STUDIA PHILOSOPHICA Vol. 68/2009

Formen der Irrationalität

Formes d'irrationalité

Redaktion: Anton Hügli
Rédaction: Curzio Chiesa

Schwabe

STUDIA PHILOSOPHICA

VOL. 68/2009

JAHRBUCH DER
SCHWEIZERISCHEN PHILOSOPHISCHEN GESELLSCHAFT

ANNUAIRE DE LA SOCIÉTÉ SUISSE DE PHILOSOPHIE

SCHWABE VERLAG BASEL

FORMEN DER IRRATIONALITÄT

FORMES D'IRRATIONALITÉ

REDAKTION / RÉDACTION
ANTON HÜGLI / CURZIO CHIESA

SCHWABE VERLAG BASEL

Publiziert mit Unterstützung der Schweizerischen Akademie
der Geistes- und Sozialwissenschaften

Publié avec l'aide de l'Académie suisse des sciences humaines et sociales

Gesamtherstellung: Schwabe AG, Druckerei, Muttenz/Basel
Printed in Switzerland
ISBN 978-3-7965-2596-4

www.schwabe.ch

Inhalt / Table des matières

Formen der Irrationalität
Formes d'irrationalité

Willensschwäche und Selbsttäuschung
Acrasie et mauvaise foi

Würdigung / Mémoire

Buchbesprechungen / Comptes rendus

Studia philosophica 68/2009

Formen der Irrationalität

Vorwort

Eine der großen Herausforderungen der Philosophie als Sachwalterin der Vernunft ist die Tatsache, dass Menschen sich nicht immer rational oder gar irrational zu verhalten pflegen. Zu dem von Menschen hervorgebrachten Irrationalen gehören Phänomene wie Irrtum und das Versäumnis der Irrtumsvermeidung, Selbsttäuschung, Wunschdenken und Willensschwäche. Diese Phänomene sind von unterschiedlicher Natur: Während man von Irrtum und Täuschung sagen kann, dass sie uns unterlaufen oder dass wir ihnen erliegen, und sie sich aufzulösen pflegen, sobald wir sie erkennen, scheinen Phänomene wie Willensschwäche und Selbsttäuschung nicht einmal auf konsistente Weise beschreibbar, geschweige denn erklärbar zu sein; denn was sind das für mentale Zustände, in denen einer etwas glaubt, was er nicht glaubt, oder genau das tut, was er für falsch hält? Dieses Problem verfolgt die Philosophie seit ihren Anfängen und es ist in jüngerer Zeit – nicht zuletzt wohl im Zuge der erhöhten Rationalitätsansprüche der analytischen Philosophie – erneut wieder virulent geworden. Es fordert uns einmal mehr heraus, uns klarer zu werden sowohl über die normativen Rationalitätsansprüche, die wir – berechtigterweise – erheben, wie auch über unser Vermögen, diesen Rationalitätsansprüchen nachzukommen.

Die Schweizerische Philosophische Gesellschaft hat sich von dieser – für unser menschliches Selbstverständnis grundlegenden – Thematik herausfordern lassen und sie im Jahr 2008 unter dem Titel «Formen der Irrationalität» zum Gegenstand eines an der Universität Bern durchgeführten Symposions gemacht, mit Beitragenden und eingeladenen Referenten nicht nur aus der Philosophie, sondern auch aus den von dieser Frage nicht weniger angesprochenen Disziplinen wie der Psychologie, der Ökonomie und der rationalen Entscheidungstheorie.

Der 68. Band der *Studia Philosophica* nimmt diese Thematik auf. Die Beiträge in diesem Band sind überarbeitete Fassungen von Symposionsvorträgen.

Aus Anlass des 25. Todestages von Hans Kunz hat sich Hans Saner in einem Festvortrag vor der Hans-Kunz-Gesellschaft mit dem Verhältnis zwischen Hans Kunz und Karl Jaspers auseinandergesetzt. Wir haben zur Würdigung von Hans Kunz den Vortrag von Hans Saner in diesen Band aufgenommen.

Anton Hügli und Curzio Chiesa

Studia philosophica 68/2009

Formes d'irrationalité

Préface

L'une des tâches fondamentales de la philosophie, en tant qu'activité vouée à la tutelle de la raison, concerne la délimitation du concept de rationalité. Mais la philosophie doit également essayer de comprendre les attitudes et les comportements irrationnels des êtres humains. Or l'irrationalité se présente de plusieurs manières : par exemple, l'erreur, plus ou moins volontaire, l'auto-illusion, le désir velléitaire et la faiblesse de la volonté. À la différence du non rationnel, qui se situe en dehors du champ de la rationalité, l'irrationalité est un manque ou une privation de rationalité, qui se manifeste à l'intérieur du champ du rationnel. C'est pourquoi le rationnel constitue la norme et le critère qui permettent de définir et de comprendre les tendances irrationnelles. Comment la philosophie envisage-t-elle le champ complexe de l'irrationalité ?

Pour essayer de répondre à ce genre de questions, la Société suisse de philosophie a consacré son Symposium 2008, qui s'est déroulé à l'Université de Berne, au thème des « Formes de l'irrationalité ».

Les contributions ont porté sur plusieurs aspects du problème et ont touché des domaines variés, de la psychologie à l'économie, à la théorie de la décision. Le Volume 68 des *Studia philosophica* publie une sélection des contributions qui ont été données au symposium.

A l'occasion du 25ème anniversaire de la mort de Hans Kunz, Hans Saner a exposé, dans son allocution d'ouverture devant la Hans-Kunz-Gesellschaft les rapports entre Kunz et Jaspers. A la mémoire de Hans Kunz, nous avons repris dans ce volume la conférence de Hans Saner.

Anton Hügli et Curzio Chiesa

Irrationalitäten in Theorien der Rationalität
Irrationalités dans les théories de la rationalité

Studia philosophica 68/2009

Georg Brun

Wer hat ein Problem mit irrationalen Präferenzen? Entscheidungstheorie und Überlegungsgleichgewicht*

Decision theory explicates norms of rationality for deriving preferences from preferences and beliefs. Empirical studies have found that actual preferences regularly violate these norms, launching a debate on whether this shows that subjects are prone to certain forms of irrationality or that decision theory needs to be revised. It has been claimed that such a revision is necessitated by the fact that normative uses of decision theory must be justified by a reflective equilibrium. The paper discusses three points. First, the debate over the impact of empirical studies on decision theories is only meaningful with respect to a decision theory that includes not only a formal system but also a theory of application. Second, differences in the concepts of reflective equilibrium appealed to are a source of confusion in the debate on rationality. Third, the assumption that normative uses of decision theory are justified by reflective equilibrium is not sufficient ground for arguing that the empirical studies call for a revision of decision theory. Such an argument must rely on substantive claims about rationality, preferences and beliefs.

Im Rahmen der Kontroverse, die als *rationality debate* oder *rationality wars*[1] bekannt ist, wurde unter anderem darüber gestritten, wie man die Tatsache interpretieren soll, dass es regelmäßig empirisch nachweisbare Präferenzen gibt, die der klassischen Entscheidungstheorie widersprechen (Abschnitt 2): Zeigen diese Resultate, dass die Entscheidungstheorie keine haltbare Ra-

* Für Diskussionen und Kommentare danke ich Urs Allenspach, Christoph Baumberger, Gertrude Hirsch Hadorn, Anna Kusser, Hans Rott, Neil Roughley und Peter Schaber.
Dieser Beitrag basiert auf Forschung mit Unterstützung von: Staatssekretariat für Bildung und Forschung SBF/COST: Europäische Zusammenarbeit auf dem Gebiet der wissenschaftlichen und technischen Forschung; Forschungsprojekt ClimPol des ETH-Bereichs.

1 Edward Stein: *Without good reason. The rationality debate in philosophy and cognitive science* (Oxford: Clarendon Press, 1996). Richard Samuels, Stephen Stich, Michael Bishop: *Ending the rationality wars. How to make disputes about human rationality disappear*, in *Common sense, reasoning, and rationality*, hg. von Renée Elio (Oxford: Oxford University Press, 2002) S. 236-268.

tionalitätstheorie ist? Oder zeigen sie, dass Menschen irrational sind? Die Literatur zu dieser Kontroverse ist sehr verzweigt und ich gehe nur auf zwei Aspekte ein, die meines Erachtens bisher nicht angemessen diskutiert wurden. Erstens möchte ich darauf hinweisen, dass die Debatte um die Relevanz der empirischen Präferenzenforschung für die Entscheidungstheorie nur sinnvoll geführt werden kann, wenn man über eine «Anwendungstheorie» verfügt, das heißt, eine Theorie, die regelt, wie die empirischen Resultate auf das formale System der Entscheidungstheorie zu beziehen sind (Abschnitt 3.1). Zweitens wende ich mich der klassischen Verteidigungsstrategie zu, die geltend macht, dass die Entscheidungstheorie eine normative Theorie ist und deshalb nicht durch empirische Befunde zu widerlegen ist. Ich diskutiere zwei Fragen: Wie lässt sich mit Bezug auf die Methode des Überlegungsgleichgewichts rechtfertigen, dass die Entscheidungstheorie normativ verwendet wird? Schließt eine solche Rechtfertigung Entscheidungstheorien aus, deren Rationalitätsnormen dazu führen, dass sich empirisch regelmäßig irrationale Präferenzen nachweisen lassen? (Abschnitt 3.2) In Abschnitt 4.1 wird zunächst das auf Goodman und Elgin zurückgehende Konzept des Überlegungsgleichgewichts eingeführt, das, wie mir scheint, am ehesten als epistemische Rechtfertigungsmethode für normativ verwendete Theorien zu verteidigen ist. Auf dieser Grundlage werde ich dann argumentieren, dass die Methode des Überlegungsgleichgewichts sowohl ausschließt, dass unsere Urteile über die Rationalität von Präferenzen generell revidiert werden müssen, als auch, dass unsere Präferenzen automatisch als rational gelten. Um für oder gegen die These zu argumentieren, dass regelmäßig empirisch nachweisbare irrationale Präferenzen gegen die Entscheidungstheorie sprechen, reicht es nicht aus, sich auf die Methode des Überlegungsgleichgewichts zu berufen. Dafür sind substanzielle Thesen über Rationalität, Präferenzen und Überzeugungen erforderlich (Abschnitt 4.2). Vorher muss aber kurz erklärt werden, in welcher Weise sich die klassische Entscheidungstheorie auf Präferenzen bezieht und in welchem Sinne sie diese als rational oder irrational beurteilt (Abschnitt 1).

1. (Ir)rationale Präferenzen in der Entscheidungstheorie

In der These, dass Präferenzen manchmal aufgrund entscheidungstheoretischer Überlegungen als irrational gelten müssen, geht es um eine ziemlich spezifische Form der Irrationalität. Ich charakterisiere sie deshalb zuerst etwas genauer und unterscheide sie von anderen Formen, die der Vorwurf der Ir-

rationalität oder Unangemessenheit von Präferenzen annehmen kann. Es geht mir dabei nicht darum, eine bestimmte Version der Entscheidungstheorie zu verteidigen, sondern um die methodologische Frage, unter welchen Bedingungen eine Entscheidungstheorie Präferenzen als irrational kritisieren kann.

Den Begriff der Präferenz verstehe ich im Folgenden so, wie er in der klassischen Entscheidungstheorie von grundlegender Bedeutung ist: ein Subjekt S zieht von zwei sich ausschließenden Optionen a und b die eine der anderen vor. Was unter einer «Option» zu verstehen ist, kann im Moment offen gelassen werden (ich komme in Abschnitt 3.1 darauf zurück); mögliche Interpretationen sind zum Beispiel Handlungsweisen, Gegenstände, Sachverhalte oder mögliche Weltzustände. Solange man sich auf die Präferenzen eines Individuums beschränkt, ist es üblich, die Referenz auf das präferierende Subjekt wegzulassen und einfach zu sagen: «a wird gegenüber b präferiert» (symbolisch: aPb). Dazu kommt noch die Möglichkeit, gegenüber zwei Optionen indifferent zu sein (aIb) und die Beziehung der schwachen Präferenz ($aRb =_{df} aPb \vee Ib$).[2]

Präferenzen können in verschiedener Hinsicht als irrational oder sonst wie unangemessen gelten. Aus entscheidungstheoretischer Perspektive geht es um die Frage, wie sich die verschiedenen Präferenzen eines Subjekts zueinander verhalten, ob sie «zusammenpassen». Etwas genauer gesagt, befasst sich die Entscheidungstheorie mit der Frage, wie aus gegebenen Präferenzen und Überzeugungen weitere Präferenzen abzuleiten sind. Es geht also weder darum, welche Präferenzen ein Subjekt überhaupt, das heißt, unabhängig von seinen anderen Präferenzen und Überzeugungen, haben sollte, noch um die Frage, wie ein Subjekt zu seinen Präferenzen kommt, und auch nicht um die Frage, wie man die Präferenzen konkreter Subjekte empirisch untersucht (im Folgenden: «Präferenzenforschung»). Insofern die Entscheidungstheorie sich nicht nur mit Präferenzen befasst, sondern auch mit Überzeugungen, schließt sie eine Theorie der theoretischen Rationalität ein, beziehungsweise setzt sie voraus.

Wenn ich von «Entscheidungstheorie» spreche, beziehe ich mich in erster Linie auf die klassische Theorie des erwarteten subjektiven Nutzens. Die übliche Terminologie verschleiert, dass die eben gegebene Charakterisierung auch auf diese Theorie passt. Das hängt einerseits damit zusammen, dass diese Theorie davon ausgeht, dass Überzeugungen Grade haben und

[2] Meist wird so definiert: $aIb =_{df} aRb \wedge bRa$ und $aPb =_{df} aRb \wedge \neg(bRa)$.

Präferenzen über subjektiven Nutzen modelliert werden und andererseits damit, dass Entscheidungen oft direkt auf Handlungen, nicht nur auf handlungsleitende Präferenzen bezogen werden.

Die These, dass gewisse Präferenzen aus entscheidungstheoretischer Sicht irrational sind, muss zunächst in zwei Richtungen abgegrenzt werden. Erstens ist entscheidungstheoretische Irrationalität nur eine von vielen Hinsichten, in denen Präferenzen als irrational oder anderswie unangemessen kritisiert werden können. Bei entscheidungstheoretischer Irrationalität von Präferenzen geht es um deren Verhältnis zu anderen Präferenzen und zu Überzeugungen. Andere Formen der Kritik an Präferenzen richten sich, um nur zwei Beispiele zu nennen, darauf, dass sie nicht zu höherstufigen Präferenzen passen oder dass verschiedene Aspekte derselben Präferenz, ihre bewusste Einschätzung, ihre motivationale Kraft und ihr Befriedigungspotenzial, nicht zusammenpassen.[3]

Zweitens setzt die Rede von irrationalen Präferenzen voraus, dass sie auch rational sein können und nicht arational sind. Was irrational ist, genügt den Normen der Rationalität nicht; was arational ist, ist außerhalb des Anwendungsbereichs von Rationalitätsnormen. Dass Präferenzen irrational sein können, ist deshalb erläuterungsbedürftig, weil, oft mit Bezug auf Hume,[4] die Auffassung vertreten wird, dass Präferenzen sich nicht sinnvoll als rational oder irrational klassifizieren lassen: Rationalität beziehe sich immer nur auf die Wahl der Mittel, aber niemals auf die Ziele, und Präferenzen seien nichts anderes als relative Ziele. Dies kann aber nicht bedeuten, dass alle Präferenzen arational sind. Erstens ist es gerade die Pointe des instrumentellen Verständnisses der Rationalität, dass es rational oder irrational sein kann, bestimmte Mittel zur Realisation gegebener Ziele zu präferieren. Und zweitens können auch aus instrumentalistischer Sicht nicht alle Präferenzen als arational gelten. Vielmehr sind viele Ziele selbst instrumentell (jemand möchte einen Marathon absolvieren, um fit zu bleiben).

Welche Präferenzen als rational gelten, bemisst sich aus entscheidungstheoretischer Perspektive daran, ob sie bestimmte formale Bedingungen erfüllen. Es ist umstritten, welches diese Rationalitätsbedingungen im Ein-

3 Harry G. Frankfurt: *Freedom of the will and the concept of a person*, in *Journal of philosophy* 68 (1971) S. 5-20; Anna Kusser: *Dimensionen der Kritik von Wünschen* (Frankfurt a.M.: Athenäum, 1989).

4 David Hume: *A Treatise of Human Nature* (Oxford: Oxford University Press, 1978) II.iii.3, S. 415.

zelnen sind und wie sie genau formuliert werden sollen.[5] Eine erste Klasse betrifft die Struktur der Präferenzen einer Person zu einer bestimmten Zeit. Prominente Anforderungen dieser Art sind Vollständigkeit und Transitivität. Vollständigkeit meint, dass in einer Menge von Optionen zwischen zwei beliebigen Optionen immer eine Präferenz-Beziehung (im Sinne von R) besteht. Eine strenge Variante der Transitivitätsforderung ist Transitivität von R (für alle Optionen a, b, c gilt $aRb \wedge bRc \rightarrow aRc$), eine schwächere ist Azyklität: es gibt keine Folge von Optionen $a_1, \ldots, a_n$, so dass $a_1P\ldots Pa_nPa_1$. Weitere Rationalitätsbedingungen betreffen das Wählen von Optionen aufgrund von Präferenzen und somit das Bilden von Präferenzen aufgrund anderer Präferenzen und Überzeugungen. Eine grundlegende Anforderung ist, dass nur Optionen wählbar sind, zu denen es keine präferierte Alternative gibt. Weiter geht die als *property alpha* bekannte Anforderung, dass eine wählbare Option wählbar bleibt, wenn die Menge der Optionen so verkleinert wird, dass diese Option noch zur Verfügung steht.

Es gibt verschiedene Argumentationslinien zur Verteidigung solcher Forderungen. Zum einen können konzeptionelle Argumente ins Feld geführt werden. Es ist kaum zu bezweifeln, dass aPa aus begrifflichen Gründen ausgeschlossen ist. Davidson argumentiert beispielsweise auch, dass Transitivität eine Voraussetzung dafür ist, jemandem überhaupt Präferenzen zuzuschreiben.[6] Eine solche Auffassung ist im vorliegenden Zusammenhang jedoch insofern uninteressant, als sie ausschließt, dass die Präferenzenforschung zeigen könnte, dass es Subjekte mit irrationalen Präferenzen gibt. Angebliche Beispiele von irrationalen Präferenzen wären schlicht keine Beispiele von Präferenzen. Eine andere Strategie sind Argumentationen, die sich darauf stützen, dass intransitive Präferenzen ermöglichen, eine sogenannte «Geldpumpe» *zu* konstruieren, indem man einer Person jeweils gegen einen minimalen Geldbetrag den Tausch einer Option gegen eine andere anbietet, die diese Person präferiert.[7] Setzt man voraus, dass eine Person bereit ist, für

5 Die Literatur zur Diskussion über die im Folgenden erwähnten Rationalitätsbedingungen ist enorm umfangreich. Einen Einstieg bieten z.B. Kap. 5-7 in Paul Anand, Prasanta K. Pattanaik, Clemens Puppe: *The handbook of rational and social choice. An overview of new foundations and applications* (Oxford: Oxford University Press, 2009).

6 Donald Davidson: *Hempel on explaining action*, in *Essays on actions and events* (Oxford: Oxford University Press, 1980) S. 261-275, hier S. 273.

7 Das Argument setzt voraus, dass Optionen mit Geldbeträgen kombiniert werden können. Vgl. Sven Ove Hansson: *The structure of values and norms* (Cambridge: Cambridge University Press, 2001) S. 29-30.

den Tausch von a gegen b einen minimalen Geldbetrag zu zahlen, wenn sie a gegenüber b vorzieht, so führen zyklische Präferenzen dazu, dass diese Person unendlich lange Optionen tauschen und dafür Geld bezahlen wird. Sind die Präferenzen zum Beispiel aPb, bPc, cPa, so wird eine Person, die über a verfügt, diese Option der Reihe nach gegen c, b und wiederum a tauschen und dafür drei Mal bezahlen. Nun kann man sich entweder unmittelbar auf eine Intuition berufen und geltend machen, dass Präferenzen irrational sind, wenn sie zu einem solchen Effekt führen, wogegen eingewendet wird, dass Ausbeutbarkeit nicht automatisch Irrationalität bedeutet. Oder man greift zu einer pragmatischen Argumentation, zum Beispiel, dass Personen mit Präferenzen, die eine Geldpumpe ermöglichen, ihre Ziele nicht erreichen können, wenn sie ihren Präferenzen entsprechend handeln, und macht geltend, das zeige, dass die betreffenden Personen irrationale Präferenzen haben.

Da die Entscheidungstheorie Präferenzen nicht nur in Beziehung zu anderen Präferenzen setzt, sondern auch zu Überzeugungen, insbesondere über die Wahrscheinlichkeit verschiedener Situationen, können sich Präferenzen auch als irrational erweisen, weil sie auf falschen Überzeugungen über Wahrscheinlichkeiten beruhen. Diese Form des Irrationalitätsvorwurfs ist in der Präferenzenforschung besonders häufig anzutreffen. Eines der bekanntesten Beispiele ist der *Spielerfehlschluss:* Der Spieler zieht es vor, auf Rot zu setzen, nachdem die Kugel mehrmals hintereinander auf einem schwarzen Feld gelandet ist, weil er überzeugt ist, dass eine rote Zahl wahrscheinlicher als eine schwarze ist, wenn viele schwarze Zahlen vorausgegangen sind. Um zu zeigen, dass diese Präferenz irrational ist, ist eine mit der Geldpumpe verwandte Argumentationsstrategie üblich. Das sogenannte *Dutch book argument* macht geltend, dass Personen, deren subjektive Überzeugungsgrade gegen die Wahrscheinlichkeitsrechnung verstoßen, ein System von Wetten akzeptieren würden, das sie in jeder möglichen Situation verlieren lässt. Das kann als offensichtlich irrational gelten und wiederum wird argumentiert, dass eine Person mit solchen Präferenzen ihre Ziele nicht erreichen kann, wenn sie aufgrund ihrer Präferenzen handelt.

Nun sind diese Argumente und die mit ihnen verteidigten formalen Rationalitätsbedingungen aus verschiedenen Gründen umstritten. Ich beschränke mich auf den Einwand, dass empirische Resultate gegen die Rationalitätsbedingungen der klassischen Entscheidungstheorie sprechen.

2. *Resultate der empirischen Präferenzenforschung*

Das von der Entscheidungstheorie geprägte Bild der Rationalität ist seit längerem umstritten. In Teilen der Debatte spielt die Interpretation empirischer Befunde eine zentrale Rolle. Studien fördern Resultate zutage, die nicht ohne weiteres mit klassischer Entscheidungstheorie zu vereinbaren sind.[8] Zur Illustration seien zwei typische Beispiele solcher Studien skizziert.

In einer Studie wurden den Versuchspersonen jeweils zwei Produkte (Autos, Restaurants usw.) zur Wahl angeboten, die so beschrieben sind, dass sie sich in zwei Dimensionen unterscheiden. Beispielsweise isst man im Restaurant a günstiger als in b, aber b bietet das bessere Essen als a. Nun fügt man eine dritte Option a^- hinzu, die leicht schlechter als a abschneidet, zum Beispiel ein Restaurant, in dem man gleich gut wie in a isst, nur etwas teurer, aber doch noch günstiger als in b. Option a^- wird nun zwar nicht gewählt, aber in dieser neuen Situation wird a häufiger b vorgezogen als in der ersten Situation ohne a^-. Das kann als eine Verletzung von *property alpha* rekonstruiert werden. Eine Standardinterpretation ist der sogenannte Anziehungseffekt: a^- wirkt als «Köder» für a, weil diese Option einer Person, die zwischen a und b schwankt, einen zusätzlichen Grund liefert, a vorzuziehen.[9]

Während es in der genannten Studie nur um Präferenzen geht, spielt in vielen Experimenten zusätzlich oder ausschließlich die Einschätzung von Wahrscheinlichkeiten eine Rolle. Eine Variante des Spielerfehlschlusses ist folgender Effekt:[10] Den Versuchspersonen wird gesagt, dass in einer Stadt alle Familien mit sechs Kindern erhoben wurden. In 72 Familien ist die genaue Reihenfolge der Kinder *Mädchen, Junge, Mädchen, Junge, Junge, Mädchen.* 80% der Befragten schätzen, dass es eine kleinere Anzahl Familien gibt, in denen die exakte Reihenfolge der Kinder *Junge, Mädchen, Junge, Junge, Junge, Junge* ist. Als Erklärung bieten die Autoren die Heuristik der Repräsentativität an. Sie gehen davon aus, dass Versuchspersonen solche Probleme mit Hilfe von Faustregeln lösen, im vorliegenden Fall mit der Regel, dass eine Abfolge umso wahrscheinlicher ist, je mehr sie charakte-

8 Sammlungen mit klassischen Studien sind z.B. Daniel Kahneman, Paul Slovic, Amos Tversky (Hg.): *Judgment under uncertainty. Heuristics and biases* (Cambridge: Cambridge University Press, 1982); Sarah Lichtenstein, Paul Slovic (Hg.): *The construction of preference* (Cambridge: Cambridge University Press, 2006).

9 Joel Huber, John W. Payne, Christopher Puto: *Adding asymmetrically dominated alternatives. Violations of regularity and the similarity hypothesis*, in *The Journal of Consumer Research* 9 (1982) S. 90-98.

10 Kahneman, op. cit. (Fn. 8) S. 3-20, 32-47.

ristische Eigenschaften der Grundgesamtheit (etwa gleich viele Jungen wie Mädchen, unregelmäßige Abfolge von Jungen und Mädchen) repräsentiert. Diese Heuristik führt, wie das Beispiel zeigt, auch zu regelmäßig auftretenden normabweichenden Resultaten.

Die Autoren solcher Studien machen geltend, dass diese empirisch feststellbaren Abweichungen von entscheidungstheoretischen Prinzipien Effekte sind, nicht bloße Fehler. Es sind nicht irgendwelche, sondern definierte, replizierbare Abweichungen, die nicht ausschließlich durch «Störfaktoren», wie Unaufmerksamkeit, Müdigkeit oder Missverständnisse erklärt werden können. Das heißt allerdings nicht, dass die Präferenzenforschung zeigt, dass Versuchspersonen *generell* gegen entscheidungstheoretische Prinzipien verstoßen, auch wenn dies aus rhetorischen Gründen gelegentlich so dargestellt wird.[11] Insofern den Prinzipien der Entscheidungstheorie widersprechende Präferenzen als irrational kritisiert werden können, weist die Präferenzenforschung also nach, dass irrationale Präferenzen systematisch auftreten, auch wenn sie aufs Ganze gesehen nicht die Regel sind. (Dieses Resultat nenne ich im Folgenden kurz «systematisch irrationale Präferenzen».)

Was zeigen nun die Resultate der Präferenzenforschung für die Entscheidungstheorie?[12] Aus Studien wie den erwähnten sind verschiedene, sich gegenseitig nicht ausschließende Vorwürfe abgeleitet worden. Erstens wird darauf hingewiesen, dass die Entscheidungstheorie Voraussetzungen macht, die empirisch meist nicht erfüllt sind. Subjekte verfügen beispielsweise einfach nicht über vollständige und transitive Präferenzordnungen. Wendet man in solchen Fällen die Entscheidungstheorie an, ist mit falschen Resultaten zu rechnen. Dies führt zum zweiten Vorwurf, dass die Entscheidungstheorie nicht die richtigen Präferenzen als rational auszeichnet. In diesem Punkt unterscheiden sich die Beispiele, die sich auf Urteile über Wahrscheinlichkeiten beziehen, von solchen, die nur die Struktur von Präferenzen betreffen. Während es kaum kontrovers ist, dass die Wahrscheinlichkeitstheorie ein Rationalitätskriterium liefert, ist das bei Prinzipien wie Vollständigkeit und Transitivität von Präferenzen weniger klar. (Eine Möglichkeit, diesem Unterschied Rechnung zu tragen, wird in Abschnitt 4.1

11 Samuels et al., op. cit. (Fn. 1).

12 Nebenbei sei bemerkt, dass sich ein großer Teil der Debatte um die empirische Präferenzenforschung weniger auf die Irrationalität selbst bezieht, als auf deren Erklärung durch kognitive Modelle (z.B. Heuristiken) und deren evolutionstheoretischen Hintergrund. Zum Folgenden vgl. Hans Rott: *Seltsame Wahlen* (in diesem Band S. 43-63).

erwähnt.) Drittens weisen Präferenzenforscher darauf hin, dass die Entscheidungstheorie nicht als Anleitung zum Bilden rationaler Präferenzen taugt, zum Beispiel, weil der Aufwand, die Theorie anzuwenden, in vielen Fällen mehr kosten würde, als was auf dem Spiel steht.

Bevor wir uns der resultierenden Debatte zuwenden, lohnt es sich zu fragen, weshalb viele Entscheidungstheoretiker sich durch die empirischen Befunde nicht aus der Ruhe bringen lassen. Zwei wohlbekannte Punkte sind unmittelbar einschlägig (ein weiterer folgt in Abschnitt 3.1).

Erstens können gewisse gängige Methoden zum Ermitteln von Präferenzen die angewandte Entscheidungstheorie für die genannten Unstimmigkeiten blind machen. Werden die Präferenzen eines Subjekts aus seinem Verhalten abgeleitet (*revealed preferences*) und monetarisiert, werden Unvollständigkeit und Intransitivität unsichtbar, weil jeder Option ein Geldbetrag zugeordnet wird und die Relation «>» auf Zahlen vollständig und transitiv ist. Eine direkte Bestimmung von Präferenz-Relationen, etwa durch Befragung (*stated preferences*), zeigt aber in manchen Fällen, dass formale Rationalitätsbedingungen nicht erfüllt sind. Beispielsweise wenn eine Person für zwei Optionen überhaupt keine Präferenzrelationen angeben will, weil sie einen Vergleich für unmöglich oder für verfehlt hält.

Zweitens werden in der Literatur zwei unterschiedliche Begriffe der Präferenz verwendet. Zum einen werden Präferenzen inhaltlich bestimmt, zum Beispiel als Dispositionen zu Wahlverhalten oder als mentale Zustände. Andererseits wird, besonders in der ökonomischen Literatur, der Begriff der Präferenz formal durch Axiome bestimmt, die garantieren, dass Präferenzen formale Rationalitätsbedingungen erfüllen.[13] Solange man ausschließlich von einem solchen formal bestimmten Begriff ausgeht, kann es, wie bereits angemerkt, keine empirischen Befunde geben, die zeigen, dass jemand den fraglichen Rationalitätsbedingungen widersprechende Präferenzen hat. Der Anschein des Gegenteils muss daher rühren, dass Präferenzen nicht korrekt ermittelt oder mit etwas anderem verwechselt worden sind.

[13] Vgl. Paul Anand: *Foundations of rational choice under risk* (Oxford: Clarendon Press, 1993) S. 100-102.

3. Die Debatte zwischen Entscheidungstheorie und empirischer Präferenzenforschung

In Anlehnung an Formulierungen von Thagard[14] kann man die Debatte zwischen Entscheidungstheorie und Präferenzenforschung durch drei Antworten auf die Frage «Wer hat ein Problem mit irrationalen Präferenzen?» strukturieren. Die erste Antwort ist, dass die Präferenzenforscher für die Diskrepanzen zwischen Entscheidungstheorie und Präferenzenforschung verantwortlich sind; würden sie alles für die Anwendung der Entscheidungstheorie Relevante berücksichtigen, zeigte sich, dass «die Leute», also die Versuchspersonen, rational sind. Die zweite Möglichkeit ist, dass die Entscheidungstheoretiker Urheber der Schwierigkeiten sind, weil sie inadäquate Rationalitätsnormen aufstellen. Oder «die Leute» haben ein Problem; sie halten sich nicht an die entscheidungstheoretischen Normen und sind also irrational. Ich werde kurz den ersten Punkt diskutieren und mich dann auf die Kontroverse zwischen der zweiten und der dritten Position konzentrieren.

3.1 Probleme bei der Anwendung des entscheidungstheoretischen Kalküls

Wie könnte man die Behauptung begründen, die Anwendung der Entscheidungstheorie in der empirischen Forschung sei dafür verantwortlich, dass den Versuchspersonen irrationale Präferenzen zugeschrieben werden? Eine Möglichkeit wäre, den einzelnen Studien konkrete methodologische Fehler nachzuweisen. Darauf gehe ich hier nicht ein. Stattdessen möchte ich einen grundlegenden Problembereich ansprechen. Wenn von «Entscheidungstheorie» die Rede ist, kann damit verschiedenes gemeint sein. Einerseits geht es um eine formale Theorie, die durch bestimmte Axiome und mathematische Strukturen gekennzeichnet ist. Nennen wir das den «entscheidungstheoretischen Kalkül». Damit man diesen Kalkül anwenden, das heißt auf Präferenzen und Überzeugungen von Personen oder auf Aussagen darüber beziehen kann, muss man über Regeln verfügen, die angeben, unter welchen Bedingungen eine solche Zuordnung adäquat ist. Das ergibt sich nicht einfach aus dem Kalkül als solchem, sondern erfordert eine «Anwendungstheorie», die Fragen wie die folgenden beantwortet:[15] Was repräsentieren

[14] Paul Thagard: *Computational philosophy of science* (Cambridge, MA: MIT Press, 1988) S. 123.

[15] Vgl. Anand, op. cit. (Fn. 13) S. 108.

die nichtlogischen Zeichen P, I, R, a, b usw. des entscheidungstheoretischen Kalküls? Das heißt insbesondere: Als welche Art von Gegenstand werden Präferenzen aufgefasst und wie werden sie individuiert? Sind Präferenzen beispielsweise mentale Zustände, Verhaltensdispositionen oder Aussagen darüber? Welche Art von Gegenständen gelten als Optionen? Werden diese zum Beispiel als Handlungsweisen, Konsumgüterbündel, mögliche Welten oder sprachliche Repräsentationen davon aufgefasst? Wie werden Optionen individuiert und welcher Bereich von Optionen wird vorausgesetzt? Auf dieser Grundlage müssen dann Bedingungen dafür angegeben werden, dass gegebene Präferenzverhältnisse durch präferenzenlogische Formeln adäquat wiedergegeben werden. Unter welchen Bedingungen ist es beispielsweise adäquat, die Präferenzen einer Person mit den drei Formeln aPb, cIb und aPc wiederzugeben? Nur wenn man solche Fragen beantworten kann, macht es überhaupt Sinn, davon zu sprechen, dass Personen Präferenzen haben, die aus entscheidungstheoretischer Perspektive rational (oder irrational) sind. Deshalb muss man im Zusammenhang mit der Frage, ob empirische Befunde der Entscheidungstheorie widersprechen, die Entscheidungstheorie als entscheidungstheoretischen Kalkül plus «Anwendungstheorie» auffassen.[16]

Tversky erläutert diesen oft vernachlässigten Punkt am Beispiel der Paradoxie von Allais:[17] Einer Person werden vier Optionen angeboten. Sie kann je eine aus der Gruppe a, b und eine aus der Gruppe c, d wählen:

		Gewinn:		
		p = 10%	p = 89%	p = 1%
Situation 1:	Option a	$ 1 000 000	$ 1 000 000	$ 1 000 000
	Option b	$ 5 000 000	$ 1 000 000	–
Situation 2:	Option c	$ 1 000 000	–	$ 1 000 000
	Option d	$ 5 000 000	–	–

Das Paradox geht so: Die meisten Personen haben die Präferenzen aPb und dPc. Situation (1) unterscheidet sich aber nur in der mittleren Kolonne von Situation (2) und in dieser Kolonne gibt es keinen Unterschied zwischen a und b, respektive c und d. Also ist diese Kolonne für die Entscheidung irre-

16 Die Situation ist analog zu derjenigen in der deduktiven Logik. Vgl. Georg Brun: *Die richtige Formel. Philosophische Probleme der logischen Formalisierung* (Frankfurt a.M.: Ontos, 2004).

17 Amos Tversky: *A critique of expected utility theory: Descriptive and normative considerations,* in *Erkenntnis* 9 (1975) S. 163-173.

levant und kann ignoriert werden, womit sich (1) und (2) nicht mehr unterscheiden und Konsistenz verlangt, dass man entweder in beiden Situationen die erste (a bzw. c) oder in beiden die zweite Option (b bzw. d) vorzieht.

Nun kann man einwenden, dass der Kalkül des subjektiven erwarteten Nutzens nicht vorschreibt, was in diesen Situationen als subjektiver erwarteter Nutzen gilt. Welche Aspekte der verschiedenen Situationen sind für die Präferenzen relevant? Obige Argumentation unterstellt, dass es nur um den Nutzen der auf dem Spiel stehenden Geldbeträge geht. Man könnte aber auch noch die Enttäuschung einbeziehen, die sich allenfalls einstellt, wenn das Resultat des Spiels bekannt ist.[18] Wer Option b wählt und nichts gewinnt, wird sich ärgern, weil ihm a einen sicheren Gewinn geboten hätte. Wer mit d nichts gewinnt, kann sich damit trösten, dass er mit c sehr wahrscheinlich auch nichts gewonnen hätte. Das mit Option b mit 1% Wahrscheinlichkeit erzielte Resultat ist also nicht einfach nur «Gewinn = $ 0», sondern «die todsichere Chance verpassen, $ 1 000 000 zu gewinnen». Wer so argumentiert, macht geltend, dass im Allais-Paradox der entscheidungstheoretische Kalkül nicht richtig angewendet wird, weil die Formel aPb ∧ dPc zusammen mit der obigen Tabelle keine adäquate Repräsentation der Präferenzen der Versuchspersonen ist.

In ähnlicher Weise lässt sich ein anderes Standardbeispiel aus der Diskussion um transitive Präferenzen analysieren:[19] Ein Gastgeber bietet dem wohlerzogenen Alf eine Frucht an: «Such Dir eine Frucht aus, ich nehme die andere.» In Situation (1) kann Alf zwischen einem großen Apfel und einer Orange wählen, in Situation (2) zwischen einem kleinen Apfel und einer Orange, in Situation (3) zwischen einem großen Apfel und einem kleinen Apfel. Alf wählt in (1) den großen Apfel, in (2) die Orange und in (3) den kleinen Apfel. Damit scheint er die nicht transitiven Präferenzen aPb, bPc, cPa (a: großer Apfel, b: Orange, c: kleiner Apfel) auszudrücken.[20] Diese Analyse kann man zurückweisen, indem man geltend macht, dass Alf transitive Präferenzen hat, wenn man berücksichtigt, wie er die Optionen auffasst. Zum Beispiel kann man die Option *großer Apfel* in Situation (1) auch als *großen Apfel nehmen und dem Gastgeber die Orange überlassen* oder *großer Apfel ohne Unhöflichkeit* beschreiben und in Situation (3) als

18 Diese Idee wird in der *regret theory* systematisch verfolgt.

19 Das Beispiel taucht in verschiedenen Varianten auf, z.B. in Anand, op. cit. (Fn. 13) S. 67.

20 Diese Interpretation setzt voraus, dass Alfs Wahl eine strikte (P), keine schwache Präferenz (R) ausdrückt.

großen Apfel nehmen und dem Gastgeber den kleinen überlassen oder *großer Apfel mit Unhöflichkeit.* Dann ist aber die obige Beschreibung von Alfs Präferenzen inkorrekt und müsste durch etwas wie aPb, bPc, cPd (a: großer Apfel ohne Unhöflichkeit, b: Orange, c: kleiner Apfel, d: großer Apfel mit Unhöflichkeit) ersetzt werden. Mit anderen Worten, ob eine Verletzung des Transitivitätsprinzips vorliegt, hängt (unter anderem) von der Individuierung der Optionen ab. Gelten Optionen mit gleicher Frucht als identisch, verletzen Alfs Präferenzen das Prinzip der Transitivität. Fasst man den Umstand, welche Früchte verbleiben, als Teil der Option auf, verschwindet das Problem.

Diese Strategie zur Verteidigung der Entscheidungstheorie ist allerdings gefährlich. Es droht, dass jede nicht schon aus begrifflichen Gründen inkonsistente Konfiguration von Präferenzen so interpretiert werden kann, dass sie der Entscheidungstheorie nicht widerspricht.[21] Dann hätte die Entscheidungstheorie aber keinen empirischen Gehalt. Dieses Resultat lässt sich vermeiden, wenn man beachtet, dass es zwar möglich sein mag, beliebige Präferenzenmuster so zu beschreiben, dass sie zum entscheidungstheoretischen Kalkül passen; aber das heißt nicht, dass man solche Beschreibungen auch dann angeben kann, wenn man eine Anwendungstheorie voraussetzt. Kurz: Auf dem Prüfstand stehen Entscheidungstheorien mit einer Anwendungstheorie, nicht bloß entscheidungstheoretische Kalküle.

