principled moral reason application* and that of *moral*, possibly particularist, *explanation*. The distinction between those levels generates, I claim, an informative and normatively strong account of default reasons. It solves the problem of statistical generalizations and reformulates the pragmatics of moral relief for moral assessors. These two moral realms correspond to the two following metaethical principles, the rule of social norms (SN) and a pragmatic constraint (PC).

SN Normally, if no justifying reasons are *actually given*, assessors are morally justified in judging moral agents on grounds of principled moral reasons such as the Ross-type prima facie values.

PC Explanation for the principled moral judgements presented in SN can be asked for in *any one case* but *not in all cases*. These explanations, however, might work in a particularist fashion.

Firstly, I will explain SN and make clear why moral assessors are justified in judging moral agents on behalf of those principles. Secondly, I go on to show how these principles are *regulated* by moral reasons even though these reasons might not fit into the principled dress. I conclude with some remarks on why that account is a version of the default reason theory.

The basic idea is that in the course of life, we make many moral commitments and perform many morally relevant actions, often more than one at a time. We simply have no time to discuss and justify them all. Therefore, we judge people using moral principles and hold them responsible on equal grounds. Some very weak kind of normativity in these principles is guaranteed by various forms of sanctions that give agents at least prudential reason to act according to those principles. Of course, merely sanction-based principles will lead to nothing but a system of non-moral coercion. Other people force agents, by dint of sanctioning, into a certain, not necessarily moral, behaviour. Whether the sanctioned action is also morally wrong would, then, be a matter of contingency. Therefore the pragmatic constraint is needed to morally *regulate* those principles. Moral rightness, on this account, is not reducible to what is approved of and sanctioned by society.

By stipulation, I take the word "normally" to say something about the relative *extensional* frequency of an event. This does not reflect its full meaning. As Pekka Väyrynen points out there are things which are not extensionally generalizable but which we still consider being normally true (Väyrynen 2003, p. 53). "Turtles normally live to a grand old age" is true, even though, in fact, turtles usually die young due to falling prey to predators. That fact, however, does not infringe the truth of the initial proposition about the normal longevity of turtles.

Moral assessors are justified in assessing others on grounds of strict moral principles because they, quite generally, have no access to the relevant contributory reasons generating moral rightness and wrongness. Earlier I showed that the relevant counterexamples to Ross' pro tanto reasons were easily recognizable as such. As I said, no intricate reflection was needed to see that there is reason to lie when lying would save a life. Unlike the agents, the assessors of other people's actions are not in such a fortunate position concerning the relevant information. In assessing other people's actions assessors are usually only confronted with the deed done and not with all the circumstantial reasons surrounding that deed. Therefore, it is rational for assessors to judge on grounds of moral principles. Moral agents, however, usually being in the epistemically fortunate posi-

tion of knowing a great deal about the circumstantial features of the particular situation on which they are called upon to act, are not morally justified in acting on these principles. I think it would be highly counterintuitive to claim that moral agents should act on certain principles in a situation *S* in spite of knowing that doing so would run counter to the moral reasons present in *S*. At this point the situation looks like a predicament: Assessors of actions are morally justified in judging and sanctioning on the basis of principles, whereas moral agents do not have that option. And the conclusion to draw would be that these principles are still no more than non-moral coercions, for moral agents could both act morally right and still be sanctioned for their actions.

The second principle tries to solve this dilemma. It opens up space to check the validity of those socially applied principles without substituting them with situation-specific explanations. It tries to present an account of how these sanction-based, social principles are *regulated*, though not *eliminated*, by moral reasons. My argument for this is, again, pragmatic. On a plausible assumption, the overall moral rightness of *each and every* action is foremostly determined by the contributory reasons in a specific situation. However, as I already stressed, in the course of life we lack the opportunity to discuss and justify all our moral commitments and actions. Therefore, because situation-specific contributory reasons determine overall moral rightness and wrongness, it follows that *any case* is, as it were, set up to be explained in terms of the contributory. But, on the pragmatic assumption, we can't explain them *all*. If it were always possible to ask for explanations and make an exception out of *each and every* case, those principles' favouring nature would devolve into being merely statistical. Then the normative status of these principles, *normative* default reasons, would vanish. But since we can, in the rush of our lives, only *sometimes* ask for justification in moral matters we are usually normatively bound by the principled moral claims people have on us. And, as I explained, moral assessors do have good reason to judge our actions on those grounds.

So, which are the principles that are morally justified? My suggestion is that a principle is justified as long as it admits of so few exceptions that all the occurring exceptions can be explained on grounds of the pragmatic constraint I gave. If a principle has too many exceptions, the excessive need for exceptional explanations would violate the pragmatic constraint which is that, in everyday life, there is only room for so many explanations. Only those social norms that do not violate the pragmatic constraint qualify as moral principles. The boundaries, however, are faded. After all, it is difficult to say from when a pile is a pile.

One might object that the question of which principles hold sway in a certain society is still a matter of contingency. After all, if the world had been different, justified principles of our world could well be different in other worlds. For that reason, the theory cannot defend any of the Rossian prima facie values. True, the principles that actually do suffice for assessing agents' actions are contingent; it

depends on the shape of the specific world. But that there is at all moral reason for assessors to judge agents on grounds of principles is, as far as I can see, necessary. My thesis is not a piece of normative ethics. It does not tell a story about which principles are actually right, much rather it aims at analysing the nexus between moral principles, whatever they might be, and situation-sensitive moral explanations. After all, the theory explains why principles like the golden rule do not suffice for application in our world and principles like "thou shall not kill" do. The former simply admits of too many counterexamples. Therefore, the need for justification would be excessive and would, hence, violate the pragmatic constraint.

Both principles together state a version of default moral principles. Moral principles, on this account, are "general moral reasons unless undermined". These principles are, justifiably, taken as principles by moral assessors, but can sometimes be undermined by special explanations.

References

Cullity, Garrett: "Particularism and Moral Theory: Particularism and Presumptive Reasons". Aristotelian Society: Supplementary Volume, 76 (1), 2002. S. 169 - 190

Dancy, Jonathan: Ethics without principles. Clarendon Press, Oxford, 2004

Dancy, Jonathan: Moral reasons. Blackwell, Cambridge, 1993

Ross, William David: The Right and the Good. Oxford University Press, 1930

Shope, Robert K.: "Prima Facie Duty". The Journal of Philosophy, 62 (11), 1965. S. 279-287

Väyrynen, Pekka: "Particularism and Default Reasons". Ethical Theory and Moral Practice, 7(1), 2004. S. 53-79

Reflexive Stabilität interessenbasierter Moralbegründung

Michael von Grundherr
michael.von.grundherr@parmenides-foundation.org
Parmenides Stiftung und Ludwig-Maximilians-Universität, München

Abstract/Zusammenfassung

Prudential arguments for the justification of moral norms are attractive – especially in modern pluralistic societies –, as they need only weak premises. Opponents argue, however, that they are reflectively unstable. They claim that once a participant of the moral practice has understood and accepted the core argument of prudential justification, he will not accept the moral practice anymore. I will suggest a methodological differentiation that helps to weaken this objection: interest based arguments are only suited for the justification, not for an analysis or explanation of the moral practice, or so I will argue. In a weaker form, however, the threat of reflective instability persists and must be accommodated by the justificatory argument.

Interessenbasierte Argumente der Moralbegründung sind gerade in modernen pluralistischen Gesellschaften attraktiv, da sie mit sehr sparsamen Prämissen auskommen. Allerdings sind sie kontrovers. Viele Argumente der Gegner einer solchen Theorie kristallisieren sich um den Vorwurf der reflexiven Instabilität: Die Theorie dekonstruiere das, was sie begründen wolle, nämlich die moralische Praxis. Ich argumentiere – in Auseinandersetzung mit der Diskussion um den interessenbasierten Kontraktualismus – für eine methodologische Differenzierung, die hilft, diesen Vorwurf als prinzipiellen Einwand zu entkräften: Interessenbasierte Argumente sollte man nur als Rechtfertigung, nicht aber als Erklärung einer moralischen Praxis verstehen. In einer abgeschwächten Form bleibt die Gefahr der reflexiven Instabilität aber bestehen und muss von Fall zu Fall konstruktiv in das Begründungsargument einbezogen werden.

Die Herausforderung der reflexiven Instabilität

Jede Theorie, die moralische Normen und Sanktionen damit begründet, dass sie langfristig im Eigeninteresse aller Mitglieder einer moralischen Gemeinschaft sind, muss das Problem der reflexiven Instabilität lösen.

Interessenbasierte Begründung

Paradigmatisch steht für solche Begründungstheorien der interessenbasierte Kontraktualismus, wie ihn Hobbes (1651), Gauthier (1986) oder zuletzt Stemmer (2000) vertreten. Sie argumentieren, dass der Zwang zu normkonformem Verhalten, den eine moralische Gemeinschaft durch Sanktionen wie sozialen Druck oder Ausschluss aus der gesellschaftlichen Kooperation ausübe, deswe-

gen gerechtfertigt sei, weil jeder einzelne so besser gestellt sei als in einer Gemeinschaft, die diesen Zwang nicht ausübe. Ohne weitere Qualifikation folgt daraus, dass der einzige rechtfertigende und motivierende Grund für die Einrichtung der moralischen Praxis Konformität mit dem Eigeninteresse ist. Zudem gilt auf den ersten Blick auch, dass die Mitglieder der moralischen Gemeinschaft den moralischen Regeln folgen, weil sie andernfalls Sanktionen fürchten. Moralische Gefühle oder ein moralisches Gewissen sind aus dieser Sicht selbst Sanktionen – aus gesellschaftlicher Sicht wünschenswerte psychologische Mechanismen, die den einzelnen zur Einhaltung der Normen zwingen. Das selbe gilt für einen großen Teil der moralischen Praxis, etwa moralische Erziehung, moralische Vorwürfe oder moralisches Lob. Diese Konsequenzen führen dazu, dass interessenbasierte Theorien der Moralbegründung sehr kontrovers sind.

In der eben vorgestellten Lesart beantwortet die interessenbasierte Moralbegründung zwei Fragen, die jede vollständige Theorie einer moralischen Praxis beschäftigen: a) die Frage nach einer Erklärung der moralischen Praxis und b) die Frage nach der Rechtfertigung dieser Praxis. Ein dritter wichtiger Fragenkomplex, der sich damit beschäftigt, was in einzelnen Handlungssituationen moralischer Weise zu tun ist, ist nicht Teil dessen, was ich eine Theorie der moralischen Praxis nennen möchte – er ist vielmehr eine theoretische Reflexion, die Teil dieser Praxis ist.

Auf die erste Frage antwortet der explanatorische Teil der Moraltheorie. Er erklärt aus einer Beobachterperspektive zum Beispiel, welche kognitiven Prozesse moralischen Urteilen zugrunde liegen, was moralische Akteure motiviert, etwa moralische Gefühle oder sozialer Druck, was moralische Urteile bedeuten, welchen Regeln die moralische Begründungspraxis folgt und welche Prämissen sie verwendet. Die interessenbasierte Begründung in obiger Lesart vertritt in diesem Punkt eine streng sanktionistische Auffassung: Moralische Akteure denken ihr zufolge direkt oder indirekt über die Folgen von Sanktionen nach, wenn sie moralische Entscheidungen treffen, sie handeln moralisch, um die negativen Folgen von Sanktionen zu vermeiden, ihre Urteile sind letztlich in Urteile über Sanktionen übersetzbar und der moralische Diskurs ist in letzter Analyse eine sachliche Auseinandersetzung über die faktische Wirksamkeit von Sanktionen.

Auf die zweite Frage antwortet hingegen der normative Teil einer Moraltheorie. Er fragt nicht, wie Personen sich de facto in einer moralischen Praxis verhalten, sondern beschäftigt sich damit, ob es gute Gründe für die Akzeptanz dieser Praxis gibt. Er richtet sich zum Beispiel an den moralischen Skeptiker, der fragt, warum er denn überhaupt moralisch sein und die Regeln der moralischen Gemeinschaft, in der er lebt, akzeptieren solle. Die oben skizzierte Variante der interessenbasierten Moralbegründung antwortet hier, dass das Leben jedes einzelnen (nach dessen eigenen Maßstäben und Präferenzen beurteilt) durch die moralische Praxis besser werde. Üblicherweise argumentiert diese Theorie mit einer hypothetischen Situation, in der es die moralische Praxis nicht gibt und zeigt,

dass es für jedes Mitglied der jetzigen moralischen Gemeinschaft aus dem jeweils eigenen Interesse heraus gute Gründe gäbe, die Einführung der moralischen Praxis zu befürworten. Konsequenter Weise, so das Argument, könne man nun, da die moralische Praxis in dieser gewünschten Form bestehe, diese nicht ablehnen. Freilich kann es sein, dass diese Diskussion zu dem Schluss kommt, dass die moralische Praxis in ihrer jetzigen Form nicht ideal ist, um das gute Leben jedes einzelnen zu fördern. Dieses Ergebnis gibt einen Anstoß dazu, eine bessere Variante der moralischen Praxis zu entwerfen und für eine Reform zu plädieren.

Peter Stemmers äußerst konsequente Variante einer interessenbasierten Begründung der moralischen Praxis eignet sich gut als Illustration einer solchen Doppeltheorie. Stemmer gibt als Grundherausforderung für seine kontraktualistische Theorie an: Die „Grundschwierigkeit" liege „darin, dass wir nicht klar haben, von welcher Art das moralische Müssen, bzw. Nicht-Dürfen ist." (Stemmer 2000, 5). Aus seiner Sicht ist die Frage, ob Moral begründbar sei, nicht unabhängig davon zu beantworten, wie sie zu verstehen sei.[1] Seine Strategie zielt darauf, eine Analyse der moralischen Praxis zu finden, die zeigt, dass es notwendigerweise vernünftig ist, moralisch zu handeln. Seine Lösung besteht darin, das moralische Müssen konsequent als sanktionskonstituiert zu analysieren, moralische Motivation also als kanalisiertes Eigeninteresse zu verstehen: „Das moralische Müssen [...] ist ein durch Sanktionen künstlich geschaffenes Müssen" (Stemmer 2000, 118). Moralische Sanktionen erfüllen nach Stemmer allerdings die Zusatzbedingungen, dass sie gerechtfertigt, da von freien Individuen wählbar und tatsächlich gewählt, sind. Deswegen sei es vernünftig, die Sanktionen für die moralischen Regeln einzuführen. Sie in jeder Situation zu befolgen sei hingegen schlicht deswegen rational, weil andernfalls nachteilige Sanktionen drohten.

Es gelingt Stemmer, eine stimmige Theorie zu entwickeln, die frei vom Ballast „einer metaphysischen und gegebenenfalls religiösen Vorstellungswelt" (Stemmer 2000, 143) ist. Auf der anderen Seite handelt sie sich wie alle vergleichbaren Ansätze aber das Problem der reflexiven Instabilität ein.[2]

1 Ausführlicher schreibt er: „Die Frage nach dem Vernünftigsein moralischen Handelns wird [...] schnell zu der Frage führen, von welcher Art die moralischen Forderungen sind und von welcher Art das moralische Müssen ist, das in ihnen zum Ausdruck kommt, wo es herkommt und was ihm seine verpflichtende Kraft gibt." (Stemmer 2000: 10).

2 Stemmer ist sich dieses Problems nach meinem Verständnis bewusst, nimmt es aber in Kauf in plädiert für eine teilweise revisionistische Sicht auf einige Teil unserer jetzigen moralischen Praxis.

Reflexive Instabilität

Christine Korsgaard fordert in Anlehnung an Bernard Williams (1985, 101f.) von einer Moraltheorie:

> A normative moral theory must be one that allows us to act in the full light of knowledge of what morality is and why we are susceptible to its influences, and at the same time to believe that our actions are justified and make sense. (Korsgaard 1996, 17)

Eine Moraltheorie, die diese Forderung erfüllt, nenne ich dem verbreiteten Sprachgebrauch entsprechend *reflexiv stabil*. Reflexive Stabilität ist eine unvermeidbare Anforderung an jede normative Moraltheorie. Eine solche Theorie möchte ja Gründe für das Befolgen bestimmter Regeln geben oder moralische Forderungen aufstellen. Ist sie nicht reflexiv stabil, gibt sie dem Skeptiker, der fragt, warum er überhaupt moralisch sein solle, stattdessen gute Gründe, die moralische Praxis und damit auch alle moralischen Forderungen abzulehnen. Ein Beispiel für eine instabile Theorie ist laut Korsgaard eine naive evolutionäre These. Diese behauptet, moralische Haltungen und Gefühle seien einfach eine Art starker Instinkt, der in der Evolution entstanden sei, und ließen sich dadurch rechtfertigen, dass sie das Überleben der Spezies Mensch förderten. Jemand, der von dieser Theorie überzeugt sei, könne und werde sich fragen, welchen Grund er selbst habe, in einer Situation moralisch zu handeln, wenn ihm das ein großes Opfer abverlange. Wie alle Instinkte könnten ja auch die moralischen zu Handlungen motivieren, die für das eigene Wohlergehen fatal seien. Ohne weiteres sei das langfristige Überleben der Spezies, so Korsgaard, kein besonders guter Grund, der diese Opfer, wie sehr sie sich uns auch aufdrängen mögen, prinzipiell aufwiege. Vielmehr gebe ein Verständnis dieser Theorie einen Grund, die eigenen moralischen Intuitionen sehr kritisch zu hinterfragen. (Korsgaard 1996, 14f.)

Die interessenbasierte Theorie scheint auf den ersten Blick in ähnlicher Weise reflexiv instabil zu sein: Muss nicht jemand, der von ihr überzeugt ist und sein Selbstbild entsprechend angepasst hat, konsequenter Weise nur noch nach seinem eigenen Vorteil handeln und Lücken in den Sanktionen ausnutzen? Die Theorie sagt ihm ja, dass der einzige Grund, die moralischen Regeln zu befolgen, andernfalls zu befürchtende Sanktionen seien. Wenn aber diese Sanktionen unvollständig sind – sollte er dann nicht das Spiel der moralischen Praxis *zum Schein* mitspielen und immer dann, wenn es für ihn in Summe besser ist, einfach die Regeln verletzen?

Korsgaards These und den Vorwurf gegen die interessenbasierte Moralbegründung rekonstruiere ich mit Hilfe der oben eingeführten Unterscheidung zwischen zwei Teilen einer Theorie der moralischen Praxis folgendermaßen: Die Theorie T einer moralischen Praxis MP enthalte einen explanatorischen Teil T_E und einen rechtfertigenden Teil T_R. T ist reflexiv instabil und daher abzulehnen, wenn ein Mitglied von MP, das von T_E überzeugt ist,

a. durch das Verständnis von T_E eine schwächere Motivation hat, MP entsprechend zu handeln oder Distanz zu den moralischen Intuitionen einnimmt und
b. bei Akzeptanz des in T_E vorausgesetzten Selbstverständnisses T_R nicht als Begründung für MP akzeptieren kann.

Angewendet auf Korsgaards Beispiel heißt das:

a. durch das Verständnis der Analyse der moralischen Motive als Instinkte (T_E) nimmt ein Mitglied einer moralischen Gemeinschaft Distanz zu ihnen ein und
b. bei Akzeptanz der These, dass die eigenen moralischen Antriebe nur evolutionär sinnvoll, dem eigenen Wohl aber oft abträglich seien, reicht zur Rechtfertigung der Teilnahme an der moralischen Praxis der Hinweis auf das Wohl der Spezies (T_R) nicht aus.