Mit der Wahrscheinlichkeitstheorie verhält es sich analog. Man muss unterscheiden zwischen der mathematischen Theorie der Wahrscheinlichkeit (z.B. Kolmogoroff-Axiome und damit beweisbare Sätze), und der Interpretation, die bestimmt, wie diese Theorie auf konkrete Situationen angewendet werden kann. Dass sich daraus wichtige Differenzen ergeben, geht aus der Kontroverse um subjektivistisches und frequentistisches Verständnis der Wahrscheinlichkeit hervor. Das spielt auch in der Debatte um die Entscheidungstheorie eine wichtige Rolle, wie vor allem Vertreter einer an der Evolutionstheorie orientierten Psychologie betont haben.[22]

In der *rationality wars*-Debatte spielt die Frage, mit welcher Theorie sich adäquate von inadäquaten Anwendungen des entscheidungstheoretischen Kalküls unterscheiden lassen, eine untergeordnete Rolle. Das ist problematisch, weil sich nur im Rahmen einer Anwendungstheorie sinnvoll darüber

21 Tversky, op. cit. (Fn. 17) S. 171; Anand, op. cit. (Fn. 13) S. 103-106; John Broome: *Can a Humean be moderate?*, in *Ethics out of economics* (Cambridge: Cambridge University Press, 1999) S. 68-87.

22 Z.B. Gerd Gigerenzer: *Rationality for mortals. How people cope with uncertainty* (Oxford: Oxford University Press, 2008).

streiten lässt, ob die Präferenzenforschung die Irrationalität der Versuchspersonen oder die Inadäquatheit der Entscheidungstheorie zeigt. Trotzdem wende ich mich nun diesem Streitpunkt zu, da eine angemessene Behandlung der Probleme bei der Anwendung entscheidungstheoretischer Kalküle hier nicht geleistet werden kann.

3.2 Irrationales Entscheiden vs. inadäquate Entscheidungstheorie

Eine klassische Strategie, die Entscheidungstheorie gegen die Vorwürfe der Präferenzenforschung zu verteidigen, analysiert die Kritik als Resultat eines Konflikts zwischen normativen und nicht normativen Entscheidungstheorien. Zuerst wird zwischen normativen, deskriptiven und präskriptiven Theorien unterschieden.[23] Normative Theorien von Präferenzen befassen sich mit der Frage, welche Präferenzen ein rationales Subjekt haben *sollte*, wenn es gewisse Präferenzen und Überzeugungen bereits hat. Deskriptive Theorien fragen, welche Präferenzen Menschen faktisch haben. Präskriptive Theorien geben praktische Regeln an, an denen man sich orientieren kann, wenn man möglichst den Anforderungen der normativen Theorie genügende Präferenzen haben möchte. Auf dieser Grundlage kann man geltend machen, dass die Entscheidungstheorie eine normative Theorie ist, die sich auf Normen der Rationalität bezieht. (Sie setzt voraus, dass andere Normen, etwa moralische, politische und ästhetische, erfüllt oder in den Präferenzen ausgedrückt sind). Die empirische Präferenzenforschung dagegen ist eine deskriptive Theorie. Ihre Befunde zeigen somit, dass Menschen systematisch irrationale Präferenzen haben.[24] Diese Resultate sprechen nicht direkt gegen die Entscheidungstheorie, so wird weiter argumentiert, weil sie eine normative, keine deskriptive oder präskriptive Theorie ist. Dass sie nicht deskriptiv ist, bedeutet zwar nicht, dass sie keinen empirischen Gehalt hätte, weil man ja untersuchen kann, ob empirisch feststellbare Präferenzen ihr genügen, wohl aber, dass sie keine prognostische Theorie ist und auch keine Theorie darüber, welche Präferenzen die meisten Menschen haben. Dass Menschen Präferenzen haben, die die Entscheidungstheorie für irrational erklärt, kann demnach nicht gegen sie vorgebracht werden. Dass die Entscheidungstheorie

23 Vgl. Jonathan Baron: *Thinking and deciding* (Cambridge: Cambridge University Press, [4]2008) Kap. 2.

24 Stein, op. cit. (Fn. 1) S. 4 hat dafür den Ausdruck «standard picture of rationality» geprägt.

nicht präskriptiv ist, bedeutet, dass sie gar nicht den Anspruch erhebt, eine praktische Anleitung für den Erwerb rationaler Präferenzen zu geben. In dieser Hinsicht verhält es sich mit der Entscheidungstheorie nicht anders als mit der klassischen deduktiven Logik, die zwar angibt, welche Folgerungen aus gegebenen Sätzen akzeptiert werden müssen, aber wenig Hilfe beim praktischen Ziehen von Schlüssen bietet und auch nicht beschreibt, was Personen tatsächlich folgern.

Diese Argumentation wird von vielen Autoren nicht akzeptiert. Aus ihrer Sicht müssen aus den deskriptiven Befunden, nach denen systematisch von den entscheidungstheoretischen Erfordernissen abweichende Präferenzen auftreten, ganz andere Schlüsse gezogen werden. Sie machen geltend, dass die empirische Forschung nicht nur die Entscheidungstheorie als deskriptive Theorie widerlegt, sondern auch zeigt, dass die Entscheidungstheorie als normative Theorie zum inakzeptablen Resultat führt, dass Menschen in systematischer Weise irrational sind. Das wird mit zwei unterschiedlichen Argumentationslinien geltend gemacht. Zum einen wird vorgebracht, es sei absurd, anzunehmen, Menschen seien systematisch irrational. Diese Argumentsweise wird im Folgenden genauer untersucht. Zum anderen wird geltend gemacht, dass die kognitiven Mechanismen, die aus Sicht der klassischen Entscheidungstheorie zu Irrationalitäten führen, andere erwünschte Konsequenzen haben, zum Beispiel effizient sind oder dem Menschen in seiner phylogenetischen Entwicklung adaptive Vorteile gebracht haben. Mit einer solchen evolutionstheoretischen Perspektive wird dann oft der Vorschlag verbunden, einen anderen Rationalitätsbegriff – manchmal *ecological rationality* genannt – zu verwenden, der sich nicht wie in der klassischen Entscheidungstheorie an formalen Gesichtspunkten wie Konsistenz, Kohärenz und allgemeiner Maximierung von Präferenzenverwirklichung orientiert, sondern daran, dass Entscheidungen mit Hilfe von Mechanismen getroffen werden, die Menschen im Kontext der für ihre Evolution relevanten Bedingungen praktische Vorteile verschafft haben.[25]

25 Explizit z.B. in Gigerenzer, op. cit. (Fn. 22) S. 18-19. Auch Aussagen wie «[…] rationality is a tool for helping organisms to reach their real-world goals, not necessarily to conform to rational norms.» (Valerie M. Chase, Ralph Hertwig und Gerd Gigerenzer: *Visions of rationality*, in *Trends in Cognitive Sciences* 2 [1998] S. 206-214, hier S. 207) sind in diesem Sinne zu verstehen. Zur Kontroverse vgl. Till Grüne-Yanoff: *Bounded rationality*, in *Philosophy Compass* 2 (2007) S. 534-563, und Keith E. Stanovich, Richard F. West: *Evolutionary versus instrumental goals. How evolutionary psychology misconceives human rationality*, in *Evolu-*

Das führt zu Unklarheiten in der Literatur, weil oft nicht zu entscheiden ist, wie verschiedene Rationalitätsbegriffe unterschieden werden, falls überhaupt. Vorschläge, den entscheidungstheoretischen Rationalitätsbegriff umzudeuten oder zu ersetzen, gehören nicht zur Argumentationslinie, die ich hier analysiere, auch wenn sie oft mit Resultaten aus der Präferenzenforschung begründet werden.

Beschränken wir uns auf den entscheidungstheoretischen Begriff der Rationalität, so resultiert als Streitfrage, ob wir die Kritik akzeptieren müssen, dass sich manche unserer Präferenzen als systematisch irrational erweisen, oder ob wir vielmehr eine neue Entscheidungstheorie brauchen, die den tatsächlichen Präferenzen besser Rechnung trägt. Kurz: Sind irrationale Präferenzen ein Problem für die Entscheidungstheorie oder für Menschen mit irrationalen Präferenzen?

Die Diskussion um diese Streitfrage ist vielschichtig.[26] Ich grenze in vier Schritten die Aspekte ab, welche ich diskutieren werde. Erstens muss man der Verteidigung der Entscheidungstheorie in einem Punkt recht geben: Die klassische Entscheidungstheorie ist eine normative Theorie. So wird sie von den meisten Befürwortern und Kritikern verstanden. Eine rein deskriptive Erforschung von Präferenzen und Entscheidungen ist natürlich auch ein respektables Projekt, aber ein anderes. Die Unterscheidung zwischen normativ und deskriptiv, die hier unterstellt wird, betrifft den Gebrauch einer Theorie und kann nicht unbedingt an der Theorie als strukturierter Menge von Sätzen festgemacht werden. Beispielsweise wird man dieselbe Aussage «Es gibt keine Folge von Optionen a_1, …, a_n, so dass $a_1 P \ldots P a_n P a_1$» in deskriptivem Gebrauch als die Behauptung interpretieren, dass Personen azyklische Präferenzen *haben*, in normativem Gebrauch hingegen als die Forderung, dass rationale Präferenzen azyklisch sein *sollen.* Die Behauptung, die Entscheidungstheorie sei in diesem Sinne eine normative Theorie, setzt lediglich voraus, dass man zwischen normativen und deskriptiven Fragen unterscheiden kann («Welche Präferenzen *sollte* ich als rationales Subjekt haben?» vs. «Welche Präferenzen *habe* ich tatsächlich?»). Sie macht geltend, dass die Entscheidungstheorie den Anspruch erhebt, normative

tion and the psychology of thinking, hg. von David Over (Hove: Psychology Press, 2003) S. 171-230.

26 Als Übersicht: Baron, op. cit. (Fn. 23) und Patrick Rysiew: *Rationality disputes. Psychology and epistemology*, in *Philosophy Compass* 3 (2008) S. 1153-1176. Der Disput ist nicht zuletzt durch forschungspolitische Motive und die entsprechende Rhetorik geprägt; vgl. Samuels et al., op. cit. (Fn. 1).

Fragen adäquat zu beantworten. Es wird aber weder vorausgesetzt, dass sich die Sätze der Entscheidungstheorie als solche in normative und deskriptive einteilen lassen, noch dass die korrekte Beantwortung deskriptiver Fragen bei der Entwicklung oder Begründung der Entscheidungstheorie keine Rolle spielt. Im Folgenden werde ich mich auf den Aspekt konzentrieren, ob die Ergebnisse der Präferenzenforschung gegen die *normative* Verwendung der Entscheidungstheorie sprechen.

Nun ist es zweitens so, dass die Entscheidungstheorie faktisch nicht nur normativ verwendet wird, sondern auch für Prognosen und Erklärungen. Unter der Annahme, dass die betreffenden Personen rational sind, werden aufgrund von bekanntem Verhalten Rückschlüsse auf ihre Präferenzen und Überzeugungen gezogen und aus bekannten Präferenzen und Überzeugungen Voraussagen über ihre weiteren Präferenzen und ihr zukünftiges Verhalten abgeleitet. Wollte man bestreiten, dass die Befunde der Präferenzenforschung gegen diese Verwendungsweise der Entscheidungstheorie sprechen, müsste man den empirischen Studien methodische Probleme bei der Anwendung der Entscheidungstheorie nachweisen. Da deskriptive Verwendungen der Entscheidungstheorie nicht zu meinem Thema gehören, werde ich diese Problematik nicht aufgreifen und nur den Aspekt berücksichtigen, dass man sich auch fragen muss, ob und inwiefern die Inadäquatheit der Entscheidungstheorie als deskriptive Theorie gegen ihre normative Verwendung spricht.

Drittens ist nun klar zu sehen, dass der Hinweis, die klassische Entscheidungstheorie sei eine normative Theorie, als Verteidigung nicht sonderlich weit trägt. Ihre Kritiker können nämlich zugeben, dass die Entscheidungstheorie auf eine normative Verwendung zugeschnitten ist, aber darauf bestehen, dass damit noch nichts darüber gesagt ist, unter welchen Bedingungen es gerechtfertigt ist, die Entscheidungstheorie so zu verwenden, und schließlich geltend machen, dass empirische Befunde bei der gesuchten Rechtfertigung eine entscheidende Rolle spielen. So verstanden lautet die Herausforderung der Präferenzenforschung: Die normative Verwendung der Entscheidungstheorie ist mindestens dann nicht gerechtfertigt, wenn sie zum Resultat führt, dass Personen systematisch irrationale Präferenzen haben.

Die bisherigen Überlegungen können in den folgenden zwei Fragen zusammengefasst werden: Unter welchen Bedingungen ist es gerechtfertigt, die Entscheidungstheorie normativ zu verwenden? Schließt eine solche Rechtfertigung Entscheidungstheorien aus, denen gemäß wir systematisch irrationale Präferenzen haben? Ich werde mich nun viertens darauf beschränken zu diskutieren, welche Antwort auf diese beiden Fragen sich

ergibt, wenn man den Vorschlag aufgreift, dass die normative Verwendung der Entscheidungstheorie mit der Methode des Überlegungsgleichgewichts gerechtfertigt werden kann.

4. Überlegungsgleichgewicht

Das Überlegungsgleichgewicht ist eine Methode zur Rechtfertigung von Theorien, die eine Abstimmung zwischen Theorie und vortheoretischen akzeptierten Urteilen, Kategorien, Methoden, Standards und Zielen ins Zentrum stellt. Im Zusammenhang mit der Entscheidungstheorie verspricht das Konzept des Überlegungsgleichgewichts einiges: eine Rechtfertigung für die normative Verwendung dieser Theorie ohne einen fundamentalistischen Dogmatismus und eine Erklärung für die Möglichkeit eines entscheidungstheoretischen Pluralismus ohne subjektivistischen Relativismus. Hier ist nicht der Ort für eine grundsätzliche Verteidigung des Überlegungsgleichgewichts als Rechtfertigungsmethode.[27] Ich setze das im Folgenden voraus und beschränke mich auf die Frage, was sich über die Rolle der empirischen Befunde aus der Präferenzenforschung folgern lässt, wenn man die Methode des Überlegungsgleichgewichts auf die Entscheidungstheorie anwendet.

Die Idee, in diesem Zusammenhang mit dem Überlegungsgleichgewicht zu argumentieren, ist nicht neu. Zu Beginn der 1980er Jahre wurde dazu in *The Behavioral and Brain Sciences* eine ausgedehnte Debatte geführt, die später unter anderem von Thagard, Stich und Stein fortgesetzt wurde und schließlich dazu geführt hat, dass viele Autoren die Idee aufgegeben haben, die Entscheidungstheorie mit Hilfe eines Überlegungsgleichgewichts zu rechtfertigen.[28] Mir scheint aber, dass diese Debatte zu einseitig geführt wurde und einige Missverständnisse enthält. Dies hängt unter anderem daran, dass verschiedene Konzepte des Überlegungsgleichgewichts im Umlauf sind. Gegen das auf Goodman und Elgin zurückgehende Konzept des Überlegungsgleichgewichts, das ich diskutieren werde, sind die in der Literatur diskutierten Einwände meines Erachtens nicht durchschlagend.

[27] Vgl. die in Fn. 29 angegebenen Texte von Daniels und Elgin.

[28] Jonathan L. Cohen: *Can human rationality be experimentally demonstrated?* und *The controversy about irrationality*, in *The Behavioral and Brain Sciences* 4 (1981) S. 317-370; 6 (1983) S. 487-533; Thagard, op. cit. (Fn. 14); Stephen P. Stich: *The fragmentation of reason. Preface to a pragmatic theory of cognitive evaluation* (Cambridge, MA: MIT Press, 1990); Stein, op. cit. (Fn. 1).

4.1 Überlegungsgleichgewicht als Rechtfertigungsmethode

Das Konzept des Überlegungsgleichgewichts ist in der Literatur in verschiedenen Varianten entwickelt worden. Die wichtigsten Traditionsstränge verlaufen von Goodman über Rawls zu Daniels und zu Elgin.[29] Im Folgenden versuche ich nicht, den Konzeptionen der verschiedenen Autoren gerecht zu werden, sondern erläutere zuerst ein Standardverständnis und gehe dann auf einige Eigenheiten von Elgins Theorie ein, auf die ich mich stützen werde.

Ausgangspunkt ist die Beobachtung, dass man einzelne Entscheidungen und somit Beziehungen zwischen Präferenzen als rational rechtfertigt, indem man zeigt, dass sie durch die Entscheidungstheorie sanktioniert werden. Andersherum rechtfertigt man die Entscheidungstheorie dadurch, dass man zeigt, dass sie Entscheidungen sanktioniert, die wir tatsächlich als rational beurteilen. Der zentrale Gedanke ist nun, dass eine solche wechselseitige Übereinstimmung nicht als problematische Zirkularität aufgefasst werden muss, sondern so gedeutet werden kann, dass gerade die wechselseitige Abstimmung für die Rechtfertigung von Entscheidungstheorie *und* von einzelnen Entscheidungen zentral ist.

In einer ersten Annäherung kann man sagen, dass ein Überlegungsgleichgewicht vorliegt, wenn man die Übereinstimmung von Entscheidungstheorie und Urteilen über rationale Entscheidungen als Resultat eines Prozesses folgender Art rekonstruieren kann: Zu Beginn verfügen wir über Urteile über rationale Entscheidungen, die wir mit mehr oder weniger Sicherheit vertreten möchten. Sie können als explizite Aussagen verfügbar sein oder sich auch nur in sprachlichen oder nonverbalen Handlungsweisen ausdrücken, etwa darin, dass wir bestimmte Entscheidungen oder Erklärungen für Entscheidungen als rational akzeptieren oder als irrational zurückweisen. Weiterhin können solche Rationalitätsurteile einzelne konkrete Entscheidun-

[29] Die wichtigsten Texte sind: Nelson Goodman: *Fact, fiction, and forecast* (Cambridge, MA: Harvard University Press, [4]1983) Kap. III.2-3; John Rawls: *A theory of justice. Revised edition* (Cambridge, MA: Belknap Press, 1999); John Rawls: *The independence of moral theory*, in *Collected papers* (Cambridge, MA: Harvard University Press, 1999) S. 286-302; Norman Daniels: *Justice and justification. Reflective equilibrium in theory and practice* (Cambridge: Cambridge University Press, 1996); Catherine Z. Elgin: *With reference to reference* (Indianapolis: Hackett, 1983) Kap. X; Catherine Z. Elgin: *Considered judgment* (Princeton: Princeton University Press, 1996). Eine Übersicht bietet Susanne Hahn: *Überlegungsgleichgewicht(e). Prüfung einer Rechtfertigungsmetapher* (Freiburg i.Br.: Alber, 2000) Teil B.

gen betreffen, oder auch sehr allgemein sein, wie beispielsweise das Urteil, dass eine Präferenzenordnung, die zu einer Geldpumpe führen kann, irrational ist. Auf dieser Grundlage werden allgemeine Prinzipien gesucht, aus denen die besagten Urteile abgeleitet werden können. Dabei sind auch die üblichen Tugenden wissenschaftlicher Theorien, zum Beispiel Einfachheit und Genauigkeit, zu berücksichtigen. Beispiele aus dem Bereich der Entscheidungstheorie sind einerseits Axiome, die die Struktur von Präferenzen betreffen, und etwa deren Transitivität fordern. Andererseits müssen auch die Regeln dazu gerechnet werden, die die Anwendung des entscheidungstheoretischen Kalküls leiten, weil sonst nicht sinnvoll davon gesprochen werden kann, dass Prinzipien und vortheoretische Urteile übereinstimmen (vgl. Abschnitt 3.1). Um Prinzipien und Urteile zur Übereinstimmung zu bringen, müssen im Allgemeinen Anpassungen vorgenommen werden. Prinzipien werden geändert, wenn sie Urteilen oder anderen Prinzipien widersprechen, die wir weniger bereit sind, aufzugeben, und Urteile werden geändert, wenn wir widersprechende Prinzipien oder Urteile nicht aufgeben wollen. Dieser Prozess ist wechselseitig, weil dabei nicht nur zu gegebenen Urteile passende Prinzipien formuliert werden, sondern auch Urteile anhand der Prinzipien beurteilt und allenfalls korrigiert werden. Und er ist iterativ, weil das Resultat bereits vorgenommener Systematisierungen und Anpassungen wiederum die Grundlage für die nächsten Systematisierungen und Anpassungen ist.

Das ist allerdings erst der Kern des Überlegungsgleichgewichts («enges Überlegungsgleichgewicht» genannt[30]). Eine erste wesentliche Erweiterung (zum «weiten» Überlegungsgleichgewicht) besteht darin, im Abstimmungsprozess zusätzlich zu Urteilen und Prinzipien weitere relevante Theorien zu berücksichtigen. Bei der Entscheidungstheorie sind naheliegende Beispiele für solche «Hintergrundtheorien» Wahrscheinlichkeitstheorie, Handlungstheorie und Theorie des Geistes. Damit gehen auch Argumente wie beispielsweise das (in Abschnitt 1) erwähnte pragmatische Argument gegen mögliche Geldpumpen in den Abstimmungsprozess ein. Um eine Übereinstimmung zwischen Urteilen, Prinzipien und Hintergrundtheorien zu erreichen, können alle drei Elemente angepasst werden. Grundsätzlich kann jedes an einem Überlegungsgleichgewicht beteiligte Element revidiert werden und es ist nicht ausgeschlossen, dass die angestrebte Übereinstimmung durch unterschiedliche Modifikationen erreicht werden kann und also verschiedene Überlegungsgleichgewichte möglich sind. Welche Revisionen vorgenommen

30 Dieser Begriff wird in der Literatur unterschiedlich verwendet; vgl. Abschnitt 4.2.

werden, hängt auch davon ab, welche Ziele mit der Theoriebildung verfolgt werden und welches Gewicht unterschiedlichen Eigenschaften der Theorie, zum Beispiel Einfachheit, Implementierbarkeit oder möglichst breite Anwendbarkeit, gegeben wird. So bemerkt beispielsweise Aldred, dass eine Theorie, die systematisch Enttäuschung berücksichtigt (vgl. Abschnitt 3.1), zwar gewisse vortheoretische Urteile wahrt, die der klassischen Theorie des erwarteten subjektiven Nutzens widersprechen, dafür aber weniger allgemein angewendet werden kann.[31] Typischerweise stehen aber gewisse Hintergrundtheorien kaum zur Debatte. Das erklärt, weshalb die Irrationalität von Präferenzen, die durch Verstoß gegen die Wahrscheinlichkeitstheorie zustande kommen (z.B. Spielerfehlschluss), wesentlich weniger umstritten ist als etwa die Irrationalität intransitiver Präferenzen. In der Debatte um rationale Präferenzen ist die Wahrscheinlichkeitstheorie eine relativ revisionsresistente Hintergrundtheorie, während das präferenzenlogische Transitivitätsaxiom zur Vordergrundtheorie gehört und somit seine Rechtfertigung gerade zur Debatte steht.

Damit ist das weite Überlegungsgleichgewicht eine pluralistische, aber nicht relativistische Form der Rechtfertigung. Es ist weder garantiert, dass es ein eindeutig bestimmtes Überlegungsgleichgewicht gibt, noch dass mehrere existieren. Und es ist jedenfalls nicht so, dass sich beliebige Anfangsverpflichtungen, Systematisierungen und Hintergrundtheorien in ein Überlegungsgleichgewicht bringen lassen. Das macht die Vielfalt der bisher entwickelten Entscheidungstheorien verständlich und in dieser Hinsicht ist die Rekonstruktion der Theoriebildung als Versuch, ein Überlegungsgleichgewicht herzustellen, epistemologisch nicht revisionistisch.

Elgin hat anschließend an Goodmans ursprüngliche Vorschläge ein Konzept des Überlegungsgleichgewichts entwickelt, das sich gegenüber dem geschilderten Standardverständnis durch Erweiterungen, Änderungen und Radikalisierungen auszeichnet. Ich nenne vier Punkte: Der erste betrifft den Anwendungsbereich. Das Überlegungsgleichgewicht ist ein Konzept der Rechtfertigung, das nicht nur auf Logik oder Moraltheorie (wie bei Goodman bzw. Rawls) angewendet werden kann, sondern auf jede wissenschaftliche Theorie. Das hat zur Konsequenz, dass das Überlegungsgleichgewicht ein holistisches Konzept der Rechtfertigung ist. Das Überlegungsgleichgewicht, das eine Rechtfertigung der Entscheidungstheorie darstellt, ist auch relevant für die Rechtfertigung der entsprechenden Hintergrundtheorien. Die

[31] Jonathan Aldred: *The money pump revisited,* in *Risk, Decision and Policy* 8 (2003) S. 59-76, hier S. 60-61.

Bezeichnung «Hintergrundtheorie» ist somit irreführend. «Im Hintergrund» erscheint eine Theorie nur aus der Perspektive einer anderen Theorie, deren Rechtfertigung gerade im Vordergrund steht.[32] Eine zweite Erweiterung besteht darin, dass die in der Standarddarstellung eingebaute Beschränkung auf propositionale Elemente (Urteile, Prinzipien und Theorien) aufgehoben wird. An einem Überlegungsgleichgewicht sind auch Kategorien (z.B. komparative oder metrische Wahrscheinlichkeitsbegriffe), Verfahren (z.B. Methoden zum Erfassen von Präferenzen), Standards (z.B. statistische Signifikanzniveaus) und Ziele, denen eine Theorie dienen soll, beteiligt. Deshalb spricht Elgin nicht von vortheoretischen «Urteilen», sondern allgemeiner von «Ausgangsverpflichtungen» (*antecedent commitments*), die eine gewisse anfängliche Haltbarkeit (*initially tenability*) haben, das heißt, für mehr oder weniger gut gesichert erachtet werden.[33] Eine dritte Erweiterung betrifft die relevanten epistemischen Subjekte. Nicht nur die Ausgangsverpflichtungen des Theoretikers, der gerade die Theorie entwickelt, sind bei der Entwicklung des Überlegungsgleichgewichts zu berücksichtigen, sondern auch diejenigen anderer Personen, denen aber mehr oder weniger Gewicht zukommen kann. Das bedeutet aber nicht, dass das resultierende Überlegungsgleichgewicht soziologisch oder psychologisch zu deuten wäre. Es besteht zwischen Urteilen und Theorien und bezeichnet nicht den Zustand einer Person oder Personengruppe.[34] Besonders wichtig ist die vierte, auf Goodman zurückgehende Erweiterung, die zwei Ebenen der Rechtfertigung unterscheidet.[35] Die Elemente eines Systems im Überlegungsgleichgewicht sind dadurch gerechtfertigt, dass sie ein System im Gleichgewicht bilden. Das System ist dadurch gerechtfertigt, dass es reflektiert ist, das heißt in Ausgangsverpflichtungen verankert.[36] Das bedeutet insbesondere, dass man nicht alle Ausgangsverpflichtungen gleichzeitig aufgeben kann, obschon

32 Mit diesem holistischen Aspekt hängt ein wichtiges Argument für die epistemologische Bedeutung des Überlegungsgleichgewichts zusammen: In allen gängigen Theorien der epistemischen Rechtfertigung spielen logische Beziehungen eine zentrale Rolle, und es scheint mir aus den Gründen, die ich in Brun, op. cit. (Fn. 16) Kap. 3 erläutert habe, ausgeschlossen, die normative Verwendung der Logik anders als mit einem Überlegungsgleichgewicht zu rechtfertigen.

33 Elgin: *Considered Judgement*, op. cit. (Fn. 29) S. 13, 105.

34 Gegen die Interpretation von Thagard, op. cit. (Fn. 14) S. 130-131.

35 Nelson Goodman: *Sense and certainty* in *Problems and projects* (Indianapolis: Bobbs-Merrill, 1972) S. 60-68.

36 So deutet Elgin den Ausdruck *reflective equilibrium*. Die übliche deutsche Bezeichnung «Überlegungsgleichgewicht» legt andere Deutungen nahe.

jede einzelne revidiert werden kann. Somit reduziert sich die Idee des Überlegungsgleichgewichts nicht auf Kohärenz.[37] Gleichzeitig wird die gängige Form des Fundamentalismus vermieden, weil es keine *bestimmten* Sätze gibt, die nicht rein inferentiell gerechtfertigt sind. damit kann auch der gelegentlich erhobene Vorwurf, die Methode des Überlegungsgleichgewichts biete nur eine Reorganisation von Vorurteilen, zurückgewiesen werden.[38]

Die Rolle der Ausgangsverpflichtungen lädt noch einen weiteren Einwand ein: Die Methode des Überlegungsgleichgewichts lasse die Rechtfertigung falscher Systeme mit falschen Ausgangsverpflichtungen zu. Stich und Nisbett haben das Beispiel der Spieler vorgebracht, die sich passende Prinzipien der Wahrscheinlichkeit zurechtlegen, so dass sie reklamieren können, ihre Spielerfehlschluss-Urteile seien in einem Überlegungsgleichgewicht mit ihren Prinzipien und also epistemisch gerechtfertigt.[39] Nun verfängt ein solcher Einwand allenfalls gegen ein enges Überlegungsgleichgewicht, aber nicht gegen die hier erläuterte Konzeption, weil die Prinzipien des Spielerfehlschlusses zurückgewiesen werden können, wenn sie in Konflikt mit Hintergrundstheorien kommen.[40] Diesen Umstand nützen Stich und Nisbett selbst aus, wenn sie gegen den Spieler Argumente vorbringen, die sich auf Wahrscheinlichkeitstheorie und auf den fehlenden kausalen Zusammenhang zwischen den Ergebnissen von zwei Münzwürfen stützen.[41]

Die größte Schwierigkeit, mit der sich das Konzept des Überlegungsgleichgewichts meines Erachtens konfrontiert sieht, ist die Tatsache, dass «Gleichgewicht» eine Metapher ist, für die wir nicht über eine angemessene Explikation verfügen. In dieser Hinsicht ist für das Konzept des Überlegungsgleichgewichts noch mehr Arbeit zu leisten als für die traditionelleren Metaphern des Fundaments und der Kohärenz, weil der Begriff des Über-

37 Das ist der entscheidende Unterschied zu Thagards Modell *from the descriptive to the normative*, in Thagard, op. cit. (Fn. 14) S. 133.

38 So z.B. Richard B. Brandt: *The concept of rational belief*, in *The Monist* 68 (1985) S. 3-23.

39 Stephen P. Stich, Richard E. Nisbett: *Justification and the psychology of human reasoning*, in *Philosophy of science* 47 (1980) S. 188-202; Stich, op. cit. (Fn. 28).

40 Elgin: *Considered Judgement*, op. cit. (Fn. 29) S. 118-119.

41 Thomas Bartelborth: *Begründungsstrategien. Ein Weg durch die analytische Erkenntnistheorie* (Berlin: Akademie, 1996) S. 50-51. Stich nimmt diesen Einwand nicht ernst und erklärt die Differenz zwischen engem und weitem Überlegungsgleichgewicht als eine Frage von «Schnickschnack» («bells and whistles», Stich, op. cit. [Fn. 28] S. 84).

legungsgleichgewichts, so wie er hier verwendet wird, wesentlich reicher ist als etwa «fundiert» und «kohärent». Damit ist noch fraglicher, wie viel davon in einer formalen Theorie geleistet werden könnte.

4.2 Die Rolle von empirischen Befunden

Wir können nun die beiden Fragen aufgreifen, welche Rolle empirische Befunde bei der Rechtfertigung der normativen Verwendung der Entscheidungstheorie spielen, insbesondere, ob dabei systematisch irrationale Präferenzen ausgeschlossen sind. Für die weitere Diskussion ist es entscheidend, zwei Arten von empirischen Befunden zu unterscheiden, solche, die Rationalitätsurteile über Präferenzen betreffen, und solche, die die Präferenzen selbst betreffen: Welche Präferenzen beurteilen die Versuchspersonen als rational und welche haben sie tatsächlich?[42]

Dass tatsächlich gefällte Rationalitätsurteile für die Rechtfertigung der Entscheidungstheorie eine entscheidende Rolle spielen, ist gerade eine der Pointen des Überlegungsgleichgewichts. Der normative Gebrauch der Entscheidungstheorie ist unter anderem dadurch gerechtfertigt, dass die Theorie eine angemessene Menge vortheoretischer Rationalitätsurteile wahrt. Das ist nur schon deshalb erforderlich, weil sonst die Theorie einfach das Thema wechseln würde. Würde eine Theorie unsere Rationalitätsurteile allgemein für falsch erklären, wäre nicht einzusehen, warum wir sie als eine Theorie der Rationalität von Präferenzen akzeptieren sollten. Beispielsweise wäre eine Theorie, gemäß der *aPb gdw. a ist einfacher als b zu realisieren* gilt, keine Entscheidungstheorie, sondern eher eine Theorie der Bequemlichkeit. Das schließt aus, dass die Entscheidungstheorie unsere Urteile über die Rationalität von Präferenzen generell für falsch erklären kann, heißt aber nicht, dass der normative Gebrauch der Entscheidungstheorie vollständig parasitär auf den berücksichtigten vortheoretischen Rationalitätsurteilen ist. Die Rechtfertigung einer Theorie durch ein Überlegungsgleichgewicht setzt keineswegs voraus, dass alle Ausgangsverpflichtungen gewahrt werden. Es ist vielmehr damit zu rechnen, dass Rationalitätsurteile revidiert werden, aus

42 In der Literatur zum Überlegungsgleichgewicht wird leider oft auf der Seite des vortheoretischen «Inputs» nicht zwischen *x* und *Urteil über x* unterschieden, also beispielsweise zwischen der Praxis des Schließens und den Urteilen über die Gültigkeit von logischen Schlüssen. Vgl. dazu Hahn, op. cit. (Fn. 29) unter dem Stichwort «Praxis».

Gründen der Systematisierung oder weil sie besser gesicherten Prinzipien, Hintergrundtheorien oder auch anderen Rationalitätsurteilen widersprechen. Die Methode des Überlegungsgleichgewichts lässt auch zu, dass wir zum Schluss kommen, dass regelmäßig auftretende Rationalitätsurteile revidiert werden müssen (so beschreibt Savage, dass er seine vortheoretischen Urteile in Allais-Paradox-Situationen regelmäßig korrigiert[43]). Es gibt keine klar vorab spezifizierbare Grenze der Revidierbarkeit. Ein Überlegungsgleichgewicht kann erreicht werden, indem eine relativ kleine Anzahl vortheoretischer Urteile revidiert werden oder indem nur relativ wenige, grundlegende Urteile bewahrt werden. Zwei Faktoren haben einen zentralen Einfluss. Erstens kann man, wie ich unten diskutieren werde, geltend machen, dass substanzielle Überlegungen zur Rationalität von Präferenzen ergeben, dass sich die Revision von Rationalitätsurteilen innerhalb bestimmter Grenzen bewegen muss. Zweitens können die Ziele der Theoriebildung wesentlich mitbestimmen, in welchem Ausmaß vortheoretische Urteile revidiert werden. Für die historische Entwicklung der Entscheidungstheorie hat sicherlich das Ziel einer formalen Theorie eine bestimmende Rolle gespielt. Es liefert einen Grund, Rationalitätsurteile zu revidieren, die sich einer mathematischen Behandlung von Präferenzen und Entscheidungen entgegenstellen.

Die empirischen Resultate, die ich hier im Auge habe, nennen nun nicht Urteile über die Rationalität von Präferenzen, sondern Präferenzen, die Personen faktisch haben. Die Herausforderung an die Adresse der Entscheidungstheorie war, dass es unzulässig sei, gewisse faktischen Präferenzen für systematisch irrational zu erklären. Erstens kann man festhalten, dass damit nicht die absurde Anforderung vorgebracht wird, dass die Entscheidungstheorie alle faktischen Präferenzen als rational gelten lassen sollte. Diese Forderung wäre nicht mit der Idee verträglich, die Entscheidungstheorie normativ zu verwenden, weil es dann keine entscheidungstheoretische Kritik an Präferenzen geben könnte. Der Versuch, alle Präferenzen als rational einzustufen, würde damit enden, dass Präferenzen arational sind, weil nur rational sein kann, was auch irrational sein kann. Zweitens kann man sich fragen, ob aus den Befunden der Präferenzenforschung zusammen mit den im vorangehenden Absatz genannten Argumenten Einschränkungen für die Entscheidungstheorie abgeleitet werden können. Damit dies möglich wäre, müsste ein geeigneter Zusammenhang zwischen den Präferenzen, die eine Person faktisch hat, und ihren Rationalitätsurteilen bestehen. Das ist keine

43 Leonard J. Savage: *The foundations of statistics* (New York: Dover, 1972) S. 102-103.

triviale Frage, wie ein Vergleich mit der Moral zeigt, wo Diskrepanzen zwischen Verhalten und Urteilen an der Tagesordnung sind. Aber selbst wenn man annimmt, dass Menschen genau diejenigen Präferenzen haben, die sie als rational beurteilen, folgt aus der Argumentation im letzten Absatz lediglich, dass die Entscheidungstheorie eine angemessene Menge der Präferenzen, die Personen faktisch (nicht) haben, als (ir)rational sanktionieren muss, aber nicht, dass eine Entscheidungstheorie, die zur Folge hat, dass sich empirisch systematisch irrationale Präferenzen nachweisen lassen, nicht mit einem Überlegungsgleichgewicht gerechtfertigt werden könnte.

Epistemische Überlegungen zum Überlegungsgleichgewicht haben also bisher kein Argument dafür geliefert, dass die Befunde der Präferenzenforschung gegen die Rechtfertigung der normativ verwendeten Entscheidungstheorie sprechen, und ich sehe auch nicht, wie sich ein solches Argument allein mit Bezug auf die Methode des Überlegungsgleichgewichts sollte konstruieren lassen. Das heißt nun nicht, dass man akzeptieren muss, dass wir systematisch irrationale Präferenzen haben, aber wer dieses Resultat zurückweisen will, muss mit substanziellen Thesen über die Rationalität von Präferenzen argumentieren.

Welche Art substanzieller Thesen Grundlage für ein solches Argument sein könnten, lässt sich aus der Diskussion ablesen, die an Hempels Arbeit zum normativ-deskriptiven Doppelcharakter der Theorie des rationalen Handelns anschließt.[44] Ausgehend von der Beobachtung, dass wir denselben entscheidungstheoretischen Kalkül sowohl normativ als auch deskriptiv, für Prognosen und Erklärungen, verwenden, kann man argumentieren, dass das nur möglich ist, wenn sich Menschen typischerweise rational verhalten. Doch folgt letzteres, wie wir eben gesehen haben, nicht schon aus dem Überlegungsgleichgewicht zwischen entscheidungstheoretisch hergeleiteten und faktisch gefällten Rationalitätsurteilen, selbst wenn wir annehmen, dass faktische Präferenzen und Rationalitätsurteile übereinstimmen. Dass sich Menschen typischerweise rational verhalten, ist eine substanzielle, keine rein methodologische These. Das zeigt sich deutlich in einer Analyse von Spohn, die drei Typen von Argumenten für diese These unterscheidet:[45] Ge-

44 Carl Gustav Hempel: *Rational action*, in *The philosophy of Carl G. Hempel. Studies in science, explanation, and rationality*, hg. von James H. Fetzer (Oxford: Oxford University Press, 2001) S. 311-326.

45 Wolfgang Spohn: *Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein?*, in *Ethische Norm und empirische Hypothese*, hg. von Lutz H. Eckensberger, Ulrich Gähde (Frankfurt a.M.: Suhrkamp, 1993) S. 151-196. Spohn geht es um Rationalitätstheorie im Allgemeinen, nicht nur um Entschei-

gen eine Rationalitätsbedingung spricht erstens, wenn Menschen sich nicht (oder nicht ohne weiteres) daran halten können, und zweitens, wenn sie sich typischerweise nicht daran halten. Drittens spricht für eine Rationalitätsbedingung, dass Menschen sie typischerweise erfüllen. Solche Argumente sind nicht deduktiv, sondern nur Plausibilisierungen und können also bei wahren Prämissen zu falschen Konklusionen führen. Dass damit substanzielle Thesen über die Rationalität (von Präferenzen und Überzeugungen) verbunden sind, zeigt sich nur schon darin, dass diese Argumentationsstrategien bei anderen normativen Theorien nicht plausibel sind. Bei der Moral etwa ist zwar der Grundsatz «sollen bedingt können» einschlägig, aber wir gehen nicht davon aus, dass Verhalten und moralische Bewertung typischerweise übereinstimmen. Und es macht auch Sinn, von Normen, etwa des heiligenmäßigen Lebens, zu sprechen, die zu erfüllen Menschen faktisch kaum schaffen.[46] In den drei genannten Argumentationsstrategien zeigt sich vor allem, dass wir das Verhalten anderer Menschen unter der Annahme interpretieren, dass sie typischerweise rational sind (aber nicht unbedingt, dass sie moralisch oder heilig sind), obschon Fehlleistungen möglich sind, beispielsweise wenn relevante und bekannte Informationen übersehen oder Fehlüberlegungen angestellt werden. Diese Rationalitätsunterstellung kann und muss natürlich präzisiert und begründet werden. Was anstelle des vagen Platzhalters «typischerweise» eingesetzt werden soll («in der überwiegenden Mehrzahl der Fälle»? «solange keine empirischen Belege für das Gegenteil vorliegen»?) ist in der Literatur umstritten.[47] Im Moment ist nur relevant, ob damit systematisch auftretende Irrationalitäten ausgeschlossen sind. Spohns weitere Argumentationslinie nimmt wesentlich Bezug darauf, dass Rationalität darin besteht, dass Präferenzen und Überzeugungen durch Präferenzen und Überzeugungen begründet und über diese Begründungsbeziehung verursacht sind.[48] Sie ist damit tief in Hintergrundtheorien, besonders Handlungstheorie und Philosophie des Geistes, verankert. Auch wenn man dieser konkreten Argumentationsweise nicht folgt, zeigt sich doch, dass eine Antwort auf die

dungstheorie. Beim zweiten Typ habe ich die Bedingung weggelassen, dass die betreffenden Personen gegen Belehrungen resistent sind, weil diese Bedingung die Argumentationsweise zwar stärker, aber auch normativ macht, da sie ein Rationalitätsurteil einführt, das wegen des angestrebten Überlegungsgleichgewichts grundsätzlich berücksichtigt werden muss.

46 *Pace* Spohn, ibid. S. 171.

47 Vgl. Paul Thagard, Richard E. Nisbett: *Rationality and charity*, in *Philosophy of Science* 50 (1983) S. 250-267.

48 Spohn, op. cit. (Fn. 45) S. 176.

Frage, in welchem Maße Rationalitätsurteile bei der Theoriebildung revidiert werden können, von substanziellen Thesen zur Rationalität abhängt und nicht schon durch die Rechtfertigung der normativ verwendeten Entscheidungstheorie durch ein Überlegungsgleichgewicht gegeben ist.

Dieser Analyse steht entgegen, dass die Diskussion im Anschluss an Cohens Artikel in *The Behavioral and Brain Sciences* über weite Strecken anders geführt wurde.[49] Cohen hat geltend gemacht, mit Bezug auf die Methode des engen Überlegungsgleichgewichts könne gezeigt werden, dass empirische Studien nicht nachweisen können, dass Menschen (normal begabte, erwachsene Laien) systematisch irrational sind. Befunde, die das angeblich nachweisen, zeigten vielmehr, dass die betreffende Rationalitätstheorie inadäquat ist. Der Kern seiner Argumentation lässt sich für die Entscheidungstheorie so rekonstruieren: Unsere Präferenzen werden durch eine Entscheidungskompetenz gebildet und der Maßstab für eine korrekte Beschreibung dieser Kompetenz ist ein enges Überlegungsgleichgewicht aufgrund unserer Urteile über Entscheidungen. Aber die Entscheidungstheorie wird mit demselben Überlegungsgleichgewicht gerechtfertigt. Also muss eine adäquate Entscheidungstheorie mit einer korrekten Beschreibung der Entscheidungskompetenz übereinstimmen. Somit zeigen systematisch irrationale Präferenzen entweder, dass die ermittelten Präferenzen nicht das Produkt der Entscheidungskompetenz waren, weil diese durch Fehlleistungen überlagert wurde, oder dass die Entscheidungstheorie inadäquat ist oder dass die entsprechenden Studien methodologische Mängel haben; sie können aber niemals zeigen, dass die Entscheidungskompetenz ein irrationales Resultat liefert.[50]

Aus den Ausführungen weiter oben ergeben sich unmittelbar zwei Einwände. Erstens erfordert die Rechtfertigung der normativ verwendeten Entscheidungstheorie ein weites Überlegungsgleichgewicht, weil dabei auch Hintergrundtheorien berücksichtigt werden müssen. Zweitens setzt Cohens Argument einen anderen als den in Abschnitt 4.1 eingeführten Begriff des engen Überlegungsgleichgewichts voraus. Bei Cohen ist es Teil einer Rechtfertigung durch ein enges Überlegungsgleichgewicht, dass vortheoretische Urteile gegenüber systematisierenden Prinzipien soweit Priorität haben, dass systematische Revisionen ausgeschlossen sind.

[49] Siehe Fn. 28.

[50] Cohen, op. cit. (Fn. 28) S. 317-318, 321-323. Eine andere Analyse findet sich in Stein, op. cit. (Fn. 1).

Das ist noch keine Widerlegung von Cohens Position, sondern macht vor allem zwei zentrale Punkte in seiner Auffassung sichtbar: Cohen vertritt die These, dass Menschen über eine Kompetenz des Entscheidens verfügen, und er verwendet einen auf Kompetenzen zugeschnittenen Begriff des engen Überlegungsgleichgewichts.[51] Im ersten Punkt bezieht er sich auf Chomskys Unterscheidung zwischen grammatischer Kompetenz (impliziter Kenntnis von Sprachregeln) und Performanz (faktischem Sprachgebrauch) eines Sprechers.[52] Cohens Begriff des engen Überlegungsgleichgewichts zeichnet sich dadurch aus, dass Ausgangsverpflichtungen auf partikuläre Urteile beschränkt sind, dass Prinzipien nur mit dem Ziel einer systematischen Beschreibung entwickelt werden und dass Revisionen der zugrundegelegten Urteile sich darauf beschränken, Performanz-Fehler zu beseitigen, das heißt, unter ungünstigen Umständen zustande gekommene Fehlurteile zu korrigieren und allfällige Inkonsistenzen zwischen Urteilen zu eliminieren.[53] Weitergehende Änderungen der Urteile im Namen der Systematisierung der Prinzipien oder aufgrund von Argumenten, die sich auf Hintergrundtheorien beziehen, sind nicht vorgesehen. Dieser Begriff des engen Überlegungsgleichgewichts ist wesentlich enger als der oben eingeführte Begriff. Cohen baut somit substanzielle Annahmen über den Begriff der Rationalität und die Aufgabe einer Rationalitätstheorie in seine Argumentation mit dem Überlegungsgleichgewicht ein: Rationalität ist eine Kompetenz und die Rationalitätstheorie soll diese Kompetenz adäquat beschreiben.

Das entscheidende Problem mit Cohens Auffassung ist: Was Rawls für die Moral geltend gemacht hat,[54] gilt auch für die Entscheidungstheorie, sie hat nicht das Ziel, eine Kompetenz zu beschreiben. Eine Entscheidungstheorie kann sinnvollerweise darauf abzielen, Bedingungen für bessere Entscheidungen zu formulieren, als sie unsere Kompetenz liefert.[55] Und das

51 Cohen identifiziert seine Charakterisierung des engen Überlegungsgleichgewichts explizit mit Goodmans Position und dem engen Überlegungsgleichgewicht bei Rawls und Daniels (Cohen, ibid. S. 317, 320). Hier ist nicht der Ort, um Vorbehalte gegen diese Interpretation, insbesondere von Goodman, zu diskutieren. Vgl. die Beiträge von Margalit/Bar-Hillel, Zabell und Daniels/Smith, in Cohen, ibid.

52 Noam Chomsky: *Aspects of the theory of syntax* (Cambridge, MA: MIT Press, 1969) S. 3-4.

53 Cohen, op. cit. (Fn. 28) S. 320-323.

54 Rawls, *Theory of justice*, op. cit. (Fn. 29) S. 43.

55 Für die folgende Argumentation vgl. Thagard/Nisbett op. cit. (Fn. 47) und Daniels/Smith, in Cohen, op. cit. (Fn. 28) S. 490-491.

bedeutet, dass eine so verstandene Entscheidungstheorie damit rechnen kann, dass sich systematisch irrationale Präferenzen empirisch nachweisen lassen. Die Grammatik der Sprachen, die wir tatsächlich sprechen, kann hingegen, wenn wir darunter die Beschreibung unserer Sprachkompetenz verstehen, nicht sinnvollerweise als eine Theorie «besserer» grammatischer Strukturen verstanden werden. Sprachverbesserungsprojekte gibt es natürlich auch, aber sie beschreiben nicht die Kompetenz für eine natürliche Sprache. Eine Konsequenz ist, dass es keinen Sinn macht, aufgrund von theoretischen Überlegungen zum Schluss zu kommen, dass gewisse Grammatikalitätsurteile kompetenter Sprecher als solche problematisch sind, während das analoge Ergebnis bei Präferenzstrukturen sehr wohl möglich ist. Es wäre absurd, zu behaupten, kompetente Sprecher des Deutschen würden systematisch den Fehler machen, «besser» statt «güter» als richtigen Komparativ von «gut» zu beurteilen oder zu argumentieren, aus Gründen der Einfachheit sollten alle Komparative regelmäßig gebildet werden. Hingegen ist die Behauptung, der in Abschnitt 2 beschriebene Anziehungseffekt zeige irrationale Präferenzen, nicht sinnlos, obschon sie natürlich bestritten werden kann. Und auch Geldpumpenargumente sind zwar umstritten, aber nicht schon deshalb irrelevant, weil es sinnlos ist, beispielsweise aufgrund einer pragmatischen Theorie des Handelns und Präferierens, für entscheidungstheoretische Normen zu argumentieren.

Man kann dieses Ergebnis auch dahingehend deuten, dass es Cohen um einen anderen als den für die Entscheidungstheorie relevanten Begriff der Rationalität geht.[56] Cohens «Rationalität» ist die Rationalität des *animal rationale*, eine natürliche Fähigkeit zum Bilden von Präferenzen und Überzeugungen, die allen Menschen, genauer, allen normalbegabten erwachsenen Laien, zukommt und die er als eine Kompetenz im Sinne Chomskys versteht. In der Entscheidungstheorie geht es dagegen um Rationalität in einem anspruchsvolleren Sinne, wobei sich die Rationalitätsideale verschiedener Entscheidungstheorien durchaus unterscheiden können. Rationalität ist für diese normativ verwendeten Theorien jedenfalls ein Ideal, das zwar von Menschen entwickelt wird, dem nachzuleben aber allenfalls selbst den Entscheidungstheoretikern nicht leicht fällt.[57] Weil sich die normativ verwendete Entschei-

[56] Eine solche Diagnose ist implizit in Thagard/Nisbett, ibid. S. 251, die Rationalität als Übereinstimmung mit den besten zurzeit verfügbaren normativen Standards auffassen, und das ist klar kein Kompetenz-Begriff.

[57] Eine ähnliche Unterscheidung zwischen Rationalität als «meeting-the-minimal-standards» und «meeting-the-maximal-or-ideal-standards» trifft Robert Hanna:

dungstheorie mit Rationalität in diesem Sinne beschäftigt, ist Cohens Begriff des engen Überlegungsgleichgewichts für deren Rechtfertigung unangemessen. Und weil es unterschiedlich anspruchsvolle Rationalitätsideale gibt, ist auch die Dichotomie zwischen einer allgemeinen Rationalitätsannahme und einer allgemeinen Irrationalitätsannahme zurückzuweisen.[58] Vielmehr ist damit zu rechnen, dass Menschen bestimmte Aspekte eines Ideals erfüllen, andere hingegen nicht.

5. Fazit

Insgesamt ziehe ich folgendes Fazit: Man kann nicht allein unter Bezug auf die Methode des Überlegungsgleichgewichts argumentieren, dass systematisch irrationale Präferenzen gegen die Entscheidungstheorie sprechen. Es kann lediglich gefordert werden, dass nicht alle unsere Rationalitätsurteile revidiert werden können und dass unsere faktischen Präferenzen nicht schon als rational gelten, nur weil wir sie haben. Sofern eine gewisse Einheit von normativ (als Rationalitätsbeurteilung) und deskriptiv (für Prognosen und Erklärungen) verwendeter Entscheidungstheorie gefordert wird, muss das mit substanziellen Thesen über Rationalität, Präferenzen und Überzeugungen begründet werden, beispielsweise, indem man argumentiert, dass ohne Rationalitätsunterstellungen anderen Menschen keine Präferenzen und Überzeugungen zugeschrieben werden können. Es sollte also in der Debatte um die «richtige» Entscheidungstheorie darum gehen, welche empirischen Befunde in welchem Maße die normative Theorie prägen sollen, und nicht darum, ob der normative Gebrauch der Entscheidungstheorie den deskriptiven ignorieren kann oder darin aufgehen sollte. Damit diese Diskussion sinnvoll geführt werden kann, muss man zusätzlich berücksichtigen, dass empirische Befunde erst auf dem Hintergrund einer Theorie über die Anwendung entscheidungstheoretischer Formalismen überhaupt relevant werden.

Rationality and logic (Cambridge, MA: MIT Press, 2006). Hanna lässt allerdings keine Abstufungen zwischen maximal und minimal zu und vermengt überdies minimale Rationalität mit dem Gegensatz zu Arationalität. Wenn aber z.B. nicht normal begabte Menschen minimale Standards nicht erfüllen, sind sie irrational, nicht arational.

58 Gegen Hanna, ibid. S. 128. Grandy diagnostiziert diesen Fehler der falschen Dichotomie bei Cohen, op. cit. (Fn. 28) S. 494.

Studia philosophica 68/2009

Hans Rott

Seltsame Wahlen. Zur Rationalität vermeintlicher Anomalien beim Entscheiden und Schlussfolgern

This paper discusses a number of apparent anomalies in rational choice scenarios, and their translation into the logic of everyday reasoning. Three classes of examples that have been discussed in the context of probabilistic choice since the 1960s (by Debreu, Tversky and others) are analyzed in a non-probabilistic setting. It is shown how they can at the same time be regarded as logical problems that concern the drawing of defeasible inferences from a given information base. I argue that initial appearances notwithstanding, these cases should not be classed as instances of irrationality in choice or reasoning. One way of explaining away their apparent oddity is to view certain aspects of these examples as making particular options salient. The decision problems in point can then be solved by 'picking' these options, although they could not have been 'chosen' in a principled way, due to ties or incomparabilities with alternative options.

1. Rationales Wählen

Die klassische Theorie der rationalen Wahl stellt ein Paradigma für rationales Handeln dar. Es sind die Präferenzen einer Person, die ihre Handlungen bestimmen. Eine rationale Person entscheidet sich für diejenige Option, die ihre Präferenzen maximiert. Dabei sind die Präferenzen inhaltlich nicht bestimmt. Sie sind in der Regel nicht egoistisch-gewinnmaximierend im Sinne des Entscheiders zu verstehen, sondern können durchaus auch moralische Maximen und soziale Güter berücksichtigen. Im Folgenden symbolisiere σ eine sogenannte *Auswahlfunktion*, die für jede Menge S von potentiell zur Auswahl stehenden Optionen – für jedes *Menü*[1] S – die Teilmenge der optimalen, also rationalerweise wählbaren Elemente bestimmt. Die Menge

[1] Der Terminus *menu* ist in der englischen Fachliteratur gebräuchlich und wird von dort übernommen. «Menü» soll hier gerade nicht zu einer vorgegebenen, fixierten Speisenfolge analog sein, sondern zu einer Speisekarte, die ein Angebot beschreibt, aus dem der Gast auswählen kann. Im Deutschen kann man etwa an die Menüs in Computerprogrammen denken.

$\sigma(S)$ ist also die Menge derjenigen Optionen im Menü S, die für eine durch σ charakterisierte Person bestmögliche Lösungen des Wahlproblems darstellen würden. Wichtig ist, dass eine Funktion σ für viele verschiedene Menüs definiert ist. Mit welchem Menü S die Person in Wirklichkeit konfrontiert wird, ist kontingent und für das Folgende unwichtig. Insofern repräsentiert σ eine Menge von *potentiellen* Wahlen oder, anders gesagt, die Wahl*dispositionen* der Person. Für eine rationale Person gilt nach dem eingangs Gesagten also: $\sigma(S) = \{x \text{ in } S\text{: es gibt kein } y \text{ in } S \text{ mit } x<y\}$.[2] Für eine rationale Person sind genau die bezüglich ihrer Präferenzrelation $<$ maximalen Elemente wählbar. Wichtig ist, dass die Präferenzrelation nicht von der (durch S repräsentierten) Wahlsituation abhängt. Die Wahldispositionen einer Person heißen *rationalisierbar*, wenn eine feste Präferenzrelation existiert, mit der die Auswahlen dieser Person als präferenzmaximierend erklärt werden können.

Intuitiv wird man sich die Wahlsituation als vergleichsweise flüchtig vorstellen: Welche Möglichkeiten der Person in einer gewissen Situation offenstehen, so wie sie abstrakt in einer Menge S zusammengefasst sind, ist mehr oder weniger kontingent, zufällig zu denken. Hingegen ist es Sache des eher beständigen, stabilen Charakters, was man unter welchen Umständen als wählbar empfindet.[3] In diesem Bild ist das Wählen und Handeln des Menschen durch seinen Charakter und durch die vorliegende Wahlsituation *determiniert*. Ein rationales Wesen *muss* so handeln, dass (oder: als ob) es seine Präferenzen maximiert(e). Denn die Theorie der rationalen Wahl postuliert, dass es eine quer über verschiedene mögliche Wahlsituationen stabile, entscheidungsleitende Präferenzrelation besitzt.[4] Demnach ist es in seinem Handeln gleichsam durch seine eigenen Präferenzen programmiert, seine Handlungen *realisieren* und *offenbaren* seine Präferenzen. Man könnte nun denken, diese Konzeption widerspräche der Idee der Freiheit des Menschen. Die Natur des Menschen ist demnach nämlich mit der eines

2 *Rationalisierbarkeit* wird nach dieser Konzeption mit *Relationalisierbarkeit* gleichgesetzt. Ich nehme hier asymmetrische Relationen $<$ als primitiv (d.h. als nicht abgeleitet von einer symmetrischen Relation $\leq$) an.

3 Vgl. Guido Löhrer: *Charakterstabilität und diachrone Kohärenz. Zurechenbarkeit im Prozess moralischen Umdenkens,* in *Zeitschrift für philosophische Forschung* 60 (2006) S. 46-71.

4 Die Unabhängigkeit der Präferenzen oder Wertschätzung jeder einzelnen Option von der Menge der gerade verfügbaren Optionen ist hier entscheidend. Daniel McFadden (*Economic Choices*, in *American Economic Review* 91 [2001] S. 351-378, hier S. 356) hat in seiner Nobelpreisrede hierfür den folgenden Slogan geprägt: *Desirability precedes availability*.

Schachcomputers vergleichbar, der auch immer denjenigen Zug wählt und wählen muss, der in seiner internen Bewertung als bester beurteilt wird.[5] Dennoch scheint es nicht adäquat, hier von einer Freiheitsbeschränkung zu sprechen. Was dazu zu sagen ist, hat schon John Locke treffend in die folgenden Worte gefasst:

> *To be determined by our own judgment, is no restraint to Liberty.* This is so far from being a restraint or diminution of *Freedom*, that it is the very improvement and benefit of it; … 'tis as much *a perfection, that desire or the power of Preferring should be determined by Good,* as that the power of Acting should be determined by the *Will* and the certainer such determination is, the greater is the perfection. Nay were we determined by any thing but the last result of our own Minds, judging of the good or evil of any action, we were not free, the very end of our Freedom being, that we might attain the good we chuse. And therefore every Man is put under a necessity by his constitution, as an intelligent Being, to be determined in *willing* by his own Thought and Judgment, what is best for him to do: else he would be under the determination of some other than himself, which is want of Liberty.[6]

Was Freiheit schafft, ist, dass es sich bei den Urteilen, Gedanken oder Präferenzen um *die der Person eigenen* («our own», «his own») Urteile, Gedanken oder Präferenzen handelt. Sind die Präferenzen der Person einmal gegeben, ergibt es keinen Sinn, sich darüber zu beklagen, dass diese Sklavin ihrer Präferenzen sei. Wenn Kritik geübt werden kann, dann an den Präferenzen selbst, aber deren Ausbildung gehört nicht zum Gegenstandsbereich der Theorie der rationalen Wahl.[7]

Es ist prima facie erstaunlich, dass allein die Tatsache, dass sich die Wahldispositionen von einer zugrunde liegenden Präferenzrelation herleiten, bestimmte strukturelle Eigenheiten dieser Wahldispositionen festlegt. Wenn etwa eine Option x optimal in S ist, d.h. wenn sie in $\sigma(S)$ ist, und x durch die Einschränkung des Menüs S auf ein Teilmenü S' nicht ausgeschlossen wird, dann ist x auch in $\sigma(S')$. Wenn eine Option y optimal in einem Menü S

5 Der Vergleich mit dem Schachcomputer ist noch vergleichsweise schmeichelhaft. Kant sprach von der «Freiheit eines Bratenwenders …, der auch, wenn er einmal aufgezogen worden, von selbst seine Bewegungen verrichtet» (*Kritik der praktischen Vernunft*, A 174/Akademie-Ausgabe V 97).

6 John Locke: *An Essay Concerning Human Understanding*, hg. von Peter H. Nidditch (Oxford: Oxford University Press, 1975 [orig. 1690]), Abschnitt II.xxi.48, S. 264.

7 Vgl. hierzu Georg Brun: *Wer hat ein Problem mit irrationalen Präferenzen? Entscheidungstheorie und Überlegungsgleichgewicht* in diesem Band, S. 11-41.

und ebenso optimal in einem anderen Menü S' ist, dann ist y auch optimal in der Vereinigungsmenge $S \cup S'$. Diese beiden Sätze gelten allein aufgrund der Existenz einer rationalisierenden Präferenzrelation, völlig unabhängig davon, welche Form und welchen Inhalt diese Relation hat. Wie tief die Beziehungen zwischen den Eigenschaften von Auswahlfunktionen über endlichen Bereichen und Eigenschaften von rationalisierenden Präferenzen gehen, fasst folgendes wohlbekanntes Theorem zusammen.[8] Alle Menüs sind hierbei als endlich vorausgesetzt.

(a) σ ist rationalisierbar genau dann, wenn σ durch die folgendermaßen definierte «offenbarte» Präferenzrelation rationalisierbar ist:
$y < x$ genau dann, wenn gilt: x ist in, aber y ist nicht in $\sigma(\{x,y\})$
«Basispräferenzen»

(b) σ ist rationalisierbar genau dann, wenn es (I) und (II) erfüllt:
(I) Wenn $S \subseteq S'$, dann $S \cap \sigma(S') \subseteq \sigma(S)$ Sens Bedingung α
(II) $\sigma(S) \cap \sigma(S') \subseteq \sigma(S \cup S')$ Sens Bedingung γ

(c) σ ist transitiv rationalisierbar genau dann, wenn es (I), (II) und (III) erfüllt:
(III) Wenn $S \subseteq S'$ und $\sigma(S') \subseteq S$, dann $\sigma(S) \subseteq \sigma(S')$ Aizermans Bedingung

(d) σ ist modular[9] rationalisierbar genau dann, wenn es (I) und (IV) erfüllt:
(IV) Wenn $S \subseteq S'$ und $\sigma(S') \cap S \neq \emptyset$, dann $\sigma(S) \subseteq \sigma(S')$
Sens Bedingung β+

Die Modularität impliziert zusammen mit der Asymmetrie die Transitivität der Relation $<$. Intuitiv liegt Modularität genau dann vor, wenn alle Elemente hinsichtlich ihrer Wünschbarkeit vergleichbar sind. In finiten Bereichen ist dies gleichwertig damit, dass den Optionen numerische Werte zugeordnet werden können, nach denen sich die Präferenzrelation richtet.

Die hiermit charakterisierte klassische Theorie der rationalen Wahl ist philosophisch plausibel motiviert und formal von bestechender Eleganz. Leider hat sie auch Probleme, die schon früh in ihrer Entwicklung erkannt wurden. Ein bekanntes Beispiel stammt von Luce und Raiffa (1957). Man vergleiche zwei alternative Szenarien eines Restaurantbesuchs. Im ersten Szenario habe das Restaurant nur Lachs *(A)* und Steak *(B)* auf der

8 Vgl. Amartya K. Sen: *Choice Functions and Revealed Preference*, in *Review of Economic Studies* 38 (1971) S. 307-317 (auch in A. K. S.: *Choice, Welfare and Measurement* [Oxford: Blackwell, 1982] S. 41-53) und Hervé Moulin: *Choice Functions Over a Finite Set – A Summary*, in *Social Choice and Welfare* 2 (1985) S. 147-160.

9 Eine Relation $<$ heißt modular, wenn $x<y$ impliziert, dass für beliebiges z entweder $x<z$ oder $z<y$ gilt.

Speisekarte, und ein Besucher entscheide sich für Lachs. Obgleich dieser Besucher eigentlich Steak bevorzugt, nimmt er davon Abstand, weil ihn die Beschränktheit der Speisekarte befürchten lässt, dass das Lokal das Steak verderben könnte. Im zweiten (alternativen, nicht sukzessiv anschließenden) Szenario liege eine Speisekarte vor, die außer Lachs und Steak auch noch Schnecken *(C)* und Froschschenkel *(D)* offeriert. In dieser Situation entscheidet sich der Besucher für Steak. Obgleich er Schnecken und Froschschenkel nicht mag, zeigt deren Nennung auf der Karte, dass es sich um ein gutes Restaurant handeln muss, dem die kunstgerechte Zubereitung eines Steaks durchaus zugetraut werden kann.

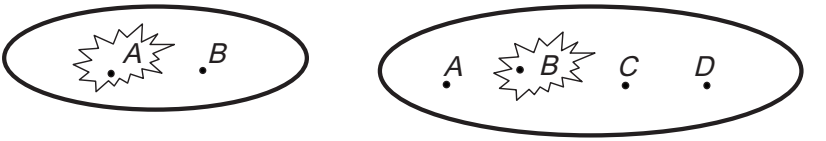


Abb. 1: Das Beispiel von Luce und Raiffa

Intuitiv deutet nichts an solchen Szenarien darauf hin, dass der Besucher irrational ist. Doch verletzen diese Wahldispositionen Sens Bedingung α und Aizermans Bedingung: Option *B* wird im zweiten Szenario ausgewählt, nicht jedoch im ersten, restringierten Szenario, obwohl es dort doch auch zur Wahl angeboten wird. Option *A* wird im ersten Szenario ausgewählt, verliert im zweiten Szenario aber diesen Status, obwohl alle optimalen Lösungen des letzteren Szenarios auch im ersten Szenario zur Verfügung stehen. Der Grund für dieses seltsame Wahlverhalten kann darin gesehen werden, dass das Menü im Beispiel nicht nur die zur Verfügung stehenden Alternativen auflistet, sondern auch selbst einen *informativen Wert* hat: Die Speisekarten in den jeweiligen Szenarien transportieren die (nicht notwendigerweise zutreffende) Information, dass das Restaurant ein eher mäßiges bzw. ein eher ambitioniertes ist. Und diese Umstände scheinen zu einer Verletzung der Bedingungen der klassischen Theorie rationaler Wahl zu führen. Gerade die durch die Speisekarte vermittelte Information nährt allerdings Zweifel an der Adäquatheit der Repräsentation des Beispielfalls. Während man im ersten Szenario Steak-in-einem-mäßigen-Restaurant angeboten bekommt, handelt es sich im zweiten Szenario um Steak-in-einem-ambitionierten-Restaurant, und Analoges gilt natürlich für den Lachs. Die Speisekarten sind eigentlich also gar nicht vergleichbar.

2. *Wählen und Schlussfolgern*

Eine ähnliche Problemlage bietet Rott (2004), wo der Fokus nicht auf dem Auswählen von Gütern oder Handlungsoptionen, sondern auf der rationalen Herausbildung von Meinungen oder Überzeugungen liegt.[10] Im dort ausgeführten Beispiel geht es um die Vergabe einer (und nur einer) Stelle für Metaphysik in einem Philosophie-Institut. Es gibt vier verbliebene Bewerberinnen und Bewerber: Amanda Andrews *(A)* ist eine ausgezeichnete Metaphysikerin, Bernice Becker *(B)* eine sehr gute Metaphysikerin mit substantiellen Logik-Kompetenzen, Carlos Cortez *(C)* ein brillanter Logiker, der auch etwas Metaphysik betrieben hat, und schließlich David Donaldson *(D)*, der allgemein für den offensichtlichen Sieger des Wettbewerbs gehalten wird. Diese Informationen seien öffentlich bekannt. Dazu erhalten wir aber nun exklusive – und überraschende – Informationen vom Vorsitzenden der Berufungskommission, von dem man weiß, dass er sich in dieser Sache nicht irrt, nicht lügt, keine Scherze macht und auch sonst nicht in die Irre führen möchte. Drei alternative (wieder nicht: aufeinander folgende) Szenarien sollen nun betrachtet werden. In Szenario 1 gibt der Vorsitzende bekannt, dass die Stelle an *A* oder *B* gehen wird. Auf diese Information hin glauben wir, dass *A* die Stelle bekommen wird, die ja für Metaphysik ausgeschrieben war und für die *A* deshalb besser qualifiziert scheint als *B*. In Szenario 2 sagt der Vorsitzende, dass *C* die Stelle erhalten wird, woraufhin wir unmittelbar schließen, dass *C* die Stelle erhalten wird. In Szenario 3 erfahren wir vom Vorsitzenden, dass die Stelle entweder an *A* oder an *B* oder an *C* gehen wird. Hier nun setzt eine etwas komplizierte Überlegung ein. Aus der Tatsache, dass *C* ein ernsthafter Kandidat ist, folgern wir, dass eine Expertise in Logik bei der Vergabe der Stelle eine gewisse Rolle spielt. Dennoch halten wir *C* nicht für den besten Kandidaten. Aber jetzt können wir zu dem Schluss kommen, dass *A*'s Vorsprung in Metaphysik auf *B* zu gering ist, um *B*'s Logik-Kompetenzen aufzuwiegen.[11]

[10] Hans Rott: *A Counterexample to Six Fundamental Principles of Belief Formation,* in *Synthese* 139 (2004) S. 225-240.

[11] Rott (op. cit.) gibt eine numerische Unterfütterung des Beispiels, welches diese Situation für eine vergleichsweise große Bandbreite von Parametern plausibel macht. Kritische Diskussionen dieses Beispiels finden sich jetzt in Brian Hill: *Towards a «Sophisticated» Model of Belief Dynamics. Part II: Belief Revision,* in *Studia Logica* 89 (2008) S. 291-323; Robert Stalnaker: *Iterated Belief Revision,* in *Erkenntnis* 70 (2009) S. 189-209; Horacio Arló-Costa, Arthur P. Pedersen: *Social Norms, Rational Choice and Belief Change*, in *Science in Flux: Belief Revision*

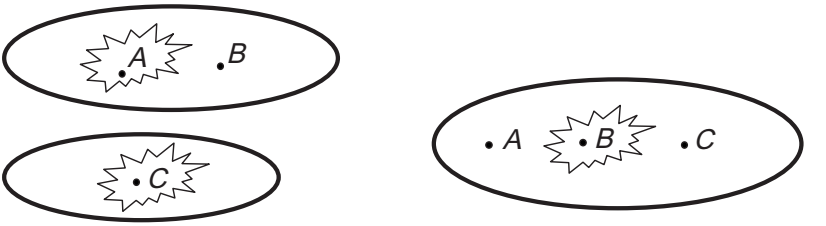


Abb. 2: Das Beispiel der Berufungskommission

Man kann sich durch einen Vergleich von Szenario 1 und Szenario 3 überzeugen, dass – ähnlich wie im Beispiel von Luce und Raiffa – Sens Bedingung α und Aizermans Bedingung verletzt sind. Auch die Diagnose erscheint zunächst analog. Dass der Logiker Cortez noch in der Konkurrenz ist, zeugt von einer gewissen Relevanz des Fachs Logik für die Besetzung der Stelle, ähnlich wie Schnecken und Froschschenkel, obgleich selbst nicht erste Wahl, von der Qualität des Restaurants zeugen. Doch gibt es hier einen wichtigen Unterschied. Während es in Luce und Raiffas Beispiel tatsächlich um eine echte Wahl ging, die zu treffen war («Welche Bestellung soll getätigt werden?»), haben wir im Fall des Berufungsverfahrens nichts zu entscheiden. Wir werden lediglich informiert, und zwar in einer Weise, die unseren Vormeinungen widerspricht (wie alle anderen waren wir ja davon ausgegangen, dass Donaldson die Stelle bekommen würde). Es handelt sich hier nicht um ein praktisches, sondern um ein theoretisch-kognitives Problem, in dem die Präferenzen nicht Wertschätzung oder Nutzen repräsentieren, sondern Grade von Plausibilität. Dass wir etwa in Szenario 1 Andrews gegenüber Becker bevorzugen, heißt, dass wir es dort für *plausibler* oder *eher möglich* halten, dass jene als dass diese die Stelle bekommen wird. Die Präferenzrelation gibt hier doxastische, keine evaluativen oder volitiven Einstellungen wieder.

Es stellt sich die Frage, ob auch auf solche Arten von Präferenzen die Theorie der rationalen Wahl sinnvoll angewendet werden kann. Diese Frage wurde spätestens in den 1990er Jahren beantwortet. Die Theorie der rationalen Wahl lässt sich hier sehr gut in Anschlag bringen und erzeugt, je nach Stärke der Bedingungen, die man an die Rationalität von Wahlen stellen will, eine Palette von Logiken, die aus der einschlägigen Literatur bereits bekannt und unabhängig motiviert sind. In welchem Sinne kann man hier

in the Context of Scientific Inquiry, hg. von Erik J. Olsson (Dordrecht: Springer, im Erscheinen).

von einer Logik reden? Das letzte Beispiel fortsetzend, wollen wir die Sätze «*A* wird die Stelle bekommen», «*B* wird die Stelle bekommen» usw. mit *«a»*, *«b»* usw. abkürzen. Wir werden die Situation nun so darstellen, dass aus variierenden Informationen des Kommissionsvorsitzenden verschiedene Schlussfolgerungen zu ziehen sind, und diese Schlussfolgerungen als logische *Inferenzen* bezeichnen. Zusammengefasst stellen sich die Szenarien nun folgendermaßen dar. In der Anfangssituation, in der noch keine zusätzliche Information des Kommissionsvorsitzenden vorliegt, glauben wir, dass *D* die Stelle bekommen wird:[12]

$$d \in \mathit{Inf}(\mathsf{T}) \qquad \text{und} \qquad \neg a, \neg b, \neg c \in \mathit{Inf}(\mathsf{T})$$

In den verschiedenen Szenarien ändern sich die Inferenzen aufgrund der neuen Informationen entsprechend:

Szenario 1	$a \in \mathit{Inf}(a \vee b)$	und	$\neg b, \neg c, \neg d \in \mathit{Inf}(a \vee b)$
Szenario 2	$c \in \mathit{Inf}(c)$	und	$\neg a, \neg b, \neg d \in \mathit{Inf}(c)$
Szenario 3	$b \in \mathit{Inf}(a \vee b \vee c)$	und	$\neg a, \neg c, \neg d \in \mathit{Inf}(a \vee b \vee c)$

«Inf» bezeichnet hier eine Inferenzfunktion, die jedem Satz die Menge seiner Schlussfolgerungen oder Konsequenzen zuordnet.[13] Wir betrachten *Inf* als eine logische Operation. Doch ist auffällig, dass eine Verstärkung der Prämisse im klassischen Sinne nicht zu einer Vermehrung der Konsequenzen führt. Die Prämisse $a \vee b$ des ersten Szenarios ist stärker, d.h. inhaltsreicher als die Prämisse $a \vee b \vee c$ des dritten Szenarios, doch fehlt die Konsequenz b, die man im dritten Szenario ziehen kann, im ersten Szenario. Sie ist hier sogar durch die Konsequenz $\neg b$ zu ersetzen. Wir können damit eine *erste zentrale Feststellung über das Schlussfolgern in alltäglichen* (nicht logischen oder mathematischen) *Kontexten* treffen:

[12] Im Folgenden stehen die Zeichen «¬» und «∨» für die Negation «nicht» und die Disjunktion «oder», «T» (das «Verum») ist die logische Satzkonstante, die unter allen möglichen Umständen wahr ist. Später werden wir auch die Zeichen «∧» und «→» für die Konjunktion «und» und das materiale Konditional «wenn … dann» verwenden.

[13] Auch endliche Prämissenmengen können durch vorhergehende Konjunktionsbildung von *Inf* verarbeitet werden, für unendliche Prämissenmengen ist die Erweiterung aber nicht trivial.

> Das Schlussfolgern im Alltag ist *nichtmonoton*; das heißt, wenn ein Satz ϕ aus einem anderen Satz ψ nach den Regeln der klassischen Logik folgt, garantiert dies im allgemeinen nicht, dass die Schlussfolgerungen von ϕ in den Schlussfolgerungen von ψ enthalten sind.[14]

Diese Feststellung drückt eigentlich nichts Neues aus und ist aus den Diskussionen über Induktion und induktive Logik wohlbekannt. Doch wurde ihre volle Reichweite erst in den letzten 30 Jahren erkannt, in denen sogenannte nichtmonotone Logiken zu einem eigenen Forschungsgebiet avancierten.[15] Man könnte nun vermuten, dass mit dem Wegfall der Eigenschaft der Monotonie jegliche Regelmäßigkeit des Schlussfolgerns verlorengeht. Dies ist aber nicht der Fall. Die zweite zentrale Feststellung über das Schlussfolgern in alltäglichen Kontexten lautet nämlich:

> Der Verlust der Monotoniebedingung für das Schlussfolgern im Alltag bedeutet kein *Anything goes* in diesem Schlussfolgern; vielmehr können viele substantielle logische Gesetze erhalten bleiben.

Es ist der Stand der Forschungen zur nichtmonotonen Logik, dass das alltägliche Schlussfolgern auch im Angesicht des Scheiterns der Monotoniebedingung beträchtliche Regelmäßigkeiten aufweist. Die folgenden logischen Gesetze dienen nicht nur als Beispiele hierfür, sondern sind für unsere weitere Argumentation von besonderem Interesse. Sie wurden in der Literatur zum nichtmonotonen Schließen als besonders plausible und in hohem Maße erwünschte Bedingungen für alltägliches Schließen verstanden.

(Kumulative Monotonie)	Wenn $\psi \in Inf(\phi)$, dann $Inf(\phi) \subseteq Inf(\phi \wedge \psi)$
(Oder)	$Inf(\phi) \cap Inf(\psi) \subseteq Inf(\phi \vee \psi)$
(Konditionalisierung)	$Inf(\phi \wedge \psi) \subseteq Cn(Inf(\phi) \cup \{\psi\})$

[14] Solange wir bei endlichen Prämissenmengen bleiben und diese mit ihren Konjunktionen identifizieren, heißt dies auch, dass eine Teilmenge X einer Prämissenmenge Y Konsequenzen haben kann, die Y selbst nicht hat. Im Übrigen beachte man beim Gebrauch der Variablen, dass a, b, c usw. für bestimmte Sätze in unseren Beispielen stehen, während ϕ und ψ Platzhalter für beliebige Sätze sind.

[15] Vgl. Wolfgang Spohn: *Induktion*, in *Logik in der Philosophie*, hg. von Wolfgang Spohn, Peter Schroeder-Heister, Erik J. Olsson (Heidelberg: Synchron Wissenschaftsverlag der Autoren, 2005) S. 137-159 zur Induktion sowie Matthew L. Ginsberg (Hg.): *Readings in Nonmonotonic Reasoning* (Los Altos, CA: Morgan Kaufmann, 1987) und David Makinson: *Bridges from Classical to Nonmonotonic Logic* (London: King's College Publications, 2005) zu nichtmonotonen Logiken.

In der Notation dieser Bedingungen verwenden wir weiterhin die alltägliche Inferenzoperation *Inf* und setzen davon eine monotone, im wesentlich klassisch zu denkende Hintergrundlogik *Cn* ab.[16] So besagt unsere Bedingung der Nichtmonotonie nun, dass aus $\phi \in Cn(\psi)$ nicht $Inf(\phi) \subseteq Inf(\psi)$ folgt.