Der Vorwurf gegen die interessenbasierte Begründung lautet nun präziser: Seien der explanatorische und der rechtfertigende Teil der interessenbasierten Theorie TI_E und TI_R für sich genommen schlüssig, dann gelten trotzdem die folgenden beiden Punkte:

a. Die Akzeptanz von TI_E führt dazu, dass man die Natur moralischer Gefühle und aller anderen Teile der moralischen Praxis als Zwangsmechanismen durchschaut und deswegen Distanz zu ihnen einnimmt und
b. bei Akzeptanz der These, dass der einzige Grund für moralische Handlungen andernfalls zu befürchtende Sanktionen seien, reicht zur Rechtfertigung der Teilnahme an der moralische Praxis (MP) der Hinweis darauf, dass MP dem moralischen Akteur insgesamt Vorteile gegenüber einer Situation ohne eine moralische Praxis biete, nicht aus. Wenn ein von TI_E beschriebener Akteur TI_R nachvollzieht, sieht er zugleich immer einen (noch besseren) Grund, die oben beschriebene nur scheinbar moralische Haltung einzunehmen.

Punkt b. ist weitgehend äquivalent mit einem anderen häufig geäußerten Kritikpunkt. Das interessenbasierte kontraktualistische Argument TI_R zeigt immer, dass es für jeden einzelnen vorteilhafter ist, wenn sich *alle* an die moralischen Regeln halten, als wenn *keiner* die Regel befolgt. Unter der Annahme, dass sich eine Begründung an alle richten muss bzw. dass die Regeln *einstimmig* akzeptiert werden müssen, liefert es jedem einen entscheidenden Grund, die moralischen Regeln ohne Ausnahme zu akzeptieren, also eine echte moralische Haltung einzunehmen. Denn: eine scheinbar moralische Haltung einzunehmen, kann keine allgemeine Zustimmung finden, da sie Nachteile für die anderen Mitglieder der moralischen Gemeinschaft mit sich bringt. Wenn ein Akteur aber die Randbedingung der *öffentlichen Rechtfertigbarkeit* oder *einstimmigen Akzeptierbarkeit* nicht akzeptiert, dann scheitert TI_R daran. Denn dann ist es aus seiner Sicht immer die beste Option, dass alle anderen die Regel immer befolgen und er das nur dann tut, wenn er durch Sanktionen gezwungen wird. Und der von TI_E beschriebene Akteur kann keinen Grund haben, die Bedingung der Einstimmigkeit zu akzeptieren.[3]

3 Zur Frage der Einstimmigkeit siehe ausführlicher von Grundherr (2007, 72f., 114f.).

Moderne interessenbasierte Begründungs-Theorien verstehen sich typischerweise als Aufklärungstheorien, d.h. es ist gewünscht, dass ihr explanatorischer Teil T_E von den moralischen Akteuren verstanden und letztlich als aufgeklärtes Selbstbild akzeptiert wird. Damit akzeptieren diese Theorien den ersten Punkt zum Teil, würden vermutlich aber abstreiten, dass es zu einer Distanzierung von den moralischen Einstellungen kommt – und vertreten damit eine sehr kontroverse These. Stemmer zum Beispiel möchte, dass man das moralische Müssen tatsächlich als sanktionskonstituiertes Müssen betrachtet.

Den zweiten Punkt sehen die kontraktualistischen Theoretiker und versuchen, seine Schärfe abzumildern. Stemmer behandelt ihn unter der Überschrift „Unrechttun im Verborgenen". Er argumentiert, eine Reihe von Effekten führe dazu, dass Sanktionen auch in vermeintlich unbeobachteten Handlungssituationen wirkten: Die Entdeckung könne nicht mit Sicherheit ausgeschlossen werden, internalisierte Sanktionen wirkten unabhängig von der Beobachtung anderer, sekundäre Sanktionen, die die Ausbildung bestimmter Dispositionen erzwängen, wirkten breit über viele Situationen hinweg und nicht zuletzt sei es meist eine kluge Faustregel, einfach immer moralisch zu handeln (Stemmer 2000, 189f.). Trotzdem „liegt in der begrenzten Wirksamkeit der moralischen Sanktionen eine Schwäche der nach kontraktualischen Baugesetzen errichteten Moral", so Stemmer (2000, 189). Fatal sei das für die Theorie keineswegs, denn jede künstlich geschaffene Institution sei unvollständig, müsse deswegen aber noch lange nicht dysfunktional sein; man denke nur an das Rechtssystem. Diese Theorie ist konsequent, aber zu einem nennenswerten Teil revisionistisch.

Auch Gauthier beschäftigt sich ausführlich mit dem Problem des unbeobachteten Regelbruchs. Er kann es abschwächen, in dem er eine relativ umständliche Theorie der moralischen Dispositionen und damit letztlich einen komplizierteren Sanktionsmechanismus einführt. (Gauthier 1986, 157-189)

Keine dieser Reparaturen kann aber das systematische Problem des zweiten Punktes beheben, dass nämlich der Grund dafür, nicht immer moralisch zu handeln, für das von der explanatorischen Theorie beschriebenen Individuum ceteris paribus besser ist, als der Grund, es immer zu tun. Die Theorie steht vor dem Dilemma, dass sie immer das Bild des moralischen Akteurs als eines eigeninteressierten Maximierers, der wegen der andernfalls erwarteten Sanktionen moralisch handelt, ein Stück weit zurücknehmen muss, oder aber eine genuin moralische Haltung im strengen Sinn nicht reflexiv stabil rechtfertigen kann. Auch wenn man zeigen kann, dass das in bestimmten Anwendungskontexten de facto aus pragmatischen Gründen kein relevantes Problem für die soziale Ordnung ist – das prinzipielle Problem bleibt bestehen, solange man die Theorie als eine Analyse unserer moralischen Praxis und des begrifflichen Rahmens, in dem wir moralische Diskurse führen, versteht.

In den schließenden Kapiteln von „Morals by Agreement", widmet sich Gauthier explizit dem Problem, dass der homo oeconomicus, den er in seiner Theo-

rie zur Rechtfertigung moralischer Regeln und Sanktionen verwendet, Moral immer nur als ein notwendiges Übel ansehen werde: „Economic man lacks the capacity to be truly the just man. [...] Given the opportunity to use morality as an instrument of domination, he unhesitatingly does so [...].“ (Gauthier 1986, 328) Tatsächliche Menschen hingegen könnten sich wirklich an Moral binden: „because we are [...] not economic men and women, we can be constrained“ (Gauthier 1986, 317). Er argumentiert damit in eine ähnliche Richtung wie mein folgender Vorschlag.

Die Strategie der methodologischen Differenzierung

Ich werde dafür argumentieren, dass eine interessenbasierte Theorie der Moralbegründung eine Strategie der methodologischen Differenzierung wählen sollte, um diesem Problem zu begegnen.

Trennung von Erklärung und Rechtfertigung

Ich halte es für einen Irrtum, dass wir als Mitglieder moderner Gesellschaften vor allem eine bessere und ganz neue Art der Erklärung der moralischen Praxis brauchen. Sicher ist eine „folk theory“ der Moral, die deren Basis in göttlichen Geboten sieht, vermutlich fehlerhaft und vor allem in modernen Gesellschaften nicht mehr überzeugend. Nur ist diese Theorie eher ein philosophischer Strohmann als ein tatsächlich wichtiger Gegenspieler. Die moderne Psychologie, die Soziologie und auch die philosophische Metaethik liefern bei aller Unvollständigkeit bereits gute explanatorische Theorien der moralischen Praxis und entwickeln sich schnell weiter. Die aus dem interessenbasierten Kontraktualismus abgeleitete explanatorische Theorie ist mit großer Sicherheit auf diesem Feld weit abgeschlagen. Was wir aber brauchen, ist eine neue Theorie moralischer Rechtfertigung die mit neuen Problemen der sozialen Interaktion in modernen Gesellschaften umgehen kann und frei ist von den nicht mehr allgemein akzeptierten Instanzen wie Gott oder einer reinen Vernunft. Hier kann eine interessenbasierte Theorie ihre Stärken ausspielen.

Die Strategie der methodologischen Differenzierung greift diese Ideen auf und argumentiert, eine metaethisch reflektierte interessenbasierte Begründung der moralischen Praxis enthalte keine explanatorische Theorie TI_E. Dann kann die reflexive Instabilität in der oben definierten Form nicht auftreten. Denn wenn es keinen explanatorischen Teil der Theorie gibt, dann kann es kein Verständnis einer solchen explanatorischen Theorie geben, das die moralische Motivation schwächt oder dem Begründungsteil der Theorie Überzeugungskraft nimmt. Um eine komplette Theorie zu erhalten, muss man dann die beste verfügbare unabhängige explanatorische Theorie T_E* ergänzen: zum Beispiel den

begrifflichen Teil aus der Metaethik, den empirischen aus der Moralpsychologie und Soziologie.

Könnte man nicht einwenden, es sei ein Erfordernis theoretischer Sparsamkeit, die kontraktualistische Theorie für möglichst viele Teile der Theorie der moralischen Praxis zu verwenden und eine vielleicht implizit immer schon angelegte oder zumindest nahe liegenden interessenbasierte Erklärung der moralischen Praxis anzuschließen? Dieser Einwand greift nicht, denn man würde die Theorie damit einfacher als nötig machen – und auch bei weitem einfacher als möglich. Denn für die etablierten (kontraktualistischen) interessenbasierten Theorien ist die moralische Praxis eine Black Box und es gibt kein theoretisches Potenzial, auf das man durch die Kombination mit einer anderen Theorie verzichtet. Dass im Kern einer *metaethisch reflektierten* interessenbasierte Begründung keine explanatorische Theorie TI_E enthalten ist, zeigt sich sehr deutlich daran, dass sie typischer Weise im Rahmen einer hypothetischen Situation argumentiert. In dieser Situation geht es – nach Abzug aller schmückenden Rahmenerzählungen über Verträge und Urgesellschaften auf der grünen Wiese – um die Frage, ob man eine vollkommen fiktive Sanktionsmaschine einführen sollte (vgl. dazu Stemmer (2002, 94f.) und von Grundherr (2007, 59ff.)). Gezeigt ist damit, dass es für jedes Mitglied der Gesellschaft aus dem eigenen Interesse heraus Gründe gibt, ein funktionales Äquivalent dieser Maschine zu akzeptieren. Das kann jede Art der sozialen Praxis sein, die in denselben Situationen und mit demselben Effekt wie diese Maschine auf die handlungswirksame Motivation der Mitglieder der Gemeinschaft einwirkt. Ob es vor allem innere oder äußere Sanktionen sind, welche Rolle Erziehung und Sozialisation spielen und wie die moralischen Akteure die Sanktionen erleben, muss diese Theorie vollkommen offen lassen. Die interne Funktion der Black Box Sanktionsmaschine bleibt ein blinder Fleck dieser Theorie.

Damit hängt die reflexive Stabilität einer vollständigen Theorie der moralischen Praxis davon ab, mit welcher externen explanatorischen Theorie T_E* man die interessenbasierte Rechtfertigungstheorie TI_R kombiniert. Korsgaards „knowledge of what morality is and why we are susceptible to its influences" (1996, 17), das diese explanatorische Theorie liefern muss, ist viel reicher und empirisch adäquater als die krude aus der kontraktualistischen Begründung abgeleitet Theorie. T_E* muss zwar zeigen, dass die moralische Praxis funktional äquivalent zu einer rechtfertigbaren Sanktionsmaschine ist, dass die Deliberation eines moralischen Akteurs und damit des Adressaten dieser Theorie aber eben gerade nicht in der Abwägung von Vor- und Nachteilen aus den Sanktionsmechanismen besteht. TI_R kann dann wiederum zeigen, dass genau das gut ist und nur eine genuin moralische Praxis aus Sicht des eigenen Interesses jedes Mitglieds der moralischen Gemeinschaft gerechtfertigt werden kann. Das durchzuführen ist ein eigenes

aufwändiges Forschungsprogramm, deswegen skizziere ich nur zur Illustration kurz die Grundzüge eines Ansatz, der das versucht.

Die Kombination aus T_E* und TI_R gewinnt zum Beispiel dann an Stabilität, wenn T_E* zeigt, dass stabile moralische Gefühle persönlichkeitskonstitutiv sind und selbst zum Wohlergehen des einzelnen beitragen. Die moralische Praxis erscheint dann nicht aufgezwungen, sondern natürlich, die Mitglieder der moralischen Gemeinschaft werden keine Distanz zu der moralischen Praxis in Form ihrer moralischen Gefühle einnehmen. Damit ist der erste Aspekt der reflexiven Instabilität (a.) ausgeschlossen.[4]

Der zweite (b.) ist ausgeschlossen, wenn die Akzeptanz von T_E* das Ergebnis der interessenbasierten Rechtfertigung stützt. Dazu kann eine solche Theorie plausibler Weise argumentieren, dass uns die Art der Deliberation, die wir laut einer als explanatorisch missverstandenen interessenbasierten Begründungstheorie üblicherweise verwenden sollten, gar nicht zur Verfügung steht. Diese Theorie hatte behauptet, der einzige Grund, moralische Regeln einzuhalten, bestehe in den andernfalls zu erwartenden Sanktionen. Plausibler ist eine Theorie, die beschreibt, dass es für uns selbstverständlich ist, soziale Regeln zu befolgen, dass es uns wichtig ist, gemeinsame Ziele zu verfolgen, und dass wir unsere Handlungen nicht als sinnvoll erachten, weil sie unser eigenes Interesse befriedigen, sondern weil sie gegenüber unserer Gemeinschaft (öffentlich) gerechtfertigt werden können.

Das ist empirisch recht plausibel: Schon Kinder ab dem einem Alter von etwa drei Jahren suchen laut verschiedener Studien aktiv nach sozialen Regeln und sind intrinsisch dadurch motiviert, sich so zu verhalten „wie man es hier macht“ (vgl. z.B. Tomasello u. a. 2009, 40-44). Thompsons handlungstheoretischer Ansatz könnte ein guter Ausgangspunkt sein, um diese Beschreibung eines sozialen Akteurs begrifflich zu fassen. Im Sinne von Thompson könnte man argumentieren, der typische Grund dafür, etwas zu tun, bestehe darin, dass es Teil einer in einer Praxis eingebetteten Aktivität sei, nicht in einer Mittel-Zweck-Argumentation. Wir überlegen nicht nach dem Muster: ‚ich tue X, weil es zum gewünschten Y führt', sondern nach dem Schema: ‚ich tue X, weil ich Y tue (und man Y eben so macht)'.[5]

---

4 Gauthiers Charakterisierung des „liberal individual“ (1986, 330ff.) ist zum Beispiel eine solche Theorie.

5 Thompson fasst sein Programm wie folgt zusammen: „It is, I want to suggest, only because we are to start with the kind of thing of which you can say something like ‚She's doing A because she's doing B' that we can be or become the sort of which you might say ‚She's doing A because she wants to do B.'“ (2008, 92). Später gibt er folgende allgemeine Begriffsbestimmung von intentionaler Handlung: „X's doing A is an intentional action (proper) under that description just in case the agent can be said, truly, to have done something else *because he or she was doing A*.“ (Thompson 2008, 112) Die übliche und primäre Erklärung dafür, dass man etwas tut, ist demnach also, dass es Teil einer anderen Hand-

Wenn das eine gute Beschreibung der Deliberation eines moralischen Akteurs ist, dann tritt die Problematik dass jemand, der diese Beschreibung als Selbstbild akzeptiert, immer einen besseren Grund hat, die scheinbar moralische Haltung einzunehmen, nicht auf. Denn erstens denkt er in moralischen Entscheidungssituationen nicht über Sanktionen und Lücken in Sanktionen nach – Sanktionen tauchen in seiner Deliberation typischerweise gar nicht auf. Die scheinbar moralische Haltung ist aus einem solchen Selbstverständnis heraus schlichtweg keine Optionen für einen psychisch gesunden Menschen – wer sich so verhält, gilt als Psychopath, aber nicht als ein kompetentes Mitglied der moralischen Gemeinschaft. Aber auch das Problem, dass TI_R nicht ausreicht, um die genuin moralische Haltung zu begründen, weil die Forderung der Einstimmigkeit oder öffentlichen Rechtfertigbarkeit für den moralischen Akteur nicht nachvollziehbar ist, trifft nicht. Vielmehr ist diese Forderung für ihn selbstverständlich.

Für die interessenbasierte Rechtfertigung sind diese Argumente wiederum wichtige Randbedingungen. Denn sie kann nur empfehlen, Sanktionen zu akzeptieren, die auch wirklich funktionieren, nämlich eine vollständige moralische Praxis, zu der die Förderung moralischer Gefühle, soziale Sanktions-Mechanismen, Erziehung und moralische Bildung gehören. Dann rechtfertigt TI_R die *genuin* moralische Haltung als *beste* Wahl aus Sicht des Eigeninteresses und zwar für jedes Mitglied der Gemeinschaft, das sich selbst entsprechend der explanatorischen Theorie T_E* versteht. Denn nur die funktioniert auch so *als ob* wir eine Sanktionsmaschine hätten.

Diese vorgeschlagene Theoriekombination zeigt, dass etwas, was wir ohnehin tun und was uns selbstverständlich ist, *zudem* aus Sicht des eigenen Interesses jedes einzelnen gut begründet werden kann. Gerade weil moralische Akteure sehr tief in die moralische Praxis ihrer Gemeinschaft eingetaucht sind, ist die interessenbasierte Rechtfertigung als Korrektiv aus einer externen Perspektive wichtig. Der interessenbasierte Rechtfertigungsteil dieser kombinierten Theorie trägt auf diese Weise zur reflexiven Stabilität der Gesamttheorie bei.

Grenzen der Begründbarkeit – ein Problem in der Wirtschaftsethik

In sehr vielen Fällen tatsächlicher moralischer Gemeinschaften, lässt sich eine Kombination aus sozialen Praktiken und moralischen Haltungen durch TI_R kontingenter Weise reflexiv stabil rechtfertigen. Das eben angeführte Argument weist aber darauf hin, dass T_E^* in gewissem Umfang eine Prämisse für TI_R ist und damit ein schwächeres Problem der reflexiven Instabilität bestehen bleibt: In bestimmten Fällen wird TI_R in Kombination auch mit einer guten explanatorischen Theorie T_E* die Haltung des nur scheinbar Moralischen rechtfertigen.

---

lung ist, und nicht, dass man ein bestimmtes Ziel erreichen oder einen Wunsch befriedigen möchte.