Das oben diskutierte Beispiel der Berufungskommission verletzt, als Übung im Schlussfolgern betrachtet, sämtliche dieser Bedingungen. Erstens ist – entgegen Oder – $Inf(a \vee b) \cap Inf(c)$ keine Teilmenge von $Inf(a \vee b \vee c)$, zweitens ist – entgegen Konditionalisierung – $Inf(a \vee b)$ keine Teilmenge von $Cn(Inf(a \vee b \vee c) \cup \{a \vee b \vee \neg c\})$. Sowohl der Satz $a \vee c$ als auch der Satz $\neg b$ kann als Beleg für beides dienen. Drittens ist zwar $a \vee b$ in $Inf(a \vee b \vee c)$, aber es ist – entgegen Kumulativer Monotonie – $Inf(a \vee b \vee c)$ keine Teilmenge von $Inf(a \vee b)$. Dies belegen beispielhaft die Sätze b und $\neg a$.[17]

Man kann nun sagen, die erwähnten logischen Bedingungen werden *gerade deshalb* verletzt, *weil* zugrunde liegende Prinzipien der rationalen Wahl verletzt werden. Die Verletzung von Oder und Konditionalisierung liegt daran, dass Sens Bedingung α verletzt ist: B ist eine beste Option in $\{A,B,C\}$ und natürlich auch eine Option in $\{A,B\}$, doch B ist keine beste Option in $\{A,B\}$. Die Verletzung der Kumulativen Monotonie hingegen ist auf eine korrespondierende Verletzung von Aizermans Bedingung zurückzuführen: Alle besten Optionen von $\{A,B,C\}$ liegen innerhalb von $\{A,B\}$, jedoch ist A eine beste Option von $\{A,B\}$, die keine beste Option von $\{A,B,C\}$ ist. Seltsame Wahlen sind also die Grundlage von seltsamen Inferenzen.

Wie kann diese Intuition, dass logische Bedingungen sich aus Bedingungen rationaler Wahl ergeben, substantiiert werden? Man kann Wahlen auf zwei verschiedenen Ebenen ansetzen. Auf der semantischen Ebene besteht das im Alltag verwendete plausible Schließen in der *Wahl bester möglicher Welten*: Wenn eine Person die Prämisse ϕ als Information zur Verfügung hat,[18] so versucht sie die plausibelsten, d.h. nächstliegenden Welten zu finden, in denen ϕ wahr ist. Auf der syntaktischen Ebene handelt es sich um eine *Wahl «schlechtester» Meinungen*: Wenn eine Person die Prämisse ϕ als Information zur Verfügung hat, so versucht sie die unplausibelsten, d.h. die am schwächsten begründeten Sätze zu finden, die zur Implikation der

16 Vgl. Kapitel 1 in Hans Rott: *Change, Choice and Inference – A Study of Belief Revision and Nonmonotonic Reasoning* (Oxford: Oxford University Press 2001).

17 Man beachte hier, dass $(a \vee b \vee c) \wedge (a \vee b)$ mit a∨b bzgl. *Cn* äquivalent ist und deshalb dieselben Schlussfolgerungen wie letzteres erlauben sollte.

18 Diese Formulierung soll hier und im Folgenden immer heißen, dass diese Prämisse *die ganze* Information enthält, die die Person zur Verfügung hat. Dies ist sehr wichtig für das Verständnis nichtmonotonen Schließens.

Falschheit von ϕ beitragen. In beiden Lesarten wird «Plausibilität» nach den doxastischen Wahlfunktionen der Person beurteilt.[19] Ein Satz ψ kann aus der Prämisse ϕ genau dann gefolgert werden, wenn ψ in allen plausibelsten ϕ-Welten wahr ist bzw. wenn $\phi \rightarrow \psi$ nicht zu den unplausibelsten *Cn*-Folgerungen von $\neg\phi$ gehört.[20] Auf der Grundlage dieser Anwendung der Theorie rationaler Wahl kann man zeigen, dass die logischen Bedingungen Oder und Konditionalisierung, die vor dem Hintergrund anderer, basalerer Bedingungen äquivalent sind, genau Sens Eigenschaft α entsprechen, während die Bedingung der Kumulativen Monotonie genau der Aizermanschen Bedingung entspricht. Es gibt aber noch eine ganze Reihe weiterer Korrespondenzen, die nahelegen, dass logische Bedingungen für nichtmonotone Inferenzrelationen als von Regeln für rationale Wahl erzeugt aufgefasst werden können.[21] Wir müssen hier keine vollständige Liste geben, die folgenden Paarungen werden sich aber noch als relevant erweisen. Der wahltheoretischen Bedingung (II) und ihrer Verstärkung

(II$^+$) Wenn x in $\sigma(S)$ und y in $\sigma(S')$ ist, so ist x oder y in $\sigma(S \cup S')$

entsprechen in dieser Reihenfolge die logischen Bedingungen

(Sehr schwache disjunktive Rationalität)	$Inf(\phi \vee \psi) \subseteq Cn\,(Inf(\phi) \cup Inf(\psi))$
(Disjunktive Rationalität)	$Inf(\phi \vee \psi) \subseteq Inf(\phi) \cup Inf(\psi)$

[19] Wir gehen an dieser Stelle nicht davon aus, dass die Wahldispositionen der Person immer durch Präferenzen rationalisiert werden können. Wenn es solche Präferenzen gibt, dann ändern sich diese im Allgemeinen infolge des Erhalts einer Information ϕ, was aber in diesem Aufsatz nicht weiter hervorgehoben werden soll. Siehe hierzu Hans Rott: *Shifting Priorities – Simple Representations for Twenty-seven Iterated Theory Change Operators,* in *Towards Mathematical Philosophy*, hg. von David Makinson, Jacek Malinowski, Heinrich Wansing (Dordrecht: Springer, 2009) S. 269-296.

[20] Die letztere Bedingung ist intuitiv leider nur schwer verständlich. Sie kann auch durch die Bedingung ersetzt werden, dass $\phi \rightarrow \psi$ kein unplausibelstes Element von $\{\phi \rightarrow \psi, \phi \rightarrow \neg\psi\}$ ist. Zur Begründung beider Bedingungen vgl. Rott, op. cit. (Fn. 16) S. 172-181. Außerdem gibt es Brückenprinzipien, die den Begriff der «besten möglichen Welt» als theoretisch gleichwertig mit dem Begriff der «schwächsten Meinung» erweisen: Eine beste ϕ-Welt ist eine, in der ϕ und außerdem alle bis auf die schwächsten *Cn*-Konsequenzen von $\neg\phi$ wahr sind. Umgekehrt ist der Satz χ ein innerhalb einer Satzmenge M schwächster Satz, wenn χ zu M gehört und es unter den besten M-falsifizierenden Welten (d.h. möglichen Welten, in denen einer der Sätze aus M falsch ist) eine Welt gibt, in der χ falsch ist (vgl. ibid. S. 208-213).

[21] Ibid. S. 200-206.

Fast alle wahltheoretischen Bedingungen haben exakt denselben Effekt, wenn man sie auf semantischer Ebene (plausible Welten) anwendet, wie wenn man sie auf syntaktischer Ebene (unplausible Sätze) anwendet. Sens Bedingung γ bildet hier eine (fast singuläre) Ausnahme. Sehr schwache disjunktive Rationalität entsteht durch die Auferlegung von (II) auf der semantischen Ebene, während die Auferlegung von (II) auf der syntaktischen Ebene eine schwache, aber nicht die sehr schwache Form der Disjunktiven Rationalität ergibt.[22]

Ich vertrete die folgende These: *Alle Probleme rationaler Wahl lassen sich (ohne größere Schwierigkeiten) als Probleme des Schlussfolgerns wiedergeben.* Im Folgenden werde ich diese These nicht im Einzelnen ausarbeiten, doch werde ich im nächsten Abschnitt das oben diskutierte Beispiel der Bewerbung um eine Metaphysik-Stelle weiterführen und auf dreifache Weise variieren. Die praktische Problematik der Kandidatenauswahl innerhalb verschiedener möglicher Bewerberfelder kann ebenso mühelos wie im obigen Beispiel auf die Problemstellung des Schlussfolgerns aufgrund verschiedener Informationslagen über den Stand des Besetzungsverfahrens übertragen werden. Dies soll Evidenz genug für die Gültigkeit der These sein.

Zum Schluss dieses Abschnitts müssen wir jedoch noch die durch das obige Beispiel entstandene Anomalie kommentieren. Handelt es sich um einen Fall von galoppierender Irrationalität? Nein, dafür ist die Geschichte zu plausibel, und die Gedankengänge sind zu gut nachvollziehbar. Doch scheint die Erklärung aus dem Luce-Raiffa-Beispiel hier nicht brauchbar zu sein. Denn dort mag es zwar illegitim gewesen sein, dass die besondere Gestalt des Menüs Information über die Situation einschmuggelte. Im Beispiel des Berufungsverfahrens liegt die Sache aber insofern anders, als es sich hier ja gerade um ein Problem handelt, das die Ausbildung und Umformung von Meinungen oder Überzeugungen betrifft. Es geht hier explizit um die Verwertung von Information, also können wir uns nicht beklagen, dass Information «eingeschmuggelt» wird. Jede Überlegung, die wir informell beschrieben haben, müsste im Modell vor dem Hintergrund der Überzeugungen und Annahmen des Hörers nachvollziehbar sein. Die Lösung liegt woanders. Der springende Punkt ist, wie schon in Rott (2004) skizziert, dass der Hörer etwa im dritten Szenario nicht einfach die Information $a \vee b \vee c$ bekommt, sondern die Information *Der Kommissionsvorsitzende sagt mir, dass* $a \vee b \vee c$, und *dies* ist es, was die vergleichsweise komplizierte Überlegung

22 Die Bedingung der schwachen disjunktiven Rationalität ist leider wenig intuitiv: $Inf(\phi \vee \psi) \subseteq Cn(Inf(\phi) \cup \{\psi\}) \cup Cn(Inf(\psi) \cup \{\phi\})$.

in Gang setzte, die mit der Konklusion *b* endet. Hätte man die Information $a \vee b \vee c$ dadurch gewonnen, dass man alle anderen Kandidaten mit langem Gesicht hatte abreisen sehen, dann hätte man nicht auf die Wertschätzung von Logik-Kompetenzen und auf *B* als Idealkandidatin in diesem Besetzungsverfahren geschlossen. Wird die Neuinformation aber explizit mit Nennung der Informationsquelle notiert, verliert das Beispiel die der Irrationalität verdächtigte Struktur: Aus *Der Kommissionsvorsitzende sagt mir, dass* $a \vee b$ folgt eben nicht *Der Kommissionsvorsitzende sagt mir, dass* $a \vee b \vee c$. Letzten Endes gibt es also doch eine Gemeinsamkeit mit dem Luce-Raiffa-Fall: Die ursprüngliche Repräsentation der Szenarios ist möglicherweise zu einfach.

3. Neue Fälle

Im letzten Abschnitt haben wir eine weitgehende Korrespondenz zwischen Forderungen an rationale Auswahlen und Forderungen für alltägliches Schlussfolgern aufgewiesen, das die logische Eigenschaft der Monotonie nicht aufweist. Wir fanden aber auch zwei seltsame Beispiele, die Anlass zur Kritik selbst an der natürlichsten und paradigmatischsten aller Bedingungen für rationale Wahl, nämlich an Sens Bedingung α, boten. Gegen Luce und Raiffas Beispiel haben wir die Erwiderung skizziert, dass die Alternativen in der Modellierung vielleicht falsch beschrieben wurden und beispielsweise statt «Steak» adäquater «Steak-in-einem-guten-Restaurant» als Alternative hätte formuliert werden müssen. Der von uns frühzeitig als Logik-Beispiel interpretierte Fall der Stellenbesetzung weist als Fall rationalen Wählens Ähnlichkeiten mit dem Beispiel von Luce und Raiffa auf.

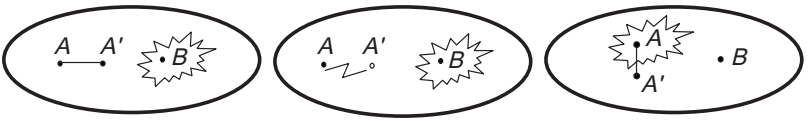
Wir wenden uns nun drei neuen Fallklassen zu. Sie wurden ursprünglich im Kontext sogenannten probabilistischen Wählens diskutiert, in dem die Häufigkeiten bestimmter Wahlen von Personen einer Referenzgruppe untersucht werden. Wir werden sehr vereinfachte, rein qualitativ präsentierte Versionen dieser Fälle betrachten. Anders als in den bisher diskutierten Fällen wird hier die zentrale Bedingung α von Sen erfüllt. Doch werden ebenfalls wichtige Bedingungen wie Sens Bedingung γ und Aizermans Bedingung verletzt. Unsere Frage wird sein, ob es sich hier um Fälle irrationaler Wahl handelt oder nicht. Gemäß der Theorie rationaler Wahl müssen solche Beispiele als irrational gelten, doch könnten sie im Umkehrschluss ja auch Anlass zu der Vermutung geben, dass die klassische Theorie der rationalen Wahl (aus bisher noch undeutlichen Gründen) selbst inadäquat ist.

Es geht (1) um den Effekt *ähnlicher Optionen*,[23] (2) um den strukturverwandten *Kompromiss-Effekt*[24] sowie (3) um den Effekt *dominierender Optionen*.[25] Alle diese Fälle betreffen Wahlen zwischen drei Optionen.

23 Gerard Debreu: *Review of R. D. Luce's Individual Choice Behavior*, in *American Economic Review* 50 (1960) S. 186-188. Diese knapp dreiseitige Rezension von R. Duncan Luce: *Individual Choice Behavior – A Theoretical Analysis* (New York: John Wiley and Sons, 1959) war für die Thematik des vorliegenden Aufsatzes bahnbrechend. Der Stein des Anstoßes, Luces Auswahlaxiom, ist es wert, hier noch einmal in einer (gelegentlich *Constant Ratio Rule* genannten) Variante wiedergegeben zu werden: Wenn A und B Elemente eines Menüs S sind und $P_S(A)$ die Wahrscheinlichkeit, dass aus dem Menü S das Element A ausgewählt wird, angibt, dann gilt $P_S(A)/P_S(B) = P_{\{A,B\}}(A)/P_{\{A,B\}}(B)$. Luces Axiom formuliert eine Art «Unabhängigkeit von irrelevanten Alternativen» (vgl. R. Duncan Luce: *Luce's Choice Axiom*, in *Scholarpedia* 3 (12): 8077, 2008). Das qualitative Analogon lautet offenbar: Wenn $S \subseteq S'$ und $\sigma(S') \cap S \neq \emptyset$, dann gilt $\sigma(S) = \sigma(S') \cap S$. Diese Bedingung ist als *Arrows Axiom* bekannt und mit der Konjunktion von (I) und (IV) äquivalent (vgl. Rott, op. cit. [Fn. 16] S. 154). – Als Beispiel verwendet Debreu ein Menü mit einer Aufnahme eines Streichquartetts von Debussy und zwei einander ähnlichen Einspielungen der Achten Symphonie von Beethoven. Für die Geschichte des Problems ähnlicher Optionen wichtig waren außerdem John Chipman: *The Foundations of Utility*, in *Econometrica* 28 (1960) S. 193-224, Amos Tversky: *Elimination by Aspects – A Theory of Choice,* in *Psychological Review* 79 (1972) S. 281-299, und Daniel McFadden: *Conditional Logic Analysis of Qualitative Choice Behavior,* in *Frontiers in Econometrics*, hg. von Paul Zarembka (New York: Academic Press, 1974) S. 105-142. In der Literatur wird das Problem weithin anhand roter und blauer Busse als nicht signifikant unterschiedener Transportmöglichkeiten diskutiert (*red bus/blue bus problem*).

24 Itamar Simonson: *Choice Based on Reasons – The Case of Attraction and Compromise Effects,* in *Journal of Consumer Research* 16 (1989) S. 158-174.

25 Joel Huber, John W. Payne, Christopher Puto: *Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis*, in *Journal of Consumer Research* 9 (1982) S. 90-98; Itamar Simonson, Amos Tversky: *Choice in Context: Tradeoff Contrast and Extremeness Aversion,* in *Journal of Marketing Research* 29 (1992) S. 281-295; Eldar Shafir, Itamar Simonson, Amos Tversky: *Reason-Based Choice*, in *Cognition* 49 (1993) S. 11-36; Eldar Shafir, Amos Tversky: *Decision Making*, in *Invitation to Cognitive Science – Thinking*, hg. von Daniel N. Osherson, Edward E. Smith (Cambridge MA: MIT Press, 1995) S. 77-109.



(i) Ähnlichkeit (ii) Kompromiss (iii) Dominanz

Abb. 3: Die neuen Fälle – Menü mit drei Elementen

Die Wahl in den dreielementigen Menüs erfolgt aufgrund bestimmter Relationen zwischen *A* und *A′*, während *B* sowohl mit *A* als auch mit *A′* in gewissem Sinn unvergleichbar ist und sozusagen isoliert steht. Die Wahlen innerhalb der Dreierkonkurrenzen werden wieder kontrastiert mit Wahlen in Zweierkonkurrenzen. Bis auf den Fall (3), in dem *A* gegenüber *A′* vorgezogen wird, sind in allen drei Fällen bei zweielementigen Menüs stets beide Optionen wählbar, sei es, weil die Optionen gleich gut (*A* und *A′* im Fall [1]), sei es, weil sie unvergleichbar sind (alle anderen paarweisen Konkurrenzen). Insbesondere gilt in allen drei Fallklassen, dass in einem aus *A* und *B* allein bestehenden Menü beide Elemente gewählt werden können.

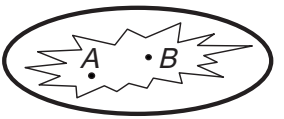


Abb. 4: Die neuen Fälle – Menü mit zwei Elementen

Zusammen mit der eindeutigen Bevorzugung von *A* (Fall ([3]) bzw. *B* (Fälle [1] und [2]) in dreielementigen Menüs ist dies vom Standpunkt der klassischen Theorie rationaler Wahl aus gesehen unzulässig. Was sich in der Dreierkonkurrenz durchsetzt, sollte sich auch dann durchsetzen, wenn ein anderer (in der Dreierkonkurrenz übrigens nicht erfolgreicher) Kandidat wegfällt. Dass dies hier nicht der Fall ist, ist vorderhand seltsam.

Der Anschaulichkeit halber führen wir das obige Beispiel des Bewerbungsverfahrens weiter. Das Bewerberinnenfeld sehe nun jedoch etwas anders aus. Cortez und Donaldson seien nicht im Rennen, dafür habe sich auch Annabel Andrews, die Schwester von Amanda, beworben.[26] Wir ordnen

[26] Dass Amanda und Alice verwandt sind, schließt eine echte Konkurrenzsituation nicht aus.

ihr die Bezeichnung A' zu. Im Fall der Ähnlichkeit kann man sich Amanda und Annabel als Zwillingsschwestern vorstellen, die sich auch in ihren Kompetenzen in Metaphysik und Logik kaum wahrnehmbar unterscheiden. Im Fall des Kompromisses haben die Schwestern hingegen gegensätzliche Talente. Während Amanda eine ausgezeichnete Metaphysikerin ist, die von Logik nicht viel versteht, verhält es sich bei Annabel gerade umgekehrt.[27] Im Fall (1) sind Amanda und Annabel gut (in der Tat: extrem gut) vergleichbar, im Fall (2) kann man sie als unvergleichbar ansehen – jedenfalls solange es keinen Maßstab gibt, der Kompetenzen in Metaphysik in ein Verhältnis zu Kompetenzen in Logik setzt. In (3) besteht wieder eine gute Vergleichbarkeit, die diesmal aber asymmetrisch ist: Amanda ist klar besser für die Stelle geeignet als Annabel. Den drei Fällen gemeinsam ist die folgende Konstellation: Obwohl Amanda und Bernice eigentlich unvergleichbar sind, hebt das Auftreten von Annabel gewisse Eigenschaften von Amanda hervor, die dann eine Entscheidung zwischen Amanda und Bernice verständlich machen. Annabel selbst vermag zwar nie in die Entscheidung mit einzugreifen, wirkt aber gleichsam als Katalysator für die Entscheidungsfindung.

Es ist vielleicht schon deutlich geworden, wie wichtig das Konzept der *Unvergleichbarkeit* ist, das zwar selten thematisiert wird, aber doch entscheidend für die Interpretation der vorliegenden Fälle zu sein scheint. In allen mir bekannten Beispielfällen stehen Optionen zur Wahl, die nach mehr als einem *Kriterium* oder, wie ich auch sagen möchte, in mehr als einer *Dimension* bewertet werden.[28] Soweit ich sehe, genügt durchwegs die Betrachtung von

[27] Wir wollen in diesem Beispiel so tun, als ob gute Metaphysik ohne Logik und gute Logik ohne Metaphysik möglich sind. Es ist für alle in diesem Aufsatz betrachteten Beispiele entscheidend, dass es verschiedene Faktoren oder Kriterien gibt, die voneinander unabhängig sind und gedacht werden.

[28] Die Rede von *Dimensionen* soll suggerieren, dass die verschiedenen Hinsichten der Bewertung voneinander unabhängig sind. Im Übrigen eröffnet sich mit der Anerkennung mehrerer Kriterien oder Dimensionen ein vollkommen neuer Problembereich. Das sehr weite Gebiet der *Social choice theory* wird unmittelbar relevant, wenn man die Präferenzen oder Wertungen nach verschiedenen in einer Entscheidung involvierten Kriterien wie die Präferenzen oder Wertungen verschiedener Mitglieder einer sozialen Gemeinschaft konzeptualisiert. Die berühmten Unmöglichkeitstheoreme der *Sozial*wahltheorie von Arrow und anderen werden dann unmittelbar auch für *individuelle* Entscheidungen relevant. Vgl. Kenneth Arrow, Amartya K. Sen, Kotaro Suzumura (Hg.): *Handbook of Social Choice and Welfare*, Bd. 1 (Amsterdam: Elsevier, 2002) und José Figueira, Salvatore Greco, Matthias Ehrgott (Hg.): *Multiple Criteria Decision Analysis: State of the Art Surveys* (New York: Springer, 2005).

zwei verschiedenen Kriterien. Gemäß eines Kriteriums oder in einer Dimension sind alle Optionen vollständig *vergleichbar*: Entweder ist *A* besser als *B*, oder *B* ist besser als *A*, oder *A* und *B* sind gleich gut. Die Sachlage ändert sich, wenn zwei Kriterien anzusetzen sind. Wenn beide Kriterien zugunsten derselben Option sprechen oder wenn eines der beiden Kriterien die Optionen *A* und *B* als gleich gut bewertet, dann gibt es kein Problem, denn es ist klar, welche der Optionen vorzuziehen ist. Der interessante und für sämtliche Beispiele relevante Fall ist der, in dem die beiden Kriterien, angewandt auf zwei Optionen, in verschiedene Richtungen ziehen: *A* schneidet nach Kriterium 1 und *B* schneidet nach Kriterium 2 besser ab. In diesem Fall liegt, zumindest vorbehaltlich weiterer Erwägungen, eine Unvergleichbarkeit der Optionen *A* und *B* vor. Eine Rangordnung (im rein qualitativen Fall) oder eine Gewichtung (im quantitativen Fall) der Kriterien könnte eine Auflösung des Konflikts bringen, ansonsten aber kann kein sinnvoller Vergleich hergestellt werden. Aufgrund des Wahlverhaltens in einer Population, in dem irgendeine von zwei Optionen gewählt werden muss, ist Unvergleichbarkeit von Gleichwertigkeit der Optionen nicht zu unterscheiden; in beiden Fällen sind beide Optionen rational wählbar.

Die einschlägige Literatur hat die erwähnten Fallklassen zunächst in Gedankenexperimenten als einleuchtend charakterisiert, dann aber auch empirisch als realistisch nachgewiesen. In den empirischen Studien ergeben sich nur probabilistische Zusammenhänge, also Prozentzahlen, die aussagen, wie viele der untersuchten Probanden in den relevanten Szenarien sich faktisch für welche Option entschieden haben. Die Übertragung auf die normative Problematik individuell-rationaler Wahl ist alles andere als trivial, sie liegt aber nahe und soll hier vorgenommen werden. Auf mehrere erschwerende Faktoren ist hierbei allerdings hinzuweisen. Ein erster wichtiger Unterschied ist, dass in einer empirischen Situation eine – und *nur* eine – Wahl getroffen wird, während in der Theorie rationaler Wahl mehrere Optionen als vernünftig und in diesem Sinne zulässig gekennzeichnet werden. In einer empirischen Situation kann zweitens auch nur *eine* Palette an Optionen, d.h. ein Menü zur Auswahl angeboten werden, während es in der Theorie rationaler Wahl gerade darum geht, strukturelle Zusammenhänge zwischen verschiedenen potentiell präsentierten Menüs zu charakterisieren. In empirischen Situationen versucht man das Problem durch die Annahme stabiler individueller Präferenzen quer über verschiedene, faktisch sequentiell angebotene Wahlsituationen hinweg zu lösen, aber dies ist nur eine mehr oder weniger fragwürdige Approximation. Drittens wird in der empirischen Situation durch den spezifischen Inhalt eines Menüs häufig Information über

die Situation vermittelt (wie im Beispiel von Luce und Raiffa), ein Effekt, den die Theorie der rationalen Wahl eigentlich gerade herausfiltern möchte. Schließlich fällt es schwer, aufgrund eines empirischen Befunds zu entscheiden, ob eine Option als rational oder zulässig zu kennzeichnen ist. Es ist klar, dass es nicht genügen kann, wenn sich in einer großen Stichprobe einige wenige Probanden für diese Option entscheiden. Hierbei könnte es sich um irrationale «Ausreißer» handeln. Ab welcher Größe aber soll man sagen, dass die Zustimmung zu einer Option innerhalb der Gruppe der Probanden so groß ist, dass man nicht mehr umhin kommt, der Option eine gewisse Rationalität zuzugestehen? Oder ist die Kluft zwischen deskriptivem Befund und normativer Empfehlung tatsächlich unüberbrückbar, so dass selbst die ausnahmslose Beachtung einer Regel in einer Population überhaupt nichts über ihre Rationalität aussagt? Wir werden all diese Probleme ausblenden und allein auf der Ebene von Gedankenexperimenten bleiben, die direkt an vorhandene Rationalitätsintuitionen der Leserinnen und Leser appellieren.

Noch einmal also zurück zu den konkreten Beispielsvarianten. Im Fall (1) der *Ähnlichkeit* gibt es drei Optionen, von denen zwei, A und A', in ihren Eigenschaften einander sehr ähnlich sind und deshalb auch als gleich gut bewertet werden. B ist andersartig als A und A', in seinen Eigenschaften unvergleichbar mit diesen, was im Effekt dazu führt, dass in einer Zweierkonkurrenz zwischen A und B keine eindeutige Entscheidung möglich ist. In der Dreierkonkurrenz aber wird von einer großen Anzahl von Probanden B gewählt. Dies ist offensichtlich deshalb so, weil die Probanden keinerlei Anhaltspunkte haben, sich zwischen A und A' zu unterscheiden, und B die einzige auffällige Option mit einem erkennbaren «Alleinstellungsmerkmal» ist.

Der Fall (2) des *Kompromisses* hat eine ganz ähnliche Struktur wie, doch anderen Inhalt als Fall (1). Der einzige Unterschied ist, dass sich hier A und A' im Profil nicht gleichen, sondern im Gegenteil sehr verschieden sind und gleichsam zwei Enden eines Entscheidungsspektrums von Optionen darstellen. Ein typischer Fall wäre es hier, wie bereits erwähnt, dass es zwei relevante Eigenschaften gibt und A die erste, aber nicht die zweite Eigenschaft, während A' umgekehrt die zweite, aber nicht die erste Eigenschaft sehr gut erfüllt. B liegt in beiden Hinsichten «zwischen» den Extremen A und A'. Auf ihre je eigene Art und Weise sind alle Kandidatinnen von der Qualität her gleichermaßen (besser: unvergleichlich) gut für die Stelle geeignet. Doch besitzt B als Kompromisskandidatin ein «Alleinstellungsmerkmal» gegenüber den beiden extremeren Alternativen und wird vermutlich deshalb häufig gewählt.

Die Sachlage in (3) für *dominierende* Optionen ist etwas anders. *A* und *A′* sind diesmal sehr gut miteinander vergleichbar, und das Ergebnis der Urteilsbildung lautet, dass *A* eindeutig gegenüber *A′* zu bevorzugen sei. *B* ist wie im Fall (1) weder mit *A* noch mit *A′* gut zu vergleichen. Da man mit der Dominanz von *A* gegenüber *A′* einen «Grund» besitzt, *A* vorzuziehen, wird *A* in empirischen Situationen auch signifikant häufiger gewählt, wenn die Dreierkonkurrenz zwischen *A*, *A′* und *B* gegeben ist. Fällt der Kontrast mit *A′* hingegen weg, ist *A* gegenüber *B* nicht mehr ausgezeichnet und wird auch nicht mehr bevorzugt gewählt.

Welche Kohärenzbedingungen werden in den soeben beschriebenen Fällen verletzt? Es handelt sich erstens um die Sens Bedingung γ, oben auch als (II) gekennzeichnet. In den Fällen der Ähnlichkeit und des Kompromisses ist *A* sowohl im paarweisen Vergleich mit *A′* als auch im paarweisen Vergleich mit *B* wählbar, in der Dreierkonkurrenz wird aber nur *B* gewählt; die Option *A* geht bei der Vereinigung als wählbare «verloren». Im Fall der Dominanz erleidet die Option *B* dasselbe Schicksal. Zweitens wird die Aizermansche Bedingung (III) verletzt. Obgleich alle gewählten Elemente der Dreierkonkurrenz – *B* in (1) und (2), *A* in (3) – sich in der Menge {*A*,*B*} wiederfinden, werden nicht alle innerhalb dieser eingeschränkten Menge wählbaren Optionen – es gilt ja $\sigma(\{A,B\})=\{A,B\}$ – in der Dreierkonkurrenz gewählt.

Der Verletzung der Wahlbedingungen (II) und (III) entspricht auf logischer Ebene die folgende Situation in den Fällen (1) der Ähnlichkeit und (2) des Kompromisses. Erstens wird die Bedingung der Sehr schwachen disjunktiven Rationalität verletzt, denn während aus der Disjunktion $a \vee a' \vee b$ gefolgert werden kann, dass $\neg a$, kann man weder aus $a \vee a'$ noch aus $a \vee b$ irgendwelche substantiellen Schlüsse ziehen, so dass $Cn(Inf(a \vee a') \cup Inf(a \vee b))$ keinen Schluss auf $\neg a$ erlaubt.[29] Zweitens wird die Bedingung der Kumulativen Monotonie verletzt, denn obgleich a∨b aus $a \vee a' \vee b$ erschließbar ist, sind nicht alle Schlussfolgerungen aus $a \vee a' \vee b$ auch Schlussfolgerungen aus $a \vee b$.[30] Es ist nämlich *b* aus dem inhaltsschwächeren $a \vee a' \vee b$, nicht aber aus dem inhaltsstärkeren $a \vee b$ erschließbar. Für den Fall (3) der Dominanz gilt wiederum Analoges, nur tauschen *a* und *b* ihre Rollen.

[29] Die Situation ist tatsächlich schlimmer. Unter der beibehaltenen Voraussetzung, dass nur eine Person angestellt werden kann, folgt aus der Vereinigungsmenge von *Inf(a∨a′)* und *Inf(a∨b)* vermittels *Cn* sogar *a*.

[30] Man beachte hier, dass $(a \vee a' \vee b) \wedge (a \vee b)$ mit $a \vee b$ bzgl. *Cn* äquivalent ist und deshalb dieselben Schlussfolgerungen wie letzteres erlauben sollte.

4. Schluss

Seltsame Auswahlen sind kein Zeichen von Wankelmut. Sie sind ernst zu nehmen und sollten nicht vorschnell als irrational disqualifiziert werden. Mögliche Antworten auf die Beispiele von Luce und Raiffa (1957) und Rott (2004) wurden in den ersten beiden Abschnitten skizziert. Das Ziel unserer Diskussion waren die neuen Fälle. Auch sie erlauben meines Erachtens eine Auflösung, welche die scheinbare Irrationalität wegerklärt.

Zunächst ist zu sagen, dass die Problematik einen metaphysischen Gehalt hat. Kommt einer Option (einer möglichen Welt, einem Güterbündel, dem Resultat einer Handlung, …) ein Wert an und für sich zu? Oder ist jeder Wert vom Kontext der alternativ zur Verfügung stehenden Optionen abhängig? «Wert» soll hierbei nur im individuellen, nicht in einem sozialen oder gar über-menschlichen Sinne verstanden werden. Die Messung der Güte einer Option durch Zuordnung eines kontextunabhängigen numerischen Wertes, etwa eines bestimmten Nutzen- oder Geldbetrags, präsupponiert eine Absolutheit des Wertes, und diese Absolutheit wurde gerade durch die obigen Beispiele in Frage gestellt. Man muss sehr ernsthaft mit der Möglichkeit rechnen, dass der absolute Wert einer Option nur eine Fiktion ist, die bestenfalls eine Abstraktion darstellt aus den metaphysisch fundamentalen Wertigkeiten, welche der Option in den verschiedenen Kontexten der je zur Verfügung stehenden Alternativen zukommen.

Diese Frage kann im vorliegenden Beitrag nicht entschieden werden. Wären aber die obigen Befunde überhaupt logisch verträglich mit der Annahme rationaler Wahlen aufgrund fundamentaler, kontextunabhängiger und universelle Vergleiche ermöglichender Werte? Dies ist in der Tat der Fall. Es ist nämlich nicht zwingend, die vorliegenden Befunde, sowohl empirische Studien als auch unsere Intuitionen in Gedankenexperimenten, dahingehend zu interpretieren, dass Dinge, die in der paarweisen Konkurrenz gleichwertig sind, in einer Konstellation zu dritt plötzlich unterschiedlich wertvoll werden. Eine alternative Interpretation unterscheidet zwischen echtem Wählen und bloßem Herauspicken.[31] *Wählen* ist rationales, von Präferenzen, Werten, Wünschen usw. geleitetes Bestimmen einer Option. *Herauspicken* ist ein solches Bestimmen, insoweit es von der Leitung durch wertende Erwägungen vollkommen frei ist. Dies heißt nicht notwendigerweise, dass das Herauspicken vollkommen zufällig geschieht. Auch solche

[31] Vgl. Edna Ullmann-Margalit, Sidney Morgenbesser: *Picking and Choosing*, in *Social Research* 44 (1977) S. 757-785.

ungeleiteten Arten von Entscheidungen können durch Faktoren beeinflusst oder verursacht sein, nur sind dies dann Faktoren, denen man keine Relevanz für Fragen der Rationalität zusprechen würde. In den diskutierten Fallklassen (1)-(3) – Ähnlichkeit, Kompromiss und Dominanz – kann das distinktive Merkmal der je ausgewählten Option als ein Merkmal gesehen werden, das für gewisse Auffälligkeiten in der Dreierkonkurrenz sorgt.[32] Doch sind weder die Merkmale selbst («den anderen, einander ähnlichen Optionen unähnlich sein», «einen Kompromiss zwischen den anderen, einander entgegengesetzten Optionen darstellen», «allein eine andere Option dominieren») noch die daraus resultierenden Auffälligkeiten an sich wertsteigernd oder -mindernd. Die Auffälligkeiten helfen nur, unter Optionen, zwischen denen keine Entscheidung als rational auszuzeichnen ist, zu unterscheiden. Damit verhindern sie eine drohende Erstarrung, lösen eine Wahlblockade, wirken als *tiebreaker*.[33] So sind diese Faktoren entscheidungsrelevant. Da eine solche Auflösung prozedural rational ist, kann man diese eigentlich wertneutralen Merkmale in einem abgeleiteten Sinn möglicherweise auch als präferenzformend betrachten. Sie kommen aber nicht den Optionen an und für sich zu, sondern sind wesentlich relationale Eigenschaften, und es ist fraglich, ob sie auch nur in einem abgeleiteten Sinne wertkonstituierend genannt werden können.

32 Das hier relevante englische Wort für «Auffälligkeit» wäre *salience.*

33 Der Begriff *tiebreaker* ist zwar eingängig, aber strenggenommen ungenau, denn der Begriff *tie* setzt eine Vergleichbarkeit voraus, und für die obigen Beispiele war ja gerade die Unvergleichbarkeit mancher Optionen entscheidend.

Studia philosophica 68/2009

Urs Allenspach

(Ir-)rational würfeln[1]

Adopting a random process seems to be a rational procedure when choosing a single option from a plurality of options, none of which is inferior to any other. Yet, for those needing to make several such decisions, the result may be less than ideal. This article offers two alternatives which block this irrational outcome. The first is based on the traditional evaluative relations of (revealed) superiority and equality. This solution is procedural in the sense that it allows for criticism of single decisions only with reference to other previously-made decisions. The second alternative involves a non-traditional evaluative relation of parity. This solution is strictly structural, viz. all single decisions are definitely either rational or irrational.

1. Parität verstanden als Lösung eines Entscheidungsproblems

Wie soll man sich rational zwischen Handlungsoptionen entscheiden, worunter sich keine als schlechter herausgestellt hat als eine andere? – Die Frage scheint keine Probleme aufzuwerfen: Wenn von den Optionen keine schlechter ist als eine andere, sind sie wohl alle zulässig. Damit präsentiert sich das Rechtfertigungsproblem bereits dahingehend gelöst, dass rationalerweise eine beliebige Option ausgesucht werden darf. Jedes Zufallsverfahren führt dazu, dass eine Entscheidung gefällt wird, die nicht als irrational kritisiert werden kann.

Diese Einschätzung ist im Einzelfall wohl richtig. Die Behauptung ist jedoch falsch, wenn nicht nur einzelne Entscheidungen, sondern ganze Mengen von Entscheidungen betrachtet werden. Zu dieser Einsicht muss man angesichts eines Beispiels gelangen, das die typische Struktur aufweist

[1] Die Forschung für diesen Aufsatz wurde unterstützt durch das Staatssekretariat für Bildung und Forschung SBF/COST, Europäische Zusammenarbeit auf dem Gebiet der wissenschaftlichen und technischen Forschung, und durch das Forschungsprojekt ClimPol des ETH Bereichs. Ich danke Georg Brun und Gertrude Hirsch Hadorn für viele fruchtbare Diskussionen. Ausserdem hat mir eine Bemerkung von Peter Schaber sehr weiter geholfen, dem dafür auch gedankt sei.

für Beispiele, die in der philosophischen Literatur zu Vergleichbarkeit und Unvergleichbarkeit, zu Kommensurabilität und Inkommensurabilität verhandelt werden.[2]

Angenommen, man soll eine verhältnismäßig geräumige Stadtliegenschaft (A) und ein schmuckes Häuschen am See (B) in der Hinsicht darauf vergleichen, in welche/s man eher einziehen möchte. So kann man gewiss zum Schluss kommen, dass keine Option schlechter ist als die andere. Die Angebote sind vergleichsweise verschieden, was die Vergleichbarkeit beeinträchtigt. Wenigstens aber diese Einschätzung ist möglich, dass keine Option der anderen an Attraktivität in der relevanten Hinsicht unterlegen ist.

Die gleiche Antwort ist denkbar, wenn man das Häuschen am See und die Stadtliegenschaft angeboten bekommt, wobei in die Stadtliegenschaft noch ein Dachfenster eingebaut wird (A^+). Das ist zwar eine minimale Verbesserung, im Vergleich mit dem derart verschiedenen Angebot am See ändert das Dachfenster aber nichts an der relativen Beurteilung.

Jede der beiden Entscheidungen – B aus $\{A^+, B\}$ und A aus $\{A, B\}$ – ist an und für sich rational. Beide Entscheide zusammen müssen irrational sein, können sie doch dazu führen, dass wir in einer Stadtwohnung ohne Dachfenster zu wohnen kommen, obwohl wir ceteris paribus dieselbe Wohnung mit Fenster hätten haben können. Dieser Fall tritt ein, wenn wir uns zwischen B und A^+ erst rationalerweise für B entscheiden, uns danach rationalerweise A für B andrehen lassen.

Wie sich zeigen wird, tritt diese Form der Irrationalität überdies auf, ohne dass ein Transitivitätsgesetz verletzt oder eine Form von Zyklizität vorliegen würde. Darin unterscheidet sich der Fall vom bekannten Money-pump-Argument.

Eine scheinbar ganz andere Sache: Seit einigen Jahren ist eine Debatte darüber im Gang, welche evaluativen Vergleichsrelationen überhaupt existieren. Insbesondere Chang[3] hat bezweifelt, dass die traditionelle Aufzählung

2 Vgl. z.B. Ronald B. De Sousa: *The Good and the True*, in *Mind* 83/332 (1974), S. 534-551, hier S. 544-545; Derek Parfit: *Reasons and Persons* (New York: Oxford University Press, 1986) S. 430-431; James Griffin: *Well-Being* (New York: Oxford University Press, 1986) S. 81; Ruth Chang: *The Possibility of Parity*, in *Ethics* 112 (2002) S. 659-688, hier S. 668; Erik Carlson: *Incomparability and Measurement of Value*, in *The Good, the Right, Life and Death*, hg. von Kris McDaniel, Jason R. Raibley, Richard Feldman, Michael J. Zimmerman (Aldershot: Ashgate, 2006) S. 19-43, hier S. 19-20.

3 Ruth Chang: *Introduction*, in: *Incommensurability, Incomparability, and Practical Reason*, hg. von Ruth Chang (Cambridge, MA: Harvard University Press,

von «besser als», «schlechter als» und «gleich gut wie» (oder: «äquivalent») erschöpfend ist. Vielmehr gebe es konzeptuellen Raum für genau eine weitere, von allen anderen disjunkte Wertrelation, Parität (im Original: «parity»).

Chang wirbt für ihre neue Wertrelation zum einen mit Beispielen, zum anderen mit Argumenten, letztere wiederzugeben allerdings einen eigenen Aufsatz erfordern würde. Für das Folgende ist lediglich von Belang, dass zuerst gezeigt werden muss, dass Optionen existieren, für die eine gewisse Vergleichshinsicht zwar relevant ist, sie in dieser Hinsicht aber traditionell nicht vergleichbar sind.

Am Beispiel der Liegenschaften: Gewiss kann man beide Angebote in der Hinsicht darauf beurteilen, wie es wäre, in ihnen zu wohnen (Hinsicht W). Es liegt hier kein begrifflicher Irrtum vor, wie wenn man fragte, wie es wäre, in einer Primzahl zu wohnen oder einer Grippe.

Jedoch scheint möglich, dass A und B bzw. B und A^+ derart unterschiedliche Eigenschaften aufweisen, dass sie im Sinne der traditionellen evaluativen Relationen unvergleichbar sind. Sowohl «A (A^+) ist besser als B in Hinsicht W», als auch «A (A^+) ist schlechter als B in Hinsicht W», oder « A (A^+) ist gleich gut wie B in Hinsicht W» wäre dann falsch.

Gerade dadurch, dass A und B bzw. A^+ und B unvergleichbar sind, kann man in das zuvor geschilderte Problem schlittern, ohne ein Transitivitätsgesetz zu verletzen oder zyklisch zu handeln. Die einzige positive Beziehung, die vorliegt, ist die Überlegenheit von A^+ gegenüber A, und damit lässt sich weder ein Zyklus erzeugen noch gegen die Transitivität verstoßen.

Sind zwei Optionen unvergleichbar, was die traditionellen evaluativen Relationen betrifft, so ist doch nicht ausgeschlossen, dass sie in einem weiteren Sinn vergleichbar sind. Etwa in dem Sinn, dass A (A^+) und B sich zwar nicht als gleich gut präsentieren, sich jedoch in ein und *derselben Liga befinden*,[4] wobei sich weisen muss, wie das genau zu verstehen ist.

Jedes materiale Beispiel für Unvergleichbarkeit oder gar Parität, das je gegeben wurde, kann angezweifelt werden.[5] Allerdings dürfte auch klar

1997) S. 1-34; Ruth Chang: *The Possibility of Parity*, in *Ethics* 112 (2002) S. 659-688.

4 Im Original: *on a par*. Ob die Übersetzung von «on a par» mit «in derselben Liga wie» adäquat ist, hängt von der konkreten Paritätsdefinition ab, die verwendet wird. Es sind in der Literatur verschiedene Möglichkeiten besprochen worden, wie die Paritätsrelation definiert werden könnte. Nicht auf jede passt der Ligenbegriff. Er passt jedoch gut auf die Definition, die im Folgenden vorgebracht wird.

5 Typischerweise werden die Beispiele dadurch angegriffen, dass versucht wird zu zeigen, dass es sich bei den vermeintlichen Fällen von unvergleichbaren Optionen

sein, dass es, technisch betrachtet, kein Problem darstellt, Relationen zu definieren, die unvollständig sind,[6] und es wäre eher überraschend, wenn in unserer Umgangssprache keine Hinsichten existierten, die zwar evaluativ-komparativen Raum bieten, der aber traditionell unvollständig besetzt ist. Ist dies gegeben, besteht auch die Möglichkeit, dass weitere Wertrelationen[7] einen Teil dieses Raumes ausfüllen.[8]

Ein ernsthafteres Problem besteht darin zu zeigen, wie Parität aufgewiesen wird und welche Funktion ihr zukommt. Man könnte auch sagen, zu zeigen, wo die Relevanz von Parität liegt.

Es ist eher unwahrscheinlich, dass eine solch unorthodoxe evaluative Relation empirisch direkt erfragt werden kann. Bezogen auf das Beispiel müsste dies bedeuten, dass Probanden in Interviews tatsächlich äußerten, sie hielten weder A (A^+) für besser als B (in der Hinsicht W) noch B für besser als A (A^+), und es träfe auch nicht zu, dass die beiden Optionen genau gleich gut seien. Vielmehr verhalte es sich so, dass sich die Optionen in derselben Liga befänden. – Ließe sich Parität explizit aufweisen, gäbe es über ihre Existenz vielleicht weniger Streit.

Im Folgenden wird denn auch nicht die Möglichkeit eines *direkten* Aufweisens von Parität besprochen,[9] sondern eine *indirekte* Methode. Es wird

in Wahrheit um Fälle vager traditioneller Vergleichbarkeit handelt. Vgl. John Broome: *Is Incommensurability Vagueness?*, in *Incommensurability, Incomparability, and Practical Reason*, hg. von Ruth Chang (Cambridge, MA: Harvard University Press, 1997), S. 67-89, oder Nicolas Espinoza: *The small improvement argument*, in *Synthese* 165 (2008) S. 127-139.

6 Die Wirtschaftswissenschaften verwenden seit langem einen solchen Begriff, das Pareto-Prinzip der Verteilzustände. Von zwei Verteilzuständen ist der eine genau dann Pareto-mindestens-so-gut wie der andere, wenn darin jedes Mitglied der Gesellschaft mindestens so gut gestellt ist wie im anderen. Das lässt viel Raum für Pareto-unvergleichbare Verteilzustände.

7 Anders als bei Chang wird hier nicht vorausgesetzt, dass Parität die einzige nicht traditionelle Wertrelation ist.

8 Um ein ernsthaftes materiales Beispiel zu geben: In Georg Brun, Gertrude Hirsch: *Ranking policy options for sustainable development*, in *Poiesis&Praxis* 5 1 (2008) S. 15-31, wird die Möglichkeit besprochen, dass Policy options in Hinsicht auf nachhaltige Entwicklung in einer Paritätsrelation stehen. Brun und Hirsch favorisieren dabei den Paritätsbegriff von Wlodek Rabinowicz: *Value Relations*, in *Theoria* 74 (2008) S. 18-49. Dieser Paritätsbegriff ist weder mit dem von Chang identisch noch mit demjenigen, der hier eingeführt wird.

9 Die eben karikierte Interview-Methode ist gewiss nur eine unter vielen direkten Methoden zum Aufweis von Präferenzen.

gezeigt, wie Parität durch Wahlverhalten indirekt aufgewiesen oder eben *aufgedeckt* werden kann.[10]

Darüber hinaus – und damit zurück zu den irrationalen Sequenzen von Entscheidungen auf der Basis rationalen Wahlverhaltens – wird plausibilisiert, dass sich Parität als Lösung eines Entscheidungsproblems konstituiert. Parität kann als eine Relation verstanden werden, deren Funktion es ist, das Zufallsverfahren als rationales Verfahren zu legitimieren bzw. diskreditieren, wenn die Optionen, die zur Auswahl stehen, unvergleichbar sind.

So betrachtet ist Parität eine Rechtfertigungsinstanz, die vor dem Hintergrund eines spezifischen Problems angerufen werden kann. Dem Problem, wie auf der Basis einer rationalen Vorauswahl rationale Entscheidungen zu treffen sind.

Im Gegensatz zu einer orthodoxeren Lösung, die ebenfalls besprochen wird, ist beim Paritätsansatz für jede Entscheidung von Anfang an klar, ob sie vor dem Hintergrund möglicher Sequenzen rational ist. Die Rationalität und Irrationalität einzelner Entscheidungen aus einer Menge von Entscheidungen ist unabhängig davon, welche anderen Entscheidungen gefällt werden.

2. *Rationalisierung von Wahlfunktionen durch aufgedeckte Präferenzen*

Sei X im Folgenden eine endliche Menge von unabhängigen Handlungsoptionen und 2^X die Potenzmenge von X. Für Ökonomen, Sozialwahl- oder Entscheidungstheoretiker ist eine Wahlfunktion eine Abbildung $c:2^X \supseteq B \rightarrow 2^X$, $\emptyset \neq S \mapsto T \subseteq S$, also eine Abbildung, die mehr oder weniger vielen Mengen von Handlungsoptionen eine Teilmenge ihrer selbst zuordnet. Aus jeder Menge von Handlungsoptionen, die ihr eingespeist wird, wählt sie gleichsam einige Optionen aus.

Im Folgenden soll gelten, dass jede Wahlfunktion $c(\cdot)$ auf alle Teilmengen von Handlungsoptionen definiert ist, also $B = 2^X \setminus \emptyset$.[11] Überdies soll die Wahlfunktion immer mindestens eine Option auswählen, also $c(S) \neq \emptyset$ für alle $S \in B$.

Eine Wahlfunktion $c(\cdot)$ gibt im Allgemeinen also mindestens eine Option, nicht im Allgemeinen aber eine einzige Option aus, wenn auf eine Menge

[10] Die Methode, die hier verwendet wird, ist unter dem Namen «Revealed preference theory» bekannt.

[11] In den Beispielen mögen solche Mengen S von Optionen fehlen, für die trivial ist oder aus dem Kontext hervorgeht, wie aus ihnen ausgewählt wird.

$S \in B$ von Handlungsalternativen aus X angewandt, sondern eine Menge von Optionen. Wir wollen davon sprechen, dass eine Wahlfunktion im Allgemeinen die Lösung eines *Wahlproblems* darstellt – sie wählt aus jeder Menge $S \in B$ die Teilmenge $T \subseteq S$ der wählbaren Optionen aus –, nicht aber die Lösung eines *Entscheidungsproblems*, verstanden als Wahl eines einzelnen Elements aus jeder dargebotenen Menge von Optionen.

Nicht jede Wahlfunktion verdient das Prädikat *rational*. Im Optimierungsparadigma des Komparativismus liegt Rationalität genau dann vor, wenn die Zuordnungsregel, die $c(\cdot)$ definiert, durch Optimierung auf einer zweistelligen Relation auf X gestiftet sein könnte. Wenn, mit anderen Worten, $x \in S$ genau dann in c(S) ist, falls eine zweistellige Relation $\geq$ auf X existiert, so dass $x \in \max(S, \geq)$. Was genau unter «$\max(S,\geq)$» zu verstehen ist, hängt davon ab, mit welcher Variante des Optimierungsparadigmas man es zu tun hat. Gängig sind die Varianten *Maximalisierung* und *Maximierung*. In der Variante Maximierung umfasst $\max(S,\geq)$ die $\geq$-Maxima von S, d.h. $\max(S,\geq)=\{x \in S; \forall y \in S[x \geq y]\}$. In der Variante Maximalisierung handelt es sich bei $\max(S, \geq)$ hingegen um die $\geq$-maximalen Elemente von S, d.h. $\max(S,\geq)=\{x \in S; \neg \exists y \in S[y > x]\}$, wobei $>$ als strikter Teil[12] von $\geq$ zu verstehen ist, von dem üblicherweise ein symmetrischer Teil[13] $\sim$ unterschieden wird.

Die Rationalität des Optimierungsparadigmas tritt verschieden qualifiziert auf. Die Qualifizierung erfolgt durch die Einführung mehr oder weniger starker Bedingungen an die Eigenschaften der binären Relation $\geq$.

Die schwächste Eigenschaft, die von $\geq$ typischerweise verlangt wird, ist *Reflexivität*. Reflexivität zu fordern, stellt eine interpretativ-konzeptuelle Notwendigkeit dar: Man möchte $\geq$ als *schwache Präferenzrelation* verstehen, d.h., «$x \geq y$» soll ausdrücken, dass x für mindestens so attraktiv gehalten wird wie y. Von der schwachen Präferenz sind die starke Präferenz – «x ist attraktiver als y» – und die Indifferenz zu unterscheiden – «x und y sind gleich attraktiv». Wie immer die Präferenztheorie inhaltlich genau ausgefüllt wird, dem Begriff der Präferenz scheint folgendes eigen zu sein:

0) Schwache Präferenz ist reflexiv
1) Starke Präferenz ist asymmetrisch
2) Indifferenz ist reflexiv und symmetrisch
3) Starke Präferenz und Indifferenz sind disjunkt.

[12] Formal: $\forall x,y[x>y \leftrightarrow_{Def} x \geq y \wedge \neg(y \geq x)]$.

[13] Der symmetrische Teil ist folgendermaßen definiert:
$\forall x,y[x \sim y \leftrightarrow_{Def} x \geq x \wedge y \geq x]$.

Fordert man von ≥ Reflexivität, also 0), folgen überdies 1)-3) für den strikten Teil > von ≥, interpretiert als starke Präferenz, bzw. den symmetrischen Teil ~ von ≥, interpretiert als Indifferenz.[14]

Weitaus stärkere Bedingungen an die rationalisierende Relation ≥ stellt die Mikroökonomie, die sich mit der Wahl von Konsumgüterbündeln vor dem Hintergrund bestimmter Preis- und Vermögensrestriktionen beschäftigt. Sie fordert von ≥ Reflexivität, Transitivität und Vollständigkeit, d.h., ≥ soll eine sogenannt *vollständige Quasiordnung* sein.[15] Vollständigkeit und Transitivität zu fordern hat methodologisch-pragmatische Gründe. Die Wirtschaftswissenschaft möchte die weit ausgebauten Apparate der Analysis und der Statistik nutzen. Zu diesem Zweck müssen die Präferenzen jedoch in quantifizierter Form vorliegen, d.h., es muss für jede Präferenzrelation ≥ eine Funktion $f: X \rightarrow \mathbb{R}$ existieren, so dass $x \geq y \leftrightarrow f(x) \geq f(y)$. Solange X endlich ist, höchstens aber abzählbar unendlich, sind Vollständigkeit und Transitivität hinreichend für die Existenz einer solchen Funktion.[16] Ist X überabzählbar unendlich, ist überdies gefordert, dass ≥ stetig ist.[17]

Hier ist ein Beispiel (1) für eine Wahlfunktion, die sogenannt *vollständig-quasiordnungs-rational* ist. Sei X={a, b, c} und gelte c({a, b, c}) = {a}, c({a, b}) = {a}, c({a, c}) = {a} und c({b, c}) = {b, c}. Diese Wahlfunktion c(·) ist vollständig-quasiordnungs-rational, weil sie durch Optimierung auf der vollständigen Quasiordnung ≥ = {(a, a), (b, b), (c, c), (a, c), (a, b), (b, c), (c, b)} induziert sein könnte. Das heißt, eine Handlungsoption x aus einer Menge S von solchen, wird genau dann ausgewählt, gelangt also genau dann in c(S), wenn sie in demjenigen Teil von ≥ optimal ist, der nur Handlungsoptionen aus S enthält.

Die Bedingungen dafür, dass ein Wahlverhalten, ausgedrückt durch eine Wahlfunktion, mehr oder weniger qualifiziert rational ist, wurden in der zweiten Hälfte des vergangenen Jahrhunderts systematisch axiomatisiert.

14 Vgl. Sven Ove Hansson: *Preference Logic*, in *Handbook of Philosophical Logic*, Bd. 4, hg. von D. Gabbay, F. Guenther (Dordrecht: Kluwer, ²2001) S. 319-393, hier S. 322.

15 Für vollständige Quasiordnungen darf im Englischen «Weak orderings» als sehr gebräuchlich gelten.

16 Aussagen wie diese heissen in der Messtheorie «Repräsentationstheoreme». Für das vorliegende Repräsentationstheorem vgl. David H. Krantz, R. Duncan Luce, Patrick Suppes, Amos Tversky: *Foundations of Measurement*, Bd. I (Mineola, NY: Dover, 2007) S. 39.

17 Vgl. Andreu Mas-Colell, Michael D. Whinston: *Microeconomic Theory* (New York: Oxford University Press, 1995) S. 47.

Der Fall, in dem c(·) durch vollständige Quasiordnungen rationalisiert werden und überdies gilt, dass $B = 2^X \setminus \emptyset$ und $c(S) \neq \emptyset$, ist von Sen bereits 1971 weitgehend erschöpfend behandelt worden.[18] In anderen Feldern wird noch immer nach Axiomen geforscht.[19]

Formuliert werden die Axiome mit Hilfe des Begriffs der *aufgedeckten* Präferenz, der auf Samuelson zurückgeht:[20] Die Handlungsoption x gilt gegenüber der Handlungsoption y als durch eine Wahlfunktion *aufgedeckt schwach präferiert* (xRy), falls x gewählt wird und auch y zur Auswahl gestanden ist. Formal: $\forall x,y \in X[xRy \leftrightarrow_{Def} \exists S \in B[x \in c(S) \wedge y \in S]]$. Der strikte Teil von R, P, heißt «*aufgedeckt starke Präferenzrelation*».

Ein Axiom, das sich direkt mit R formulieren lässt, stammt von Richter und lautet: $\forall x \in X \forall S \in B[x \in S \wedge \forall y \in S[xRy] \rightarrow x \in c(S)]$.[21] In Worten, wenn es Handlungsoptionen auszuwählen gilt und eine Option existiert, die gegenüber allen Optionen, die zur Auswahl stehen, aufgedeckt schwach präferiert wird, so muss diese Option gewählt werden. Jede beliebige Wahlfunktion, die dieses Axiom erfüllt, ist zumindest unqualifiziert rational.[22] Das heißt, es gibt zumindest irgendeine zweistellige Relation $\geq$, so dass die Wahlfunktion der Maximierung auf $\geq$ entspricht.[23]

Zurück zur Überlegung, dass eine Wahlfunktion nicht im Allgemeinen für jedes $S \in B$ eine einzelne optimal attraktive Option ausgibt, sondern lediglich eine Teilmenge $T \subseteq S$ von gleichermaßen optimalen Optionen. Angenommen, unser Wahlverhalten werde über eine noch stärkere Relation als eine vollständige Quasiordnung rationalisiert, nämlich eine antisymmetrische, vollständige Quasiordnung, eine sogenannte *Kette*. Dann weist c(S) allerdings für jedes $S \in B$ genau ein Element auf. Das heißt, in diesem Spezialfall stellt die Lösung eines Wahlproblems zugleich die Lösung eines Entscheidungsproblems dar. Werden Entscheidungen vor dem Hintergrund

18 Vgl. Amartya K. Sen: *Choice Functions and Revealed Preference*, in *The Review of Economic Studies* 38/3 (1971) S. 307-317.

19 Vgl. z.B. Kfir Eliaz, Efe A. Ok: *Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences*, in *Games and Economic Behavior* 56 (2006) S. 61-86, wovon im Folgenden noch die Rede sein wird.

20 Vgl. Paul A. Samuelson: *A Note on the Pure Theory of Consumer's Behaviour*, in *Economica* 5/17 (1938) S. 61-71.

21 Vgl. Marcel K. Richter: *Rational Choice*, in *Preferences, Utility, and Demand*, hg. von John S. Chipman, Leonid Hurwicz, Marcel K. Richter, Hugo F. Sonnenschein (New York: Harcourt Brace Jovanovich, 1971) S. 29-58, hier S. 33.

22 Ibid.

23 R selbst ist eine solche Relation.

einer Kette rationalisiert, so entsteht nie das Problem, dass aus einer Menge von wählbaren Optionen eine einzige erwählt werden muss.

Sei nun eine *Entscheidungsfunktion* eine Abbildung $d:2^X \supseteq B \rightarrow X$, $\emptyset \neq S \mapsto x \in S$. So existiert offenbar für jede Wahlfunktion $c(\cdot)$, die über eine Kette rationalisiert worden ist, eine Entscheidungsfunktion $d(\cdot)$ mit $\{d(S)\}=c(S)$, für alle $S \in B$.

3. Rationalisierung von Entscheidungsfunktionen über rationale Wahlfunktionen

Das unerfreuliche Beispiel der Liegenschaften motiviert, nicht nur Wahl-, sondern auch Entscheidungsfunktionen Rationalitätsbedingungen zu unterwerfen. Die Rationalitätsbedingungen sollten insbesondere verhindern, dass eine Sequenz von rationalen Entscheidungen zu einem irrationalen Ergebnis führt.

Eine erste Rationalitätsbedingung regelt den Zusammenhang zwischen Wahlfunktion und Entscheidungsfunktion. Gewiss ist keine rationale Entscheidung zu erwarten, die nicht auf einer rationalen Wahl beruht. Vielmehr handelten wir *irrational*, wenn wir uns bezüglich einer Menge S von Optionen für eine Option entschieden, für die es keine Wahlfunktion $c(\cdot)$ gibt, so dass die Option in der Wahlmenge c(S) zu finden ist. Die Wahl kann als Mechanismus verstanden werden, welcher den Entscheidungsspielraum einschränkt. Schließt dieser Wahlmechanismus eine Option rationalerweise aus, so dürfen wir auch keine rationale Entscheidung zugunsten dieser Option erwarten.

Die erste Rationalitätsbedingung an Entscheidungsfunktionen $d(\cdot)$ lautet deshalb:

(I) $d(S) \in c(S)$, für eine rationale Wahlfunktion $c(\cdot)$

Eine zweite Rationalitätsbedingung betrifft Sequenzen von Entscheidungen, worunter das Beispiel mit den Liegenschaften einen Spezialfall darstellt. Angenommen, wir haben eine Sequenz $d(S_i)$, $S_i \in B$, $1 \leq i \leq n \in \mathbb{N}$, von Entscheidungen zu treffen, wobei wir Option $d(S_i)$ für Option $d(S_{i+1})$ aufgeben. So möchten wir vermeiden, dass wir uns an irgendeinem Punkt für eine Option entscheiden, die schlechter ist als eine Option, die früher zur Wahl gestanden hat – geschweige denn als eine Option, für die wir uns früher entschieden hatten und die wir dann aufgegeben haben. Wir müssten uns, wie im Beispiel

der Liegenschaften, den Vorwurf der Irrationalität gefallen lassen, wenn wir eine Entscheidungsregel anwendeten, die es erlaubte, dass wir nach einer Folge von Entscheidungen mit einer Option dastehen, die schlechter ist als eine, die wir verschmäht haben.

Die zweite Rationalitätsbedingung an Entscheidungsfunktionen $d(\cdot)$ lautet deshalb:

(II) Falls $d(S_i) \in S_{i+1}$ und $d(S_1) = x_1 \wedge d(S_2) = x_2 \wedge \ldots \wedge d(S_n) = x_n$, dann $\forall x \in S_1 [x_n \in c(\{x, x_n\})]$

Die Option, mit der wir uns am Ende einer Entscheidungskette wiederfinden, sollte sich als mindestens so gut wie jede, auf jeden Fall aber nicht schlechter als irgendeine Option aufdecken lassen, mit der wir auch hätten enden können.

Leicht könnte formal gezeigt werden, was vollkommen intuitiv ist: Wenn eine Entscheidungsfunktion auf der Basis einer Wahlfunktion definiert wird, die in Hinsicht auf eine Kette rational ist, dann erfüllt diese nicht nur Rationalitätsbedingung I, sondern auch II. Wie bereits zuvor argumentiert, fällt die Wahlfunktion in diesem Fall quasi mit einer Entscheidungsfunktion zusammen.

Unter der plausiblen Annahme, dass ziemlich häufig Entscheidungssituationen auftreten, in denen Wahlfunktionen über eine schwächere Struktur rationalisiert werden als über Ketten, müssen wir uns mit dem Gedanken anfreunden, dass nicht jede rationale Wahlfunktion mit genau einer Entscheidungsfunktion korrespondiert. Im Fall von (nicht antisymmetrischen) vollständigen Quasiordnungen finden wir uns leicht in einer Situation, in der unsere Wahlfunktion mit verschiedenen Entscheidungsfunktionen verträglich ist.

Ziehen wir etwa das vorne eingeführte Beispiel 1 heran, so erfüllen genau folgende zwei Entscheidungsregeln die Rationalitätsbedingung I:

$$- d_1(S) = \begin{cases} x, \text{ falls } c(S) = \{x\} \\ b, \qquad \text{sonst} \end{cases}$$

$$- d_2(S) = \begin{cases} x, \text{ falls } c(S) = \{x\} \\ c, \qquad \text{sonst} \end{cases}$$

Überdies gilt für dieses Beispiel, dass alle (beide) Entscheidungsregeln, welche die Bedingung (I) erfüllen, auch die Bedingung (II) erfüllen. Das ist eine ungeheuer angenehme Eigenschaft, da wir uns in solchen Fällen für ein *beliebiges* $x \in S$ entscheiden können, wenn immer die vorgeschaltete Wahl-

funktion c(·) für S kein eindeutiges Ergebnis liefert. Mit anderen Worten, wir können einen Würfel einsetzen, eine Münze werfen oder ein beliebiges anderes Zufallsverfahren wählen, um d(·) auf der Grundlage von c(·) zu definieren, ohne je Gefahr zu laufen, eine Anwendung dieser Entscheidungsregel lasse uns irrational handeln.

Das Ergebnis, das für dieses Beispiel gewonnen werden konnte, kann – auch das ist intuitiv zugänglich – verallgemeinert werden. Es kann leicht formal bewiesen werden, dass jede Entscheidungsregel, welche die Bedingung I erfüllt, auch die Bedingung II erfüllt, falls ihr eine vollständig-quasiordnungs-rationale Wahlfunktion zugehört. Die einzigen Fälle, die nicht bereits durch c(·) entschieden sind, betreffen äquivalente Optionen, und die können per Zufall entschieden werden.

4. Zufallsverfahren determinieren nicht allgemein rationale Entscheidungsfunktionen

Die Probleme beginnen, wenn die Bedingungen an die rationalisierende Präferenzrelation weiter abgeschwächt werden, zum Beispiel in Hinsicht auf Vollständigkeit. Betrachten wir also Wahlfunktionen c(·) die durch eine reflexive, transitive, nicht notwendig aber vollständige Präferenzrelation ≥ auf X, also eine Quasiordnung, rationalisiert werden können.

Hier ist ein Beispiel (2) für eine solche Wahlfunktion: Sei X = {a, b, c} und gelte c({a, b, c}) = {a, c}, c({a, b}) = {a}, c({a, c}) = {a, c} und c({b, c}) = {b, c}. c(·) ist rational, ja sogar reflexiv-transitiv-rational, da sie durch Maximalisierung auf der Basis der Quasiordnung ≥ = {(a, a), (b, b), (c, c), (a, b)} entstanden sein könnte.[24]

Unter den Entscheidungsregeln, welche die Rationalitätsbedingung I erfüllen, findet sich folgende:

$$d(S) = \begin{cases} x, \text{ falls } c(S) = \{x\} \\ c, \text{ falls } c(S) = \{a, c\} \\ b, \qquad\qquad \text{sonst} \end{cases}$$

Es gilt $d(\{a, c\}) = c$, $d(\{b, c\}) = b$, aber $b \notin c(\{a, b\}) = \{a\}$, weshalb die Entscheidungsregel d(·) die Rationalitätsbedingung II verletzt, uns also irrational entscheiden bzw. handeln lässt, wenn wir uns nach ihr richten.

[24] Die Wahlfunktion ist auch vollständig-rational, aber eben nicht vollständig-transitiv-rational, also nicht vollständig-quasiordnungs-rational. Die aufgedeckte Relation R macht die Wahlfunktion vollständig-rational. R ist aber nicht transitiv.

Dieses Beispiel zeigt, dass wir uns, unter den vielen Entscheidungsfunktionen, welche die Bedingung I erfüllen, rationalerweise *nicht* für eine beliebige entscheiden können, wenn die zugrunde liegende Wahlfunktion lediglich quasiordnungs-rational ist, nicht aber vollständig-quasiordnungsrational.

Damit ist klar, dass wir in solchen Fällen nicht im Allgemeinen auf ein Zufallsverfahren zurückgreifen dürfen, um zwischen Optionen zu entscheiden, die aus einem Wahlverfahren als gleichermassen optimal hervorgegangen sind. Zumindest dürfen wir dies nicht, wenn wir rational handeln wollen.

5. Zwei Methoden zur Erzeugung rationaler Entscheidungsfunktionen

Im Folgenden werden zwei Methoden vorgestellt, mit denen sich jede Festlegung auf eine irrationale Entscheidungsfunktion unterbinden lässt.

Die erste Methode stützt sich allein auf die herkömmlichen evaluativen Relationen, Überlegenheit bzw. strikte Präferenz und Äquivalenz bzw. Indifferenz. Die Möglichkeit, auf der Basis einer Wahlfunktion rational zu entscheiden, ist keinerlei weiteren strukturellen Einschränkungen unterworfen. Allerdings muss gefordert werden, dass bereits gefällte Entscheide im künftigen Entscheidungsverhalten Niederschlag finden. Die Definition der Entscheidungsfunktion erhält so eine prozedurale Dimension: Die Eigenschaften einer rationalen Entscheidungsfunktion $d(\cdot)$ sind nicht unabhängig von der Reihenfolge, in der die Werte von $d(\cdot)$ festgelegt werden. Oder anders: Liegt eine irrationale Entscheidungsfunktion $d(\cdot)$ vor, so lassen sich irrationale Entscheidungen nur in Abhängigkeit der Reihenfolge benennen, in der die Werte von $d(\cdot)$ bestimmt wurden.

Die zweite Methode ist frei von solch prozeduralen Aspekten. Sie restringiert die Entscheidungsfunktionen durch eine zusätzliche strukturelle Dimension – eine Paritätsrelation. In diesem Fall schlägt sich die Reihenfolge in keiner Weise auf die Eigenschaften rationaler Entscheidungsfunktionen nieder. Liegt eine irrationale Entscheidungsfunktion vor, so umfasst diese unabhängig irrationale Entscheidungen.

5.1 Die Anwendbarkeit des Zufallsverfahrens wird prozedural eingeschränkt

Im Detail funktioniert die erste Methode so, dass zwar ein Zufallsverfahren angewandt werden darf, jedoch nicht bedingungslos. Wann immer nach der Anwendung des Zufallsverfahrens eine weitere Entscheidung ansteht, müssen die «transitiven Folgen» des ersten Zufallsverfahrens herangezogen werden, um festzustellen, welche weiteren Entscheidungen durch die Anwendung des Zufallsverfahrens unter Berücksichtigung von Transitivität gefallen sind. Formal bedeutet dies nichts anderes, als die Rationalitätsbedingung II als logische Schlussregel (II‘) zu verwenden:

(II‘) $\{d(S_i) \in S_{i+1},\ d(S_1)=x_1,\ d(S_2)=x_2,\ldots,\ d(S_n)=x_n\} \vdash$
$\forall x \in S_1[x_n \in c(\{x, x_n\})]$

Am Beispiel der Liegenschaften lässt sich zeigen, wie mit dieser Schlussregel in verschiedenen Situationen zu verfahren wäre. Angenommen, $d(c(\{A^+, B\})) = d(\{A^+, B\})$ soll entschieden werden. Weder im Fall von $d(\{A^+, B\}) = A^+$ noch im Fall von $d(\{A^+, B\}) = B$ lässt sich mit II‘ ein Widerspruch herbeiführen. Es ist also rational, hier das Zufallsverfahren anzuwenden. Sei ferner vorausgesetzt, die Würfel fielen für A^+ (Fall 1). Steht in der Folge die Entscheidung $d(c(\{A, B\})) = d(\{A, B\})$ an, so sind erneut beide Fälle möglich, ohne dass sich mit II‘ ein Widerspruch erzeugen ließe. Das Zufallsverfahren ist also erneut zugelassen. Fallen die Würfel hingegen für B (Fall 2), so verbietet sich für $d(\{A, B\})$ das Zufallsverfahren, denn führt dieses zu A, so folgt mit $d(\{A^+, B\}) = B$, $d(\{A, B\}) = A$ und Schlussregel II‘, $A \in c(\{A^+, A\})$. Das steht im Widerspruch zur Wahlfunktion $c(\cdot)$, auf der die Entscheidung basiert. Es muss also $d(\{A, B\}) = B$ gelten, wonach keine Widersprüche auftreten.

Hat man es mit einer Situation zu tun, in der nicht eine rationale Entscheidungsfunktion zu erzeugen ist, sondern in der eine irrationale Entscheidungsfunktion vorliegt, so lassen sich nicht im Allgemeinen spezifische Fehlentscheidungen eruieren. Jeder Fehlentscheid ist ein solcher lediglich in Abhängigkeit einer Reihenfolge, in der die Entscheide getroffen wurden.[25]

[25] Dies ist eine substantielle Behauptung. Sie ist in der sogenannten Poset-Theorie formal bewiesen worden, der Theorie der antisymmetrischen Quasiordnungen (Partially ordered sets, posets). Vgl. William T. Trotter: *Combinatorics and Partially Ordered Sets* (Baltimore: Johns Hopkins University Press, 1992) S. 16. Dieser Beweis lässt sich leicht für Quasiordnungen verallgemeinern, indem äquivalente Optionen zu Äquivalenzklassen zusammengefasst und durch einen Repräsentanten dieser Klasse vertreten werden.

Liegt im Beispiel der Liegenschaften $d(c(\{A^+, B\})) = B$ und $d(c(\{A, B\})) = A$ vor, so ist jeder dieser beiden Entscheide lediglich unter der Bedingung irrational, dass die andere Entscheidung (vorher) gefällt worden ist.

Dass immer eine rationale Entscheidungsfunktion existiert, lässt sich formal beweisen.[26] Es wird also nie so sein, dass eine Wahlfunktion vorliegt, die über eine Quasiordnung rationalisierbar ist, auf deren Basis keine Entscheidungsfunktion definiert werden kann, welche die Rationalitätsbedingung II erfüllt.

5.2 Die Anwendbarkeit des Zufallsverfahrens wird strukturell eingeschränkt

Die zweite Methode geht auf Changs Idee zurück, Optionen könnten zwar traditionell unvergleichbar, im Sinn einer neuen Wertrelation, Parität, jedoch vergleichbar sein. Damit Parität definiert werden kann, beziehungsweise *aufgedeckte Parität*, müssen einige Vorarbeiten geleistet werden. Insbesondere muss ermöglicht werden, dass sich Optionen als traditionell unvergleichbar zueinander aufdecken lassen – solches ist in der Standardvariante der Theorie aufgedeckter Präferenzen nicht vorgesehen.

Wir haben vorne festgehalten, dass der strikte Teil, P, einer aufgedeckten schwachen Präferenz, R, als aufgedeckte starke Präferenzrelation interpretiert wird. Entsprechend wird auf der orthodoxen Linie der Theorie der aufgedeckten Präferenz der symmetrische Teil von R, bezeichnet mit «I», als aufgedeckte Indifferenz verstanden. Nun lässt sich auf verschiedene Weisen zeigen, dass diese Interpretation nicht immer adäquat ist.

Auf einer abstrakten Ebene lässt sich dies demonstrieren, indem man darauf hinweist, dass eine Wahlfunktion $c(\cdot)$, die – wie hier überall vorausgesetzt – auf alle Teilmengen von $2^X \setminus \varnothing$ definiert ist und die immer mindestens ein Element auswählt, keine unvergleichbaren Optionen aufdecken kann. Dies ergeht aus der Tatsache, dass, da $c(\cdot)$ auf allen Teilmengen definiert ist, $c(\cdot)$ insbesondere auf allen Paarmengen $\{x, y\}$ definiert ist für $x, y \in X$. Das hat zur Folge, dass für zwei beliebige Handlungsoptionen gilt: $x\,R\,y \vee y\,R\,x$. Eine solche Interpretation ist inadäquat, weil damit allein die Umstände einer Wahlsituation entscheiden, ob ein Wähler seine Unvergleichbarkeit zwischen Handlungsoptionen überhaupt ausdrücken kann. Es scheint angemessen, dass Unvergleichbarkeit auch dann ausdrückbar ist, wenn die Wahlsituation

[26] Die Behauptung folgt aus dem *Satz von Szpilrajn*, ibid. S. 17.

der Wählenden alle Paarmengen von Handlungsoptionen aufzwingt und überdies immer mindestens eine ausgewählte Option verlangt.

Auf einer konkreteren Ebene kann mit Beispielen gezeigt werden, dass mitunter Optionen ausgewählt werden, zwischen denen eine Wählende nicht indifferent ist, sondern die sie für unvergleichbar hält. Eliaz und Ok[27] bringen folgendes Beispiel, das noch deutlicher aufzeigt als es das Beispiel der Liegenschaften tut, wo die Gründe für die Unvergleichbarkeit von Optionen häufig verortet werden – in sich widersprechenden, grundlegenderen Ordnungen.

Frau Watson möchte für ihre zwei Kinder, Alice und Tom, genau eine Kopie eines Videofilms ausleihen. In Frage kommt überhaupt nur ein Film aus der Menge {a, b, c} der kindgerechten Filme. Alice und Tom haben bei ihrer Mutter folgende persönliche, explizite Präferenzen angemeldet, die es möglichst zu berücksichtigen gilt: b>c>a, b>a (Alice) bzw. a>b>c, a>c (Tom). In der Hoffnung, der Filmverleih möge alle Filme vorrätig haben, hat sich Frau Watson c({a, b, c})={a, b} zurechtgelegt, mit dem Vorsatz, zwischen a und b die Münze sprechen zu lassen. Die Wahl scheint vernünftig, würde doch, von Anfang an nur a zu wählen, Tom bevorzugen, und wählte sie ausschließlich b, gälte das gleiche für Alice. c auch noch wählen, wäre hingegen irrational, da beide Kinder b für besser erachten als c. Stellt sich dann heraus, dass doch nur a und c auf Lager sind, so trifft sie die Wahl c({a, c}) = {a, c}, erneut mit der Absicht, die definitive Entscheidung auszuwürfeln. In der veränderten Ausgangslage gehört c ausgewählt, schließlich zieht Alice c a vor. Frau Watson deckt insgesamt also die schwache Präferenz R = {(a, b), (b, a), (a, c), (c, a), (b, c), (c, a), (b, b), (c, c)} auf und damit nach traditionellem Verständnis die starke Präferenz P = {(b, c)} und die nicht transitive Indifferenz I = {(a, b), (b, a), (a, c), (c, a), (a,a), (b, b), (c, c)}.

Nun würde sich Frau Watson wohl dagegen verwahren, nicht transitive Präferenzen zu haben. Und in der Tat gibt es eine plausiblere Alternative, ihr Wahlverhalten zu deuten. Plausibler, als ihr Indifferenz zwischen a und b bzw. zwischen a und c zu unterstellen, lässt sich ihre Wahl als Ausdruck von Unvergleichbarkeit interpretieren. Sie kann sich auf den Standpunkt stellen, die sich in dieser Sache auf nicht symmetrische Weise widersprechenden Präferenzen von Alice und Tom ließen sie zu keinem positiven Urteil über die Vergleichbarkeit der besagten Optionen kommen.

[27] Vgl. Kfir Eliaz, Efe A. Ok: *Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences*, in *Games and Economic Behavior* 56 (2006) S. 61-86, hier S. 62-63.

Eliaz und Ok[28] ermöglichen das Aufdecken von Unvergleichbarkeit, indem sie zeigen, wie sich der symmetrische Teil einer aufgedeckten Präferenz in Optionspaare unterteilen lässt, zwischen denen der Wählende indifferent ist und solchen, die er für unvergleichbar hält. Der Term «$S_{y,-x}$» sei im Folgenden für alle Mengen S definiert, die x enthalten, nicht aber y, und er bezeichne die Menge $(S \cup \{y\}) \setminus \{x\}$. Gilt für zwei Optionen x, $y \in X$ und ein rationales $c(\cdot)$, dass $c(\{x, y\}) = \{x, y\}$, so sind die Optionen *aufgedeckt unvergleichbar* (I_{II}), falls ein $S \in B$ existiert, mit $x \in S \wedge y \notin S$ und

a) $x \in c(S) \wedge y \notin c(S_{y,-x})$, oder
b) $x \notin c(S) \wedge y \in c(S_{y,-x})$, oder
c) $c(S) \setminus \{x\} \neq c(S_{y,-x}) \setminus \{y\}$

Stark vereinfacht gesprochen sind zwei Optionen aufgedeckt unvergleichbar, wenn sie zwar als Paar betrachtet den Anschein erwecken, für indifferent gehalten zu werden, sobald jede Option für sich mit anderen Optionen in Verbindung gebracht wird, sich aber zeigt, dass sie sich gegenüber diesen anderen Optionen nicht wie zueinander äquivalente Optionen verhalten.[29]

Als Folge der Teilung von I muss auch «R» umgedeutet werden, denn kann «$xRy \wedge yRx$» nicht mehr allgemein als aufgedeckte Indifferenz interpretiert werden, so «xRy» nicht mehr allgemein als aufgedeckt schwache Präferenz. Eliaz und Ok bieten an, «xRy» dahingehend zu interpretieren, dass x *aufgedeckt nicht schlechter* ist *als y*.