Diese schwache Form der reflexiven Instabilität ist gerade in Fragen der Wirtschaftsethik sehr relevant. Prägnant taucht das Problem hier in der Diskussion darüber auf, ob moralkonformes Verhalten bloß Teil einer PR- oder Marketing-Strategie (also scheinbar moralisches Verhalten) ist. Es ist plausibel, dass die beste Theorie T_E* in diesem Fall zeigt, dass oben genannte moralpsychologische Mechanismen nur eingeschränkt greifen, weil es um das Verhalten von Institutionen oder Personen in selektiven Funktionszusammenhängen geht. In vielen Fällen wird man daher nur einen modus vivendi und einen rechtsförmigen institutionellen Rahmen reflexiv stabil rechtfertigen können.[6]

Fazit

Ich habe dafür argumentiert, dass der Einwand reflexiver Instabilität für eine Theorie, die ein interessenbasiertes Argument sowohl als Rechtfertigung als auch als Beschreibung der moralischen Praxis sieht, gefährlich ist. Jemand, der den explanatorischen Teil dieser Doppeltheorie akzeptiert, kann nicht überzeugt sein, dass die moralische Praxis (in nicht revidierter Form) durch den Rechtfertigungsteil der Theorie begründet werden kann. Trennt man den Rechtfertigungsteil von dem ohnehin unplausiblen explanatorischen Teil und ersetzt letzteren durch eine eigenständige explanatorische Theorie, trifft dieser Einwand nicht mehr prinzipiell.

Auch wenn in manchen Fällen, etwa im Kontext der Wirtschaftsethik, das interessenbasierte kontraktualistische Rechtfertigungsargument in Kombination mit der bestmöglichen explanatorischen Theorie reflexiv instabil werden kann und dann in manchen Fällen zu Beschränkungen rechtfertigbarer Normen oder Sanktionen führt, ist das interessenbasierte Begründungsargument in Kombination mit einer guten explanatorischen Theorie in sehr vielen der üblichen Fälle gut geeignet, die Akzeptanz einer genuin moralischen Haltung reflexiv stabil zu rechtfertigen.

Literaturverzeichnis

Gauthier, David: Morals by Agreement. Oxford University Press, Oxford, 1986

von Grundherr, Michael: Moral aus Interesse: Metaethik der Vertragstheorie. de Gruyter, Berlin/New York, 2007

6 Diese Analyse gibt eine Erklärung dafür, dass die Diskussion um die institutionenökonomisch geprägte Wirtschaftsethik Homanns (2001) notorisch kontrovers geblieben ist.

Hobbes, Thomas: Leviathan. Cambridge University Press, Cambridge, 1996, 1651

Homann, Karl: „Ökonomik: Fortsetzung der Ethik mit anderen Mitteln“. In: *Georg Siebeck*: Artibus Ingenuis: Beiträge zur Theologie, Philosophie, Jurisprudenz und Ökonomik. Mohr-Siebeck, Tübingen, 2001. S. 85–110

Korsgaard, Christine M: The Sources of Normativity. Cambridge University Press, Cambridge, 1996

Stemmer, Peter: Handeln zugunsten anderer: Eine moralphilosophische Untersuchung. de Gruyter, Berlin/New York, 2000

Stemmer, Peter: “Moralischer Kontraktualismus“. Zeitschrift für philosophische Forschung, 56(1), 2002. S. 1–21

Thompson, Michael: Life and action: elementary structures of practice and practical thought. Harvard University Press, Cambridge (MA), 2008

Tomasello, Michael/Dweck, Carol/Silk, Joan/Skyrms, Brian/ Spelke, Elizabeth S.: Why We Cooperate. MIT Press, 2009

Williams, Bernard: Ethics and the Limits of Philosophy. Fontana, London, 1993, 1985

8 Ästhetik und Religionsphilosophie

Ästhetische Erfahrung(en): über ihre Pluralität und ihre Rolle in der Philosophie der Kunst

Stefan Deines
deines@em.uni-frankfurt.de
Goethe-Universität Frankfurt am Main

Abstract/Zusammenfassung

This paper addresses the question to what extent the dimension of experience is of relevance for a philosophy of art. In the course of the argument two frequent positions are rejected: firstly the position which assumes the existence of one specific form of aesthetic experience, which characterizes every situation of art reception and which can be used to define the concept of art and be called on to measure the quality of an artwork. Secondly the position which is aware of the theoretical problems of such a notion of aesthetic experience and therefore holds that a theory of art should abandon the notion of experience altogether. In contrast to both positions this paper claims that it is crucial to take the dimension of experience into account to understand the processes of understanding and evaluating works of art; however the plurality and the respective contexts of experience in art reception have to be considered.

In diesem Beitrag werden Überlegungen zu der Frage angestellt, inwieweit die Dimension der Erfahrung für eine Philosophie der Kunst von Relevanz ist. Dabei werden zwei populäre Positionen zurückgewiesen: Einerseits eine Position, die es als die Aufgabe der Kunsttheorie ansieht, eine spezifisch ästhetische Erfahrung im Singular zu explizieren, die für alle Formen und Situationen der Kunst Relevanz besitzen und der Definition und der Bewertung von Kunst dienen soll. Andererseits eine Position, die sich den theoretischen Schwierigkeiten mit einem solchen Konzept ästhetischer Erfahrung bewusst ist und daraus den Schluss zieht, dass eine Theorie der Kunst auf den Begriff der Erfahrung insgesamt verzichten sollte. Demgegenüber wird dafür argumentiert, dass die Berücksichtigung der Dimension der Erfahrung für eine Theorie der Interpretation und der Beurteilung von Kunstwerken unabdingbar ist, dass aber der Vielfalt und dem jeweiligen Kontext der Erfahrungen in der Kunstrezeption Rechnung getragen werden muss.

Einleitung: Zum Begriff der ästhetischen Erfahrung

Obwohl der Begriff der ästhetischen Erfahrung in der philosophischen Ästhetik der letzten Jahrzehnte eine zentrale Stellung einnimmt, herrscht nicht sehr viel Einigkeit darüber, was mit diesem Begriff genau gefasst werden soll und kann: diese Uneindeutigkeit rührt teilweise daher, dass er sich aus zwei Begriffen zusammensetzt, von denen jeder für sich genommen bereits mehrere Bedeutungsaspekte in sich trägt: so lassen sich beim Begriff der Erfahrung vier verschiede-

ne (allerdings eng verzahnte) Begriffsverwendungen unterscheiden: Erstens kann mit dem Begriff der Erfahrung die gesamte Breite bewussten menschlichen Erlebens und Empfindens gemeint sein; in einem zweiten Sinn ist er stark auf Erkenntnis bezogen und meint das, was man über die Sinne oder durch die Kommunikation mit anderen in Erfahrung bringen kann; in einem dritten emphatischeren Sinn, wie er sich in prominenter Weise etwa bei Hans-Georg Gadamer findet, macht man eine Erfahrung, wenn man über den krisenhaften Weg der Enttäuschung von Erwartungen und Vorannahmen zu einer veränderten Sicht auf sich oder die Welt gezwungen wird, und in einem vierten Sinn sprechen wir von jemandem als erfahren, wenn er in seinem Leben bereits eine Anzahl von Erfahrungen in den vorangegangenen Sinnen gesammelt hat.

Der Begriff der Ästhetik besitzt ähnlich viele Facetten: Die drei, die ich hier anführen möchte, lassen sich in einer etwas vereinfachenden Lesart von jeweils einem Hauptvertreter der klassischen philosophischen Ästhetik herleiten. Bei diesen wird ‚ästhetisch' jeweils auf verschiedene paradigmatische Gegenstandsbereiche bezogen: bei Baumgarten vorrangig auf die sinnliche Wahrnehmung bzw. die unteren Erkenntnisvermögen, bei Hegel auf Gehalte und Darbietungsweisen von Kunstwerken und bei Kant bereits selbst auf bestimmte Formen der Erfahrung von Subjekten, insbesondere die durch Wohlgefallen gekennzeichnete Erfahrung des Schönen.

Aufgrund dieser Offenheit der beiden Teil-Begriffe ergeben sich nun recht verschiedene Phänomene, die mit dem Begriff der ästhetischen Erfahrung bezeichnet werden können: ob ich beispielsweise erläutern möchte, auf welche Weise ich schöne Objekte sinnlich wahrnehme, oder aber, wie Gegenstände der Kunst zu einer – um mit Gadamer zu sprechen – „Erschütterung [....] des Gewohnten" führen, die mich zu der Einsicht führen: „Du musst Dein Leben ändern",[1] scheinen sehr unterschiedliche Unterfangen zu sein.

Trotz dieser potentiellen Vielfalt verläuft die philosophische Auseinandersetzung um die ästhetische Erfahrung aber nun nicht so, dass alle verschiedenen möglichen Konzepte der ästhetischen Erfahrung expliziert werden, die den oben genannten Unterschieden Rechung tragen würden. Vielmehr drehen sich die Debatten meist um die Frage, wie *die* ästhetische Erfahrung im Singular richtig zu konzeptualisieren sei. Es wird davon ausgegangen, dass die ästhetische Erfahrung ein singuläres und einheitliches Phänomen ist, das sich durch eine bestimmte Struktur oder durch bestimmte konstitutive Elemente beschreiben und von allen anderen möglichen Erfahrungen abgrenzen lässt.

1 Hans-Georg Gadamer: „Ästhetik und Hermeneutik", 9.

I. Die Rolle ästhetischer Erfahrung in der Kunstphilosophie

In den meisten Fällen wird auf den Begriff der ästhetischen Erfahrung im Rahmen einer Kunsttheorie zurückgegriffen. Ästhetische Erfahrung, so die Überlegung, kann uns erläutern, was das Spezifische an unserer Auseinandersetzung mit Kunst ist. In paradigmatischer Weise lassen sich bei Monroe Beardsley all die systematischen Vorteile aufweisen, die das Konzept der ‚ästhetischen Erfahrung' in der Theorie der Kunst haben kann: In seiner *Aesthetic Defintion of Art* gewinnt Beardsley erstens eine Definition von Kunst: Kunstwerke sind danach all die Objekte, die mit der Intention hergestellt wurden, beim Rezipienten eine ästhetische Erfahrung hervorzurufen; er erläutert zweitens, warum wir die Auseinandersetzung mit Kunst suchen: wir haben ein Interesse an ästhetischen Erfahrungen, weil sie für uns angenehm und wertvoll sind, und er liefert drittens ein Kriterium für die ästhetische Qualität von Kunstwerken: diejenigen Werke sind gute Kunstwerke, denen es gelingt, ästhetische Erfahrung in einer gewissen Intensität hervorzurufen.[2]

Die Schwierigkeiten für eine solche definitorische und essentialistische Kopplung von Kunst und ästhetischer Erfahrung ergeben sich naturgemäß, wenn es daran geht, den Begriff der ästhetischen Erfahrung inhaltlich in einer Weise zu füllen, die es erlaubt, damit tatsächlich all die verschiedenen Werke und Situationen der Kunst und nur diese zu beschreiben. Um dies zu leisten ist versucht worden, die ästhetische Erfahrung über eine spezifische emotionale Komponente wie eine bestimmte Art der Lust oder des Wohlgefallens zu bestimmen (z.B. Andrea Kern, Jerrold Levinson), über eine bestimmte Rolle der Sinnlichkeit oder des sinnlichen Erscheinens (Jerrold Levinson, Martin Seel), über eine bestimmte Ganzheitlichkeit, Lückenlosigkeit und Intensität der Erfahrung (Monroe Beardsley, John Dewey), über eine bestimmte Zeitlichkeit ihres Vollzugs (Martin Seel, Michael Theunissen) oder über eine spezifische Form des Verstehens bzw. des Scheiterns von Verstehensprozessen (Georg Bertram, Hans Ulrich Gumbrecht, Andrea Kern, Christoph Menke, Juliane Rebentisch).

Die meisten dieser Bestimmungen sind nun an paradigmatischen Kunstformen gewonnen, an denen sie sich besonders gut demonstrieren lassen; die Übertragung auf die jeweils anderen Formen der Kunst erfolgt dagegen meist nicht reibungslos. Exemplarisch und verkürzt gesprochen haben die Theorien, die sich eher auf sinnliche Wahrnehmung stützen, gewisse Probleme mit den anästhetischen Gattungen wie der *Concept Art* oder auch der literarischen Prosa; für die Theorien, für die das Verstehen bestimmter Gehalte der Kunst im Mittelpunkt steht, stellt dagegen meistens die Musik die größte Herausforderung dar. Was geschehen kann, wenn die Schwierigkeiten der Übertragung der paradigmatischen Bestimmung auf alle Gegenstände, die gemeinhin als Kunstwerke

2 Vgl. Monroe C. Beardsley: „An Aesthetic Definition of Art".

gelten, allzu groß werden, lässt sich wiederum an Beardsley zeigen. Er zieht die Grenzen dessen, was als Kunst gelten soll, gemäß seiner Definition schlicht neu: Duchamps *Ready-Mades* sind danach einfach keine Kunstwerke; Kinderzeichnungen und Fälschungen aber unter Umständen schon. Angesichts solcher Schwierigkeiten muss man Zweifel daran hegen, dass es der Kunstphilosophie gelingen kann, die eine Form der Erfahrung zu bestimmen, die das Wesen aller Kunst ausmacht.

II. Kritik am Begriff der ästhetischen Erfahrung

Entgegen solcher Versuche, Kunst über eine bestimmte einheitliche Form, Struktur und Erlebnisqualität der Erfahrung zu definieren, findet sich in der analytischen Ästhetik eine Traditionslinie, die von George Dickie bis zu Noel Carroll reicht und die die Rede von ‚ästhetischer Erfahrung' und von ‚ästhetischer Einstellung' massiv kritisiert. Diesem Theoriestrang zufolge handelt es sich bei der vermeintlich spezifisch ästhetischen Einstellung (als einer Form von Interesselosigkeit oder Distanziertheit) um nichts weiter als um die Aufmerksamkeit für die Gehalte, Strukturen und Eigenschaften der Kunstobjekte. Und auch die ästhetische Erfahrung ist danach nichts anderes als eben der aufmerksame Nachvollzug dieser Elemente und Eigenschaften des Objekts in der Weise, in der das Werk sie dem Rezipienten darbietet. Für die Dimension der Erfahrung, die eine eigene relevante Form oder Qualität in den Rezeptionsprozess einbringen würde, haben die Theorien dieser Ausrichtung keinerlei Verwendung.

Nelson Goodman hat in diesem Geist in *Sprachen der Kunst* eine zeichentheoretische Ästhetik vorgelegt, in der Kunst und ihr Wert für uns vollständig ohne Rückgriff auf die Dimension der Erfahrungen bestimmt werden. Danach handelt es sich bei Gegenständen schlicht dann um Objekte der Kunst, wenn sie als Zeichen in bestimmten Symbolsystemen fungieren, welche sich im Normalfall durch mehrere symptomatische Eigenschaften von anderen Symbolsystemen unterscheiden. Der Wert der Kunst besteht darin, dass wir über den Weg der Interpretation der Zeichen der Kunst zu bestimmten Erkenntnissen gelangen können, und die Qualität einzelner Kunstwerke bemisst sich daran, „wie gut sie der kognitiven Zielsetzung dienen" und nützliche Mittel für das „Erfassen, Erkunden und Durchdringen der Welt" darstellen.[3]

Auch bei Goodman finden sich nun aber einige Bemerkungen über die Erfahrung von Kunstwerken, denn bestimmte Sorten von Zeichen können eine besondere Weise der Rezeption erfordern: die ästhetische Erfahrung wird so als eine spezifische Form des Verstehens bestimmt. Die für die Symbolsysteme der

3 Nelson Goodman: *Sprachen der Kunst*, 237.

Kunst symptomatischen Eigenschaften der syntaktischen Fülle sowie der syntaktischen und semantischen Dichte führen zu einer komplexen und aufwendige Prozedur der Deutung ihrer Zeichen. Wo eine Vielzahl verschiedener Eigenschaften eines Zeichens bedeutungsvoll sind, wie etwa Farbe, Linienführung, Pinselstrich und möglicherweise sogar die Struktur der Leinwand bei einem Gemälde, und wo darüber hinaus jede noch so kleine Veränderung eines dieser Aspekte zu einer Veränderung in der Bedeutung des Werkes führt, kann die Interpretation des Zeichens sehr konzentrierte Aufmerksamkeit und viel Zeit erfordern.

Ein weiteres Merkmal der Erfahrung von Werken der Kunst, das Goodman thematisiert, ist die Dimension der Emotionalität. Auch diese wird über ihre Beteiligung im Prozess des Verstehens erläutert.[4] Emotionen können danach im Verlauf der Rezeption von Werken dabei helfen, bestimmte Eigenschaften und Gehalte eines Kunstwerks zu entdecken: Emotionen sind Erkenntnismittel. Diese epistemische Rolle der Emotionalität ist aber keineswegs etwas, was für das Verstehen von Kunstwerken insgesamt charakteristisch wäre, da es Kunstwerke gibt, für die Emotionen nicht die adäquaten Erkenntnismittel darstellen, noch ist die kognitive Rolle der Emotionen auf den Bereich der Kunst beschränkt; denn Emotionen können Goodman zufolge sowohl im Alltag als auch in den Wissenschaften dieselbe Funktion übernehmen.

III. ‚Context of discovery' oder ‚context of justification'?

Obwohl Goodman für bestimmte Aspekte der ästhetischen Erfahrung im Prozess der Zeicheninterpretation einen Ort findet, entsteht der Eindruck, dass die Zeitlichkeit und die Emotionalität der Rezeption bei Goodman nicht die Rolle spielen, die ihnen tatsächlich zukommen oder die ihnen doch zumindest in bestimmten Kunsterfahrungen zukommen können. Mit Hans Reichenbachs Unterscheidung aus der Wissenschaftstheorie ließe sich sagen, dass Goodman die Aspekte der Erfahrung rein dem ‚context of discovery' zugeordnet hat, während wir dagegen zumindest für manche Fälle darauf bestehen sollten, dass die Elemente und Strukturen der Erfahrung zum ‚context of justification' gehören.

Für die Beurteilung des Kunstwerkes ist Goodman zufolge lediglich von Relevanz, was es mir zu verstehen gibt, welche Erkenntnisse es mir vermittelt bzw. welche Version von Welt es entwirft. Auf welchem Wege ich zu diesen Erkenntnissen gelange ist für die Bedeutung der Zeichen, die verstanden werden, und damit auch für den Wert und die Qualität der Werke, nicht von Belang. Genausowenig wie die Dauer und der Grad der Konzentration bei der Lektüre eines mit unleserlicher Handschrift geschriebenen Briefs für dessen Inhalt von Bedeu-

4 Ebd. 241.

tung sind, genausowenig ist der Gehalt des Kunstwerkes bei Goodman mit dem Prozess seiner Rezeption und Interpretation in substantieller Weise verbunden. Und da das Ziel und der Wert der Kunst für uns nach Goodman ausschließlich im Verstehen ihrer Gehalte liegt, ist die ästhetische Erfahrung auch keine, die wir um ihrer selbst willen suchen und schätzen, sondern nur ein Zwischenschritt auf dem Weg zu einer Erkenntnis; die wir dann allerdings sehr wohl um ihrer selbst willen schätzen können.

Ich möchte gegen eine solche Position argumentieren, dass die Strukturen und Elemente der Erfahrung sowohl für die Bestimmung des Gehalts als auch für die Beurteilung der Qualität von Kunstwerken durchaus von zentraler Bedeutung sein können. Wenn etwa die Kritzeleien von Cy Twombly, die verwischten und unscharfen Fotografie-Gemälde von Gerhard Richter oder die Präsentation eines vollständig lichtlosen Raums von James Turrell von ihren Rezipienten ein außergewöhnliches Maß an Zeit und Konzentration erfordern, um entziffert oder erkannt zu werden, dann ist diese Verzögerung in der Erfahrung selbst etwas, was von den Werken gezielt evoziert und eigens thematisiert wird, indem auf diesem Weg etwa die Brüchigkeit von Kommunikation, unser alltäglicher oberflächlicher Umgang mit dokumentarischen Fotos oder die Unzuverlässigkeit der sinnlichen Wahrnehmung vor Augen geführt werden kann. Der Verweis auf die Besonderheiten der Erfahrung in der Auseinandersetzung mit diesen Werken kann daher eine legitime Rolle im Rahmen ihrer Interpretation spielen.