Die Interpretation eines gewissen Wahlverhaltens als Ausdruck von Unvergleichbarkeit zwischen Handlungsoptionen reißt eine Lücke in den Raum X^2 aller Optionspaare. Anders als im traditionellen Verständnis, nach dem dieser Raum notwendig als vollständig geordnet aufgedeckt wird, sobald gewisse äußere Umstände gegeben sind, besteht nun die Möglichkeit, dass gewisse Optionen als unvergleichbar aufgedeckt werden.

Genauer sollte man hier jedoch sagen: als «traditionell unvergleichbar». Sie sind unvergleichbar in dem Sinn, als keine der traditionellen Vergleichsrelationen P und I aufgedeckt wird. Wie Chang plausibilisiert hat, muss dies nicht bedeuten, dass nicht eine andere, eine neue Vergleichsrelation einen Teil dieser Lücke füllte. Sie hat dafür den Begriff der Parität vorgeschlagen, hier: den Begriff des In-derselben-Liga-Seins.

28 Ibid. S. 66-67.

29 Formal und genauer: Damit $xI_{II}y$ gelten kann, muss eine Option z existieren, so dass entweder $xI_{II}z \wedge (yPz \vee zPy)$ oder $y \wedge (xPz \vee xPy)$. Ibid. S. 68. Diese Bedingung muss beachtet werden. Sie ist aber so schwach, dass jedes hier präsentierte Beispiel sie erfüllt.

Da äquivalente Optionen, wie gesehen, keine Probleme bieten, sollen im Folgenden keine solchen mehr auftreten. Vielmehr wollen wir von antisymmetrischen rationalisierenden Relationen ausgehen.

Um den Raum der traditionell unvergleichbaren Optionen wieder anzufüllen, definieren wir rekursiv Mengen von aufgedeckten Ligen. Die oberste aufgedeckte Liga einer durch R geordneten Menge X enthält genau die R-maximalen Elemente von X, also $\max(X, R) = \{ x \in X; \neg \exists y \in X [y P x] \}$. Die zweitoberste Liga enthält diejenigen Elemente von X, die maximal werden, wenn die oberste Liga aus X entfernt wird. Allgemein: Sei $X_1 = X$ und sei $X_{i+1} = X_i \backslash \max(X_i, R)$. So ist $\max(X_i, R)$ die *i-te Liga*, die durch R aufgedeckt wird.

Aufgedeckte Parität kann nun über die Zugehörigkeit zu ein und derselben Liga definiert werden: Gilt für zwei Optionen $x, y \in X$, dass sie als traditionell unvergleichbar aufgedeckt werden, also $x I_{II} y$, so wird x *in derselben Liga wie* y *aufgedeckt* ($x I_{\chi(} y$), falls $\exists i \in \mathbb{N} [x \in \max(X_i, R) \wedge y \in \max(X_i, R)]$.

Die aufgedeckte Paritätsrelation ist *irreflexiv*, *symmetrisch* und *distinkt transitiv*.[30] Hararys Idee,[31] Relationen «Paritätsrelationen» zu nennen, die irreflexiv, symmetrisch und distinkt transitiv sind, scheint in der philosophischen Debatte zu neuen evaluativen Relationen bislang unberücksichtigt geblieben zu sein. Die Irreflexivität ergibt sich daraus, dass die Optionen als traditionell unvergleichbar aufgedeckt sein müssen, jede Option gegenüber sich selbst aber als äquivalent aufgedeckt wird. Die Symmetrie und die eingeschränkte Form der Transitivität folgen leicht aus der Definition von $I_{\chi(}$.

Eine solch ligenartige Paritätsrelation weicht in mindestens einem formalen und einem inhaltlichen Punkt von der Parität ab, die Chang bespricht. Formal unterscheiden sich die beiden Ansätze dadurch, dass die hier eingeführte Paritätsrelation bis auf eine technische Einschränkung transitiv ist. Chang hingegen setzt auch für paarweise nicht äquivalente Optionen keine Transitivität voraus.

Inhaltlich unterscheiden sich die Ansätze dadurch, dass die hier eingeführte Relation keinerlei Anspruch auf Exklusivität erhebt. Während Chang darauf besteht, dass genau eine weitere evaluative Relation den Raum der traditionell unvergleichbaren Optionen strukturiert, ist es hier vielmehr so,

30 Der Terminus stammt aus Frank Harary: *A Parity Relation Partitions its Field Distinctly,* in *The American Mathematical Monthly* 68/3 (1961) S. 215-217, und die Relation ist folgendermassen definiert: Für beliebige Optionen $x,y,z \in X$, von denen keine zwei identisch sind, gilt $x I_{\chi(} y \wedge y I_{\chi(} z \rightarrow x I_{\chi(} z$.

31 Ibid.

dass sich eine weitere Strukturierung nachgerade aufdrängt. Ähnlich wie Äquivalenzrelationen, also Relationen, die reflexiv, symmetrisch und transitiv sind, zerlegt die Paritätsrelation die Menge X in eine Partition,[32] d.h. es gibt eine Menge $\{Y_1, \ldots, Y_n\}$ von Teilmengen Y_i von X, so dass

i) $\cup_i^n Y_i = X$

ii) $Y_i \cap Y_j = \varnothing$, falls $i \neq j$

iii) $\forall x,y \in X[x I_\chi y \leftrightarrow \exists i \in \{1, \ldots, n\}[x \in Y_i \wedge y \in Y_i]]$

Die aufgedeckten Ligen, $\max(X_i, R)$, bilden präzise eine solche Partition.

Es bietet sich nun an, nicht nur diejenigen traditionell unvergleichbaren Optionen in eine neue Relation zu stellen, die sich in derselben Liga befinden – in die Paritätsrelation –, sondern auch alle anderen traditionell unvergleichbaren Optionen anhand ihrer Ligenzugehörigkeit zu ordnen. Zu diesem Zweck kann eine Abbildung $l(\cdot)$ eingeführt werden, die für jede Option diejenige natürliche Zahl ausgibt, die der Klasse entspricht, der sie zugehört. Die Ordnung durch Ligenzugehörigkeit (xI_Py) ergibt sich dann durch: $\forall x,y \in X[x I_P y \leftrightarrow_{Def} x I_{II} y \wedge l(x) < l(y)]$.

Mit Hilfe der Paritätsrelation und der induzierten ligenweisen Überlegenheitsrelation lässt sich sicherstellen, dass die Rationalitätsbedingung II immer eingehalten wird, wenn die Rationalitätsbedingung I erfüllt ist. Dafür ist lediglich notwendig, das Zufallsverfahren auf diejenigen Optionen einzuschränken, die in einer möglichst hohen Liga in der Paritätsrelation stehen. Formal:

(III) $d(S) \in \{x \in c(S); \forall y \in c(S)[x \neq y \rightarrow x I_\chi y \vee x I_P y]\}$

Entscheidungsfunktionen können also auf der Basis von Wahlfunktionen noch immer per Zufallsentscheid ermittelt werden, allerdings kommen jeweils ausschließlich diejenigen Optionen in Frage, zwischen denen paarweise Parität herrscht, und zwar in der höchsten Liga.

Für das Beispiel der Liegenschaften folgt daraus, dass nur die Entscheidung zwischen dem Häuschen am See (B) und der Stadtliegenschaft mit Dachfenster (A^+) beliebig ausfallen darf, nicht aber die zwischen B und der Stadtliegenschaft (A). Letztere Entscheidung muss für B ausfallen. Für Frau Watson folgt, dass sie rationalerweise Toms Präferenz a nachgeben muss, und Alice das Nachsehen hat.

32 Ibid.

Die erhöhte Strukturierung, die sich durch den Einbezug von Ligen ergibt, hat überdies zur Folge, dass Fehlentscheidungen unabhängig von anderen Entscheidungen festgemacht werden können. Anders als bei der Lösung durch die Berechnung von transitiven Folgen, ist $d(c(\{A, B\})) = A$ immer irrational, wenn als zusätzliche Entscheidungsstruktur diejenige Paritätsrelation herangezogen wird, die hier definiert worden ist. Die Entscheidung ist irrational, weil A und B zwar traditionell unvergleichbar sind, A sich aber in einer tieferen Paritätsklasse befindet als B, weshalb B A zwingend vorgezogen werden muss.

6. Schluss

Einige abschließende Bemerkungen: Der Umstand, dass I_P asymmetrisch und transitiv ist und I_χ vereinigt mit I reflexiv und transitiv, könnte diese Lösung dem Verdacht aussetzen, es handle sich bei den scheinbar neuen evaluativen Relationen lediglich um eine Ausdehnung der traditionellen Relationen der aufgedeckten strikten Präferenz bzw. der aufgedeckten Indifferenz. Dieser Verdacht kann aber mit dem Hinweis ausgeräumt werden, dass Überlegenheit durch Ligenzugehörigkeit und strikte Präferenz bzw. Parität und Indifferenz auf ganz verschiedene Weisen aufgedeckt werden. Das Wahlverhalten, das entscheidend ist für das Vorliegen der jeweiligen Beziehung, unterscheidet sich, wie dargelegt, in allen vier Fällen. Deshalb kann auch keiner dieser Fälle in einen anderen kollabieren. Es handelt sich um vier unterschiedliche Fälle von aufgedeckten Relationen.

Ein zweiter Einwand könnte lauten, dass weder Parität noch Überlegenheit durch Ligenzugehörigkeit echte zweistellige Relationen sind. Schließlich lassen sich die Wahrheitsbedingungen für eine Aussage der Form «$xI_\chi y$» bzw. «$xI_P y$» nicht angeben, ohne Rückgriff auf weitere Optionen. Die Beobachtung ist zwar richtig, der Einwand aber haltlos. Es ist nicht ungewöhnlich, dass das Zutreffen einer zweistelligen Relation nicht nur von intrinsischen Eigenschaften der in Relation stehenden Objekte abhängt, sondern vom Verhältnis dieser Objekte zu weiteren Objekten. Insbesondere hindert diese *Kontextabhängigkeit* nicht, die Relation als zweistellig zu betrachten. Entscheidend für Wohldefiniertheit und Stelligkeit einer Relation ist, dass sich alle Variablen im Definiendum ihrer Definition unterscheiden und dass das Definiens keine weiteren freien Variablen enthält.[33]

[33] Vgl. Patrick Suppes: *Introduction to Logic* (Belmont: Wadsworth, 1957) S. 156.

Studia philosophica 68/2009

PETER SCHULTE

Was ist instrumentelle Irrationalität?

In this paper, I start from the observation that there are obvious instances of instrumental irrationality, i.e. cases where subjects act knowingly against their strongest preferences. This observation raises an important question: Which facts determine the 'strength' of preferences? I consider a standard answer to this question – 'revealed preference theory' – which turns out to be unsatisfactory. Then I turn to a more promising alternative: the 'higher order theory' of preference strength. But this proposal also faces a major problem, the 'problem of authority'. I argue that this difficulty can be overcome, but only if we are prepared to make some significant normative assumptions – assumptions which imply, among other things, that instrumentalism – the view that instrumental rationality is all there is to rationality – cannot be maintained.

1. Einleitung

Wir schreiben das Jahr 1863. In der Nähe von Gettysburg, Pennsylvania, liefern sich Konföderierte und Unionstruppen eine erbitterte Schlacht. Tex, ein Soldat der Konföderierten, hat gerade eine schwere Verletzung am Bein erlitten und wird ins Lazarett gebracht. Der Sanitätsoffizier sieht sofort, dass Tex kaum eine Überlebenschance hat, wenn das Bein nicht amputiert wird. Er teilt Tex seine Diagnose mit. Tex ist mit der Operation einverstanden, obwohl er weiß, dass keine Narkosemittel zur Verfügung stehen. Doch als der Sanitätsoffizier die Amputationssäge hervorholt, schüttelt Tex den Kopf. «Nein, tu das nicht», sagt er und versucht, den Offizier wegzustoßen.[1]

Was geht in Tex vor? Es bieten sich hier, wie mir scheint, drei Interpretationsmöglichkeiten. *Erstens* könnte es sein, dass Tex sich über seine eigenen Präferenzen getäuscht hat, als er sich mit der Operation einverstanden erklärte: Er dachte, dass ihm eine hohe Überlebenschance wichtiger sei als die Vermeidung von Schmerzen, erkennt aber in dem Moment, in dem er die

[1] Dies ist die ausgeschmückte Version eines Beispiels von Christine Korsgaard. Vgl. Christine M. Korsgaard: *The Normativity of Instrumental Reason*, in *Ethics and Practical Reason*, hg. von Garrett Cullity, Berys Gaut (Oxford: Clarendon Press, 1997) S. 215-245. Das Beispiel findet sich auf S. 237-238.

Amputationssäge vor sich sieht, dass er in Wirklichkeit lieber die Schmerzen vermeiden möchte. *Zweitens* wäre es möglich, dass sich Tex' Präferenzen kurzfristig verändert haben: Als er der Operation zustimmte, zog er eine hohe Überlebenschance der Vermeidung von Schmerz vor, aber wenige Minuten später hat sich – angesichts der Amputationssäge – sein Präferenzprofil verändert; nun hat die Schmerzvermeidung für ihn Priorität. Aber es gibt noch eine *dritte* Möglichkeit. Es könnte sein, dass Tex die ganze Zeit über eine stärkere Präferenz für die höhere Überlebenschance hat, aber im entscheidenden Moment *instrumentell irrational handelt*, also nicht das tut, was im Lichte seiner Präferenzen insgesamt am besten wäre. Der Anblick der Amputationssäge löst bei ihm eine Angstreaktion aus, die dazu führt, dass er 'wider besseres Wissen' versucht, die Operation zu verhindern. Handlungen dieser Art werden traditionell als Fälle von 'Akrasie' oder 'Willensschwäche' bezeichnet.

Welche dieser drei Interpretationen – *Selbsttäuschung*, *Veränderung* oder *Willensschwäche* – ist korrekt? Es ist im Prinzip denkbar, dass die erste Interpretation zutrifft. Vielleicht wurde Tex kurz vor der Operation bewusst, dass ihm sein eigenes Leben eigentlich gar nicht so viel bedeutet, wie er immer dachte. Ähnliches gilt für die zweite Interpretation. Es ist nicht ausgeschlossen, dass Tex im Augenblick vor der Operation ein Erlebnis hatte, dass seine Präferenzen radikal veränderte und aus ihm einen Menschen machte, der zwischen Leben und Tod keinen gewichtigen Unterschied sieht. Doch diese beiden Deutungen wirken sehr weit hergeholt. Weit plausibler erscheint die dritte Alternative: Tex ist ein gewöhnlicher Mensch, dem sein eigenes Überleben sehr wichtig ist, und dies ändert sich auch nicht in dem Moment, in dem er die Amputationssäge sieht. Tex handelt einfach *unvernünftig*; seine Angst ist so stark, dass er nicht das tut, was – wie er selbst weiß – für ihn das Beste wäre. *Prima facie* ist seine Handlung ein typischer Fall von Willensschwäche.[2]

[2] Man könnte hier einwenden, dass Tex möglicherweise aus einem 'inneren Zwang' heraus handelt, und dass Handlungen aus innerem Zwang nicht unter die klassische Definition von Willensschwäche fallen, nach der Willensfreiheit eine notwendige Bedingung für Willensschwäche ist (vgl. Gary Watson: *Skepticism about Weakness of Will*, in *The Philosophical Review* 86 [1977] S. 317). Doch dieser Einwand ist hier nicht relevant. Für meine Zwecke ist es unerheblich, ob Tex' Handlung frei oder unfrei ist; allein auf die *instrumentelle Irrationalität* seiner Handlung kommt es mir an.

Auch in weniger extremen Situationen scheint Willensschwäche weit verbreitet zu sein, und oft (wenn auch nicht immer)[3] sind willensschwache Handlungen zugleich Fälle von instrumenteller Irrationalität. Das gilt für zahlreiche Standardbeispiele aus der Literatur, etwa für die Wissenschaftlerin, die einen Aufsatz fertigschreiben will, aber dennoch das Fußballspiel im Fernsehen zu Ende schaut; für den Übergewichtigen, der abnehmen will und trotzdem ein Schokoladenkeks nach dem anderen verspeist; oder für den Schüchternen, der ein unangenehmes Telefongespräch tagelang hinauszögert. Alle diese Personen entscheiden sich für Handlungen, die *relativ zu ihren eigenen Präferenzen* suboptimal sind. Die Realität instrumentell irrationaler Handlungen ist also kaum zu bestreiten.

Trotzdem erscheint das Phänomen der instrumentellen Irrationalität vielen Theoretikern mysteriös. Die Fragen «Wie ist instrumentell irrationales Handeln möglich?» und insbesondere «Wie ist Willensschwäche möglich?» sind Gegenstand zahlreicher philosophischer Erörterungen. Das Problem wird traditionell darin gesehen, dass instrumentelle Irrationalität oder Willensschwäche *prima facie* mit bestimmten, intuitiv plausiblen Prinzipien der Handlungstheorie inkompatibel zu sein scheint, und dass deshalb die augenscheinliche Realität dieser Phänomene Rätsel aufgibt.[4] Ich denke jedoch, dass vorher noch eine grundlegendere Frage beantwortet werden muss, die in der Forschung bislang eher vernachlässigt wurde: Was ist instrumentelle Irrationalität eigentlich genau? Welche Tatsachen müssen vorliegen, damit eine Handlung als «instrumentell irrational» bezeichnet werden kann?

Das ist die Frage, der ich in den folgenden Abschnitten nachgehen werde. In Abschnitt 2 skizziere ich eine erste, einfache Antwort, die aber in einem entscheidenden Punkt unbefriedigend ist, da sie den Begriff der *Präferenzstärken* unanalysiert lässt. In Abschnitt 3 stelle ich kurz die klassische Konzeption von Präferenzstärken vor, die «Theorie der offenbarten Präferenzen» («revealed preference theory»), die sich aber – zumindest für unsere Zwecke – als ungeeignet erweist. Ich gehe daher in Abschnitt 4 auf eine alternative Konzeption

3 Willensschwäche wird gewöhnlich definiert als (freies) Handeln ‘gegen besseres Urteil’, d.h. gegen das Urteil, dass eine andere (dem Akteur offen stehende) Handlung besser wäre. Der Ausdruck ‘besser’ kann dabei auch in einem moralischen Sinn gemeint sein. Der ‘willensschwache’ Akteur handelt dann in einer Weise, die er selbst für ‘moralisch suboptimal’ hält. Solche Handlungen sind nicht zwangsläufig Instanzen von instrumenteller Irrationalität.

4 Ein exemplarischer Fall ist hier Donald Davidson: *How is Weakness of the Will Possible?*, in *The Essential Davidson*, hg. von Donald Davidson (Oxford: Clarendon Press, 2006) S. 72-89.

ein, die zur Bestimmung von Präferenzstärken auf höherstufige Präferenzen zurückgreift. Ich komme zu dem Schluss, dass sich die alternative Analyse verteidigen lässt – allerdings nur dann, wenn man bereit ist, bestimmte normative Voraussetzungen zu akzeptieren. (Analoges gilt für eine weitere theoretische Alternative, die Präferenzstärken über hypothetische Entscheidungen unter Idealbedingungen definiert, vgl. Abschnitt 5.) Diese Überlegungen haben interessante Konsequenzen. Insbesondere zeigt sich, so meine weitergehende These, dass die normativen Voraussetzungen, die eine adäquate Analyse von instrumenteller Irrationalität mit sich bringt, mit einem reinen Instrumentalismus inkompatibel sind (vgl. Abschnitt 6).

2. Was ist instrumentelle Irrationalität?

Kehren wir noch einmal zum Fall von Tex zurück. Tex hat zwei Handlungsoptionen: Er kann den Sanitätsoffizier daran hindern, das Bein zu amputieren (Option h) oder ruhig liegenbleiben und die Operation über sich ergehen lassen (Option h*). Er kennt die voraussichtlichen Folgen beider Handlungsweisen. Wenn er h wählt, wird er mit großer Wahrscheinlichkeit sterben (Konsequenz T); wenn er sich für h* entscheidet, wird er zwar starke Schmerzen ertragen müssen, aber mit großer Wahrscheinlichkeit überleben (Konsequenz L). Warum ist es für Tex instrumentell rational, h* zu wählen? Die einfache Antwort lautet: weil ihm (wie den meisten Menschen) sein Überleben sehr viel wichtiger ist als die kurzfristige Vermeidung von Schmerz. Oder, anders formuliert: weil seine Präferenz für L erheblich stärker ist als seine Präferenz für T.

Dieses Resultat lässt sich verallgemeinern: Für die Frage, ob eine bestimmte Handlung instrumentell rational oder irrational ist, sind die Präferenzstärken des handelnden Subjekts entscheidend. (Die Wissenschaftlerin hat eine stärkere Präferenz für eine erfolgreiche Karriere als für einen gemütlichen Fernseh-Fußballabend; der Übergewichtige hat eine stärkere Präferenz dafür, abzunehmen, als dafür, weiter Kekse zu essen usw.) Eine präzise Definition instrumenteller Irrationalität muss allerdings auch die Wahrscheinlichkeitseinschätzungen des Subjekts (die sogenannten ‘subjektiven Wahrscheinlichkeiten’) berücksichtigen. Schließlich sähe selbst im Fall von Tex die Situation anders aus, wenn die Beinamputation seine Überlebenschancen nur minimal verbessern würde (und Tex dies wüsste) – unter solchen Umständen würden wir Tex’ Entscheidung nicht unbedingt als irrational beurteilen. Ein Begriff, der diese Komplikationen berücksichtigt,

ist der entscheidungstheoretische Begriff des erwarteten Nutzens («expected utility»). Im erwarteten Nutzen einer Handlung sind Präferenzstärken und subjektive Wahrscheinlichkeiten miteinander verrechnet. Damit können wir instrumentelle Rationalität und Irrationalität nun exakt definieren: Instrumentell rational sind Handlungen, die den erwarteten Nutzen des handelnden Subjekts maximieren; alle anderen Handlungen sind instrumentell irrational.

Die Details der entscheidungstheoretischen Konzeption des erwarteten Nutzens (und die Kontroversen um diese Details) sind für unsere Diskussion nicht relevant. Entscheidend ist, dass auch für den erwarteten Nutzen einer Handlung die Präferenzstärken eine entscheidende Rolle spielen. Die Frage nach der Natur instrumenteller Irrationalität kann also erst dann als beantwortet gelten, wenn der Begriff der «Präferenzstärken» geklärt ist. Was bedeutet es, dass eine Präferenz «stärker» ist als eine andere? Welche Tatsachen determinieren zum Beispiel, dass Tex' Präferenz für ein langes Leben stärker ist als seine Präferenz für Schmerzvermeidung? Wie die Diskussion in den nächsten Abschnitten zeigen wird, sind diese Fragen nicht leicht zu beantworten.

3. Präferenzstärken und die Theorie der offenbarten Präferenzen

Ist es tatsächlich so, dass der Begriff der Präferenzstärken Probleme aufwirft? Auf den ersten Blick scheint dies nicht der Fall zu sein. Schließlich haben John von Neumann und Oskar Morgenstern eine Analyse des «subjektiven Nutzens» entwickelt, die bis heute von der großen Mehrheit aller Entscheidungstheoretiker akzeptiert wird,[5] und der subjektive Nutzen von Sachverhalten (oder 'Gütern') ist im Grunde nichts anderes als ein Maß für die Präferenzstärken des Subjekts. (Je stärker die Präferenz einer Person S für einen Sachverhalt f ist, desto höher ist – *per definitionem* – der subjektive Nutzen, den f für S hat.) Wenn die Von-Neumann-Morgenstern-Analyse allgemein anerkannt ist, dann scheint es so, als ob unser Problem schon gelöst ist.

Doch dieser Eindruck täuscht. Der Von-Neumann-Morgenstern-Analyse liegt nämlich (zumindest in ihrer ursprünglichen Form) eine problematische Annahme zugrunde, die in der Literatur gewöhnlich als «Theorie der offenbarten Präferenzen» («doctrine of revealed preference») bezeichnet wird.

5 Vgl. John von Neumann, Oskar Morgenstern: *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944) S. 15-31.

Nach dieser Auffassung «offenbaren» sich die Präferenzen einer Person in ihrem Wahlverhalten, genauer: Das Wahlverhalten einer Person *legt fest*, welche Präferenzen sie hat und wie stark diese Präferenzen sind. Wenn sich jemand für eine bestimmte Option entscheidet, dann *bedeutet das*, dass diese Option – relativ zu seinen Meinungen und Präferenzen – optimal ist.[6]

Die Theorie der offenbarten Präferenzen impliziert, dass instrumentelle Irrationalität unmöglich ist. Wenn Tex versucht, die Amputation seines Beines zu verhindern, dann bedeutet das, dass ihm – zumindest in diesem Moment – die Vermeidung von Schmerz wichtiger ist als eine hohe Überlebenschance. Dasselbe gilt, *mutatis mutandis*, für die Wissenschaftlerin, den Übergewichtigen und den Schüchternen (vgl. Abschnitt 1); sie alle tun das, was im Lichte ihrer gegenwärtigen Präferenzen optimal ist – auch wenn sie das selbst nicht zugeben würden.

Wer die «Theorie der offenbarten Präferenzen» vertritt, muss bestreiten, dass es 'willensschwache' Handlungen gibt. Er muss alle Fälle, die gewöhnlich als Instanzen von Willensschwäche betrachtet werden, anders analysieren – z.B. als Fälle von Selbsttäuschung oder abrupter Präferenzveränderung. Diese alternativen Interpretationen sind aber in der Regel *psychologisch höchst unplausibel*. Wir haben daher gute Gründe, die «Theorie der offenbarten Präferenzen» abzulehnen. Donald Hubin resümiert die Situation treffend, wenn er schreibt:

> [T]his account of 'utility' [the doctrine of revealed preference; PS] has little to recommend it anyway. Its starkly behaviorist credentials – perhaps attractive to an earlier generation of philosophers – do not hold out much attraction now even if our objective is a predictive or explanatory theory of intentional human behavior.[7]

Wenn wir dieser Argumentation folgen, müssen wir zugestehen, dass die Stärke von Präferenzen nicht durch das Wahlverhalten des handelnden Subjekts determiniert wird. Auch (einfache) Handlungsdispositionen scheiden unter diesen Umständen als determinierende Faktoren aus, und zwar aus

6 Die «Theorie der offenbarten Präferenzen» hat insbesondere unter Wirtschaftswissenschaftlern viele Anhänger. Der Ökonom Kelvin Lancaster z.B. schreibt: «Action indicates preference. That is, given a choice, a consumer will actually take the collection [das 'Güterbündel'; PS] he prefers.» Und er fügt hinzu: «[T]he choice actually made by a consumer in any given situation gives him the greatest welfare he can attain in the situation.» Vgl. Kelvin Lancaster: *Introduction to Modern Microeconomics* (Chicago: Rand McNally, 1969) S. 186-187.

7 Donald C. Hubin: *The Groundless Normativity of Instrumental Rationality*, in *The Journal of Philosophy* 98 (2001) S. 451.

analogen Gründen.[8] Wenn z.B. gelten würde: «S's Präferenz für f ist stärker als seine Präferenz für f* gdw. er dazu *disponiert* ist, f gegenüber f* vorzuziehen», dann wäre instrumentelle Irrationalität wiederum unmöglich, da niemand etwas tun kann, das seinen Dispositionen (im philosophischen Sinn) zuwiderläuft. Die Frage danach, was die Stärke von Präferenzen festlegt, muss auf andere Weise beantwortet werden.

Bevor ich zu den alternativen Analysen komme, noch zwei kurze Anmerkungen. Erstens könnte man einwenden, dass meine Kritik an der «Theorie der offenbarten Präferenzen» verfehlt ist, weil ihre Vertreter sie nicht als Analyse, sondern als *stipulative Definition* des Präferenzbegriffs verstehen. Dieser Einwand ist nicht unberechtigt; mir scheint in der Tat, dass die «Theorie der offenbarten Präferenzen» manchmal so verstanden wird. Doch das ist kein Problem für meine Position. In diesem Aufsatz geht es um die Analyse von instrumenteller Irrationalität und Willensschwäche, d.h. um Fragen der *philosophischen Psychologie*,[9] und ein stipulativer Präferenzbegriff – d.h. ein Begriff, der gar nicht den *Anspruch* erhebt, psychologisch realistisch zu sein – ist für die Beantwortung solcher Fragen irrelevant. (Die Beobachtung, dass eine willensschwache Handlung in einem speziellen, stipulativen Sinn von «Präferenz» meine «Präferenzen» optimal befriedigt, trägt zur philosophischen Analyse von Willensschwäche wenig bei.)

Zweitens könnte man den Eindruck haben, dass ich ein wichtiges Element der klassischen Entscheidungstheorie – die Von-Neumann-Morgenstern-Analyse des subjektiven Nutzens – zu leichtfertig über Bord werfe. Dies wäre allerdings ein Missverständnis. Ich möchte nicht bestreiten, dass die Von-Neumann-Morgenstern-Analyse eine wichtige Einsicht enthält. Sie zeigt, dass man subjektive Nutzenwerte (und damit auch Präferenzstärken) auf einer Intervallskala angeben kann, wenn man Entscheidungen zwischen verschiedenen 'Lotterien' berücksichtigt.[10] Ich denke, dass eine adäquate Analyse von Präferenzstärken dieser Erkenntnis Rechnung tragen muss.

8 'Raffinierte' dispositionale Analysen sind von diesem Einwand nicht betroffen; vgl. Abschnitt 5.

9 Genauer: um Fragen der philosophischen Psychologie und der Handlungstheorie. Ich denke aber, dass diese beiden Gebiete ohnehin so eng zusammengehören, dass sie kaum sinnvoll getrennt behandelt werden können.

10 Vgl. von Neumann/Morgenstern, op. cit. (Fn. 5) S. 18. Von Neumann und Morgenstern verwenden allerdings eine andere Terminologie: Sie sprechen von «combinations» oder «combined events». Der Ausdruck «Lotterien» kommt erst später auf, vgl. etwa R. Duncan Luce, Howard Raiffa: *Games and Decisions* (New York: Dover Publications, 1989) S. 21-22.

Was ich hier behaupte, ist lediglich, dass die *ursprüngliche* Von-Neumann-Morgenstern-Analyse – die die «Theorie der offenbarten Präferenzen» voraussetzt – nicht haltbar ist; eine modifizierte Version dieser Analyse (die auf einer der Theorien basiert, die in den nächsten Abschnitten diskutiert werden) halte ich dagegen für sehr aussichtsreich.

4. Präferenzstärken und Präferenzen höherer Stufe

Nach der «Theorie der offenbarten Präferenzen» ist immer diejenige Präferenz am stärksten, die sich *de facto* durchsetzt. In einem Sinn von 'Stärke' ist das sicher korrekt: Dadurch, dass Tex' Präferenz für Schmerzvermeidung in der entscheidenden Situation die Oberhand behält, erweist sie sich in einem rein kausalen Sinn als 'stärkste Präferenz'. Doch das ist nicht der Stärkebegriff, auf den es in unserer Diskussion ankommt. In Abschnitt 2 habe ich instrumentell irrationale Handlungen (grob) als diejenigen Handlungen charakterisiert, deren Ausführung den stärksten Präferenzen des Akteurs zuwiderläuft. In dieser Charakterisierung steht der Ausdruck 'stärkste Präferenz' für das, was dem Akteur am *wichtigsten* ist – für das, was er *'eigentlich'* oder *'wirklich'* will.

Solche Formulierungen suggerieren eine alternative Antwort auf die Frage, wodurch (nicht-kausal verstandene) Präferenzstärken festgelegt sein könnten. Sie verweisen auf eine Unterscheidung, die aus der zeitgenössischen Willensfreiheitsdebatte geläufig ist: die Unterscheidung zwischen Präferenzen (Wünschen), die handlungswirksam sind, und Präferenzen (Wünschen), die den 'eigentlichen' Willen des Akteurs konstituieren. Diese begriffliche Differenzierung bildet die theoretische Grundlage für den sogenannten «Real Self View», dessen bekanntester Vertreter Harry G. Frankfurt ist.[11]

Ein Standardbeispiel, an dem Frankfurt seine Theorie illustriert, ist der Fall des Drogenabhängigen, der sich Heroin spritzt, obwohl er 'eigentlich' von seiner Sucht loskommen möchte.[12] Nach Frankfurts Analyse hat der

11 Vgl. Harry G. Frankfurt: *Freedom of the Will and the Concept of a Person*, in *The Importance of What We Care About*, hg. von H. G. F. (Cambridge: Cambridge University Press, 1988) S. 11-25. Die Bezeichnung «Real Self View» stammt von Susan Wolf, vgl. Susan Wolf: *Freedom Within Reason* (Oxford: Oxford University Press, 1990) S. 23ff.

12 Frankfurt bezeichnet diese Figur als «unwilling addict», vgl. op. cit. S. 17-18.

Drogenabhängige zwei Präferenzen[13] erster Stufe: die Präferenz dafür, Heroin zu nehmen (P_1), und die Präferenz dafür, kein Heroin zu nehmen (P_2). Die handlungswirksame Präferenz ist offensichtlich P_1. Doch der Drogenabhängige besitzt auch Präferenzen zweiter Stufe (d.h. Präferenzen, die Präferenzen erster Stufe zum Gegenstand haben): Er möchte, dass statt P_1 die Präferenz P_2 handlungswirksam wird. P_2 ist die Präferenz, mit der sich der Drogenabhängige *identifiziert*, die in einem starken Sinne *'seine eigene'* Präferenz ist. Man kann P_2 daher auch als die 'eigentliche' oder 'wahre' Präferenz des Drogenabhängigen bezeichnen.

Auf dieser Grundkonzeption baut Frankfurt eine Theorie der Willensfreiheit und der moralischen Verantwortlichkeit auf, die hier aber nicht weiter diskutiert werden soll. Hier kommt es nur auf die Konzeption selbst an, da sie eine alternative Interpretation für Präferenzstärken nahelegt: Eine Präferenz P ist danach genau dann stärker als eine Präferenz P*, wenn der Akteur eine Präferenz (zweiter Stufe) dafür hat, dass P sich in Entscheidungssituationen gegenüber P* durchsetzt. Instrumentell irrational wären folglich diejenigen Entscheidungen und Handlungen, bei denen die Präferenzen, die durch Präferenzen zweiter Stufe gestützt sind, durch andere ('ungestützte') Präferenzen übertrumpft werden.[14]

Diese Interpretation lässt sich *prima facie* gut auf die Beispiele anwenden, die in Abschnitt 1 diskutiert wurden. Tex entscheidet sich angesichts der Amputationssäge gegen die Operation, d.h. seine Präferenz für Schmerzvermeidung setzt sich gegenüber seiner Präferenz für die Operation durch. Die Zusatzannahme, dass Tex dabei – auf der zweiten Stufe – wünscht, dass seine Präferenz für die Operation die Oberhand behalten soll, scheint psychologisch plausibel. Wenn Tex ein Akteur mit normalem Präferenzprofil ist, dann ist ihm sein Überleben wichtiger als die Vermeidung von (relativ kurz andauernden) Schmerzen: Er möchte im entscheidenden Moment 'stark bleiben' und die Operation ruhig über sich ergehen lassen; er ist wütend

[13] Frankfurt spricht von «Wünschen» («desires») erster und zweiter Stufe. Da ich den Ausdruck «Präferenz» in genau dem Sinne verwende, in dem Frankfurt das Wort «Wunsch» gebraucht, gibt meine Formulierung Frankfurts Position adäquat wieder.

[14] Die Idee, dass Präferenzen höherer Stufe zur Analyse von Willensschwäche verwendet werden können, stammt von Jeffrey (vgl. Richard C. Jeffrey: *Preference among Preferences*, in *The Journal of Philosophy* 13 [1974] S. 377-391) und wurde in jüngerer Zeit von Bigelow, Dodds und Pargetter verteidigt (vgl. John Bigelow, Susan Dodds, Robert Pargetter: *Temptation and the Will*, in *American Philosophical Quarterly* 27 [1990] S. 39-49).

über sich selbst, als er dann doch den Sanitätsoffizier zurückweist; und er macht sich später Vorwürfe, dass er angesichts der Amputationssäge schwach geworden ist.

Ähnliches gilt für den Fall der Wissenschaftlerin. Sie hat eine Präferenz dafür, das Fußballspiel zu Ende zu schauen, und sie hat eine Präferenz dafür, den Fernseher auszuschalten und sich an den Schreibtisch zu setzen. Sie möchte, dass sich die zweite Präferenz durchsetzt, aber sie «kann sich nicht aufraffen»; die Präferenz, die letztlich handlungswirksam wird, ist die erste Präferenz. Auch hier liefert die an Frankfurt orientierte Interpretation von Präferenzstärken eine intuitiv plausible Situationsbeschreibung.

Um es noch einmal zu betonen: Diese Interpretationen sind nicht *zwingend*. In beiden Fällen wäre es möglich, dass der jeweilige Akteur sich über sich selbst täuscht oder dass sich seine Präferenzen von einem Moment auf den anderen grundlegend ändern. Mein Punkt ist lediglich dieser: Es ist unplausibel, anzunehmen, dass *nie* Fälle der beschriebenen Art vorliegen. Im Gegenteil, instrumentell irrationales Handeln – Handeln gegen seine 'eigentlichen' Präferenzen – scheint ein weit verbreitetes Phänomen zu sein.

5. Probleme mit höherstufigen Präferenzen

Gegen Frankfurts Theorie sind im Laufe der Jahre eine Reihe von Einwänden vorgebracht worden. Nicht alle dieser Einwände sind für unsere Diskussion relevant, aber es gibt zwei 'klassische' Probleme des Ansatzes, an denen wir nicht vorbeikommen: das *Regressproblem* und das *Autoritätsproblem*.

Das Regressproblem wird von Frankfurt selbst angesprochen. Der Ausgangspunkt für dieses Problem ist die Beobachtung, dass es neben Präferenzen erster und zweiter Stufe auch Präferenzen dritter Stufe geben kann – d.h. Präferenzen, die sich auf Präferenzen zweiter Stufe beziehen. Selbst die Existenz von Präferenzen vierter, fünfter und sechster Stufe ist nicht *a priori* auszuschließen: «There is no theoretical limit to the length of the series of desires of higher and higher orders», wie Frankfurt schreibt.[15] Damit stellt sich die Frage, welche dieser höherstufigen Präferenzen nun eigentlich festlegen sollen, was die 'wahren' Präferenzen des Akteurs sind. Darauf zu beharren, dass es allein auf die Präferenzen zweiter Stufe ankommt, wäre willkürlich.

[15] Frankfurt, op. cit. (Fn. 11) S. 21.

Dieses Problem scheint mir nicht unüberwindbar zu sein. Warum sollten wir nicht sagen, dass es auf die Präferenzen der *höchsten* Stufe ankommt, die der Akteur *de facto* besitzt? Im Normalfall werden dies Präferenzen der zweiten Stufe sein, in seltenen Fällen Präferenzen höherer Stufen. Vielleicht müssen noch Zusatzbedingungen formuliert werden, wie z.B. die Anforderung, dass die Präferenzen der höchsten vorhandenen Stufe nicht miteinander in Konflikt stehen dürfen, oder dass der Akteur sich entschieden hat, die höchststufigen Präferenzen keiner weiteren Prüfung zu unterziehen.[16] Doch wenn diese Voraussetzungen erfüllt sind – d.h., wenn eindeutige, unhinterfragte Präferenzen der höchsten Stufe *n* vorliegen – scheint es *prima facie* recht plausibel, die Präferenzen der Stufe *n* als 'Regress-Stopper' anzusehen.

Es gibt aber noch ein zweites Problem mit höherstufigen Präferenzen, das weitaus gravierender ist. Wenn wir sagen, dass die höherstufigen Präferenzen festlegen, was die 'eigentlichen Präferenzen' eines Akteurs sind oder was dieser Akteur 'wirklich will', dann schreiben wir ihnen eine bestimmte Art von *Autorität* zu. Die Präferenzen höherer Stufe – so die These – 'sprechen für den Akteur'; sie bestimmen, welche seiner Präferenzen 'intern' und welche 'extern' sind. Doch es ist fraglich, ob Präferenzen höherer Stufe dies wirklich leisten können. Gary Watson schreibt:

> The problem with [Frankfurt's] response is not that there is a regressive ascent up the hierarchy, or that people are not that complex, but simply that higher-order volitions are just, after all, desires, and nothing about their level gives them any special authority with respect to externality. If they have that authority they are *given* it by something else.[17]

Watson weist darauf hin, dass auch Präferenzen (Wünsche) höherer Stufe zunächst nur gewöhnliche Präferenzen sind. Was sie von Präferenzen erster Stufe unterscheidet ist lediglich ihr *Gehalt*, d.h. die Tatsache, dass sie sich auf andere Wünsche beziehen, aber es ist nicht klar, warum dieser Umstand ihnen einen besonderen normativen Status verleihen sollte. Die Rede von

16 Vgl. Harry G. Frankfurt: *Identification and Wholeheartedness*, in *The Importance of What We Care About*, hg. von H. G. F. (Cambridge: Cambridge University Press, 1988) S. 159-176.

17 Gary Watson: *Free Action and Free Will*, in *Mind* 46 (1987) S. 149. Dieses Problem wird in der Literatur oft als das «identification problem» bezeichnet, vgl. etwa Laura Ekstrom, *Free Will. A Philosophical Study* (Boulder: Westview Press, 2000) S. 76-77.

Präferenzen «höherer Stufe» (oder von «Präferenzhierarchien») *suggeriert* zwar Autorität, aber eine sachliche Basis dafür ist auf den ersten Blick schwer zu finden.

Die Stärke dieses Einwands wird noch deutlicher, wenn man die Analyse instrumenteller Irrationalität betrachtet. In Abschnitt 3 wurden diejenigen Entscheidungen und Handlungen als instrumentell irrational klassifiziert, bei denen die Präferenzen, die durch Präferenzen höherer Stufe gestützt sind, durch andere ('ungestützte') Präferenzen übertrumpft werden. Tex' Entscheidung gegen die Operation gilt danach als instrumentell irrational, weil er sich (auf der zweiten Stufe) wünscht, dass seine Präferenz *für* die Operation (und für die Chance auf ein langes Leben) sein Handeln bestimmt. Wenn wir eine Entscheidung als «instrumentell irrational» charakterisieren, dann stellt dies aber keine vollkommen neutrale Beschreibung dar; wir sagen damit zugleich, dass der Entscheidungsprozess «nicht so abgelaufen ist, wie er ablaufen sollte». Wir konstatieren mit unserem Urteil ein *Defizit:* Wir stellen fest, dass die Entscheidung des Akteurs bestimmte normative Anforderungen nicht erfüllt.

Doch was ist die Basis für diese negative Beurteilung? Wenn man die Entscheidung von Tex rein deskriptiv betrachtet, könnte man sagen: Tex entscheidet sich für eine Handlungsweise, die (voraussichtlich) einige seiner Präferenzen befriedigt (in erster Linie die Präferenz für Schmerzvermeidung und die von ihr abgeleitete Präferenz für das Nicht-Stattfinden der Operation) und zugleich einige andere seiner Präferenzen frustriert (die Präferenz für ein langes Leben, die von ihr abgeleitete Präferenz für das Stattfinden der Operation und die Präferenz zweiter Stufe, die diese Präferenzen stützt). Wenn Tex sich stattdessen für die Operation entschieden hätte, sähe die Situation grundsätzlich nicht anders aus, nur dass die Präferenzen frustriert würden, die *de facto* (d.h. in unserem Beispiel) befriedigt werden, und umgekehrt. Warum beurteilen wir Tex' Entscheidung im ersten Fall negativ («instrumentell irrational»), im zweiten Fall positiv («instrumentell rational»)?

Die bloße *Anzahl* der befriedigten bzw. frustrierten Präferenzen kann nicht entscheidend sein. Abgesehen davon, dass zweifelhaft ist, ob sich Präferenzen eindeutig zählen lassen,[18] kann das Beispiel von Tex so reformuliert

[18] Angenommen ich habe eine starke Präferenz für Schmuck aus Jade, weiß aber, dass zwei verschiedene Arten von Gestein – Jadeit und Nephrit – als «Jade» bezeichnet werden. Habe ich nun *eine* Präferenz (für Schmuck aus Jade) oder *zwei* Präferenzen (für Schmuck aus Jadeit und für Schmuck aus Nephrit)? Mir scheint, dass es auf diese Frage keine sinnvolle Antwort gibt.

werden, dass die Mehrzahl der Präferenzen durch die *irrationale* Handlungsweise befriedigt wird. Dazu müssen wir nur annehmen, das Tex – zusätzlich zu seiner Präferenz für Schmerzvermeidung – zwei weitere (schwache) Präferenzen hat, die gegen eine Operation sprechen: eine Abneigung gegen den Anblick von Blut und eine Abneigung gegen die Geräusche, die eine Amputationssäge gewöhnlich verursacht. Seine Präferenzen zweiter Stufe stützen aber nach wie vor die Präferenzen, die für eine Operation sprechen. Unsere Analyse aus Abschnitt 3 liefert auch für dieses Szenario das Ergebnis, dass Tex' Entscheidung gegen die Operation instrumentell irrational ist. Ich denke, dass dies durchaus unseren Intuitionen entspricht. Die Frage ist nur: Was ist die *Rechfertigung* für unser Urteil? Was spricht dafür, sich bei der normativen Beurteilung von Tex' Entscheidung an den Präferenzen höherer Stufe zu orientieren? Höherstufige Präferenzen unterscheiden sich, wie gesagt, nicht grundlegend von anderen Präferenzen. Die Behauptung, es komme auf Tex' höherstufige Präferenzen an, «weil dies die Präferenzen sind, die sich auf andere Präferenzen beziehen», scheint willkürlich. Genauso gut könnte man behaupten, es komme auf Tex' Präferenz für Schmerzvermeidung an, «weil das die Präferenz ist, die sich auf Schmerzen bezieht», was offensichtlich absurd wäre. Die Frage nach der Autorität höherstufiger Präferenzen stellt also auch unsere Analyse instrumenteller Irrationalität vor große Probleme.

6. Die Autorität höherstufiger Präferenzen: ein Lösungsvorschlag

Frankfurts Drogensüchtiger will, dass sich seine Präferenz für ein drogenfreies Leben in den relevanten Entscheidungssituationen durchsetzt. Tex hat den Wunsch, dass seine Präferenz für eine hohe Überlebenschance handlungswirksam wird. Die Wissenschaftlerin möchte, dass ihr Handeln durch ihre Präferenz für eine erfolgreiche akademische Karriere bestimmt wird.

Wenn man diese und andere Beispiele betrachtet, die gewöhnlich in der Literatur diskutiert werden, dann wird schnell ein Muster deutlich: In fast allen Fällen stützen die Präferenzen zweiter Stufe diejenigen Präferenzen, die am *langfristigen Wohlergehen* des Akteurs orientiert sind. Für das Wohlergehen des Drogensüchtigen wäre der Verzicht auf Drogen offensichtlich förderlicher als der Drogenkonsum, für Tex' Wohlergehen wären die Operation und die damit verbundenen hohen Überlebenschancen optimal, und zum Wohlergehen der Wissenschaftlerin würde eine erfolgreiche Karriere voraussichtlich mehr beitragen als ein ausgedehnter Fernsehabend. Analo-

ges gilt für die Beispiele des Übergewichtigen und des Schüchternen, die in Abschnitt 1 vorgestellt worden sind.

Meine Vermutung ist nun, dass diese Beobachtung den Schlüssel zur Lösung des Autoritätsproblems liefert: Wir schreiben den Präferenzen höherer Stufe nur deshalb eine besondere Autorität zu, *weil sie mit hoher Zuverlässigkeit am langfristigen Wohlergehen des Akteurs orientiert sind*. Der normative Status, den höherstufige Präferenzen besitzen, ist also *abgeleitet*; er beruht darauf, dass diese Präferenzen im Normalfall 'vernünftige' Präferenzen erster Stufe stützen.

Der Ausdruck «Wohlergehen» ist hier in seinem gewöhnlichen, alltagssprachlichen Sinn zu verstehen. Das Wohlergehen eines Akteurs schließt alles ein, was – objektiv betrachtet – *gut für ihn* ist: Gesundheit, langes Leben, hedonisches Glück, bedeutsame menschliche Beziehungen (z.B. Liebe, Freundschaft), die Entwicklung von Talenten und Fähigkeiten, und so weiter. Was genau in diese Liste aufgenommen werden soll, ist eine interessante normative Frage, die im Rahmen dieses Aufsatzes aber nicht diskutiert werden kann.[19] Wichtig für unsere Diskussion ist in erster Linie nur, dass die Autorität höherstufiger Präferenzen anscheinend mit Rückgriff auf Tatsachen über (objektiv verstandenes) Wohlergehen erklärt werden muss. Dies ist eine überraschende Konsequenz, die für die Diskussion um instrumentelle Rationalität große Bedeutung hat. Denn wenn eine adäquate Theorie instrumenteller Irrationalität (die auch die Möglichkeit instrumentell irrationalen Handelns zulässt) auf Tatsachen über Wohlergehen – also *substantielle normative Tatsachen* – zurückgreifen muss, dann bedeutet das, dass bestimmte rationalitätstheoretische Positionen nicht haltbar sind. Dies ist ein Punkt, den ich in Abschnitt 6 noch einmal aufgreifen werde.

Zunächst möchte ich jedoch drei kritische Fragen diskutieren, die meinen Lösungsvorschlag betreffen. Die erste Frage lautet: Wie kann dieser Vorschlag mit Fällen von *inverser Akrasie* umgehen, d.h. mit Fällen, in denen die willensschwache Handlung des Akteurs 'besser' ist als die alternative Handlung, die der Akteur 'eigentlich' bevorzugt?[20] Betrachten wir ein Beispiel. Alice möchte unbedingt abnehmen; sie will (auf der zweiten Stufe), dass

19 Zur objektiven Konzeption von Wohlergehen («objective list theory»), vgl. Roger Crisp: *Well-Being*, in *The Stanford Encyclopedia of Philosophy (Winter 2006 Edition)*, hg. von Edward N. Zalta (2006), URL = <http://plato.stanford.edu/archives/win2006/entries/well-being/>, insbesondere Abschnitt 4.3.

20 Zu inverser Akrasie vgl. (u.a.) Nomy Arpaly, Timothy Schroeder: *Praise, Blame and the Whole Self*, in *Philosophical Studies* 93 (1999) S. 161-188.

sich ihre Präferenz für das Durchhalten der Diät gegenüber ihrer Präferenz für nährstoffreiches Essen durchsetzt. Als sie ein saftiges Steak vorgesetzt bekommt, greift sie aber dennoch – unter heftigen Selbstvorwürfen – zu. Ihre Handlung ist eine klare Instanz von instrumenteller Irrationalität. Doch Alice ist nicht übergewichtig; im Gegenteil, wenn Alice weiter abnehmen würde, wäre ihre Gesundheit in Gefahr. (Alice weiß dies, aber ihre Gesundheit ist ihr nicht wichtig, deshalb spielt dieser Aspekt bei ihrer Entscheidung nur eine untergeordnete Rolle.)

In diesem Fall stützen die höherstufigen Präferenzen des Akteurs eine Präferenz, die nicht an dessen Wohlergehen orientiert ist. Wenn die Autorität höherstufiger Präferenzen aber von ihrer Verbindung zu 'wohlergehensförderlichen' Präferenzen erster Stufe abhängt, stellt sich die Frage, ob die höherstufigen Präferenzen in Fällen wie diesen überhaupt Autorität besitzen. Wenn sie aber keine Autorität besitzen, ist zweifelhaft, ob man Alices Handeln noch als «instrumentell irrational» bezeichnen kann.

Ich denke, dass der hier vorgestellte Lösungsvorschlag für das Autoritätsproblem mit Szenarien dieser Art zurechtkommt. Zunächst einmal ist klar, dass wir Alices Handeln *insgesamt* gutheißen: Wir bewerten es als positiv, dass ihre Willensschwäche sie vor der Magersucht bewahrt. Dennoch wäre es falsch zu sagen, dass Alices Präferenzen höherer Stufe keine Autorität besitzen: Die Autorität höherstufiger Präferenzen beruht (laut meinem Vorschlag) nicht darauf, dass sie *in jedem Einzelfall* 'wohlergehensförderliche' Präferenzen stützen, sondern darauf, dass sie dies im Normalfall – also mit hoher Wahrscheinlichkeit – tun. Diese 'derivative' Autorität besitzen auch Alices Präferenzen, und das ist der Grund, warum wir ihr Verhalten trotz allem als «instrumentell irrational» charakterisieren. Unsere normative Kritik bezieht sich dabei ausschließlich auf den Entscheidungsprozess (bei dem «nicht alles so abgelaufen ist, wie es ablaufen sollte») und fällt gegenüber der positiven Bewertung der Entscheidung selbst kaum ins Gewicht; dennoch sollte man m.E. nicht bestreiten, dass es *eine* (*untergeordnete*) *Hinsicht* gibt, in der wir Alices Verhalten kritisieren.

Diese Überlegungen führen direkt zur zweiten Frage: Ist die 'derivative Autorität', die ich höherstufigen Präferenzen zuschreibe, wirklich eine Art von *Autorität?* Meine Antwort lautet: Ja – in einem schwachen Sinn. Der normative Status, den ich höherstufigen Präferenzen zuspreche, ist vergleichbar mit dem Status, der 'rechtfertigenden Faktoren' im Rahmen des erkenntnistheoretischen Reliabilismus zukommt. Die zentrale These des (Prozess-) Reliabilismus lautet (vereinfacht): Eine Meinung ist genau dann gerechtfertigt, wenn sie durch einen Prozess erzeugt wurde, der zuverlässig – d.h.

mit großer objektiver Wahrscheinlichkeit – wahre Meinungen produziert.[21] Die Eigenschaft des Gerechtfertigt-Seins kann also mit der (historischen) Eigenschaft identifiziert werden, *durch einen zuverlässigen Prozess erzeugt worden zu sein*. Der normative Status von epistemischer Rechtfertigung ist auch hier derivativ: Er beruht auf der Tatsache, dass Meinungen mit den Eigenschaften des Gerechtfertigt-Seins im Normalfall wahr (also 'epistemisch wertvoll') sind. Meine Erklärung für die Autorität höherstufiger Präferenzen folgt demselben Muster. Man könnte den Ansatz daher auch als 'Autoritätsreliabilismus bezüglich höherstufiger Präferenzen' bezeichnen.

Es bleibt die dritte und letzte Frage. Ich bin in diesem Abschnitt davon ausgegangen, dass unsere höherstufigen Präferenzen mit hoher Zuverlässigkeit an unserem langfristigen Wohlergehen orientiert sind. Das ist eine Voraussetzung, die von der Alltagspsychologie (d.h. von unserer *Common-Sense*-Theorie des menschlichen Geistes) gestützt wird. Was aber, so die Frage, wenn diese 'Zuverlässigkeitshypothese' sich dennoch als falsch erweisen sollte? Die Alltagspsychologie ist nicht über jeden Zweifel erhaben. Auch wenn es *prima facie* unwahrscheinlich scheint, könnten empirische Untersuchungen zeigen, dass in Wirklichkeit keine zuverlässige Verbindung zwischen den höherstufigen Präferenzen und dem langfristigen Wohlergehen eines Akteurs besteht. In diesem Fall, so scheint es, hätten höherstufige Präferenzen nicht einmal derivative Autorität. Ich denke, dass dies tatsächlich aus meinem Vorschlag folgt, glaube aber nicht, dass dies problematisch ist. Falls sich unsere *Common-Sense*-Annahme als falsch erweisen sollte, wäre auch die Kritik, die wir mit Urteilen über die instrumentelle Irrationalität von Handlungen zum Ausdruck bringen, hinfällig; es wäre unter diesen Umständen schlicht ungerechtfertigt, zu behaupten, dass 'willensschwache' Entscheidungen Prozesse sind, «die nicht so ablaufen, wie sie ablaufen sollten». Die Falsifizierbarkeit der Zuverlässigkeitshypothese stellt somit kein Problem für meinen Ansatz dar.

7. Reflektierte Präferenzen statt Präferenzen höherer Ordnung?

Im letzten Abschnitt habe ich dafür argumentiert, dass das Autoritätsproblem nur mit Hilfe substantieller normativer Annahmen lösbar ist. Dies wirft die Frage auf, ob es nicht vielleicht eine alternative Analyse von (nicht-kausalen)

21 Vgl. Thomas Grundmann: *Analytische Einführung in die Erkenntnistheorie* (Berlin: De Gruyter, 2008) S. 265.

Präferenzstärken gibt, die nicht auf Präferenzen höherer Stufe zurückgreift. Mit Hilfe dieser Analyse – so die Idee – könnten die substantiellen normativen Annahmen, die vielen Theoretikern ein Dorn im Auge sind (vgl. Abschnitt 6), eventuell vermieden werden.

Welche alternative Analyse bietet sich hier an? Der aussichtsreichste Kandidat scheint mir eine *modifizierte dispositionale Theorie* zu sein, nach der die Präferenz P eines Akteurs genau dann stärker als die Präferenz P* ist, wenn sie sich *unter Idealbedingungen* gegen P* durchsetzen würde. Die 'Idealbedingungen' können unterschiedlich spezifiziert werden; in der Regel enthalten sie die Anforderung, dass der Akteur sich alle relevanten Informationen, über die er verfügt, deutlich vor Augen führt und bei seinen Überlegungen einen 'kühlen Kopf' bewahrt, manchmal kommt noch die Bedingung hinzu, dass der Akteur sich einer 'kognitiven Psychotherapie' unterzieht, die pathologische Wünsche und Abneigungen eliminiert.[22] Die Details der Analyse sind hier aber nicht ausschlaggebend.

Dass ein Akteur instrumentell irrational handelt, bedeutet nach der modifizierten dispositionalen Theorie, dass sich bei ihm in einer Entscheidungssituation die Präferenz P* gegenüber der Präferenz P durchsetzt, obwohl sich unter Idealbedingungen P gegen P* durchsetzen würde. Wenn wir die Fälle von Tex, der Wissenschaftlerin, dem Übergewichtigen und dem Schüchternen betrachten, scheint eine solche Analyse nicht von vornherein unplausibel (vorausgesetzt, dass die Idealbedingungen adäquat ausbuchstabiert werden).

Dennoch ist die Theorie (mindestens) zwei schwerwiegenden Einwänden ausgesetzt. Zum einen gibt es plausible Gegenbeispiele gegen die vorgeschlagene Analyse.[23] Zum anderen – und das ist hier entscheidend – steht die dispositionale Theorie vor dem gleichen Problem wie die Theorie der höherstufigen Präferenzen: Sie muss erklären, warum hypothetischen Entscheidungen unter Idealbedingungen eine besondere Autorität zukommt.

22 Vgl. dazu Richard B. Brandt: *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979) S. 10-12, 110-129.

23 Betrachten wir z.B. den Fall von Egon, dem sensiblen Egoisten. Egon ist sich bewusst, dass er viel Geld für humanitäre Zwecke spenden würde, wenn er sich alles, was er über die Hungernden in Afrika weiß, deutlich vor Augen führen würde. Doch Egon möchte sein Geld lieber behalten und unterlässt es daher, sich die relevanten Tatsachen zu vergegenwärtigen. Ist Egon instrumentell irrational? Ich denke nicht. Egons Verhalten ist zwar *unmoralisch*, aber durchaus nicht unvernünftig. (Dies ist die modifizierte Version eines Beispiels von Gibbard; vgl. Allan Gibbard: *Wise Choices, Apt Feelings* [Oxford: Clarendon Press, 1990] S. 21).

Der Ausdruck «Idealbedingungen» hat zwar normative Konnotationen, streng genommen ist er aber nicht mehr als eine Abkürzung für «die Bedingungen B, C, D, E, …», wobei B, C, D, E, … für *rein deskriptive Sachverhalte* stehen. Die Frage ist nun: Warum sollen wir annehmen, dass die Entscheidungen, die der Akteur unter den Bedingungen B, C, D, E, … treffen würde, seine 'wahren' oder 'eigentlichen' Präferenzen reflektieren? Warum kommt diesen hypothetischen Entscheidungen eine besondere Autorität zu?

Wenn dieses Problem lösbar ist, dann – so meine Vermutung – auf dieselbe Weise wie das analoge Problem für die Theorie höherstufiger Präferenzen: Die Autorität von hypothetischen Entscheidungen unter Idealbedingungen könnte darauf beruhen, *dass solche Entscheidungen mit hoher Wahrscheinlichkeit am langfristigen Wohlergehen des Akteurs orientiert sind.* Doch damit wäre der Versuch, normative Annahmen mit Hilfe der dispositionalen Theorie zu vermeiden, eindeutig gescheitert.

8. Zusammenfassung: Abschied vom Instrumentalismus

Der Ausgangspunkt dieses Aufsatzes war die Beobachtung, dass Menschen manchmal instrumentell irrational handeln: Sie treffen hin und wieder Entscheidungen, die ihren 'stärksten' Präferenzen – d.h. dem, was sie 'eigentlich' wollen – zuwiderlaufen. Standardbeispiele aus der Debatte um Willensschwäche zeigen, dass die Existenz solcher Fälle kaum bestritten werden kann.

Das stellt die Theoretiker, die sich mit instrumenteller Rationalität befassen, vor ein Problem: Sie müssen spezifizieren, welche Tatsachen die Stärke von Präferenzen festlegen. Die «Theorie der offenbarten Präferenzen», die oft zur Analyse von Präferenzstärken verwendet wird, scheidet aus, da sie keinen Raum für instrumentelle Irrationalität lässt. Doch welche Alternativen gibt es?

Die beste Alternative scheint eine Analyse zu sein, die Präferenzstärken mit Hilfe höherstufiger Präferenzen definiert. Doch diese Theorie lässt (in ihrer ursprünglichen Version) eine wichtige Frage offen: Sie erklärt nicht, woher die Autorität höherstufiger Präferenzen stammt. Ich habe dafür argumentiert, dass es eine plausible Erklärung für diese Autorität gibt: Sie beruht auf der Tatsache, dass höherstufige Präferenzen mit hoher Zuverlässigkeit am langfristigen Wohlergehen des Akteurs orientiert sind. Diese Erklärung setzt allerdings voraus, dass es in Bezug auf das Wohlergehen des Akteurs objektive Tatsachen gibt: Bestimmte Handlungen sind gut für den Akteur,

andere schlecht, und diese Tatsachen sind nicht (vollständig) durch die Präferenzen des Akteurs determiniert. Mit anderen Worten: Die vorgeschlagene Erklärung nimmt auf normative Tatsachen Bezug – Tatsachen über die *substantielle Rationalität* von Handlungen.[24]

Wenn meine Argumentation korrekt ist, hat dies gravierende Konsequenzen für die Debatte um instrumentelle Rationalität. Aus der hier vorgeschlagenen Analyse folgt nämlich, dass der Instrumentalismus – wie er z.B. von Gauthier, Hubin, Schroeder und anderen 'Humeanern' vertreten wird[25] – keine kohärente Position ist. Die These, dass instrumentelle Rationalität die einzige Form von Rationalität ist, scheint mit einer plausiblen Analyse von instrumenteller Rationalität (und Irrationalität) nicht vereinbar zu sein.[26]

[24] Mit dieser Formulierung möchte ich mich nicht auf einen Realismus bezüglich normativer Tatsachen festlegen. Tatsächlich glaube ich, dass eine expressivistische (oder 'quasi-realistische') Analyse von Normativität ingesamt plausibler ist als ein klassischer Realismus, vgl. Peter Schulte: *Zwecke und Mittel. Eine expressivistische Analyse instrumenteller Rationalität* (Paderborn: mentis Verlag, noch nicht erschienen).

[25] Vgl. David Gauthier: *Morals by Agreement* (Oxford: Oxford University Press, 1986), S. 32-38; Hubin, op. cit. (Fn. 7); Mark Schroeder: *Slaves of the Passions* (Oxford: Oxford University Press, 2007).

[26] Korsgaard, op. cit. (Fn. 1), kommt ebenfalls zu dem Schluss, dass instrumentelle Rationalität substantielle Rationalität voraussetzt. Ich halte ihre Argumentation jedoch nicht für überzeugend, da sie implizit von der problematischen «Narrow Scope»-Interpretation des instrumentellen Prinzips ausgeht; vgl. dazu die Kritik von John Broome: *Normative Requirements*, in *Normativity*, hg. von Jonathan Dancy (Oxford: Blackwell, 2000) S. 97-98.

Studia philosophica 68/2009

Hartmut Westermann

Der göttliche *intellectus*: ein irrationales Konzept von Überrationalität? Zu Lorenzo Vallas Boethius-Kritik

The subject of this paper is the relation of human ratio and divine intellectus as understood by Boethius and Valla. In Consolatio Philosophiae Boethius constructs the notion of divine intellectus and uses it to identify the inferior human ratio. Lorenzo Valla in De libero arbitrio disqualifies Boethius' approach as irrational. At first glance, Valla's critique of Boethius seems to make the very mistake Valla is up to criticise in Boethius. Just like Boethius himself, Valla runs the risk of hubris. Just like Boethius, when reasoning about the divine intellectus Valla seems to extend human reasoning far beyond the limited ratio. I contend that when presupposing a sharp distinction between ratio and intellectus – as both Boethius and Valla do – statements about the divine intellectus can only be rationally constructed as statements on the level of meta-language. Along these lines, I suggest an interpretation of Valla's critique that is not subject to the very same objection he raises against Boethius.

Unter Rückgriff auf die Bestimmung des Menschen als *zōon logon echōn* in der *Politik* des Aristoteles[1] subsumiert die klassische Anthropologie den Menschen unter die Gattung der Lebewesen und differenziert ihn *qua* der ihm proprietär zukommenden Rationalität in spezifischer Weise von Tieren und Pflanzen. Entsprechend ist es das dem Menschen eigene rationale Kognitionsvermögen, welches diesen allererst zum Menschen macht. Wird der Mensch als das *animal rationale* gedeutet, so bieten sich traditionell zwei signifikante Bezugsgrößen an, von denen die menschliche Rationalität zwecks eigener Profilierung abgehoben werden kann: der Instinkt des Tieres einerseits und der Intellekt Gottes andererseits. Aufgrund der jeweiligen Abgrenzungsfunktion gegenüber der Rationalität des Menschen sind zwar sowohl der Instinkt des Tieres wie auch der Intellekt Gottes als Formen der Nichtrationalität zu verstehen, doch bleibt ein wichtiger Unterschied zu beachten: Wird der tierische *instinctus* der menschlichen *ratio* gewöhnlich

1 Vgl. Aristoteles: *Politik* I 2, 1253a9-10.

untergeordnet, so betrachtet man den göttlichen *intellectus* als ein Kognitionsvermögen, welches das menschliche bei weitem übersteigt – der Intellekt Gottes als Nichtrationalität im Sinne von Überrationalität.

Im ersten Abschnitt dieses Textes werde ich referieren, wie der göttliche *intellectus* in der *Consolatio Philosophiae* des Boethius inhaltlich gefasst und als Negativfolie zur Charakterisierung der menschlichen *ratio* genutzt wird. Im zweiten Abschnitt lege ich dar, wie Lorenzo Valla in seinem Dialog *De libero arbitrio* das von Boethius vertretene Konzept von Überrationalität als ein hybrides und damit selbst irrationales Unterfangen kritisiert. Im dritten Abschnitt wird zunächst Vallas Kritik am Konzept göttlicher Überrationalität durch die Formulierung eines Selbstapplikationseinwands ihrerseits kritisch diskutiert, ehe ich im vierten Abschnitt mit der These schließe, dass die scharfe Differenzierung zwischen menschlicher Rationalität und göttlicher Überrationalität eine Art von Metatheologie provoziert, deren Ziel darin liegt, einer solchen «Theologie der Differenz» die eigene Irrationalität aufzuzeigen.

1. Die Konzeption göttlicher Überrationalität in Boethius' Consolatio Philosophiae[2]

Den Begriff des göttlichen *intellectus* konzipiert Boethius im fünften Buch seiner *Consolatio Philosophiae*. Den Problemhintergrund bildet dabei die Frage nach dem Verhältnis von menschlicher Freiheit (*libertas arbitrii*) und göttlicher Allwissenheit (*sapientia Dei*): Kann man die anthropologische Annahme, der Mensch sei in seinen Handlungen und Entscheidungen frei, und die theologische Annahme, Gott verfüge über ein allumfassendes Wissen, zugleich für wahr halten?

[2] Die ersten beiden Abschnitte dieses Textes gehen auf einen längeren Aufsatz zu Vallas *De libero arbitrio* zurück, in dem ich die These vertrete, dass Valla anhand der Überprüfung, wie sich die *libertas arbitrii* zur *sapientia Dei*, zur *potentia Dei* und zur *bonitas Dei* verhält, auf eine drohende Inkonsistenz des tradierten christlichen Gottesbegriffs aufmerksam macht (vgl. Hartmut Westermann: *Lorenzo Valla:* De libero arbitrio*. Die Freiheit des Menschen im Angesicht Gottes*, in *Des Menschen Würde: entdeckt und erfunden im Humanismus der italienischen Renaissance*, hg. von Rolf Gröschner, Stephan Kirste, Oliver W. Lembcke [Tübingen: Mohr Siebeck 2008], 113-139). Demgegenüber fokussiert der vorliegende Text auf die (von Boethius proponierte und von Valla problematisierte) Unterscheidung zwischen der menschlichen *ratio* und dem göttlichen *intellectus*, um die Frage nach der Irrationalität einer solchen Differenzierung zu diskutieren.

Dass zwischen der *libertas arbitrii* und der *sapientia Dei* eine erhebliche Spannung besteht, lässt sich durch die Formulierung von insgesamt sechs Thesen zeigen, von denen die erste die essentiellen Prädikate des tradierten christlichen Gottesbegriffs zum Ausdruck bringt, während die übrigen fünf jedenfalls den Anschein erwecken, unmittelbar von der jeweils vorhergehenden ableitbar zu sein. Die sechste und letzte These wird zur Konsequenz haben, dass dem Menschen – unter der anfangs getroffenen Annahme einer Allwissenheit Gottes – sowohl die Handlungs- als auch die Willensfreiheit abgesprochen werden muss.

These 1: Zum Begriff Gottes gehört essentiell, dass er nicht nur allmächtig und allgütig, sondern auch allwissend ist.

These 2: Wenn Gott allwissend ist, so weiß er nicht nur um alles Vergangene und Gegenwärtige, sondern auch um alles Zukünftige.

These 3: Wenn Gott um alles Zukünftige weiß, so weiß er auch um die künftigen Entscheidungen und Handlungen der Menschen.

These 4: Wenn Gott um die künftigen Entscheidungen und Handlungen der Menschen weiß, so können die Menschen sich nur für das entscheiden, von dem Gott zuvor schon weiß, dass sie sich dafür entscheiden werden, und sie können nur so handeln, wie Gott zuvor schon weiß, dass sie handeln werden.

These 5: Wenn die Menschen sich nur für das entscheiden können, von dem Gott zuvor schon weiß, dass sie sich dafür entscheiden werden, und wenn sie nur so handeln können, wie Gott zuvor schon weiß, dass sie handeln werden, dann sind sie in ihren Entscheidungen und in ihren Handlungen festgelegt.

These 6: Wenn die Menschen in ihren Entscheidungen und in ihren Handlungen festgelegt sind, so sind sie nicht frei darin, sich für das Eine oder für das Andere zu entscheiden und das Eine oder das Andere zu tun.

Folgt man dieser Problemskizze, so ergibt sich allein aus der Allwissenheit Gottes, dass die Annahme einer menschlichen Willens- und Handlungsfreiheit aufgegeben werden muss.[3] Um eine solche Konsequenz zu vermeiden,

[3] Vgl. Boethius: *Consolatio Philosophiae* V, 3.p., 5-14: «Es scheint mir im höchsten Grade ein Gegensatz und Widerspruch zu sein, dass Gott alles im voraus kenne und zugleich irgendein freier Wille sei; denn wenn Gott alles voraussieht und auf keine Weise irren kann, so muss mit Notwendigkeit eintreffen, was die Vorsehung als künftig vorausgesehen hat. Deshalb, wenn sie von Ewigkeit nicht

lässt Boethius die als Dialogfigur auftretende personifizierte *Philosophia* eine Problemlösung entwickeln, die zwar die – den tradierten christlichen Gottesbegriff artikulierende – Ausgangsthese akzeptiert, doch bereits These 2, wonach ein allwissender Gott um alles Vergangene, alles Gegenwärtige und ebenso um alles Zukünftige wisse, für derart präzisierungsbedürftig hält, dass diese ohne eine solche Klärung nicht vertretbar erscheint.

Als Basis für die notwendige Präzisierung dient eine scharfe Unterscheidung zwischen dem menschlichen und dem göttlichen Denken: Boethius differenziert terminologisch zwischen der spezifisch menschlichen *ratio* und dem *intellectus*, der allein Gott, nicht aber den Menschen zukommt.[4] Anders als die irrtumsanfällige menschliche *ratio*, die den Bedingungen der Temporalität unterworfen ist und ihren perspektivischen Ort in der Zeit hat, ist der infallible göttliche *intellectus* nicht zeitlich situiert. Boethius begreift Gott demnach als ein ewiges Wesen, das – im Unterschied zu den sterblichen Menschen, die in der Zeit entstehen und vergehen,– nicht aus einer zeitlich bedingten Perspektive auf die Ereignisse in der Zeit blickt, sondern die gesamten Zeitläufte von der Warte der Ewigkeit aus überschaut. Aufgrund der Einfachheit (*simplicitas*) der göttlichen Natur ist die Ewigkeit, die Boethius dem *intellectus* zuschreibt, nicht (wie die omnitemporal zu verstehende Ewigkeit der Welt) im Sinne der *perpetuitas* bzw. *sempiternitas* als eine unbegrenzte zeitliche Dauer zu begreifen, welche die temporale Sukzession ja nur ins Endlose prolongieren würde. Vielmehr muss die Ewigkeit des göttlichen *intellectus* im Sinne der *aeternitas* gedeutet werden: als radikale Verneinung temporaler Sukzession.[5] Während die zeitlich situierte mensch-

nur die Taten der Menschen, sondern auch deren Absichten und Willen vorausweiß, so gibt es keine Freiheit des Willens; denn es kann weder eine Handlung noch irgendein Wille existieren, den die göttliche Vorsehung nicht unfehlbar vorausweiß.» / «Nimium, inquam, adversari ac repugnare videtur praenoscere universa deum et esse ullum libertatis arbitrium. Nam si cuncta prospicit deus neque falli ullo modo potest, evenire necesse est, quod providentia futurum esse praeviderit. Quare si ab aeterno non facta hominum modo, sed etiam consilia voluntatesque praenoscit, nulla erit aribitrii libertas; neque enim vel factum aliud ullum vel quaelibet exsistere poterit voluntas, nisi quam nescia falli providentia divina praesenserit.»

4 Vgl. Boethius: *Consolatio Philosophiae* V, 5.p., 17-18: «Die Vernunft aber ist allein der menschlichen Art eigen, wie die Intelligenz nur der göttlichen.» / «Ratio vero humani tantum generis est sicut intelligentia sola divini.»

5 Vgl. Joachim Gruber: *Kommentar zu Boethius De consolatione philosophiae* (Berlin, New York: De Gruyter 1978) S. 378: «Der Begriff der Ewigkeit (sc. Gottes) wird von dem der Zeit aus entwickelt und von dem des *perpetuum* un-

liche *ratio* demnach notwendig auf Vergangenes, Gegenwärtiges und Zukünftiges bezogen ist, ruht der atemporale göttliche *intellectus* als punktuell Gegenwärtiges ganz in sich selbst und hat die unendliche Dauerhaftigkeit der sich sukzessive vollziehenden Weltzeit als reine Gegenwart vor sich. Damit ist für den *intellectus* nicht allein das präsent, was den Menschen zu einem bestimmten Zeitpunkt gegenwärtig ist, sondern auch das, was sich – aus der zeitlich situierten menschlichen *ratio* betrachtet – bereits ereignet hat, wie auch dasjenige, was sich erst noch ereignen und der *ratio* daher erst später zugänglich sein wird.

Angesichts dieser Differenzierung zwischen der temporalen *ratio* und dem als aeternital gedachten *intellectus* wird verständlich, weshalb Boethius These 2 der oben vorgelegten Problemskizze für präzisierungsbedürftig halten muss: Zwar weiß Gott in gewisser Weise um Vergangenes, Gegenwärtiges und Zukünftiges, doch weiß er um all dies in der Form eines zeitlosen Wissens. Er hat strenggenommen also kein Wissen von Vergangenem *als Vergangenem*, von Gegenwärtigem *als Gegenwärtigem* und Künftigem *als Künftigem*, sondern besitzt aufgrund des ihm eigentümlichen, rein präsentischen *intellectus* ein allumfassendes Wissen von dem, was der menschlichen *ratio* als Vergangenes, aktual Gegenwärtiges und Zukünftiges gegeben ist. Eine Einschränkung der Wissensinhalte, die eine unzulässige Begrenzung der *sapientia Dei* bedeuten würde, hat diese Proprietät des göttlichen Erkenntnismodus, welcher die Erkenntnisgegenstände in spezifischer Weise prägt,[6] daher nicht zur Folge: Die besondere Natur des Erkennenden, also des ewigen Gottes einerseits und der zeitlichen Menschen andererseits, prägt zwar die jeweilige Form, *wie* das zu Erkennende dem Erkennenden gegeben ist, limitiert aber das, *was* dem Erkennenden als Wissensinhalt zukommt, nur im Falle der menschlichen *ratio*, nicht im Falle des göttlichen *intellectus*.

Die weiteren Schritte der von Boethius angestrebten Problemlösung sind rasch nachgezeichnet: Es ist nicht die aeternital verstandene göttliche Allwis-

terschieden. Wie nun die Ewigkeit Gottes alle Zeit als Gegenwart umfasst, so erfasst auch sein Geist alles Geschehen als gegenwärtig.» Vgl. hierzu auch Ernst Gegenschatz: *Die Freiheit der Entscheidung in der «Consolatio philosophiae» des Boethius*, in *Boethius*, hg. von Manfred Fuhrmann, Joachim Gruber (Darmstadt: Wissenschaftliche Buchgesellschaft, 1984) S. 344-345.

6 Vgl. Boethius: *Consolatio Philosophiae* V, 4.p., 76-78: «Alles nämlich, was erkannt wird, wird nicht gemäß seiner eigenen Fähigkeit erkannt, sondern gemäß der Möglichkeit des Erkennenden.» / «Omne enim, quod cognoscitur, non secundum sui vim, sed secundum cognoscentium potius comprehenditur facultatem.»

senheit, welche die menschliche Willens- und Handlungsfreiheit ausschließt, sondern nur eine falsch verstandene Allwissenheit, die These 2 irrtümlich so deutet, dass sie – ohne Rücksicht auf die *simplicitas* Gottes – ein göttliches Vorherwissen (eine *praevidentia*) annimmt und dieses nicht als aeternital, sondern als temporal versteht. Sobald man aber über das Missverständnis, mit dem göttlichen Denken verhalte es sich im Grunde wie mit dem menschlichen, aufgeklärt ist, erkennt man, dass Gottes *intellectus* über gar keine (zeitlich situierte) *praevidentia* verfügt:

> Wenn Du also seine Voraussicht, mit der er alles erkennt, richtig einschätzen willst, so wirst Du sie nicht als Vorherwissen einer etwaigen Zukunft, sondern viel richtiger als ein Wissen von einer niemals entschwindenden Gegenwart auffassen. Daher wird es nicht Vorhersehen (*praevidentia*), sondern lieber Vorsehung (*providentia*) genannt, weil sie sich fern von den niederen Dingen aufhält und gewissermaßen vom erhabenen Gipfel der Dinge herunter alles vor sich sieht.[7]

Demnach ist alles, was sich ereignet, dem göttlichen *intellectus* zugleich präsent. Es gibt für ihn kein zeitliches Nacheinander, sondern nur ein in übertragenem Sinne räumliches Nebeneinander. Entsprechend ist das, was für die göttliche Kognition als aktual und in diesem Sinne als notwendig gegeben erscheint, gemäß der menschlichen *ratio* und für sich selbst – d.i. gemäß der eigenen zeitlichen Natur – betrachtet, keineswegs notwendig, sondern in einer Weise kontingent, die der menschlichen Freiheit den benötigten Entscheidungs- und Handlungsspielraum eröffnet.[8] Entscheidend für die angestrebte Problemlösung ist also, dass zwar die irrige Annahme einer göttlichen *praevidentia* die menschliche Freiheit ausschließt, die für Boethius stimmige Annahme einer göttlichen – mit der Ewigkeit des *intellectus* gegebenen – *providentia* aber mit der menschlichen Freiheit verträglich ist.

7 Boethius: *Consolatio Philosophiae* V, 6.p., 68-72: «Itaque si praevidentiam pensare velis, qua cuncta dinoscit, non esse praescientiam quasi futuri, sed scientiam numquam deficientis instantiae rectius aestimabis. Unde non praevidentia, sed providentia potius dicitur, quod porro a rebus infimis constituta quasi ab excelso rerum cacumine cuncta prospiciat.»