Auch die von Kunstwerken evozierten Emotionen können als Teil ihres Gehalts oder ihrer Thematik beschrieben werden. Die Aggressionen, die die Filme Lars von Triers beim Betrachter auslösen können, und die Verstörung, in die Michael Hanekes *Klavierspielerin* den Rezipienten stürzen kann, sind nicht nur Mittel zum Verstehen von bestimmten ästhetischen Eigenschaften des Werkes, sondern sie werden selbst zu Elementen des Werkes, indem durch sie die üblichen Reaktionen und Konventionen bezüglich sozialer Gewalt oder sexueller Orientierungen aufgerufen und ins Spiel gebracht werden, die das Thema dieser Filme darstellen. So kann die Erfahrung eines Kunstwerks auf dieselbe Weise zur Thematik und zum Gehalt dieses Werkes gehören wie etwa auch die wütende Reaktion des spießigen Passanten ein bedeutsamer Teil einer Performance in der Fußgängerzone einer Kleinstadt sein kann. Diese Reaktionen auf die Werke sind Elemente, die in einer Interpretation der Werke selbst eine Rolle spielen können.

IV. Relevante und irrelevante Aspekte der Erfahrung

Natürlich ist nicht jedes Element der Erfahrung des Kunstwerkes in dieser Weise relevant, sondern es sind nur bestimmte Reaktionen und Emotionen, die Ein-

gang in eine angemessene Interpretation des Werkes finden können, nämlich nur diejenigen, so könnte man sagen, die auf die richtige Weise mit dem Werk und seinen Eigenschaften in Verbindung stehen. Diese ‚richtige' Verbindung lässt sich mit zwei Bestimmungen aus Arthur Dantos Buch *The Abuse of Beauty* besser fassen, in dem er dem Schönen, dem Erhabenen, dem sinnlichen Erscheinen und anderen „toxic properties" wieder einen, wenn auch bescheidenen, Platz in der philosophischen Ästhetik zuweist, nachdem er sie gut 25 Jahre vorher in der *Verklärung des Gewöhnlichen* von ihrem Platz im Zentrum der Theorie der Kunst vertrieben hatte:[5]

Zum einen können wir mit Danto sagen, dass die relevanten Reaktionen Reaktionen auf die pragmatischen Eigenschaften des Werkes sein müssen. Pragmatische Eigenschaften des Werkes sind diejenigen, die bestimmte Emotionen oder Einstellungen beim Rezipienten auslösen.[6] Nur wenn meine Reaktionen im Rezeptionsprozess tatsächlich von den entsprechenden Eigenschaften des Kunstwerkes hervorgerufen werden, sind sie Kandidaten dafür, als relevante Elemente in einer Interpretation dieses Werkes zu fungieren: dass mir schwindlig und schummrig wird und ich meiner Wahrnehmung misstraue, sollte in diesem Sinn also banalerweise auch wirklich auf das verwirrende und unübersichtliche Arrangement der Videoinstallation zurückzuführen sein – und nicht auf meinen übermäßigen Alkoholkonsum am Abend zuvor.

Mit Dantos Unterscheidung zwischen internen und externen Eigenschaften lässt sich nun zweitens weiter erläutern, warum nicht alle auf diese Weise zustandegekommenen Reaktionen in derselben Weise für die Interpretation eines Werks von Bedeutung sind.[7] Diese Unterscheidung nutzt Danto zur Fortführung des traditionsreichen Streits um Duchamps Ready-Made *Fountain*. Vielleicht ist es möglich, so gesteht Danto zu, dass ich das Urinal als einen ästhetischen Gegenstand betrachte und mir in der Ausstellung seine Schönheit: sein Glänzen, sein reines Weiß und seine elegant geschwungene Form zu Bewusstsein kommen. Diese Schönheit sei nun aber lediglich extern, da sie für die Deutung des Gehalts des Werkes keinerlei Relevanz besitze. Das Thema, der Witz und die Kraft von *Fountain* lassen sich explizieren, ganz ohne auf diese ästhetischen Eigenschaften Bezug zu nehmen. Mit Danto müsste man daher noch präzisieren, dass die Schönheit extern ist, insofern es sich bei ihr gar nicht um eine Eigenschaft des Kunstwerkes im engeren Sinn handelt, sondern dass sie lediglich zu den Eigenschaften des ‚materiellen Gegenstückes' zählt.[8] Entsprechend müsste

5 Arthur C. Danto: *The Abuse of Beauty*, XIX; vgl. auch: Ders.: „A Future for Aesthetics".

6 „Pragmatic Properties are intended to dispose an audience to have feelings of one sort or another toward what the artwork represents." (Danto: *The Abuse of Beauty*, XV.)

7 Vgl. ebd. insb. 94-102.

8 Vgl. zu dieser Unterscheidung Arthur C. Danto: *Die Verklärung des Gewöhnlichen*, insb. 159-164.

man konstatieren: Das Urinal ist möglicherweise schön, *Fountain* aber in jedem Fall nicht.

Dass die Unterscheidung zwischen intern und extern sinnvoll ist und nicht alle Reaktionen, die von einem Kunstwerk ausgelöst werden, in gleichem Maße für seine Interpretation von Relevanz sind, sollte einem vor allem an der Konfrontation mit Werken aus anderen Kulturen und auch aus vergangenen Zeiten deutlich werden. Während beispielsweise der Bollywoodfilm oder die mittelalterliche Ikone, bei dem angestammten Zielpublikum, das sie jeweils adressieren, möglicherweise Freude, Mitgefühl oder Ehrfurcht auslösen, können meine Reaktionen darauf deutlich davon abweichen. Und zur Bestimmung des Themas und der Qualität dieser Werke sollten in einer besonderen Weise die Reaktionen der Rezipienten berücksichtigt werden, an die sie sich eigentlich wenden. Denn nur hier finden sich die Reaktionen, die von den Werken selbst kalkuliert worden sind.

Kunstwerke richten sich im Regelfall an ein Publikum, das mit bestimmten Perspektiven, Themen, Konventionen der Darstellung und Umgangsweisen mit den ästhetischen Medien und Materialien vertraut ist. Diese Vertrautheit umfasst nicht nur ein theoretisches und kunstgeschichtliches Wissen, sondern auch eine praktische Haltung, die aus einem Training und einer Einübung in bestimmte Seh- und Rezeptionsweisen hervorgeht. Durch diese Einübung in eine bestimmte Praxis der Kunstbetrachtung haben sich die Rezipienten, die von einem Werk adressiert werden, in die notwendige Ausgangsposition gebracht, bestimmte Reaktionen überhaupt zeigen und bestimmte Erfahrungen überhaupt machen zu können. So kann nur derjenige angemessen auf ein Werk klassischer Musik reagieren, der schon einen Referenzrahmen verschiedener anderer Werke klassischer Musik besitzt und eine grobe Einsicht in deren bedeutsame Elemente und Strukturen hat. Viele Zeitgenossen sind z.B. unfähig, auf Werke der abstrakten Malerei oder der neueren Musik angemessen zu reagieren, da ihnen die Vertrautheit und der Referenzrahmen fehlen, die notwendig wären, um ihre Eindrücke in bestimmte, qualifizierte und passende Erfahrungen zu verwandeln.[9]

Kunstwerke operieren mit einer Art ‚idealem Rezipienten', der bestimmte Kenntnisse hat, bestimmte Rezeptionshaltungen einnehmen kann und damit die ‚richtigen' Reaktionen auf das Werk überhaupt zeigen kann. In der Tradition der Rezeptionsästhetik ließe sich sagen, dass Kunstwerke jeweils auf einen spezifischen kulturellen und historischen „Erfahrungsraum" und „Erwartungshorizont"

9 Darauf weist z.B. Stephen Davies hin: „And one needs to perceive and understand a lots of things about music in order to recognize expressiveness in it (and to respond to what one hears with appropriate emotions). So practical is the knowledge involved that its role is not always apparent to the absorbed listener." Stephen Davies: „Rock versus Classical Music", 195.

bezogen sind.[10] Die Bedeutung und der Rang eines Werkes lassen sich danach nicht rein formal oder werkimmanent begreifen, sondern das Werk muss im Kontext seiner Rezeptionsweisen und Wirkungen und der herrschenden Traditionen und Konventionen betrachtet und beurteilt werden. In diesem Sinn konstatiert Hans Robert Jauss für Werke der Literatur:

> Ein literarisches Werk, auch wenn es neu erscheint, präsentiert sich nicht als absolute Neuheit in einem informatorischen Vakuum, sondern prädisponiert sein Publikum durch Ankündigung, offene und versteckte Signale, vertraute Merkmale und implizite Hinweise für eine ganz bestimmte Weise der Rezeption. Es weckt Erinnerungen an schon Gelesenes, bringt den Leser in eine bestimmte emotionale Einstellung und stiftet schon mit seinem Anfang Erwartungen für ‚Mitte und Ende', die im Fortgang der Lektüre nach bestimmten Spielregeln der Gattung oder Textart aufrechterhalten oder abgewandelt, umorientiert oder auch ironisch aufgelöst werden können.[11]

Die Bewertung der Qualität eines Werkes lässt sich nur mit Rücksicht auf sein Verhältnis zu den konkreten Rezeptionsweisen erläutern: mit Blick darauf, inwieweit und in welcher Hinsicht bei den Lesern oder Betrachtern in der Erfahrung des Werks die an es herangetragenen Erwartungen bestätigt oder durchkreuzt wurden.

Ausblick: Pluralität der Kunsterfahrung

Die vorangegangenen Überlegungen haben eine Art Mittelweg beschrieben, der zwischen zwei Alternativen der Kunsttheorie hindurchführt Zum einen ist bezweifelt worden, dass sich ein Begriff ästhetischer Erfahrung finden lässt, der für alle Kunstwerke und Kunstgattungen in ihrer Pluralität und Diversität in gleicher Weise Anwendung finden kann. Auf der anderen Seite ist aber auch der Ansatz abgelehnt worden, auf die Dimension der Erfahrung der Werke in der Kunsttheorie ganz zu verzichten. Ich habe dagegen versucht zu zeigen, dass bestimmte für die Deutung und die Einschätzung von Kunstwerken zentrale Aspekte verlorengehen, wenn man die Dimension ihrer Erfahrung vollständig ausblendet. Es ist danach die Aufgabe einer Theorie der Kunst, zu explizieren, in welchen Hinsichten und in welcher Weise die Erfahrung von Kunstwerken für diese Werke selbst konstitutiv sein kann; als Elemente ihrer Interpretation oder ihrer Bewertung.

Damit ist die Kunsttheorie an ein breites Feld vielfältiger und verschiedenartiger Erfahrungen in der Auseinandersetzung mir Kunst verwiesen. Viele der Elemente und Strukturen dieser Erfahrungen werden sich sicherlich auch außer-

10 So die titelgebende Terminologie des berühmten Aufsatzes von Reinhart Kosellek, die sich aber fast identisch auch schon findet in: Hans Robert Jauss: „Literaturgeschichte als Provokation der Literaturwissenschaft", 46f.

11 Ebd.

halb des Kunstkontextes im Alltag und anderen Bereichen des Lebens finden, andere sind dagegen möglicherweise typisch für die Auseinandersetzung mit bestimmten Kunstformen (zu denken wäre hier etwa an die besondere Form eines gleichzeitig involvierten wie distanzierten emotionalen Mitvollzugs, wie die narrativen und fiktionalen Gattungen ihn ermöglichen). Die Kunsttheorie sollte dabei nicht auf Vereinheitlichung und Reduzierung aus sein, sondern die Phänomene in ihrer Differenziertheit anerkennen. Indem sie verschiedene konkrete Formen und Funktionen von Erfahrung analysiert, trägt sie zum Verständnis der unterschiedlichen Verfahren und Potentiale der verschiedenen Medien, Gattungen und Werke bei und vergrößert auf diesem Weg auch unser allgemeines Verständnis davon, was Kunst ist und was sie kann.

Literaturverzeichnis

Beardsley, Monroe C.: Aesthetics: Problems in the Philosophy of Criticism. Harcourt, Brace & World, New York, 1958

Beardsley, Monroe C.: „An Aesthetic Definition of Art". In: *H. Curtler (Hrsg.):* What is Art?. Haven, New York, 1983. S. 15-29

Bertram, Georg W.: Kunst. Eine philosophische Einführung. Reclam, Stuttgart, 2005

Carroll, Noel: „Aesthetic Experience Revisited". British Journal of Aesthetics, 2/42, 2002. S. 145-168

Danto, Arthur C.: Die Verklärung des Gewöhnlichen. Eine Philosophie der Kunst. Suhrkamp, Frankfurt/M., 1991

Danto, Arthur C.: The Abuse of Beauty. Aesthetics and the Concept of Art. Open Court, Chicago/La Salle, 2003

Danto, Arthur C.: „A Future for Aesthetics". The Journal of Aesthetics and Art Criticisms, 2/51, 1993. S. 271-277

Davies, Stephen: „Rock versus Classical Music". The Journal of Aesthetics and Art Criticisms, 2/57, 1999. S. 193-204

Dewey, John: Kunst als Erfahrung. Suhrkamp, Frankfurt/M., 1988

Dickie, George: „Beardsley's Phantom Aesthetic Experience". The Journal of Philosophy, 5/62, 1965. S. 129-136

Dickie, George: „The Myth of the Aesthetic Attitude". American Philosophical Quarterly, I/1964. S. 55-65

Gadamer, Hans-Georg: Ästhetik und Hermeneutik. In: Gesammelte Werke, Bd. 8. Mohr, Tübingen, 1993. S. 1-8

Goodman, Nelson: Sprachen der Kunst. Entwurf einer Symboltheorie. Suhrkamp, Frankfurt/M., 1995

Gumbrecht, Hans Ulrich: „Epiphanien". In: *Joachim Küpper/Christoph Menke (Hrsg.)*: Dimensionen ästhetischer Erfahrung. Suhrkamp, Frankfurt/M., 2003. S. 203-222

Jauss, Hans Robert: „Literaturgeschichte als Provokation der Literaturwissenschaft". In: *Dorothee Kimmich et al. (Hrsg.)*: Texte zur Literaturtheorie der Gegenwart. Reclam, Stuttgart, 1996. S. 41-55

Kern, Andrea: Schöne Lust. Eine Theorie der ästhetischen Erfahrung nach Kant. Suhrkamp, Frankfurt/M., 2000

Levinson, Jerrold: „What is Aesthetic Pleasure". In: *Levinson, Jerrold*: The Pleasures of Aesthetics. Cornell Univ. Press, Ithaca, 1996. S. 3-10

Rebentisch, Juliane: Ästhetik der Installation. Suhrkamp, Frankfurt/M. 2003

Seel, Martin: „Über die Reichweite ästhetischer Erfahrung – Fünf Thesen". In: *Seel, Martin*: Die Macht des Erscheinens. Suhrkamp, Frankfurt/M., 2007. S. 56-66

Theunissen, Michael: „Freiheit von der Zeit. Ästhetisches Anschauen als Verweilen". In: *Theunissen, Michael*: Negative Theologie der Zeit. Suhrkamp, Frankfurt/M., 1991. S. 285-298

Musical Expression: A Wittgensteinean Account

Franz Knappik
franz.knappik@gmail.com
Ludwig-Maximilians-Universität München

Abstract/Zusammenfassung

Musical expression displays a fundamental ambivalence between a subjective side, including ineffability and the inappropriateness of paraphrases, and an intersubjective side, including our ability to argue about expressive descriptions of music. By drawing on the late Wittgenstein, I sketch an account of expression that is supposed to do justice to both sides. The intersubjective side requires, it is argued, a psychological and normative constraint of the possible ways we can hear a given piece of music. An account of these constraints can be developed from Wittgenstein's treatment of aspect-perception, if its normative dimension is highlighted: familiarity with musical cultures creates a psychological constraint on available ways of hearing; the norms which govern those cultures settle which ways of hearing are appropriate. These points may seem to reduce expression to simple symbolic relations (the 'manual view'), and therefore to ignore the subjective side of the ambivalence of expression. I argue that a normative account of expression can deal with this objection, and account for the subjective side of the ambivalence of expression, by drawing on Wittgenstein's Nachlass work in the philosophy of psychology. Wittgenstein investigates into the imponderability and unpredictability characteristic of norms that govern mental terms and related phenomena. The same features hold for musical expression; or so I argue by discussing Wittgenstein's example of a music box in MS 137. By appreciating these normative peculiarities, the Wittgensteinean account avoids being a manual view. More detail is given to the account of the subjective side of expression by introducing distinctions between two types of aspects (decipherable – undecipherable) and of understanding (transitive – intransitive); according to them, expression is accessible only through individual experience plus the relevant normative background. This does not lead, in turn, to subjectivism about the intersubjective side: In discussing Wittgenstein's views about aesthetic descriptions and explanations, I argue that these are sufficient to prevent expression from being merely subjective, although they differ importantly from usual empirical descriptions and explanations.

Musikalischer Ausdruck ist durch eine grundlegende Ambivalenz zwischen einer subjektiven und einer intersubjektiven Seite gekennzeichnet. Die subjektive Seite umfasst die Phänomene der Unsagbarkeit von Ausdruck und der Unangemessenheit von Paraphrasen; zur intersubjektiven Seite zählt unsere Fähigkeit, über expressive Beschreibungen von Musik zu diskutieren. Im Anschluss an den späten Wittgenstein skizziere ich eine Deutung von Ausdruck, die beiden Seiten Rechnung tragen soll. Die intersubjektive Seite erfordert, so argumentiere ich, eine psychologische und eine normative Einschränkung der möglichen Hörweisen eines Musikstücks. Eine Erklärung dieser Einschränkungen kann auf der Grundlage von Wittgensteins Behandlung von Aspektwahrnehmung entwickelt werden, wenn deren normative Dimension betont wird: Die Bekanntschaft mit Musikkulturen schafft dann eine psychologische Einschränkung möglicher Hörweisen; die Normen, die diese Kulturen regeln, legen fest, welche

Hörweisen angemessen sind. Es kann so scheinen, als würden diese Punkte Ausdruck auf einfache symbolische Beziehungen reduzieren (die "Manual-Theorie") und deshalb die subjektive Seite der Ambivalenz von Ausdruck ignorieren. Dagegen versuche ich zu zeigen, dass eine normative Ausdruckstheorie diesen Einwand entkräften und die subjektive Seite der Ambivalenz von Ausdruck erklären kann, indem sie sich auf Wittgensteins Nachlass-Bemerkungen zur Philosophie der Psychologie stützt. Wittgenstein untersucht dort die Unwägbarkeit und Unvorhersehbarkeit, die für Normen charakteristisch sind, welche mentale Begriffe und verwandte Phänomene regieren. Dass dieselben Merkmale auch musikalischen Ausdruck kennzeichnen, zeige ich in einer Diskussion von Wittgensteins Beispiel einer Spieluhr in MS 137. Indem die Wittgensteineanische Position diese normativen Besonderheiten berücksichtigt, kann sie der Gefahr der Manual-Theorie entgehen. Die Erklärung der subjektiven Seite entwickle ich durch Unterscheidungen zwischen zwei Arten von Aspekten (entzifferbar – nicht entzifferbar) und von Verstehen (transitiv – intransitiv) weiter, nach denen Ausdruck nur durch individuelle, normativ informierte Wahrnehmung zugänglich ist. Aus dieser Deutung folgt aber nicht umgekehrt ein Subjektivismus bezüglich der intersubjektiven Seite: Auf der Grundlage von Wittgensteins Auffassung von ästhetischen Beschreibungen und Erklärungen argumentiere ich dafür, dass diese trotz wichtiger Unterschiede zu gewöhnlichen empirischen Beschreibungen und Erklärungen hinreichen, um zu verhindern, dass Ausdruck rein subjektiver Natur ist.

Musical expression is a phenomenon that is in some ways deeply *subjective*. Individual perception is crucial to it; you could not know, say by paraphrase or testimony, the expressive meaning of a piece you have not listened to yourself. Different listeners, and even the same listener on different occasions, may perceive completely different meanings in the music. Such differences often come in shades which are too subtle to be communicated. But at the same time, what we hear in the music is not a matter of arbitrariness or ideosyncracy. There is an infinite range of things you may hear in the enigmatic variations theme of the *Allegretto* movement in Beethoven's *Seventh Symphony*, but you will hardly be able to hear it, say, as triumphantly joyful: The range of available ways of hearing is *psychologically* constrained in a non-ideosyncratic way. And even *if* a listener is able to hear the theme that way – say someone not familiar at all with Western classical music –, his understanding is criticisable as being inadequate, as really missing the music: so there is a *normative* constraint on our perception of expression, too. And this is possible only because we can characterize our understanding of musical meaning, describe the music, and argue about it. These are phenomena of an *intersubjective* dimension about musical expression which counterbalances the subjective side, and is equally important.

I take it that any adequate theory of musical expression has to account for these two dimensions, which together form what may be called a fundamental *ambivalence* about expressive musical meaning. My contention is that an account of expression which aptly fulfils this criterion can be developed by exploring Wittgenstein's late work, including his *Nachlass* manuscripts, on aspects, aesthetics, and the philosophy of psychology. In this contribution, I indicate the

shape an account of musical expression inspired by Wittgenstein may take by discussing the moves it can make in order to deal with the ambivalence of expression.

The first step is to account for the psychological and the normative constraint which are crucial for the *intersubjective* side. The relevant strand in the late Wittgenstein's work is the thematic complex commonly referred to as '*aspect-perception*'. Aspect-perception essentially means *perceiving something as something else*. The most famous example is the duck-rabbit head; there you can switch between two aspects, two ways of seeing the drawing at will. But Wittgenstein discusses other cases, too, where there is only one aspect seen in an object, and where we cannot refrain from seeing it at will.[1] Aspect-perception is important for Wittgenstein's ideas about musical expression since Wittgenstein explicitly treats our perceiving expressive features of music as a type of aspect-perception. The point that I want to emphasize about aspect-perception in this context is the relation which Wittgenstein sometimes claims to obtain between aspect-perception and *participation in a rule-governed social practice*. At one place, Wittgenstein discusses a triangle which can be seen such that now one side is its base, now another one. He remarks that only someone who is familiar with our mathematical concepts can see, and understand what it means to see, a line of the triangle as its base. The required capacity, the technique whose mastery is the "substratum of this experience" (PI ii, p. 208e),[2] includes the linguistic grasp of the relevant concepts, as well as the ability of "making certain applications of the figure" – such as drawings, explanations, descriptions – "quite freely" (PI ii, p. 208e), that is, without much reflection and without making distorting mistakes. The crucial point is *familiarity* with the practice, here the conceptual system of elementary geometry. A beginner may already understand the concepts and be able to apply them inferentially – 'This is a triangle; triangles have bases and apexes; so if this is the base, the opposite angle is the apex'. Still he is not familiar enough with them to *see* them immediately to apply, hence to perceive the aspects. So this is a kind of perception that presupposes abilities and techniques which we have to acquire by training. Perception of primary and secondary qualities of ordinary physical objects, by contrast, does not require

1 The importance of continuous aspect-perception is stressed and developed in Mulhall (1990).

2 Wittgenstein's texts are quoted using the following abbreviations: PI = Philosophical Investigations; PI ii = idem, part II; PG = Philosophical Grammar; L&C = Lectures and Conversations on Aesthetics, Psychology and Religious Belief; LC = Wittgenstein's Lectures Cambridge, 1932-1935; MS = manuscript no. in the Bergen Nachlass edition. Letters following page-numbers from PI ii indicate the position of the remark on the page. In quoting remarks from the Nachlass manuscripts, I supply my own translation and add the German original in a note.

any mastery of colour-concepts and similar terms: a child that is not able yet to apply colour-predicates can still intelligibly be said to perceive colours.

If we interpret our perception of musical expression according to this model, we should assume it to require familiarity with a musical culture. A musical culture is a complex normative practice which constitutes a *psychological* constraint on available ways of hearing: we cannot hear the Beethoven theme as triumphant because to do so would require a habit, a familiarity with a hypothetical musical language in which the theme would have that expression. Accounting for the *normative* constraint is more difficult here. For the musical culture provides only a *hypothetical* constraint which can be expressed in conditionals of the form: 'If the set of norms N is applied to it, object A is correctly seen as B rather than as C'. But what we need are *unconditional* statements in which the consequens is detached from the conditional, for example: 'It is not correct to hear the Beethoven theme as triumphant'. We get such unconditional normative constraint only if there is a matter of fact settling which norm[3] is *really* the measure the music in question should be interpreted and judged by.

Whether there is such a matter of fact or not depends on the *attitude* to the music that is at work in a particular context. Insofar as we use the music for decorative, ceremonial, and similar purposes, no unconditional normative constraints of the type described above seem to be in place. Imagine a movie scene that is accompanied by music from some highly sophisticated musical culture unknown to the average audience of the movie. Imagine further that the emotional effects created by the music as it accompanies the movie work perfectly well but have nothing to do with the original meaning of the music. In some way, this is not a problem; it would seem slightly arrogant to insist that there was a misunderstanding here. Where music is listened to, by contrast, apart from such narrow instrumental relationships, it becomes something worth exploring – a gesture whose meaning is not immediately understood and requires interpretative efforts. The situation is similar then to cases of linguistic interpretation. In principle we could attach whatever meaning we like to tokens of alien linguistic behaviour. But if we want to *understand* the other person, we need to grant authority to the norms governing the practice in which her behaviour originates. This second attitude, then, implies acknowledgement of a matter of fact as to which norm is decisive for the interpretation; wherever this attitude is prevalent, we get the unconditional form of normative constraint.

If we grant that our perception of music is shaped and made possible by the 'mastery of a technique', as is Wittgenstein's aspect-perception, we can understand why it is psychologically and normatively constrained in a non-individualistic manner. But there is a widespread objection to normative accounts of expression. They may seem to be committed to a view on which musical expres-

3 I use the expressions "norm" and "rule" interchangeably in this paper.

sion is a matter of musical *symbols*; such symbols are linked to expressive properties by *clear-cut rules* of the type that can be made explicit in a *manual*. But this view (the 'manual view') of musical expression is squarely false, for whenever music *does* employ stock figures that follow clear-cut rules, it is not expressive at all but rigid and dull.

The next thing to do, then, is to argue that our account is not a version of the manual view; in order to do so, we must turn now to the *subjective* side of the ambivalence. The first step is to identify a type of rules that is not of the clear-cut sort, thereby rejecting the equation of normative views and the manual view. Again, we can draw here on Wittgenstein. In his late work on the philosophy of psychology, he explores what might be called *the grammar of the soul*: he investigates into the characteristic features of the norms which govern the grammar of psychological terms. Their use, as well as the behaviour in which that use is embedded, is distinguished by a constitutive *indeterminacy*. The mastery of such concepts cannot be learned and applied methodically, according to a system of rules that are or can be made explicit, as in mathematics; rather we acquire them through unsystematic experience that is guided, not by manuals, but by hints and clues. Correspondingly, the application of mental terms is highly context-sensitive and based upon "imponderable evidence" ("unwägbare Evidenz"); fine shades of behaviour and perception are crucial here: "Imponderable evidence includes subtleties of glance, of gesture, of tone" (PI ii, p. 228d). Even if I am familiar enough with a person to know how she will behave in a given situation, the precise expression she displays will always be new, fresh and unpredicted: "Expression consists for us in unpredictability. If I knew exactly how he would grimace, move, there would be no facial expression, no gesture", as Wittgenstein states in MS 137 (p. 67 a).[4]

So the grammar of soul does not belong to the clear-cut type of norms. In fact Wittgenstein himself often cites musical expression as one of the phenomena within the logical realm of the mental, akin to expressive behaviour and the 'soul' of meaningful words. After the statement about unpredictability I just quoted, Wittgenstein goes on in MS 137 to discuss the example of a music box (Spieluhr), which is illuminating for our purposes. Music prima facie seems to contradict the point about unpredictability, as Wittgenstein notes: we can hear a piece over and over again, and still its expression can appear always new. Yet there is another sense in which music can be predictable or unpredictable independently of any sound recording. Music played on a music box is predictable because it sounds dull, rigid and mechanical. Just as a banal piece of music is predictable because it follows rigid rules of composition, uses stock gestures etc., the music box music does not display the expressiveness of the original

4 "Ausdruck *besteht* für uns in Unberechenbarkeit. Wüßte ich genau wie er sein Gesicht verziehen sich bewegen wird, so wäre kein Gesichtsausdruck, keine Gebärde vorhanden."

music because it does not exhibit the fine shades and modulations which are characteristic of real music performance. These nuances let the music seem always new and fresh even if we listen again and again to the same recording of it; not because we could not remember them, but because they do not fit into a rigid scheme. Music box music, by contrast, as well as badly composed, banal music, does display such relations, and hence is able also on the first listening; it is music whose structure is capable of being codified and interpreted by way of clear cut rules.

This claim is made more plausible if we acknowledge a further point of Wittgenstein's: the formal similarity between mental and musical discourse is not incidental; rather the grammatical peculiarities help, in Wittgenstein's picture, to demarcate what is soulful – persons and their behaviour as well as music, poems etc. – from the mechanical and lifeless. The music box case is interesting in this respect, too. For on Wittgenstein's view, the right way to approach that demarcation is to compare expressive gestures in behaviour and music with the fake gestures produced by robots or by music boxes: "The opposite of the soulful, however, is the machine-like" (MS 131, p. 140).[5] Wittgenstein states this link explicitly in dealing with the music box: "It would already give us a strange and profound impression if we met people who know only musical box music. Perhaps we would expect them to have a kind of gestures which we do not understand, which we would not know how to react to" (MS 137, p. 67b).[6]

So on the view we are recommending, the perception of musical expression is governed by norms of the type that is characteristic of mental terms and expressive behaviour. Next, we exploit this point in order to account for a further important feature that belongs to the *subjective* side of expression: the intimate relation between expression and perception. First, we explain why expression can be detected only via perception by introducing a distinction between two kinds of aspects that Wittgenstein himself does not draw explicitly. We call these kinds respectively *decipherable* and *undecipherable* aspects. Decipherable aspects can be *inferentially reconstructed* by someone who is not able to perceive them, e.g. because he is not familiar with the practice in question, and hence not able to non-inferentially apply the relevant concepts. Such aspects pertain to practices governed by clear-cut, explicitly statable conventions of the mathematical type, which permit us to reconstruct the aspects. Wittgenstein discusses, for example, the way we deal with technical drawings:[7] they are so com-

5 "Das Gegenteil des Seelenvollen aber ist das Maschinenhafte."

6 "Schon das würde uns einen fremden und tiefen Eindruck machen, wenn wir zu Menschen kämen, die nur Spieluhrmusik kennten. Wir würden uns vielleicht von ihnen auch eine Art Gebärden erwarten die wir nicht verstünden, auf die wir nicht zu reagieren wüßten."

7 Cf., for example, PI ii, p. 204i: "For when should I call it a mere case of knowing, not seeing? – Perhaps when someone treats the picture as a working drawing, *reads* it like a blueprint."

plicated that we cannot see their meaning at a glance unless we have much experience. We can use them nevertheless, for there are clear representational conventions; by employing a legend that states them, we can decipher the drawings. If we access the drawings inferentially, we will be slower and clumsier in using them than an expert will be who just *sees* their meaning. Still, their meaning is accessible to us.

Undecipherable aspects, by contrast, *cannot* be accessed in this indirect, inferential way precisely because they are governed by norms of the indeterminate type. Given the logical features of those norms, no legend can be compiled which would permit to reconstruct the corresponding aspects inferentially. Cases of undecipherable aspects include expressive behaviour, musical expression and other musical qualities, and what Wittgenstein calls 'meaning experience' (Bedeutungserleben) – the kind of experience we have, for example, when the words of a poem seem to be imbued with their meaning. Someone who is not familiar with the relevant practice has no access to the aspects of that type.

So the imponderable type of norms enables us to understand why musical meaning can be detected only by applying concepts non-inferentially in perception. But the subjective side of the ambivalence of musical meaning entails a stronger point: it is not sufficient that *someone* hear the expression and you accept his paraphrase as a testimony, as you could do with a newspaper article; you must hear the music yourself in order to adequately understand it. Again, an account of this phenomenon can be based on the distinction of two types of norms. In practices of the clear-cut type, equivalence relations between different meaningful expressions can be established. Take the form of elementary linguistic understanding that is typical of beginners in a foreign language. This understanding mainly consists in the mastery of the rules that are stated in grammar books and dictionaries. Someone understands an expression in this sense, Wittgenstein tells us, if he is able to replace it by another expression in the same or a different language, in particular his mother tongue. We can call this type of understanding '*transitive*'. Transitive understanding is opposed to a further type of understanding, which can be labelled '*intransitive*'[8]. Intransitive understanding does not involve a substitution; rather it consists in a grasp of what is "expressed only by these words in these positions" (PI, §531).

A paradigm example of intransitive understanding is understanding poetry. The dictionary knowledge of a language is not sufficient here; rather you have to grasp the connotative and physiognomic features of single words and expressions, and this presupposes a deep familiarity with a language, including the literary works that have shaped it. But although Wittgenstein does not make that connection himself, this is precisely the type of understanding you will get when

8 Wittgenstein himself refers to this type of understanding as "intransitive understanding" in PG, p. 79.

what you understand is determined by norms of the indeterminate species. For just as there are no manuals and legends in practices of that sort, there are no simple equivalence relations that would make substitutional understanding possible. Now whereas linguistic understanding, as we have seen, includes both types of understanding and hence of rules, the understanding of music belongs, according to Wittgenstein, exclusively to the intransitive type: he remarks that objects of intransitive understanding cannot be replaced by anything else, '[a]ny more than one musical theme can be replaced by another' (PI, §531). If you read a poem in a language you have only an elementary knowledge of, you will hardly understand anything of its poetic meaning; still you might grasp some literary meaning which you can paraphrase or translate. If someone listens to music, by contrast, which he is only familiar with in a rudimentary manner, he does not really understand anything; he will rather feel like Agamemnon when he is puzzled by the strident music of the Trojans (Ilias 10, v. 12 f.).

By drawing on the distinction between clear-cut and imponderable norms and the related ones between decipherable and undecipherable aspects and between transitive and intransitive understanding, a normative account of musical expression is able to take into account not only the intersubjective, but also the subjective side of musical expression. Yet, by way of conclusion, we must turn once more to the *intersubjective* side of the ambivalence. For what we have been saying so far might suggest that our account is committed to a one-sided stress on the subjective side: a view on which there is no way of identifying, communicating, or describing musical expression. Again, this is only an apparent consequence of the Wittgensteinean account. For the intransitive type of understanding does not preclude the possibility of characterizing what is understood. Rather, as Wittgenstein remarks in a passage where he analyzes his own response to the last variation of the above-mentioned *Allegretto* movement in Beethoven's *Seventh Symphony*, it is the music itself which challenges us to give such characterizations: in a passage that is "immensely expressive" ("ungeheuer ausdrucksvoll") (MS 130, p. 60), the "notes stimulate me to give a description" (ibid., p. 62)[9]. If we give in to this provocation and describe the music, we can say that the music expresses *something*. But at the same time, no description will ever exhaust musical meaning – "It is as if there was still infinitely much to be understood" (MS 130, p. 64)[10]. This is because our characterizations of musical expression do not state properties that are simply there; rather they could be said to trace out meaningful connections that are implicit in the complex of music-cum-normative-background. For what aesthetic descriptions and explanations do is "to *draw one's attention* to certain features, to place things side by side so as to exhibit these features" (LC, p. 38 f.). This is in line with a crucial remark

9 "Diese Töne reizen mich zu einer Beschreibung."

10 "Es ist als ließe sich hier noch ungeheuer viel *verstehen*."

about aspect-perception in general: „[...] what I perceive in the dawning of an aspect is not a property of the object, but an internal relation between it and other objects“ (PI ii, p. 212a). By “internal relations”, we should understand relations that are essentially linked to normative practices, such as relations between concepts and their criteria of application. In the case of music, there are internal relations established by our practices that enable us to place the music ‘side by side’ with many other things – other musical passages, but also gestures, stories, poems, poetic styles, discursive and rhetorical means. We do so not in order to settle once and for all the meaning of the music, but to guide our perception of it.

Therefore, aesthetic description and explanation belong to what Wittgenstein calls (in discussing Freud) “[a]n entirely new account of a correct explanation. Not one agreeing with experience, but one accepted. You have to give the explanation that is accepted. This is the whole point of the explanation” (L&C, p. 18). In arguing for a way of listening, we do not adduce evidence that might decide the case in the way evidence can prove an empirical hypothesis. We try to make the other person hear the piece the way we do, and the explanation is valid if the other person accepts it. But as should by clear by now, we are not recommending a relativistic view: whether someone is able to hear the music a certain way, whether the contextualization that is suggested by a given characterization makes sense to him, is a matter of the normative framework in which the music is appropriately evaluated and interpreted. – It is this normative framework which in the end, on the view I have sketched, explains the ambivalence of musical expression. But it does so only if its particular nature is appreciated – a nature which it shares, for Wittgenstein, which everything that is soulful.

References

Mulhall, Stephen: On Being in the World. Wittgenstein and Heidegger on Seeing Aspects. Routledge, London, 1990

Wittgenstein, Ludwig: Lectures and Conversations on Aesthetics. Psychology and Religious Belief. Edited by Cyril Barrett. Blackwell, Oxford, 1966 [L&C]

Wittgenstein, Ludwig: Philosophical Grammar. Edited by R. Rhees. Translated by A. Kenny. Blackwell, Oxford, 1974 [PG]

Wittgenstein, Ludwig: Philosophische Untersuchungen / Philosophical Investigations. Translated by G.E.M. Anscombe. Blackwell, Oxford, 2. Auflage, 1958 [PI]

Wittgenstein, Ludwig: Wittgenstein’s Nachlass. The Bergen Electronic Edition. Oxford University Press, Oxford, 2000

Wittgenstein, Ludwig: Wittgenstein's Lectures Cambridge, 1932-1935. Edited by Alice Ambrose. Blackwell, Oxford, 1979 [LC]

Anna Karenina und die anderen - Wie fühlen wir für fiktive Figuren?

Eva Weber-Guskar
eva.weber-guskar@phil.uni-goettingen.de
Georg-August-Universität Göttingen

Abstract/Zusammenfassung

Characters in novels and comics, film heroes etc. – in short: protagonists of narrative fictions can be objects of our emotions. We have pity with Anna Karenina, we fear King Kong. Some consider this to be a paradox. They say: Normally, we refer with an emotion to something that we belief to be existent. But in these cases we know that it is just fiction that means that we do not belief that it is true – but still, we experience emotions. How is this possible?

A well known solution of this paradox comes from Kendall Walton. He argues that such fictional emotions, that is emotions that refer to fiction, are just „quasi-emotions"; they are kind of imagined emotions that arise when someone imagines himself to be part of the fictional world. I do not take this solution for convincing.

In this paper I show in two steps why it is not plausible to speak of quasi-emotions here. First I distinguish different forms of emotional relation to fiction to clear up the different possible problems and to show that in many assumes cases of examples, actually the paradox does not occur. Second I show, referring to an idea of Richard Moran, that one can erode the paradox from the inside by putting in question one of the premises. Finally I develop some positive explanations of why and how we can and should take the emotions that we experience towards fiction for real emotions and not just imagined ones.

Romanpersonal, Filmhelden, Comicfiguren – kurz: Protagonisten von narrativen Fiktionen gegenüber können wir Gefühle empfinden. Wir bemitleiden Anna Karenina und fürchten uns vor King Kong. Einige sehen darin ein Paradox. Mit einer Emotion, sagen sie, beziehen wir uns normalerweise auf etwas, das wir für wahr halten. Hier aber wissen wir, dass es sich nur um Fiktion handelt, glauben also nicht daran, und entwickeln dennoch Emotionen. Wie kann das sein?

Eine bekannte Lösung für dieses Paradox der Fiktion stammt von Kendall Walton. Nach ihm sind solche fiktionalen Emotionen, das heißt auf Fiktion bezogene, nur Quasi-Emotionen. Es seien vorgestellte Emotionen, die entständen, wenn man sich selbst als Teil der fiktiven Welt vorstelle. Diese Lösung halte ich für nicht überzeugend.

Ich zeige in diesem Aufsatz in zwei Schritten, wieso es unplausibel ist, hier nur von Quasi-Emotionen zu sprechen. Zuerst unterscheide ich verschiedene Formen des emotionalen Bezugs auf Fiktion, um die verschiedenen Problemmöglichkeiten klar vor Augen zu führen und deutlich zu machen, dass in vielen vermeintlichen Beispielfällen das Paradox in Wahrheit gar nicht auftritt. Dann zeige ich, mit Bezug auf eine Idee von Richard Moran, inwiefern man das Paradox von innen her aushöhlen kann, indem man eine Prämisse in Frage stellt. Schließlich bringe ich noch konstruktive Erläuterungen dazu, warum und inwiefern man die Emotionen,

die wir gegenüber Fiktion erleben, tatsächlich als reale Emotionen, und nicht nur als vorgestellte verstehen kann und sollte.

Einleitung

Mag es heute auch nicht mehr genau das sein, was Aristoteles phobos und eleos oder Schiller Furcht und Mitleid genannt haben, so kennen wir zumindest etwas sehr Ähnliches heute noch: Wenn wir einen Roman lesen oder wenn wir einen Film oder ein Theaterstück ansehen, dann gibt es Momente, in denen wir uns fürchten, und andere, in denen wir Mitleid mit einer der fiktiven Figuren haben. Das sind klassische Beispiele für unser emotionales Involviertsein bei Kunstrezeption. Wir können uns auch empören über eine Figur und uns mit einer anderen freuen. In diesem Vortrag gehe ich der Frage nach, welchen Status solche Emotionen haben, verglichen mit denen, die wir im Leben sonst erfahren. Etwa verglichen mit der Furcht, die man als Elternteil um sein Kind hat, wenn es allein auf Reisen geht; oder verglichen mit dem Mitleid, das der Anblick eines Obdachlosen auf der Straße im Berliner Winter hervorruft.

Nicht selten wird die Ansicht vertreten, dass es sich bei diesen ästhetischen Emotionen um keine echten oder realen Emotionen handeln würde. Diese Ansicht kritisiere ich, ich bin der Meinung, man kann und sollte hier sehr wohl von echten Emotionen sprechen, die mit den sonst erlebten in wesentlicher Hinsicht vergleichbar sind. Ich stelle also zunächst diese andere Ansicht dar, insbesondere wie sie von Kendall Walton vertreten wird, anschließend diskutiere ich sie mit eigenen Überlegungen und in Rückgriff auf eine Idee von Richard Moran.

1. Emotionen gegenüber Fiktion als Quasi-Emotionen

Zunächst die andere Ansicht: Was wir Kunst, genauer, narrativer Fiktion gegenüber entwickeln, das sind überhaupt keine Emotionen. Denn das, worauf wir uns mit den Emotionen beziehen, existiert in Wirklichkeit gar nicht. Es gibt King Kong nicht wirklich, also gibt es im Kino nichts zu fürchten. Echte Diebe, die einen Reisenden ausrauben können, gibt es hingegen in dieser Welt zur Genüge. Anna Karenina hat sich nicht wirklich umgebracht, weil sie nie wirklich gelebt hat, und deshalb gibt es auch keinen Grund für Mitleid mit ihr. Der Obdachlose in der eisigen Kälte friert sehr real und mein Mitleid, das mich dazu veranlasst, ihm etwas Geld zu geben, hat mit dieser Realität zu tun. Es scheint ein großer Unterschied zu sein, ob es etwas nur als erfundene Geschichte oder in Wirklichkeit gibt. Dass wir vor diesem Hintergrund dennoch auch bei Fiktion häufig mit

Emotionen reagieren, wird deshalb oft als „Paradox der Fiktion“ diskutiert[1], das man folgendermaßen darstellen kann:

i) Ohne bestimmte Überzeugungen kann es bestimmte Emotionen nicht geben und zu den bestimmten Überzeugungen gehört die der Existenz der Person oder der Situation, auf die wir uns mit einer Emotion beziehen.
ii) Wir sind nicht von der Existenz fiktionaler Figuren und der erzählten Situationen überzeugt.
iii) Wir haben (aber) offensichtlich Gefühlserlebnisse gegenüber Fiktion.

Ein besonders einflussreicher Umgang mit diesem Paradox bestreitet die Existenz dieser Emotionen auf raffinierte Weise.[2] Er leugnet nicht, dass wir irgendwie emotional involviert sind. Doch, so ist die Idee, dabei handelt es sich nicht um *reale* Emotionen. Stattdessen, so der Ausdruck, der vor allem von Kendall Walton propagiert wird, erleben wir in solchen Fällen vorgestellte, „fiktionale“ oder „Quasi“-Emotionen.[3] Genauso, wie wir uns vorstellen, diese Anna Karenina, von der wir im Roman lesen, existiere, so stellen wir uns in solchen Fällen vor, was wir ihr gegenüber fühlen. Die Quasi-Emotionen sind Teil des Spiels, so Walton, auf das wir uns bei jeder aufmerksamen Rezeption von Fiktion einlassen. Wir haben ja nur dann Gefühle angesichts eines Films oder eines Romans, wenn wir uns davon in gewisser Weise „gefangen“ nehmen lassen. Sich derart zu involvieren heiße, sich als Teil der fiktionalen Welt zu sehen. Realisiert man in dieser fiktionalen Welt nun etwas Furchterregendes wie ein Schleim-Monster, was Waltons Beispiel ist, dann hat man die fiktionale Überzeugung, es gäbe dieses Monster. Auf dieser fiktionalen Überzeugung, oder dieser Vorstellung, kann nun ein Gefühl basieren – aber nur ein fiktionales. Insofern könne Charles, der Horrorfilmzuschauer, das Schleim-Monster nur fiktional fürchten. Das gelte auch, wenn er seinen Freunden nachher erzählt, er habe sich „wirklich gefürchtet“. Damit beschreibe er nur die Intensität. Die direkte Rede vom Fürchten ist nur eine verkürzte Redensart davon, dass man sich eigentlich nur fiktional gefürchtet hat, so Walton. Dafür spreche auch die Antwort, die Charles auf eine andere Frage geben wird. Fragt man ihn am nächsten Tag, ob er in letzter Zeit einen starken Furchtmoment hatte, wird er nicht, oder nur ironisch, auf das Filmerlebnis verweisen. Er weiß, dass man eigentlich nach etwas anderem fragt, nach realer oder „genuiner“ Furcht, die sich auf reale Gegenstände richtet. Der Zustand könne sich sehr ähneln: Schweißausbruch, Herzrasen, Adrenalinschub.

1 Anstoß zur Debatte um dieses Paradox gab insbesondere: Radford, Colin: How can we be moved by the fate of Anna Karenina? In: Proceedings of the Aristotelian Society 1975 (Supplement 49), S.67-79.

2 Auf andere Umgangsweisen mit dem Paradox der Fiktionalität gehe ich ein in: Weber-Guskar, Eva: Die Klarheit der Gefühle. Berlin/New York 2009. S.199 ff.

3 Im Zusammenhang seiner Mimesis-Theorie in: Walton, Kendall: Mimesis as Make-Believe. Cambridge 1990. Unterkapitel „Fearing Fictions”, S.195-204 und „Fearing fictionally”, S.241-249.

Aber zwei Dinge mindestens sind anders, nach Walton: Bei Quasi-Emotionen handelt man nicht der vermeintlichen Emotion entsprechend. Würde sich Charles wirklich fürchten, meint Walton, würde er aus seinem Fernsehsessel aufspringen und davonlaufen oder die Polizei rufen. Solche typischen Handlungsmotivationen aber fehlen. Und zweitens, worauf diese fehlende Handlungsmotivation schließlich zurückzuführen ist: Die Überzeugung, auf der das Gefühl basiert, ist von einer anderen Art als sonst, und so ist es der Zustand als ganzer auch, er ist nämlich nur fiktional.

2. Kritik am Ansatz der Quasi-Emotionen

Dieser Umgang mit dem Paradox erscheint mir nicht überzeugend, und zwar in zwei Hinsichten. Erstens werden in so einer Darstellung die Phänomene zu undifferenziert berücksichtigt, genau betrachtet ergibt sich das Paradox in einigen Fällen gar nicht. Zweitens kann man für die Fälle, in denen die Ausgangslage für das Paradox tatsächlich gegeben ist, das Paradox von innen her anzweifeln.

2.1 Kritik a: Undifferenzierte und unangemessene Phänomenbeschreibung

Zum ersten Punkt: Meiner Ansicht nach muss man verschiedene Formen des emotionalen Bezugs auf Fiktion viel deutlicher unterscheiden. Ich nenne drei Möglichkeiten: Zum einen können wir Gefühle derart entwickeln, als befänden wir uns selbst *in* der dargestellten Situation und würden dementsprechend wie eine der Figuren fühlen. Dann *identifizieren* wir uns mit einem der fiktionalen Charaktere. Zum anderen können wir Gefühle für eine der Figuren entwickeln, und zwar aus unserem persönlichen Charakter und Standpunkt her. In diesem Fall identifizieren wir uns mit niemand anderem, sondern *beziehen* uns nur *auf* die fiktionale Person. Die Konstellation, die Walton als sein Hauptbeispiel nimmt, ist ein Sonderfall. Hier richten sich Aktionen in einer Fiktion aus dieser „heraus“ auf den Zuschauer oder Leser zu, wie es das Schleim-Monster tut, wenn es auf die Kamera, und damit Charles, zurast. Doch das ist selten.

Die präziser gefassten Varianten lassen sich einzeln genauer betrachten und darauf prüfen, inwiefern sie als Quasi-Emotionen in Waltons Sinn anzusehen sind. Dafür ist entscheidend, dass sich der Fühlende, wie gesagt, als Teil der fiktionalen Welt sieht und sich so seine Emotionen vorstellt. Für die ersten beiden Varianten scheint mir das keine naheliegende Beschreibung. Was die erste Variante betrifft, so handelt es sich doch einfach um die Situation, Emotionen eines anderen nachzufühlen. So wie ich im Alltag nicht irgendwie direkt von den Gefühlen des anderen berührt werde, sondern durch mein lebhaftes Vorstellen der Lage des anderen sich bei mir Gefühle entwickeln, so gilt es auch gegenüber der Fiktion: Die Gefühle selbst sind originär die meinen und real, auch wenn ich sie

jemand anderem nachempfinde. Entsprechend sieht es mit Emotionen des zweiten Falls aus. Wieso sollte ich nicht real Mitleid für Anna Karenina empfinden? Es ist eine reale Emotion, denn es ist eine Reaktion meinerseits, meiner eigenen Persönlichkeit, auf etwas, von dem ich erfahre. Inwieweit die Tatsachte entscheidend ist, dass es sich um fiktive Figuren handelt, auf die wir uns in solchen Fällen als Beobachtende beziehen, darauf gehe ich unten noch einmal ein. Im Moment geht es nur darum, dass wir in diesen Fällen offensichtlich nicht derart in die fiktive Welt involviert sind, wie es für Waltons Beschreibung der Fall sein müsste.

Waltons Beschreibung passt eben nur auf sein Beispiel von Charles und dem Schleim-Monster, was ein Sonderfall ist. Hier verhält es sich anders. Denn Charles empfindet einer Figur gegenüber etwas, aber nicht als Beobachter oder Zuhörer, sondern als jemand, der in die fiktive Situation involviert ist; involviert, jedoch nicht als eine der fiktiven Figuren, sondern als individuelle, reale Person selbst. Das kann man tatsächlich für ein Paradox halten: Wie sollte Charles, der nicht an die Realität dieses Monsters (in seiner Welt) glaubt, sich davor (als Person seiner Welt) fürchten? Walton löst das Paradox wie gesagt so auf, dass er sagt, Charles glaubt nicht an das Monster in seiner realen Welt, wohl aber in einer fiktionalen Welt, und dementsprechend fürchtet er sich nicht real, sondern nur fiktional.

Doch auch diese Rede von fiktionalen Emotionen überzeugt mich nicht ganz. Ich bin skeptisch, ob dieses Phänomen, das zu dem Paradox führen soll, wirklich so auftritt. Es gibt bei einem Horrorfilm sicher nicht nur diese eine Art, sich zu fürchten. Und vielleicht gibt es sie sogar so gerade *nicht*. Als weitere Beschreibungsmöglichkeiten kommen in Betracht: Der Film, in dem Monster Menschen verfolgen und töten (King Kong), stellt eine Stimmung der Angst her, die sich auf die Zuschauer, und damit auf Charles, überträgt. Oder: Diese realistische Darstellung lässt bei Charles Gedanken daran aufkommen, dass es die Monster vielleicht wirklich geben könnte, oder ähnliche, in seiner Welt (zumindest bissige Gorillas), und vor denen fürchtet er sich („präventiv“). Wie genau man es auch beschreiben möchte, einer Tatsache muss man auf jeden Fall Rechnung tragen, die gegen Waltons Sicht der Dinge spricht: Niemand im Kino hat genau eine solche Furcht, wie er sie hätte, wenn er tatsächlich so einem Schleim-Monster gegenüberstehen würde. In dem Fall müsste man wirklich Todesangst spüren. Doch niemand in dem Kino, selbst die nicht, die geschrieben haben, werden nachher sagen, sie hätten Todesangst gehabt. Waltons Beschreibung dieses Sonderfalls führt meiner Meinung nach in die Irre. Er tut so, als könne man als Zuschauer um sein Leben fürchten. Genau das aber ist nicht der Fall. Man fürchtet nicht um sein *eigenes* reales Leben, als wäre es durch etwas in dem Film bedroht. Man fürchtet sich, wenn, dann vor dem Monster in dem Film, und zu dem gehört, dass es einem zwar Schrecken einjagen, aber nicht töten kann. Der Bezug der Emotionen ist ein anderer, was den Filmzusammenhang betrifft,

als in unserem übrigen Leben. Das will ich hier noch etwas detaillierter veranschaulichen, anhand der detailliert aufgeschlüsselten intentionalen Struktur einer Emotion.

Die von mir favorisierte Emotionstheorie ähnelt insbesondere der von Peter Goldie und Bennett Helm.[4] Danach sind Emotionen qualitativ erlebte persönliche Wertungen. Oder anders: Mit Emotionen erfassen wir erlebnishaft etwas in der Welt als für uns bedeutsam auf. Sie fallen unter den Oberbegriff der Gefühle, zu denen außerdem noch Stimmungen und Empfindungen gerechnet werden. Emotionen zeichnet aus, dass sie uns, anders als Urteile, in der Regel passiv überkommen und mit ihnen gehen Handlungsmotivationen und oft ein gewisser Ausdruck (Mimik, Gebärde, Verhalten) einher. Die spezifische Intentionalität von Emotionen ist dreigliedrig strukturiert. Sie bezieht sich erstens auf etwas, das ich allgemein *Gegenstand* nenne: z.B. die Person, auf die man wütend ist. Zweitens bezieht sie sich darauf in einer bestimmten Hinsicht: die Person, insofern sie einem absichtlich Schaden zugefügt hat, womit ein *formales Objekt* zu identifizieren ist, hier das Zufügen des Schadens. Und schließlich gibt es noch ein *Hintergrundobjekt* (oder Bedeutsamkeitsfokus), etwas, das das, was die Person getan hat, zu einem Schaden für einen macht: zum Beispiel das Fahrrad, das man dringend jeden Tag braucht, und das ein Vandale kaputt gemacht hat.

Für unseren Monsterfall bedeutet das: Begegnet Charles im Urwald einem Riesengorilla, fürchtet er sich vor diesem Gorilla (Gegenstand), insofern er ihn für gefährlich hält (formales Objekt), nämlich als eine Bedrohung seines Lebens (Fokus). Im Kino aber fürchtet Charles sich vor dem Monster (Gegenstand), einer fiktiven Figur, zwar auch insofern er in ihm eine Gefahr sieht (formales Objekt), aber nicht im Sinne einer Bedrohung für sein Leben, sondern höchstens für seine Gemütsruhe oder ähnliches – denn er jagt einen gehörigen Schrecken ein.

Und da eine Emotion über ihre Intentionalität identifiziert ist, kann man die Emotionen so außerhalb und innerhalb des Kinos unterscheiden, ohne dass man dabei auf „quasi-Titel“ zurückgreifen muss.

2.2 Kritik b: Kritik an der ersten Prämisse

Bisher habe ich also die Beschreibung der Phänomene kritisiert, die zum Paradox führen. Man kann aber auch das Paradox selbst unterhöhlen. Und damit sind wir beim zweiten Teil meiner Kritik. Diese Idee ist in einiger Hinsicht schon von Richard Moran angegangen worden.[5] Moran zeigt, dass man eine Prämisse des Paradoxes einfach bestreiten kann, so dass es sich auflöst: Emotionen haben keine Überzeugung darüber, was der Fall ist, als notwendige Bedingung. Bei-

4 Vgl. Goldie, Peter: The Emotions. Oxford 2000. Und: Helm, Bennett: Felt Evaluations. American Philosophical Quarterly 39. S.13-30.

5 Vgl. Moran, Richard: The Expression of Feeling in Imagination. In: Philosophical Review 103(1). S.75-104.

spiele dafür sind kontrafaktische Emotionen. Das sind Emotionen außerhalb der Kunstrezeption, die sich auf etwas beziehen, das nicht der Fall ist. Man denke dafür an Reue oder Bedauern darüber, dass man etwas nicht getan hat; oder an das Erschauern beim Gedanken daran, was alles zu einem vergangenen Zeitpunkt an Furchtbaren hätte geschehen können. Bei diesen Beispielen sind die zugrunde liegenden Überzeugungen auch nur Vorstellungen von Ausgedachtem, statt Überzeugungen über die aktuelle Realität, und könnten so als „fiktional" bezeichnet werden. So erscheint es unmöglich, auf diese Weise eine klare Trennlinie zwischen paradigmatischen und Sonder-Fällen von Emotionen zu ziehen. Emotionen gegenüber Fiktionen unterscheiden sich in diesem Punkt nicht von kontrafaktischen Emotionen. Rechnet man die kontrafaktischen Emotionen zu den fiktiven, irrealen, so wüsste man nicht mehr, wodurch sich die Emotionen gegenüber künstlerischer Fiktion auszeichnen sollten, was doch eigentlich erklärt werden soll. So gesehen ist es nicht hilfreich, die Überzeugung über etwas Irreales als entscheidendes Kriterium für Emotionen gegenüber Fiktion zu betonen. Zählt man die kontrafaktischen Emotionen zu den realen, dann kann man auch die auf Fiktionen gerichteten für reale halten.

Das rechtfertigt die Weise, wie ich oben leichtfertig umgegangen bin mit den anderen beiden Varianten möglicher Emotionen gegenüber Gefühle. Jetzt kann ich sagen, dass und warum es mir zunächst unproblematisch erscheint, das Nachfühlen mit einer fiktionalen Figur für vergleichbar zu halten mit dem Nachfühlen der Gefühle einer realen Person. Und Gleiches gilt für den Fall, dass ich einer fiktiven Figur eine Emotion entgegenbringe. Ob es fiktive Figuren, historische Persönlichkeiten oder Freunde sind, die mir von ihren letzten Erlebnissen in einer E-Mail schreiben, scheint mir für diesen Schritt nicht (unbedingt) relevant zu sein. Natürlich können Charakter und Intensität einer Emotion ganz verschieden sein, je nachdem, wen sie betrifft, das heißt, wie nahe mir die betreffende Person steht. Aber deshalb muss man noch nicht von einer grundsätzlich anderen Art der Emotion sprechen bzw. ihr einen „Quasi"-Titel verleihen.[6] Wir können sie wie kontrafaktische Emotionen ernst nehmen.

3. Konstruktive Erläuterung

Die Behauptung, dass Emotionen gegenüber narrativer Fiktion durchaus als echte Emotionen ernst zu nehmen sind, kann durch alltägliche Beobachtungen unterstützt werden.

6 Damit beziehe ich mich auf die Einschränkung, die Walton vornimmt. A.a.O. S.247. Auf die weiteren Unterschiede, die er zwischen echten und „Quasi"-Emotionen sieht, und inwiefern er dabei eine richtige Beobachtung hinsichtlich der Handlungsmotivation macht, gehe ich zum Schluss noch ein.

Erstens: Unser Umgang mit anderen Personen und ihren Emotionen gegenüber Kunst zeigt, dass wir sie ernst nehmen in dem Sinn, dass sie reale Emotionen der jeweiligen Person sind. Denken Sie daran: Wie reagieren Sie, wenn im Theater ihr Sitznachbar offenbar erfreut reagiert, wenn King Lear geblendet wird? (Nehmen wir an, es ist eine gute Inszenierung und keine so schlechte, die die Szenen unglaubwürdig oder lächerlich macht.) Würden Sie nicht zusammenzucken oder sich zumindest Ihren Teil denken, was für ein gefühlloser, roher Mensch das sei, wenn ihn das nicht rührt, bzw. er sogar absolut gegensätzliche Emotionen zeigt, als Sie es für angemessen halten? Oder stellen Sie sich einen anderen vor, der bei einem ironielosen rassistischen Witz auf der Bühne lacht. Sie werden ihn auch zur Verantwortung ziehen wollen für dieses sein Belustigungsgefühl. Er kann sich später nicht damit entschuldigen, er habe es so wenig wirklich lustig gefunden wie er nicht an die Realität des ganzen Theaterspiels geglaubt habe. Denn die emotionalen Reaktionen auf Fiktion finden in unserer Realität statt.[7] Sie sind ein Teil der Persönlichkeit bzw. Zeichen der Einstellungen und Dispositionen, nicht ein Teil der Fiktion.

Ein zweiter Hinweis auf die Realität (und nicht nur Vorgestelltheit) der Emotionen in Bezug auf narrative Fiktion ist auch der Umgang mit den Emotionen, wenn wir sie selbst erleben bzw. die Art des Erlebens an sich. Zum Beispiel kann man aus einem Kinofilm wirklich emotional mitgenommen oder erschöpft herauskommen. Die Emotionen fallen nicht sofort wieder von einem ab. Natürlich gibt es individuell sehr verschiedene Sensibilitäten, was das betrifft. Einer kann sofort wieder darüber reden, was ihm Lustiges auf dem Hinweg passiert ist, und ein anderer möchte erst den Film noch einmal Revue passieren lassen und sich darüber unterhalten und so seine erlebten Emotionen besser verstehen, sie ordnen oder auch nur ihnen gebührenden Raum geben. Das, was man beim Filmschauen erlebt hat, hat man wirklich so weit erlebt, dass es Spuren in einem hinterlässt. Man kann sich an die Emotionen erinnern wie an andere, in der Realität erlebte. Und während der Woche oder Wochen, die man einen Roman stückchenweise liest, trägt man auch die Emotionen mit sich herum, die man den Figuren darin entgegenbringt. Ich denke, nur wenn wir diesen Emotionen diesen Status zutrauen, können wir die Kraft und Bedeutung der Kunst verstehen, die sie für Menschen hat.

Diese Beispiele sollten also noch einmal positiv zeigen und veranschaulichen, dass wir die Emotionen, die wir bezüglich Kunst erfahren, so echt erleben, dass es irreführend wäre sie nur als vorgestellte zu bezeichnen.

Was bleibt nun aber vom Unterschied zwischen Emotionen bezüglich Fiktion und Emotionen bezüglich Realem? Wenn sie beide tatsächlich erlebt werden -

7 Für dieses Beispiel gehe ich davon aus, dass das Verhalten der beschriebenen Theaterbesucher eines aus ihren Gefühlen heraus ist, kein bewusst zynisches oder täuschendes, so dass zu unserer Reaktion darauf auch gehört, dass wir ihnen reale Gefühle unterstellen, die man kritisieren kann, und nicht nur ein Verhalten.

ist jede Emotion gegenüber fiktionalen Figuren in allem mit Emotionen gegenüber realen Personen vergleichbar?

Man könnte meinen, die Handlungsmotivationen seien andere – das ist auch eins der Charakteristika von Waltons „Quasi-Emotionen". Doch wenn man genauer hinsieht, trifft auch das strukturell nicht zu. Man muss nur die je spezifische Intentionalität berücksichtigen. Und das tun wir in der Praxis selbstverständlich. Wenn ich Mitleid mit einer realen Person habe, tu ich in der Regel, was mir möglich ist, ihr zu helfen. Wenn ich (unter sonst gleichen Umständen) Mitleid ganz ohne eine solche Handlungsmotivation habe, wird man oder ich selbst es mir nicht ganz abnehmen. Es erwartet jedoch niemand von mir, dass ich versuche, irgendwie King Lear zu helfen, indem ich auf die Bühne klettere und ihn an der Hand nehme. Dennoch glauben mir andere und ich mir selbst, dass er mir Leid tut. So ein Mitleid richtig zu verstehen heißt, es für eine reale Regung einer Person zu halten. Es heißt aber auch, dabei zu berücksichtigen, was das Objekt der Emotion ist. Der Gegenstand ist fiktional. Wer das berücksichtigt, für den ist klar, dass keine eingreifende Handlung am Platz ist – wie, um eine andere Art von Emotion und entsprechende Handlungsmotivation zu nennen, bei Reue über etwas Vergangenes auch niemand erwartet, dass ich die Unmöglichkeit versuche, Vergangenes zu verändern. Wir sind physisch von den Figuren aus Romanen, Filmen und Theaterstücken durch einen unüberwindlichen (ontologischen) Graben getrennt. Dennoch können wir uns psychologisch für sie engagieren. Das Engagement bleibt dabei selbstverständlich in den Grenzen der einen Seite des Grabens. Wir schreiten zu keiner Tat in der fiktiven Welt. Wir können aber wohl das Buch zur Seite legen und eine Weile in den Himmel sehen, wenn uns die Szenen im Roman zu nahe gehen. Das kann dann eine angemessene Handlung aus einer realen Emotion heraus sein.

Fazit

Auf Fiktion bezogene Emotionen sind also, so wollte ich zeigen, viel weniger problematisch, als meist angenommen wird. Man muss nur erstens für die verschiedenen Fälle die richtige Beschreibung anwenden und zweitens die differenzierte Intentionalität von Emotionen berücksichtigen. Dann kann man die Reichweite unseres realen emotionalen Lebens anerkennen, die uns in Verbindung mit Fiktionen setzen kann, ohne dass dabei die Emotionen selbst fiktiv würden.

Literaturverzeichnis:

Goldie, Peter: The Emotions. A Philosophical Exploration. Oxford, 2000

Helm, Bennett: “Felt Evaluations”. American Philosophical Quarterly, 39, S. 13-30

Moran, Richard: “The Expression of Feeling in Imagination”. Philosophical Review, 103(1). S. 75-104

Radford, Colin: "How can we be moved by the fate of Anna Karenina?". Proceedings of the Aristotelian Society, Supplement 49, 1975. S. 67-79

Walton, Kendall: Mimesis as Make-Believe. On the Foundations of Representational Arts. Cambridge, 1990

Weber-Guskar, Eva: Die Klarheit der Gefühle. Was es heißt, Emotionen zu verstehen. Berlin/New York, 2009

Dichter als Vordenker, Leser als Nachdenker[1]
Der kognitive Gehalt fiktionaler Werke

Wolfgang Huemer
wolfgang.huemer@unipr.it
Università degli Studi di Parma, Italien

Abstract/Zusammenfassung

Many philosophers have argued that works of fiction do not have cognitive value; poets are, as David Hume has formulated it, "liars by profession" who only pretend to formulate true descriptions. In fictional contexts, however, propositions are, according to this position, either false or do not have a truth-value at all. False propositions cannot communicate information, though; one can learn only from true propositions, or so the anti-cognitivist argument goes. In this contribution I will argue that this argument is based on a problematic conception of cognitive progress that reduces the latter to the accumulation of true propositions. I will show that literary texts can have a cognitive value *sui generis*: not by communicating information or by developing logically valid arguments that inescapably lead to a certain conclusion, but rather by describing fictional scenarios and persons rich in detail: in so doing they develop a perspective on the real world (rather than on some other fictional world) that allows the readers to draw their own conclusions. With their descriptions of fictional scenarios authors invite the readers to reflect and, in consequence, to come to new insights. In addition, the cognitive value of literature lies also in the fact that in this process the readers further develop their dialectic capacities. The argument does not intend to show, however, that the cognitive value is the only value that counts in literature.

In der Geschichte der Philosophie wurden fiktionalen Werken häufig jedweder kognitive Wert abgesprochen; Dichter seien, wie Hume es pointiert formuliert hat, „professionelle Lügner", die nur vorgeben, wahre Aussagen zu formulieren. In fiktionalen Texten enthaltene Propositionen sind, dieser Position zufolge, jedoch falsch oder haben keinen Wahrheitswert. Falschen Propositionen können aber keine Information vermitteln; lernen könne man vor von wahren Propositionen, so das Argument der Antikognitivisten. In diesem Beitrag argumentiere ich, dass dieses Argument auf einer fragwürdigen Konzeption des kognitiven Fortschrittes beruht, die diesen auf das Ansammeln von wahren Propositionen reduziert. Ich will zeigen, dass literarische Texte auf eine ihnen eigene Art Erkenntnisse vermit-

1 Aus stilistischen Gründen werde ich in diesem Aufsatz von *dem* Autor, *dem* Dichter und *dem* Leser sprechen. Dies ist in zweierlei Hinsicht unangebracht: Zum einen handelt es sich, grammatikalisch gesehen, um die männliche Form; Autorinnen, Dichterinnen und Leserinnen sind offensichtlich mitgemeint. Zum anderen wird diese Formulierung der Unterscheidung zwischen Literatur und Fiktion nicht gerecht. Nicht jeder literarische Text ist ein Werk der Fiktion, nicht jedes fiktionale Werk ein literarischer Text. Mit den Ausdrücken „Autor" und „Dichter" meine ich alle Urheberinnen und Urheber fiktionaler Werke (also auch Regisseure, Cartoonisten, etc.).

teln: nicht indem sie Informationen kommunizieren oder Argumentationslinien entwickeln, die zwingend zu einer bestimmten Konklusion führen, sondern indem sie fiktionale Szenarien und Personen detailliert beschreiben und so eine Perspektive entwickeln, die es den Leserinnen und Lesern überlässt, eigene Schlussfolgerungen zu ziehen. Autorinnen und Autoren zeichnen gewisse Linien *vor*, über die Leserinnen und Leser nach-denken, um so, in der eigenen Reflexion, zu neuen Erkenntnissen zu gelangen. Der kognitive Wert der Fiktion liegt aber auch darin, dass sie in diesem Prozess ihre kognitiven Fähigkeiten weiter entwickeln und schärfen können – was aber nicht implizieren soll, dass der kognitive der einzige Wert ist, um derentwillen Literatur und andere fiktionale Werke einen zentralen Stellenwert in unserer Gesellschaft einnehmen (sollten).

Literarische Texte und andere fiktionale Werke nehmen in unserer Gesellschaft einen nicht unerheblichen Stellenwert ein: es vergeht für die meisten von uns wohl kaum ein Tag, an dem wir uns nicht in irgendeiner Form mit Fiktionen beschäftigen, sei es bei der Lektüre eines Romans oder einer Kurzgeschichte, beim Erwägen eines hypothetischen Szenarios oder eines Gedankenexperiments, oder bei der Rezeption eines Films, einer Seifenoper, oder auch nur eines Cartoons in der Tageszeitung. Außerdem gilt es als weithin anerkannt, dass zumindest manche fiktionale Werke – man denke etwa an die „hohe“ Literatur, die Oper oder den Autorenfilm – einen wesentlichen Beitrag zur Bildung einer Person darstellen; ihre Rezeption gilt demnach nicht (nur) als Vergnügen oder entspannender Zeitvertreib, sondern als angesehene Tätigkeit, die (auch) dazu dient, den eigenen geistigen Horizont zu erweitern – nicht zuletzt deshalb verlangen wir von Schulkindern, auch fiktionale Werke zu studieren. Es scheint also unproblematisch, zumindest manchen literarischen Texten (und anderen fiktionalen Werken) kognitiven Wert zuzuschreiben.

1. Der literarische Antikognitivismus

Gerade diese Auffassung wurde von einer Reihe von Philosophen in Zweifel gezogen, was wohl auch daran liegt, dass fiktionale beziehungsweise literarische Texte eine andere Stoßrichtung verfolgen als nichtfiktionale, also etwa philosophische, journalistische oder wissenschaftliche Texte. Beide Textsorten sind einander zwar auf einer formalen Ebene insofern ähnlich, als sie Beschreibungen von Personen und Ereignissen enthalten. Diese Ähnlichkeit besteht aber nur auf der Oberfläche; bei Nähe betrachtet sieht man, dass die Beschreibungen jeweils eine andere Rolle spielen. Aussagen nichtfiktionaler Texte erheben den Anspruch, die Wirklichkeit so zu beschreiben, wie sie ist; sie sind der Wahrheit verpflichtet. Fiktionale Texte hingegen enthalten Beschreibungen von Personen, die nie gelebt, und von Ereignissen, die nie stattgefunden haben; schreibend kreieren Dichter ihre fiktionalen Welten; was die aktuale Welt betrifft, erheben sie

keinerlei Wahrheitsanspruch. Für Platon sind Dichter deshalb – genauso wie Maler – „Gaukelkünstler und Nachahmer“[2], die die Menschen täuschen wollen und deshalb aus der idealen Stadt verbannt werden sollten. Platon ist mit seinem Urteil nicht alleine, er vertritt eine lange Reihe von Philosophen, die über die Jahrhunderte hinweg an der Auffassung festgehalten haben, dass Dichter eigentlich „professionelle Lügner”[3] und dass Aussagen in literarischen Text allesamt falsch seien[4] oder gar keinen Wahrheitswert hätten[5] – dass es sich also, im strengen Sinne, nicht einmal um Behauptungen beziehungsweise Propositionen handle.

Diese Auffassung von Literatur hat zur Position des literarischen, oder allgemeiner, des ästhetischen Antikognitivismus geführt, der fiktionalen Texten jedweden kognitiven Wert abspricht. Anders als beim ethischen Antikognitivismus steht hier nicht die Frage, ob Aussagen in literarischen Texten einen Wahrheitswert haben (beziehungsweise, ob sie wahr sein können), im Mittelpunkt. Es geht vielmehr darum, ob wir von literarischen Texten relevante Erkenntnisse gewinnen oder, in anderen Worten, ob wir durch die Lektüre etwas lernen können. Da fiktionale Texte keine wahren Aussagen enthalten, so das Argument, könnten wir auch nichts von ihnen lernen, denn lernen bestehe im Aufnehmen von wahren Propositionen. Natürlich könne man von der Lektüre eines Textes neue Erkenntnisse gewinnen; so lernt der Leser von Kafkas *Der Prozeß* etwa, dass die Hauptfigur des Romans Josef K. heißt, dass das Buch cirka 200 Seiten lang ist, etc. Diese Erkenntnisse betreffen aber lediglich Fakten über den Text, und nicht solche, die durch den Text kommuniziert werden; sie gehören demnach kaum zu den relevanten Erkenntnissen, die Kognitivisten vor Augen haben, wenn sie den Erkenntniswert der Literatur verteidigen.

Manche Antikognitivisten gestehen ein, dass wir bei der Lektüre eines fiktionalen Textes gelegentlich Schlüsse auf allgemeine Wahrheiten ziehen können; so könnte etwa ein Leser von Shakespeares *Othello* zu dem Schluss kommen, *übertriebene Eifersucht könne zu irrationalen Reaktionen führen, die man später bereut*. Bei diesen Wahrheiten handle es sich allerdings um Erkenntnisse, die so

2 Platon, *Der Staat*. Übersetzt von Otto Apelt, Leipzig: Meiner, 1993, S. 393 (598[d]).

3 So etwa David Hume: *A Treatise on Human Nature*. Oxford: Clarendon Press, 1978, S. 121: „Poets themselves, tho’ liars by profession, always edeavour to give an air of truth to their fictions, and where that is totally neglected, their performances, however ingenious, will never be able to afford much pleasure“.

4 So stellt zum Beispiel Bertrand Russell fest: „The propositions in the play [*Hamlet*] are false because there was no such man“. Vgl. Russell: *An Enquiry into Meaning and Truth*. London: Allen and Unwin, 1962, S. 277.

5 So etwa Sir Philip Sidney: „Now for the poet, he never affirmeth, and therefore never lieth: for, as I take it, to lie is to affirm that to be true“. Vgl. Sidney: *A Defence of Poetry*. Oxford: Oxford University Press, 1973, S. 52).

allgemein seien, dass sie trivial werden[6]; zudem handle es sich um Alltagsweisheiten, die die Leser wohl schon kannten, bevor sie das Buch zur Hand nahmen, was die zeitraubende Lektüre des Textes (zumindest, wenn es einem um Erkenntnisgewinn geht) als überflüssig erscheinen lässt.[7]

Wenn Antikognitivisten der Literatur jedweden kognitiven Wert absprechen, so heißt das natürlich nicht, dass sie deswegen die Literatur ablehnen; viele weisen auf die ästhetischen Qualitäten literarischer Werke hin, auf ihren Unterhaltungswert, oder auf ihre Fähigkeit, ein Lebensgefühl zum Ausdruck zu bringen. Diese Haltung birgt meines Erachtens aber die Gefahr, dass Literatur so zu einem Ornament wird, zu einem Zeitvertreib, der unterhaltsam sein mag, aber letztlich unnütz ist. Das erscheint auch deshalb bedenklich, weil man so kaum erklären kann, warum literarische Texte (und andere fiktionale Werke) eine so zentrale Stellung in unser aller Leben einnehmen.

2. Der literarische Kognitivismus

Nicht nur der literarische Antikognitivismus kann auf eine Tradition bedeutender Philosophen verweisen; auch sein Gegenspieler, der Kognitivismus, blickt auf eine lange Geschichte zurück. Vertreter dieser Position betonen nicht nur, dass fiktionale Werke – oder allgemeiner: Kunstwerke – kognitiven Gehalt haben können, sondern behaupten für gewöhnlich auch, dass dieser kognitive Gehalt wesentlich zum ästhetischen Wert des Werkes beiträgt, denn die Kunst habe eine ganz besondere, nur ihr eigene Art, Erkenntnisse zu vermitteln.

Für gewöhnlich gestehen Kognitivisten gerne ein, dass literarische Werke nicht auf dieselbe Weise Wissen vermitteln wie journalistische Texte, Sachbücher oder wissenschaftliche Abhandlungen; sie enthalten keine wahren Propositionen über reale Personen oder Ereignisse, die wir eins zu eins übernehmen könnten, noch Argumente, die eine Konklusion schlüssig beweisen würden – wenn sie überhaupt Argumente enthalten, dann zumeist solche, die von einer fiktiven Figur (oder dem Erzähler) vorgetragen werden, was bedeutet, dass der Autor die Wahrheit der Prämissen nicht behauptet.

Wir könnten aber, so wird typischerweise argumentiert, Wissen erwerben, das aus dem Text hervorgeht, das also im Aufnehmen von Propositionen be-

6 Vgl. Jerome Stolnitz: „On the Cognitive Triviality of Art“, in: *British Journal of Aesthetics* 32 (1992), S. 191–200.

7 Eine weitere paradoxale Konsequenz dieser Position ist, dass wir fiktionalen Figuren gegenüber keine Emotionen empfinden können, da wir dafür von der Existenz dieser Figuren überzeugt sein müssten – während wir im Gegenteil aber von deren Nichtexistenz wissen. Für eine frühe und einflussreiche Formulierung dieses so-genannten *Paradox of Fiction*, siehe Colin Radford: „Can We Be Moved by the Fate of Anna Karenina?“, in: *Aristotelian Society Supplementary Volume* 49 (1975), S. 67–80.

steht, die nicht direkt im Text enthalten sind, aber doch, in der einen oder anderen Form, vom Text nahe gelegt werden: so weist schon Aristoteles darauf hin, dass „die Dichtkunst etwas ernsthafteres und philosophischeres"[8] sei als die Geschichtsschreibung, da letztere sich auf das Faktische, das, was wirklich gesehen sei, beschränken müsse, während Dichter das „nach den Regeln der Wahrscheinlichkeit oder Notwendigkeit Mögliche"[9] beschreiben und damit in der Darstellung konkreter Ereignisse das Allgemeine zeigen. Neuere Ansätze argumentieren, dass fiktionale Texte für gewöhnlich eine „Aussage" enthalten, die zumeist nicht explizit im Text enthalten ist, aber in einem *thematic statement* ausgedrückt werden könne, und lokalisiert den kognitiven Gehalt des Textes in dieser Aussage[10]; Andere suggerieren, dass Autoren in uns durch die Schilderung eines konkreten Einzelfalles eine Art des Verständnisses erzeugen können, das über das reine Faktenwissen hinaus geht. Wenn jemand etwa Zeuge eines Unfalls werde, so gewinnt er zwar neue Informationen über die Welt – er lernt etwa, dass sich eben hier ein Unfall ereignet hat und dass zwei Personen schwer verletzt worden sind –, wenn er aber nicht reagiert und Hilfe leistet, so würden wir sagen, er habe etwas wichtiges nicht *verstanden*; wobei es sich hier um eine Art des Verständnisses handelt, die nicht auf der Ebene des propositionalen Wissens, sondern vielmehr auf der des Anerkennens anzusiedeln ist.[11] Und gerade auf dieser Ebene, so wurde argumentiert, kann Literatur einen wichtigen Beitrag leisten.[12]

3. Die gemeinsame Grundlage: Das Sender / Empfänger Modell

Die Debatte zwischen Kognitivismus und Antikognitivismus besteht also in der Frage, ob, und wenn ja, wie es möglich ist, propositionales Wissen (oder Formen nichtpropositonalen Wissens wie Anerkennen) aus literarischen Texten zu erwerben (Kognitivismus) oder aus ihnen zu destillieren (Antikognitivismus). Das zeigt aber, dass beide Positionen mit einer Vorstellung von kognitivem Gehalt arbeiten, der sich sehr stark an der Kommunikation von wahren Propositio-

8 Aristoteles, *Poetik*. Übers. und hg. von Manfred Fuhrmann, Stuttgart: Reclam, 1982, S. 29 (1451^b).

9 Ebenda.

10 Vgl. Peter Lamarque: „Cognitive Values in the Arts: Marking the Boundaries", in: Matthew Kieran (Hrsg.), *Contemporary Debates in Aesthetics and the Philosohy of Art*, Oxford: Blackwell, 2006, 127–139. Lamarque argumentiert allerdings, dass der kognitive Wert eines literarischen Textes völlig unabhängig von dessen ästhetischem Wert sei.

11 Für die Unterscheidung von Wissen und Anerkennen, vgl. Stanley Cavell: „Knowing and Acknowledging", in: ders., *Must We Mean What We Say? A Book of Essays*, Charles Scribner's Sons, New York, 1969, S. 238–266.

12 Vgl. z.B. John Gibson: *Fiction and the Weave of Life*, Oxford: Oxford University Press, 2007.

nen orientiert – der Unterschied besteht lediglich darin, dass Antikognitivisten eine direkte Vermittlung von Information vor Augen haben, während Kognitivisten argumentieren, dass die relevanten Propositionen auch indirekt durch den Text vom Autor an den Leser vermittelt werden können. Hinter dieser Debatte steht also, wie mir scheint, eine geteilte Auffassung des Verhältnisses von Autor und Leser, die von dem Sender / Empfänger Modell geprägt zu sein scheint: der Autor „sendet" bzw. kommuniziert (direkt, oder indirekt) wahre Propositionen, die vom Empfänger aufgenommen (oder: zuerst herausdestilliert und dann aufgenommen) werden. In diesem Bild wird dem Autor eine sehr große Autorität zugestanden: er kann unseren kognitiven Fortschritt steuern indem er den Inhalt der von uns bei der Lektüre des Romanes zu lernenden Propositionen festlegt.[13] Wenn der Roman kognitiv gehaltvoll ist und der Leser ihn richtig zu lesen weiß, dann kann der Autor so festlegen, was der Leser lernt. Der Beitrag des Lesers besteht lediglich darin, dies zuzulassen oder abzublocken. Der Leser ist gleichsam eine *tabula rasa*, in die der Autor seine Propositionen einschreiben kann; dieser zeichnet gewisse Muster vor, die der Leser nachzeichnet, er denkt vor, was der Leser *nach-denkt*.

Dieses Bild mag für die Lektüre eines Zeitungsartikels oder eines Lehrbuches zutreffend sein: wenn der Leser die Autorität des Autors anerkennt, so wird er die angebotenen Informationen – gegebenenfalls nach Konsultation einer zweiten Meinung – übernehmen. Es wird aber nicht dem komplexen Zwischenspiel zwischen Autor (oder allgemeiner: Künstler) und Rezipienten gerecht, das einen zentralen Stellenwert bei unserem Umgang mit Kunstwerken zu haben scheint.

4. Vorgedachtes nachdenken versus über etwas nachdenken

Fiktionale Texte entwickeln, wie wir oben gesehen haben, Szenarien, in denen detailgetreu beschriebene Menschen (die nie wirklich gelebt haben) auf bestimmte Ereignisse (die nie stattgefunden haben) mit bestimmten Handlungen (die nie wirklich ausgeführt worden sind) reagieren. In seiner Fiktion muss jeder Autor aber auch an Aspekten der realen Welt anknüpfen; er übernimmt vertraute Elemente der Wirklichkeit, die der Leser wiedererkennen kann – ein Punkt, den

13 In dieser Konzeption erstreckt sich die Autorität des Autors bis hin zur Unfehlbarkeit: wenn Dichter keinen Wahrheitsanspruch erheben, sondern die beschriebenen Personen und Ereignisse frei erfinden, wie diese Position voraussetzt, so können sie dabei keine Fehler machen, denn man kann eine Person oder ein Ereignis nicht falsch erfinden; selbst eventuelle Ungereimtheiten müssten als Besonderheiten der vom Autor beschriebenen fiktionalen Welt akzeptiert werden. Für eine ausführlichere Diskussion über Fehler im fiktionalen Kontext vgl. mein „Gibt es Fehler im fiktionalen Kontext? Grenzen der dichterischen Freiheit", in: Otto Neumaier (Hrsg.), *Was aus Fehlern zu lernen ist – in Alltag, Wissenschaft und Kunst*. Wien–Münster: LIT Verlag, 2010 (im Druck).

auch Wolfgang Iser in seiner Definition von Fiktion betont.[14] In der Kombination dieser Aspekte, aber auch in den Handlungen der Protagonisten zeigt sich eine individuelle Perspektive auf die Welt; im Verlauf des Textes die Konsequenzen der von den Protagonisten getätigten – oder unterlassenen – Handlungen.

Der Autor bietet also dem Leser eine Perspektive auf die Welt an – auf die Welt der Protagonisten, durch die sich indirekt letztlich aber auch die des Autors zeigt; er zeigt dem Leser, wie es ist, in einer bestimmten Situation zu sein, wie sich die Welt (zum Beispiel) für einen heranwachsenden Teenager, einen Kriegsheimkehrer oder eine allein erziehende Mutter darstellt. Dies wird freilich (typischerweise) nicht direkt ausgesprochen oder argumentiert, sondern zeigt sich in den vom Autor erzählten Ereignissen.

Die Tatsache, dass er sich dabei nicht auf harte Fakten oder auf eine statistisch relevante Anzahl von Einzelfällen beruft, sondern sich auf nur einen einzigen Einzelfall beschränkt, hat zur Folge, dass er diesen mit einer Fülle von scheinbar unnötigen Details beschreiben kann, die dazu dienen, die jeweilige Perspektive noch deutlicher hervortreten zu lassen. Es zeigt aber auch, dass Dichter einen anderen Anspruch hat als Wissenschaftler: es geht eben gerade nicht darum, eine stringente, logisch gülitge Argumentation zu entwickeln, die – die Plausibilität der Prämissen vorausgesetzt – zwingend zu einer bestimmten Konklusion führt. Das Ziel ist vielmehr, die Leser dazu einzuladen, ihre Blickrichtung auf gewisse Situationen oder Ereignisse gegebenenfalls zu ändern und dann die eigenen Schlussfolgerungen zu ziehen – Schlussfolgerungen, die die Leser ziehen *können*, aber nicht *müssen*.

Der Autor macht den Lesern also ein Angebot, in eine gewisse Richtung weiter zu denken, sich eine Meinung über Situationen zu bilden, die sie vielleicht so noch nicht erlebt haben; oder die Welt aus einer Perspektive zu sehen, die sie so nicht kennen – die aber das Verhalten, die Probleme oder Haltungen bestimmter Personen oder Personengruppen nachvollziehbar macht. Der Autor zeichnet also gewisse Linien vor, und lädt dazu ein, *über* bestimmte Themen beziehungsweise Thesen *nach*zudenken. Der Leser wird implizit dazu aufgefordert, sich nicht mit der Rolle eines passiven Empfängers zu begnügen, der einfach *nach-denkt*, was andere *vor-denken*, sondern aktiv über gewisse Themen nachzudenken, zu reflektieren.

Das impliziert aber zweierlei: zum einen zeigt es, dass literarische Texte auch einen Bezug zur realen Welt haben und Aspekte dieser realen Welt thematisieren – Aspekte, die der Leser ohne den jeweiligen Text vielleicht nicht gesehen oder ausreichend gewürdigt hätte. Zum anderen wird aber durch die spezifisch literarische Form – die Schilderung eines Einzelfalles, das Fehlen eines Argu-

14 Vgl. Wolfgang Iser, *Das Fiktive und das Imaginäre: Perspektiven literarischer Anthropologie*. Suhrkamp, Frankfurt/M. 1991.

mentes – deutlich, dass der kognitiven Gehalt von literarischen Texten auf ganz besondere Weise vermittelt wird: der Autor diktiert uns nicht seine Einsichten, noch gelangt er zu seinen Konklusionen auf einem logisch zwingenden Argumentationsgang; vielmehr lädt er uns zur Reflexion ein; er stellt zur Debatte und zeigt eventuell auch gewisse Lösungsstrategien auf – er lässt erkennen, was er in dieser Situation für richtig oder angemessen hält –, kann aber letztlich nie wirklich bestimmen, zu welchen Schlussfolgerungen der Leser gelangt; denn der Leser übernimmt in dieser Konzeption eine aktive Rolle.

Der kognitive Wert der Literatur liegt also darin, dass der Leser, angeregt durch den Text, sich Aspekten der Welt zuwendet und sich Gedanken über Themen macht, die er ohne die Lektüre des Textes so wahrscheinlich nicht gehabt hätte. In diesem Prozess – besonders (aber nicht nur) dann, wenn er die vom Autor angebotenen Schlussfolgerungen ablehnt – schärft er auch seine diskursiven Fähigkeiten; er sieht sich gezwungen, zu einem bestimmten Thema Stellung zu beziehen, was ihm helfen kann, unbegründete Vorurteile zu überwinden oder zumindest ein bislang unbegründetes Vorurteil in eine reflektierte Überzeugung weiter zu entwickeln.

Das zeigt aber auch, dass ein Autor eine gewisse Verantwortung hat, was die Auswahl des Themas betrifft. Wenn ein Autor die Perspektive zum Beispiel eines Kriegsverbrechers oder eines Serienmörders beschreibt, aber auch, wenn er bestimmte historische Ereignisse zur Kulisse wählt, so lenkt er die Aufmerksamkeit der Leser in eine ganz bestimmte Richtung. Umgekehrt kommt natürlich auch den Lesern Verantwortung zu: es liegt an ihnen, sich nicht vom Autor verführen oder manipulieren zu lassen, sondern eigene Schlussfolgerungen zu ziehen.

Ich will mit diesen Ausführungen also nicht suggerieren, dass nur der Leser wichtig sei und dem Autor keinerlei Bedeutung zukomme; wir müssen nicht so weit gehen, wie so mancher französische Philosoph Ende der sechziger Jahre des letzten Jahrhunderts, der suggeriert, „Die Geburt des Lesers [sei] zu bezahlen mit dem Tod des Autors".[15] Ich will vielmehr vorschlagen, dass das Verhältnis von Autor und Leser nicht nach einem einfachen Sender / Empfänger Modell konzipiert werden sollte; es handelt sich vielmehr um ein sehr komplexes Verhältnis, und gerade in dieser Komplexität liegt der besondere kognitive Wert fiktionaler Werke, der ihnen eigen ist.

Abschließend will ich noch betonen, dass es mir auch nicht darum geht, fiktionale Werke auf deren kognitiven Wert reduzieren. Die Tatsache, dass literarische Texte kognitiv relevant sein *können* zeigt weder, dass es *alle* auch tatsächlich sind, noch, dass der kognitive der einzige Wert ist, dessentwillen die Lektüre eines guten Buches empfohlen werden sollte. Manchmal will man eben

15 Roland Bartes: „Der Tod des Autors", in: F. Jannidis, G. Lauer, M. Martinez und S. Winko (Hrsg.), *Texte zur Theorie der Autorschaft*, Stuttgart: Reclam, 2000, S. 185–193, hier S. 193.

nicht lernen, sondern sich entspannen, erfreuen, ablenken, oder etc. In meinem Beitrag geht es mir lediglich darum, zu zeigen, dass es falsch wäre, der Fiktion kategorisch jedweden kognitiven Wert abzusprechen, wie es in der Geschichte der Philosophie (auch in der der analytischen Philosophie) viel zu oft und viel zu leichtfertig geschehen ist.

Literaturverzeichnis

Aristoteles: Poetik. Übers. und hrsg. von Manfred Fuhrmann. Reclam, Stuttgart 1982

Barthes, Roland: „Der Tod des Autors". In: *F. Jannidis, G. Lauer, M. Martinez und S. Winko (Hrsg.):* Texte zur Theorie der Autorschaft. Reclam, Stuttgart, 2000. S. 185–193

Cavell, Stanley: „Knowing and Acknowledging". In: *Cavell, Stanley:* Must We Mean What We Say? A Book of Essays. Charles Scribner's Sons, New York, 1969. S. 238–266

Gibson, John: Fiction and the Weave of Life. Oxford University Press, Oxford, 2007

Huemer, Wolfgang: „Gibt es Fehler im fiktionalen Kontext? Grenzen der dichterischen Freiheit". In: *O. Neumaier (Hrsg.):* Was aus Fehlern zu lernen ist – in Alltag, Wissenschaft und Kunst. LIT Verlag, Wien–Münster, 2010 (im Druck)

Hume, David: A Treatise on Human Nature. Clarendon Press, Oxford, 1978

Iser, Wolfgang: Das Fiktive und das Imaginäre. Perspektiven literarischer Anthropologie. Suhrkamp, Frankfurt/M., 1991

Lamarque, Peter: „Cognitive Values in the Arts: Marking the Boundaries". In: *Matthew Kieran (Hrsg.):* Contemporary Debates in Aesthetics and the Philosohy of Art. Blackwell, Oxford, 2006. S. 127–139

Platon: Der Staat. Übersetzt von Otto Apelt. Meiner, Leipzig, 1993

Radford, Colin: „Can We Be Moved by the Fate of Anna Karenina?". Aristotelian Society. Supplementary Volume 49, 1975. S. 67–80

Russell, Bertrand: An Enquiry into Meaning and Truth. Allen and Unwin, London, 1962

Sidney, Sir Philip: A Defence of Poetry. Oxford University Press, Oxford, 1973
Stolnitz, Jerome: „On the Cognitive Triviality of Art“. British Journal of Aesthetics, 32, 1992. S. 191–200