8 Boethius: *Consolatio Philosophiae* V, 6.p., 103-106: «Ich werde weiterhin antworten, dass ein und dasselbe Zukünftige, wenn es auf die göttliche Erkenntnis bezogen wird, notwendig, wenn es aber nach seiner eigenen Natur gewürdigt wird, völlig frei und unabhängig erscheint.» / «Respondebo namque idem futurum, cum ad divinam notionem refertur, necessarium, cum vero in sua natura perpenditur, liberum prorsus atque absolutum videri.»

Zusammenfassend lässt sich die von Boethius verwendete Begrifflichkeit terminologisch wie folgt fassen: Der *intellectus* Gottes und die ihm eigene *providentia* zeichnen sich durch die Merkmale der (a.) Atemporalität, (b.) Universalität und (c.) Infallibilität aus. Allein durch das erste Merkmal unterscheidet sich das göttliche Denken von der (nach Boethius irrtümlichen) Annahme einer temporal situierten *praescientia* Gottes, durch alle drei Merkmale hingegen von der *praevidentia*, insofern diese den temporalen, partikulären und falliblen Zukunftsbezug der menschlichen *ratio* benennt. Durch die drei genannten Merkmale werden die Begriffe des *intellectus* und der *ratio* nicht nur wechselseitig konturiert und inhaltlich profiliert, es wird auch deutlich, dass der *intellectus* als der *ratio* deutlich überlegen und als eine spezifische Form von Nichtrationalität, nämlich als Überrationalität gedacht wird. Klar markiert wird der zwischen *intellectus* und *ratio* bestehende Rangunterschied auch in der von Boethius hierarchisch geordneten Taxonomie der Kognitionsvermögen, in welcher der *intellectus* an oberster und die *ratio* an zweiter Stelle positioniert wird, während der Vorstellungskraft (*imaginatio*) der dritte und der (auch den Tieren zugebilligten) sinnlichen Wahrnehmung (*sensus*) lediglich der vierte und letzte Platz eingeräumt wird.[9]

2. Lorenzo Valla: die Kritik göttlicher Überrationalität

Lorenzo Vallas Kritik am Begriff eines göttlichen *intellectus* erfolgt vor demselben Problemhintergrund wie die Konzeptionierung dieses Begriffs durch Boethius: In seinem Dialog *De libero arbitrio* thematisiert Valla die Frage nach dem Verhältnis zwischen menschlicher Freiheit und göttlicher Allwissenheit, und wie Boethius versucht auch Valla darzulegen, dass sich die Annahmen einer *libertas arbitrii* und einer *sapientia Dei* durchaus miteinander vereinbaren lassen. Doch entwickelt Valla einen eigenen – auf der Differenz zwischen Vorauswissen (*praescientia*) und Vorausbestimmung (*praedeterminatio*) basierenden – Lösungsansatz, den er scharf von demjenigen des Boethius abgrenzt, indem er dessen Unterscheidung von *ratio* und *intellectus* problematisiert. Die entsprechenden Ausführungen beginnen

9 Vgl. Boethius: *Consolatio Philosophiae* V, 4.p., 85-92. Zum neuplatonischen Hintergrund dieser Taxonomie vgl. Henry Chadwick: *Boethius. The Consolations of Music, Logic, Theology and Philosophy* (Oxford: Clarendon, 1981) S. 246-247.

mit einer – der Dialogfigur Antonius in den Mund gelegten – Herabsetzung aller bislang erfolgten Behandlungen des Problems, namentlich aber der des Boethius:

> In dieser Frage jedoch, über die ich mit dir zu sprechen begonnen habe, kann ich, ohne dir und anderen zu nahe treten zu wollen, absolut niemandem zustimmen. Denn was soll ich über die anderen sagen, wenn selbst Boethius, dem in der Behandlung dieser Frage von allen der erste Rang zugesprochen wird, das, was er sich vorgenommen hat, nicht erfüllen kann, sondern zu eingebildeten und erlogenen Dingen seine Zuflucht nimmt?[10]

Boethius' Position wird mit wenigen treffenden Worten referiert, der entscheidende Einwand – nicht weniger knapp – unmittelbar angefügt:

> [Boethius] sagt nämlich, dass Gott durch die Einsicht (*intelligentia*), die über dem Verstand (*ratio*) ist, und durch die Ewigkeit (*aeternitas*) alles wisse und alles gegenwärtig habe. Aber ich, der ich ein mit Verstand begabtes Wesen bin und nichts außerhalb der Zeit erkenne, wie kann ich zur Erkenntnis der Einsicht und der Ewigkeit zu kommen hoffen?[11]

Mit dieser rhetorischen Frage macht Valla deutlich, dass der Unterschied zwischen einem göttlichen, als aeternital, universal und infallibel gedachten *intellectus* auf der einen Seite und einer menschlichen, als temporal, partikulär und fallibel begriffenen *ratio* auf der anderen Seite nicht immer schon vorliegt, sondern das Produkt eines Unterscheidungsaktes darstellt. Es ist also ein Akteur vonnöten, der den Unterschied zwischen *ratio* und *intellectus* vollzieht, und dieser Akteur kann nur die menschliche *ratio* sein, die offensichtlich zwischen sich selbst als einem unterlegenen und dem *intellectus* als einem überlegenen Kognitionsvermögen differenziert. Doch stellt sich damit sogleich die Frage, ob die *ratio* zu einer solchen Differenzierungsleistung überhaupt in der Lage ist: Kann der Mensch mit den begrenzten Mitteln, die seinem zeitlichen, partikulären und irrtumsanfälligen Kognitionsvermögen

[10] Lorenzo Valla: *De libero arbitrio* § 29, 151-156: «In hac autem, de qua tecum loqui instituo, pace tua et aliorum dictum sit, nemini prorsus assentior. Nam quid de aliis dicam? Cum Boëtius ipse, cui in explicanda hac quaestione datur ab omnibus palma, quod susceperit implere non possit; sed ad quasdam res confugiat imaginarias et commentitias.»

[11] Lorenzo Valla: *De libero arbitrio* §§ 29-30, 157-161: «Ait enim Deum per intelligentiam, quae supra rationem est, et per aeternitatem omnia scire, omniaque habere praesentia. At ego ad cognitionem intelligentiae et aeternitatis, qui rationalis sum et nihil extra tempus agnosco, aspirare qui possum?»

zur Verfügung stehen, tatsächlich zwischen der spezifisch menschlichen *ratio* und einem göttlichen *intellectus*, der ewig, universal und irrtumsresistent sein soll, unterscheiden?

Vallas Dialogfiguren sind sich einig in der Verneinung dieser Frage: Wenn die menschliche *ratio* meint, sie könne aus der Perspektive der Zeitlichkeit sinnvoll über die Ewigkeit sprechen und sich selbst vom göttlichen *intellectus* abgrenzen, den sie in dieser Unterscheidung ja begrifflich in Anspruch nehmen muss, so verfällt sie einer eklatanten Selbstüberschätzung. Während eine aufgeklärte *ratio* um die eigene Limitierung weiß, stellt die Abgrenzung zwischen *ratio* und *intellectus* den hybriden Versuch des Menschen dar, in der Unterscheidung vom menschlichen auch das göttliche Denken denken zu wollen.[12] So gesehen, kann der Versuch, ein Konzept von göttlicher Überrationalität zu entwickeln, als ein seinerseits irrationales, da die begrenzten Möglichkeiten der *ratio* nicht berücksichtigendes Unterfangen demaskiert werden. Valla hält die Unterscheidung zwischen *ratio* und *intellectus* also nicht für falsch, sondern – schlimmer noch – für sinnlos, da sich das menschliche Denken gar keinen sinnvollen Begriff vom göttlichen Denken machen kann. Der Ansatz von Vallas Kritik ist, wie die folgende Bemerkung von Boethius zeigt, bereits in der *Consolatio Philosophiae* selbst angelegt: «Dabei ist besonders zu beachten, dass die höhere Kraft des Begreifens die niedere umspannt, während die niedere sich auf keine Weise zur höheren erheben kann.»[13] Allerdings verzichtet Boethius – im Unterschied zu Valla – darauf, die Skepsis hinsichtlich der Erkennbarkeit und Nachvollziehbarkeit des göttlichen *intellectus* durch die menschliche *ratio* auf die Unterscheidbarkeit beider Kognitionsvermögen und die Konzeptionierbarkeit des Begriffs des *intellectus* selbst anzuwenden.

12 Der gegenüber der *intellectus*-Konzeption erhobene Vorwurf einer philosophischen Hybris wird flankiert von der bereits im Prooemium von *De libero arbitrio* geäußerten Kritik, Boethius habe das Problem der Willensfreiheit nicht in gebotener Weise behandelt, da er ein allzu großer *philosophiae amator* gewesen sei. Vgl. auch Lorenzo Valla: *De libero arbitrio* § 108, 807–§ 109, 821, wo Aristoteles der *superbia* bezichtigt wird.

13 Boethius: *Consolatio Philosophiae* V, 4.p., 93-95: «In quo illud maxime considerandum est; nam superior comprehendendi vis amplectitur inferiorem, inferior vero ad superiorem nullo modo consurgit.» Vgl. Joachim Gruber: *Kommentar zu Boethius De consolatione philosophiae* (Berlin, New York: De Gruyter 1978) S. 403-404: «Untergeordnete Erkenntniskräfte können über die Erkenntnisse der übergeordneten Kräfte nicht urteilen, also auch nicht der menschliche Verstand über die Vernunft Gottes.»

3. Diskussion von Vallas Kritik

Folgt man Vallas Kritik, so stellt die göttliche Überrationalität keine vernünftige Bezugsgröße für eine Selbstverständigung menschlicher Rationalität dar, da eine solche Form der Nichtrationalität zwar nicht hinsichtlich ihres Gehalts, wohl aber hinsichtlich ihres Konzeptionierungsaktes als irrational – nämlich als eine zutiefst unvernünftige Überheblichkeit und Unaufgeklärtheit menschlicher Vernunft – zu bewerten ist. Doch kann Vallas Kritik am Konzept göttlicher Überrationalität – so scheint es jedenfalls auf den ersten Blick – leicht auf sich selbst angewendet werden. Demnach muss derjenige, der die Unterscheidung zwischen menschlicher *ratio* und göttlichem *intellectus* unter Sinnlosigkeitsverdacht stellt, zur Formulierung eines solchen Vorwurfs genau das tun, was er der Unterscheidung vorwirft: Er muss vom Begriff einer göttlichen Überrationalität Gebrauch machen. Denn um zu zeigen, dass die *ratio* den *intellectus* nicht zu fassen vermag, scheint es unumgänglich, den Begriff des *intellectus* selbst ins Spiel zu bringen: Man konzipiert einen Begriff des göttlichen Denkens, der so weit über dem menschlichen Denken steht, dass ihn dieses nicht zu denken vermag. Damit aber verfällt man – trotz der ostentativen Bescheidenheitsgeste – selbst der Hybris. Für den Nachweis, dass sich das göttliche Denken gar nicht denken lässt, muss man eben das göttliche Denken – in welcher Weise auch immer – denken. So gesehen, fällt der Irrationalitätsvorwurf gegenüber der Unterscheidung zwischen menschlicher *ratio* und göttlichem *intellectus* unmittelbar auf den zurück, der ihn äußert. Anscheinend stellt die Kritik an einer solchen Unterscheidung ein Argumentationspotential bereit, das nicht nur die Konzeptionierung göttlicher Überrationalität unterminiert, sondern auch deren Problematisierung.

Falls dieser Selbstapplikationseinwand triftig ist, so muss ein Disput wie der zwischen Boethius und Valla notwendig im Patt enden: Sobald der Vorwurf der Irrationalität erhoben wird, kann dieser durch den Hinweis gekontert werden, auch selbst irrational zu sein. In der Folge ergibt sich eine problematische Selbstimmunisierung des Konzepts göttlicher Überrationalität, das zwar jede Sinnkritik ihrerseits einer Sinnkritik unterziehen kann, ohne damit aber den eigenen Sinnlosigkeitsverdacht plausibel zu entkräften. Aus einer dritten Perspektive könnte entsprechend – in der Diktion des späten Wittgenstein – von einem Sprachspiel gesprochen werden, das als ganzes unter Sinnlosigkeitsverdacht steht – mit der Konsequenz, dass in einem solchen Spiel nicht nur Züge – wie der des Boethius –, sondern auch Gegenzüge – wie der Vallas – prekär erscheinen müssen. Eben diesem Sprachspiel

würde dann auch der Selbstapplikationseinwand selber zugehören, da dieser ja ebenfalls den Begriff des *intellectus* gebrauchen muss, um die Kritik an der Unterscheidung zwischen *ratio* und *intellectus* kritisieren zu können. Als weitere unerfreuliche Konsequenz droht demnach – neben dem genannten Patt und der drohenden Selbstimmunisierung – ein unendlicher Regress von beliebig iterierbaren Sinnlosigkeitsnachweisen.

Doch lässt sich Vallas Kritik am Konzept göttlicher Überrationalität auch anders fassen und gegen den eben geäußerten Selbstapplikationseinwand in Schutz nehmen: Demnach muss derjenige, der die Unterscheidung zwischen menschlicher *ratio* und göttlichem *intellectus* unter Sinnlosigkeitsverdacht stellt, zur Formulierung eines solchen Vorwurfs eben nicht genau das tun, was er der Unterscheidung vorwirft: Er muss keinen Gebrauch vom Begriff einer göttlichen Überrationalität machen, da er diesen nicht selbst – in einem objektsprachlichen Sinne – verwendet, sondern nur – in einem metasprachlichen Sinne – über diesen Begriff spricht. Die Kritik thematisiert also keineswegs selbst den *intellectus*, sondern vielmehr die Rede über den *intellectus*, die daher unter Sinnlosigkeitsverdacht gestellt werden kann, ohne dass sich dieser rückwendend auf den Verdacht selbst überträgt. Um nochmals mit Wittgenstein zu sprechen: Eine Kritik wie diejenige, die Valla gegen Boethius vorbringt, ist nicht als ein Gegenzug zu charakterisieren, der demselben Sprachspiel zugehören würde wie der Zug, auf den er reagiert. Vielmehr stellt eine solche Kritik einen Zug in einem anderen Sprachspiel dar. Dieses zweite Sprachspiel bezieht sich nun zwar auf das erste, das ihm in zeitlicher wie logischer Hinsicht vorausgeht, doch sind beide Sprachspiele – mit Blick auf ihre jeweilige Sinnhaftigkeit – streng voneinander zu unterscheiden. Entsprechend kann der Vorwurf, die Konzeption göttlicher Überrationalität sei irrational, seinerseits als durchaus rational betrachtet werden. Mit anderen Worten: Einer bestimmten Konzeption Sinnlosigkeit nachzuweisen, führt nicht zur Sinnlosigkeit dieses Nachweises. Jedenfalls solange nicht, wie es der Kritik gelingt, den eigenen Gebrauch solcher Begriffe zu vermeiden, die als sinnlos zu betrachten sind; die Thematisierung solcher Begriffe aber ist nicht nur erlaubt, sie scheint für einen gelungenen Sinnlosigkeitsnachweis sogar geboten.

4. Schluss

Der Disput, den ich anhand von Boethius und Valla vorgeführt habe, scheint symptomatisch für eine bestimmte Denktradition, die ich vorschlagsweise die «Theologie der Differenz» nennen möchte. Kennzeichnend für diese ist, dass Gott als «Der ganz Andere» (gegenüber dem Menschen) und das Denken Gottes als «Das ganz Andere» (gegenüber dem menschlichen Denken) gedacht wird. Die alternative Richtung, die bestimmte Eigenschaften sowohl Gott als auch dem Menschen zuzuschreiben bereit ist und das Denken Gottes nicht in Abgrenzung gegenüber dem menschlichen, sondern in Anlehnung an das menschliche Denken zu fassen versucht, könnte man entsprechend als eine «Theologie der Konvergenz» bezeichnen. Als erster wichtiger Vertreter der «Theologie der Differenz» wäre Xenophanes zu benennen, der bekanntlich nicht nur die tradierten Gottesvorstellungen der griechischen Mythologie als anthropomorph kritisiert, sondern auch einen eigenen Gottesbegriff entwirft, den er von anthropomorphen Zügen freizuhalten versucht: Zwar schreibt auch Xenophanes Gott – mit dem *nous* – eine bestimmte Kognitionsfähigkeit zu, doch grenzt er diese von dem – als *dianoia* bezeichneten – Kognitionsvermögen des Menschen entschieden ab.[14] Anders als die menschliche ist die göttliche Kognition nicht an die Körperlichkeit der Sinnesorgane, die Beschränktheit der Perspektiven, die Aspekthaftigkeit des Betrachteten und die prinzipielle Medialität gebunden: Das Denken Gottes vollzieht sich nicht wie das des Menschen auf eine diskursiv-dianoetische Weise, sondern intuitiv-noetisch, gleichsam als unvermitteltes Erfassen der Wirklichkeit in ihrer Gesamtheit.[15] Damit ist das göttliche Kognitionsvermögen vom menschlichen nicht nur partiell verschieden und ihm auch nicht bloß komparativisch überlegen, sondern in kategorischem Sinne unvergleichbar.[16]

14 Vgl. Xenophanes: *Fragment Diels/Kranz* 21, B 23.

15 Vgl. ibid. 21, B 24.

16 Da es nach Xenophanes prinzipiell keine Eigenschaften gibt, die Gott mit den Menschen teilen würde, kann das Gestaltungsprinzip seines Gottesbegriffs als das einer Mensch-Gott-Heteromorphie bezeichnet werden, während die mythologischen Gottesvorstellungen einer Mensch-Gott-Isomorphie folgen. Vgl. hierzu Hartmut Westermann: *Religiöse und doppelte Poesie. Götterkritik und Gottesbegriff bei Xenophanes und im Kulturentstehungsmythos des Sisyphos* (*DK 88, B 25*), in *Philosophische Anthropologie und Lebenskunst. Rainer Marten in der Diskussion*, hg. von Guido Löhrer, Christian Strub, Hartmut Westermann (München: Fink 2005) S. 81-98.

Dieser von Xenophanes begründeten «Theologie der Differenz» kann nun nicht nur Boethius zugerechnet werden, wenn er den göttlichen *intellectus* als eine von der menschlichen *ratio* nicht erreichbare Form von Überrationalität konzipiert, sondern auch Valla, wenn er Boethius' Unterscheidung von *ratio* und *intellectus* mit Hilfe eines Arguments problematisiert, welches das göttliche Denken so stark vom menschlichen abhebt, dass es von diesem gar nicht mehr gedacht werden kann. Wird die Kluft zwischen menschlicher und göttlicher Kognition auf eine solch radikale Weise gedeutet, dann ist schlicht jede Aussage von Menschen über Gott, schon allein weil sie eine Aussage von Menschen ist, als anthropomorph und damit als Gott inadäquat zu begreifen. Kurz: Sinnvolle Aussagen über Gott erscheinen – jedenfalls aus menschlicher Perspektive – ausgeschlossen. Was hingegen möglich bleibt, sind sinnkritische Aussagen, die nicht Gott, sondern Aussagen über Gott zum Gegenstand haben. In letzter Konsequenz führt die «Theologie der Differenz» also zu einer Art Metatheologie, deren Theorieambition darin liegt, der «Theologie der Differenz» die eigene Irrationalität nachzuweisen.[17]

[17] Für kritische Anregungen zu einer früheren Version dieses Textes danke ich Rafaela Hillerbrand, Wulf Kellerwessel und Guido Löhrer.

Irrationales Handeln?
Agir de manière irrationnelle ?

Studia philosophica 68/2009

Yves Bossart

Sind pyrrhonische Skeptiker irrational? Radikale Skepsis und die Grenzen der Rationalität

Pyrrhonian skepticism is the most radical and consequent form of skepticism. It is a way of life in which suspension of judgement (epochē) leads to peace of mind (ataraxia). A Pyrrhonian does not believe anything, because he has the impression that there are no beliefs for which there are sufficient reasons. It seems to him that the reasons for the truth of every single statement are no better than the reasons against it (isostheneia). How can he then make claims and argue? Are his considerations not self-refuting? Is he able to decide rationally between alternative ways of life? I try to show that these questions can be answered in favour of Pyrrhonian rationality. There may be one suspicion of irrationality left: For a Pyrrhonian, theory is a form of therapy and thus a means for successful practice. If rationality prevents him from having peace of mind, he would give up rationality for happiness.

Im ersten Teil dieses Aufsatzes werde ich erläutern, was ein pyrrhonischer Skeptiker ist, welche Überlegungen er anstellt, wie er argumentiert und woran er sich im Denken und Handeln orientiert. Im zweiten Teil lege ich dar, was ich unter Rationalität verstehen möchte, um anschließend fünf Irrationalitätsvorwürfe zu besprechen, die an den pyrrhonischen Skeptiker gerichtet werden können.

1. Pyrrhonische Skepsis

Die pyrrhonische Skepsis ist eine antike Denk- und Lebenskunst, als deren Begründer Pyrrhon von Elis (360-270 v. Chr.) gilt. Der spätantike Arzt und Philosoph Sextus Empiricus erarbeitete am Ende des 2. Jahrhunderts n. Chr. unter Rückgriff auf Aenesidemus und Agrippa eine systematische Rekonstruktion der pyrrhonischen Skepsis mit dem Titel «Grundzüge der pyrrhonischen Skepsis» (*pyrrhoneiai hypotypōseis*). Wenn ich im Folgenden von «pyrrhonischer Skepsis» spreche, beziehe ich mich damit auf die Darstellung des Sextus Empiricus.

Im Vergleich mit anderen skeptischen Positionen der Philosophiegeschichte gilt die pyrrhonische Skepsis zu Recht als umfassendste und kon-

sequenteste Form des Skeptizismus. Was den Pyrrhonismus auszeichnet, ist jedoch nicht nur seine Radikalität, sondern insbesondere auch seine Praxisbezogenheit. Das Ziel des Pyrrhonikers ist kein theoretisches, sondern ein praktisches. Die pyrrhonische Skepsis sei, so Sextus, keine Lehre (*hairesis*), die sich durch bestimmte Lehrmeinungen (*dogmata*) auszeichne, sondern in erster Linie eine Tätigkeit (*agōgē*) und eine Lebenskunst.[1] Das praktische Ziel des Pyrrhonikers ist die Glückseligkeit (*eudaimonia*), die in der Gemütsruhe (*ataraxia*) und in einer gemäßigten Erlebnisintensität (*metriopatheia*) besteht.[2] Die pyrrhonische Skepsis gleicht in praktisch-therapeutischer Ausrichtung also den beiden damals konkurrierenden hellenistischen Philosophie-Schulen: der Stoa und dem Epikureismus. Jede dieser drei intellektuellen Strömungen zielt auf eine gelingende Lebensführung und begreift die intellektuelle Praxis als Weg zur Glückseligkeit.

Nach pyrrhonischer Auffassung entsteht alles Unglück (*kakodaimonia*) durch eine Verwirrung (*tarachē*) des Gemüts. Der Ursprung dieser Gemütsunruhe liege dabei in den festen Überzeugungen der Menschen. Wer fest von etwas überzeugt ist und zu wissen glaubt, was richtig und falsch ist, der lebe ungestüm (*meta sphodron*) und strebe eifrig (*syntonos*) auf ein Ziel hin, das er für ein objektives Gut und für von Natur aus (*physei*) erstrebenswert hält.[3] Der «Dogmatiker» – wie Sextus alle Menschen nennt, die Lehrmeinungen (*dogmata*) vertreten und starke Überzeugungen haben – lebe unbefriedigt und trachte nach einem schwer zu erreichenden Ziel, um dessen Erhaltung er sich sorge, sobald er es erreicht habe. Auch werde er sich im Gespräch ereifern, da er von der Wahrheit bestimmter Thesen überzeugt sei. Er versuche diese vehement zu verteidigen, aus Angst, in seinen Meinungen widerlegt zu werden und sich dadurch vor die schmerzliche Aufgabe gestellt zu sehen, das Altvertraute liegen lassen und sich nach einer neuen Orientierungsgrundlage des eigenen Denkens und Lebens umsehen zu müssen.

Der pyrrhonische Skeptiker dagegen lebt ohne feste Überzeugungen und enthält sich jeglichen Urteils (*epochē*). Diese Überzeugungslosigkeit erweist sich als Grundlage der erstrebten Gemütsruhe (*ataraxia*). Wie aber befreit sich der Pyrrhoniker von seinen festen Überzeugungen?

1 Sextus Empiricus: *Grundzüge der pyrrhonischen Skepsis* (PH) (Frankfurt a.M.: Suhrkamp, 1985) §§ 16-17. Der altgriechische Text findet sich in: Sextus Empiricus, vol. I-IV, übers. von R. G. Bury (Cambridge, MA: Harvard University Press, 1955).

2 Sextus Empiricus: *Against the Ethics* (PE) (Cambridge, MA: Harvard University Press, 1953) § 141, S. 160-161; PH I, § 25.

3 PE, § 112.

Der Pyrrhoniker ist ein Meister der dialektischen Kunst der Entgegensetzung (*dynamis antithetikē*)[4]. Er versteht es, zu jeder glaubhaften Position ebenso glaubhafte Gegenargumente zu formulieren. Entweder zeigt er uns, dass wir für unsere Meinungen keine guten Gründe haben, oder er macht uns darauf aufmerksam, dass es für andere Positionen ebenso gute Gründe gibt. Aufgrund der anscheinenden Gleichwertigkeit (*isostheneia*) widerstreitender Argumente zieht der pyrrhonische Skeptiker es vor, sich des Urteils zu enthalten; er übt sich in doxastischer Zurückhaltung (*epochē*).

Das in den Schriften des Sextus Empiricus überlieferte *Begründungstrilemma* des Agrippa erweckt das vielleicht radikalste Misstrauen gegen Wissens- und Begründungsansprüche jeglicher Art.[5] Ausgangspunkt ist die plausible These, dass jemand, der etwas behauptet, in der Lage sein sollte, Gründe für das Behauptete anzuführen. Wer behauptet, er brauche keine Gründe anzuführen, da die Sache doch völlig klar und einsichtig sei, stiehlt sich aus der Verantwortung. Mit der Selbstevidenz ist es nämlich so eine Sache: Was als klar, unmittelbar einsichtig, gegeben und plausibel gilt, hängt jeweils davon ab, für wen und in welcher Situation dies der Fall ist. Die von Sextus angeführten zehn skeptischen Tropen des Aenesidemus und insbesondere die erste der zwei Tropen des Menodot sollen zeigen, dass die Evidenz aufgrund ihrer Abhängigkeit von Subjekt und Kontext kein allgemeingültiges Wahrheitskriterium sein und also nichts gleichsam durch sich selbst als wahr erkannt werden kann (*ouden ex heautou katalambanetai*).[6] Jede Behauptung, die mehr sein soll als eine *bloße* Behauptung, muss durch Gründe, Argumente und Überlegungen gestützt werden. Die zur Bestätigung einer These angeführten Gründe und Argumentationsprämissen sind jedoch in den Augen des Skeptikers ebenso begründungsbedürftig wie die durch sie zu begründende These.[7] Man ahnt, was droht. Wer behauptet, er könne seine Meinung begründen, dem bleiben drei unbefriedigende Alternativen: Entweder (1) er durchläuft unendlich viele Argumentationsschritte (*ho eis apeiron ekballōn*), da er für die Annahmen, auf denen seine Begründung basiert, eine weitere Begründung anführen muss, die wiederum auf Annahmen basiert, die begründet werden müssen usf., oder (2) er argumentiert zirkulär

4 PH I (Fn. 1), § 8.

5 PH I, § 164-177.

6 PH I, § 178.

7 In manchen Fällen folgt die zu begründende These nicht aus den Prämissen; in solchen Fällen reichen eine logische Analyse und der Hinweis, dass bei induktiven und abduktiven Schlüssen die Wahrheit der Prämissen die Falschheit der Konklusion nicht ausschliesst.

für seine Annahmen, indem er diese durch Argumente stützt, die dasjenige, wofür er argumentieren möchte, bereits voraussetzen (*ho diallēlos*), oder (3) er stoppt die Begründungskette bei irgendeiner Behauptung, die dadurch den Status einer unbegründeten Hypothese aufweist (*ho hypothetikos*). Wer Begründungsansprüche erhebt, manövriert sich also in ein Trilemma. Er kann nur noch wählen zwischen (1) einem infiniten Regress, (2) einer zirkulären Begründung oder (3) einem dogmatischen Abbruch.

Am Anfang der pyrrhonischen Skepsis steht die Einsicht, dass jede Position, also auch die jeweils eigene, lediglich als Teil eines unaufgelösten Widerstreits zu betrachten ist (*ho epi tēs diaphōnias*)[8] und nur relative Geltung für sich in Anspruch nehmen kann (*ho epi tou pros ti*). Jeder Versuch, mehr beanspruchen zu wollen als eine schlecht begründete Position innerhalb eines unentschiedenen Konflikts, mündet – so der skeptische Verdacht – in das skizzierte Trilemma.

Der Pyrrhoniker stellt selbst keine Thesen auf, sondern verhält sich stets reaktiv. Im Denken und in Gesprächen mit anderen zeichnet er sich – wie bereits erwähnt – durch die Fähigkeit aus, Gegenpositionen stark zu machen und die Schwächen der gegnerischen Position aufzuzeigen (*dynamis antithetikē*)[9]. Ziel dieser Entgegensetzung ist es, den Eindruck einer Gleichwertigkeit (*isostheneia*) der jeweils widerstreitenden Argumente und Thesen herbeizuführen. Aufgrund dieses Eindrucks sieht sich der Skeptiker nicht mehr in der Lage, Stellung zu beziehen. Ihm bleibt nur noch die Zurückhaltung (*epochē*) im Urteilen. An dieser Stelle zeigt sich nun die praktische Ausrichtung des Pyrrhonikers. Die Zurückhaltung ist kein Selbstzweck, sondern Bestandteil eines glücklichen, ausgeglichenen Lebens. Die erwünschte Gemütsruhe (*ataraxia*), das praktische Ziel des Pyrrhonikers, beruhe nämlich auf der Zurückhaltung und stelle sich wie von selbst ein, sobald man diese habitualisiert habe. Der Weg zum skeptischen Glück sieht also wie folgt aus: Durch Entgegensetzung soll sich der Eindruck der Gleichwertigkeit einstellen, an den sich die Zurückhaltung anschließt, welcher wiederum die gewünschte Gemütsruhe folgt. Dieses lebenspraktische Rezept darf allerdings nicht als eine Theorie des Glücks missverstanden werden. Zum einen würde ein solches Glücksversprechen nämlich der skeptischen Grundhaltung widersprechen, die auch in Fragen der gelungenen Lebensführung keine Gewissheiten anzubieten vermag, und zum anderen würde eine mit der pyrrhonischen Praxis verbundene Glücksgarantie – ebenso wie bei dogmatischen

8 Vgl. PH I (Fn. 1), § 59 (*meros tēs diaphōnias*).
9 PH I, § 8.

Lehren – ein eifriges, besorgtes, vielleicht sogar angstvolles Streben nach der Gemütsruhe auslösen, die paradoxerweise umso weniger erreicht wird, je heftiger und bedingungsloser sie gewollt wird.[10] Sextus entgeht diesem Vorwurf, indem er schreibt, die Gemütsruhe stelle sich bei geübter Zurückhaltung im Urteilen «wie durch Zufall (*tychikōs*)» ein.[11] Dadurch vermeidet er die absurde Konsequenz eines verbissenen Strebens nach der Gemütsruhe als einem Zustand des Nicht-Strebens.

Die Zurückhaltung des Pyrrhonikers ist die natürliche Konsequenz seines Eindrucks der Gleichwertigkeit. Er behauptet nicht, man *solle* etwas nur dann glauben, wenn man auch Gründe dafür habe.[12] Dies wäre aus seiner Sicht nur eine weitere unbegründbare Behauptung. Sextus behauptet an keiner Stelle, man solle sich des Urteils enthalten, wenn keine Gründe dafür sprechen. Der Eindruck, der sich aufgrund skeptischer Überlegungen einstellt, das Gefühl der Grundlosigkeit und Gleichwertigkeit aller Überzeugungen, führt beim Pyrrhoniker aufgrund psychologischer Gesetzmäßigkeiten dazu, dass er seine Meinungen aufgibt. Sein Verhalten entspricht somit zwar der Regel «Glaube nicht, wofür du keine Gründe hast». Er *folgt* jedoch nicht dieser Regel.

1.2 An den Grenzen zur Widersprüchlichkeit

Der Pyrrhoniker möchte zeigen, dass wir für *keine* unserer Überzeugungen gute Gründe haben. Damit würde er auch zeigen, dass wir kein Wissen haben, denn wer keine guten Gründe hat, der hat auch kein Wissen. Dabei ist wichtig, dass der Skeptiker keine überhöhten Anforderungen an eine gute Begründung stellt. Er sollte von einer Begründung nicht mehr verlangen als das, was auch seine Gegner bereit sind zuzugestehen. Stellt er ungewohnt hohe Ansprüche an Wissen und Rechtfertigung, so würde er das Thema wechseln, ähnlich wie jemand, der behauptet, es gäbe keine guten Ärzte, da schließlich kein Arzt in der Lage sei, jedwede Krankheit innerhalb von fünf Minuten zu heilen.

10 Darin unterscheidet sich die pyrrhonische Haltung von der Skepsis des Arkesilaos (PH I, § 233).

11 PH I, § 26.

12 Alan Bailey: *Sextus Empiricus and Pyrrhonean Scepticism* (Oxford: Clarendon Press, 2002) S. 270.

Die Anforderungen, die der Pyrrhoniker an gute Begründungen stellt, scheinen auf den ersten Blick ganz in Ordnung zu sein: Eine gute Begründungskette sollte weder zirkulär noch unendlich lang sein noch sollte sie auf Prämissen beruhen, die unbegründet sind. Diese Anforderungen scheint auch der Dogmatiker implizit zu akzeptieren. Das Unbefriedigende und Unbehagliche dabei ist allerdings, dass es, wenn der Skeptiker recht hat, *gar keine* akzeptablen Begründungen gibt, da jedem Rechtfertigungsversuch ein trilemmatisches Schicksal bevorsteht. Ist das Trilemma also etwa die Begründung dafür, dass es keine Begründung geben kann? Vollzieht der Pyrrhoniker aber nicht einen praktischen Selbstwiderspruch, indem er versucht, gute Gründe für die These anzuführen, dass es keine guten Gründe gibt? An dieser Stelle zeigt sich nun die Radikalität des pyrrhonischen Skeptikers. Er glaubt nämlich *nicht*, dass seine skeptischen Argumente besser oder stichhaltiger seien als die Argumente seines Gegners.[13] Die Gründe des Dogmatikers, die dafür sprechen, dass wir für einige unserer Überzeugungen gute Gründe haben, scheinen dem Skeptiker nicht besser und nicht schlechter zu sein als die skeptischen Gründe, die dagegen sprechen. Der Pyrrhoniker versucht lediglich, in seinem Opponenten denselben Eindruck hervorzurufen, den auch er hat, nämlich dass die Gründe *pro* und *contra* sich auch in diesem Fall die Waage halten. Da sich der Dogmatiker an Argumenten orientiert, liefert der Skeptiker Argumente. Gleichzeitig appelliert er an die für skeptische Überlegungen basalen Intuitionen des Dogmatikers. Er ahnt, dass seine skeptischen Argumente, und insbesondere das Begründungstrilemma, aus Sicht des Dogmatikers gute Gründe für einen globalen Rechtfertigungsskeptizismus darstellen. Er holt den Dogmatiker da ab, wo sich dieser befindet, indem er aufzeigt, dass unsere Rechtfertigungspraxis die vom Dogmatiker implizit akzeptierten Anforderungen an valide Begründungen schlicht nicht erfüllt. Der Skeptiker argumentiert also ausgehend von dogmatischen Prämissen und Intuitionen, ist selbst jedoch bereits einen Schritt weiter. Er wendet nämlich die skeptischen Argumente, die letztlich Argumente gegen die Validität aller möglichen Argumente sind, auch auf diese Argumente selbst an und zeigt so, dass für skeptische Auffassungen des Wissens und der Rechtfertigung letztlich keine besseren Gründe sprechen als für dogmatische. Der Pyrrhoniker ist sich der Selbstbezüglichkeit (*peritropē*) skeptischer Argumente also wohl bewusst, betont aber deren positiven *therapeutischen* Effekt, indem er sie mit der Wirkungsweise eines abführenden Medikaments vergleicht, dessen Einnahme dazu führt, dass es zusammen mit den krankheitsverursachenden

[13] PH II (Fn. 1), § 187.

Giften vom Körper ausgeschieden wird.[14] Mit den Giften sind dabei die starken Überzeugungen gemeint, welche der Skeptiker durch Gegenargumente vertreiben möchte.[15]

Der Pyrrhoniker ist ein Therapeut, der durch Argumentationen und Reden sich selbst und andere heilt (*iasthai logō*).[16] Die therapeutische Einstellung des Pyrrhonikers verlangt ein gewisses Maß an Sensibilität und Anpassungsfähigkeit im Gespräch mit dem Dogmatiker. Die antidogmatischen Argumentationen und Überlegungen sollten auf den jeweiligen Gesprächspartner abgestimmt sein. Schließlich sind bestimmte skeptische Einwände nicht für alle Dogmatiker in gleicher Weise verständlich und glaubhaft. Der versierte Skeptiker ist flexibel und beherrscht die Perspektivenübernahme. Er versetzt sich in die Lage des Gesprächspartners, zeigt Verständnis und geht zusammen mit ihm ein Stück des langen Weges zur Meinungslosigkeit. Auch Beispiele, Analogieschlüsse und Rhetorik können in gewissen Kontexten als skeptische Heilmittel dienen, die den Dogmatiker von seiner Gewissheit abzubringen vermögen.[17] Ziel ist es, einen bleibenden Eindruck der Gleichwertigkeit herbeizuführen. Welches die geeigneten Mittel dazu sind, hängt vom jeweiligen Gesprächspartner ab.

Der pyrrhonische Skeptiker scheint jemand zu sein, der bezüglich *jeder beliebigen* These den Eindruck hat, dass sich die Gründe *pro* und *contra* die Waage halten. Dies legt jedoch einen Verdacht nahe: Einem eingesessenen Pyrrhoniker müssten nämlich auch die Gründe für und gegen die These, dass sich bei allen ihm bekannten Thesen die Gründe dafür und dagegen die Waage halten, gleich stark scheinen. Hätte er aber tatsächlich diesen Eindruck, dann wären seine Eindrücke inkonsistent. In jedem einzelnen konkreten Fall scheint ihm nämlich ebenso viel dafür wie dagegen zu sprechen. Alles spricht somit für die Gleichwertigkeit, für die Isosthenie. Dann kann er aber nicht den Eindruck haben, die Gründe für die Isosthenie seien

14 PH I, § 206-207; PH II, § 187-188.

15 Medizinische Metaphern sind nicht nur bei der empirischen Ärzteschule präsent, die Sextus besuchte, sondern sind ein Charakteristikum therapeutischer Philosophiekonzeptionen: So vergleicht Heraklit das blosse Wähnen (*oiēsin*) mit einer zu heilenden Krankheit (*hieran noson*) (Hermann Diels/Walther Kranz: *Die Fragmente der Vorsokratiker* [Berlin: Weidmannsche Verlagsbuchhandlung, 1951] B 46) und Sokrates versteht seine diskursiv-philosophische Praxis als maieutische Hilfeleistung zur geistigen Katharsis oder gar als Impuls zur Selbsterkenntnis der Diskurspartner (Platon: *Charmides*, 155e-157c, und *Theaitetos*, 149a-b).

16 PH III (Fn. 1), § 280.

17 Ibid.

insgesamt nicht besser als jene gegen sie. Wir müssen dem Skeptiker also unterstellen, dass sich bezüglich der Gleichwertigkeit gerade *kein* Eindruck der Gleichwertigkeit einstellt. In einem einzigen Fall also hat auch der vorbildhafte Pyrrhoniker den Eindruck eines Ungleichgewichts – nämlich dann, wenn es um die Gründe für und gegen das Gleichgewicht der Gründe geht. Es herrscht ein Ungleichgewicht zwischen Gleichgewicht und Ungleichgewicht, zugunsten des Gleichgewichts. Dieser Eindruck eines Ungleichgewichts ist jedoch nur eine subjektive Empfindung, mit der weder ein Wissens- noch ein Begründungsanspruch einhergehen. Man sollte dem Pyrrhoniker also weder einen Widerspruch unterstellen, noch ihm Inkonsequenz vorwerfen. Er behauptet ja nicht, es spräche *tatsächlich* mehr für die Gleichwertigkeit als gegen sie. Bei ihm hat sich lediglich das unbegründete Gefühl eingestellt, es spräche mehr für als gegen die Gleichwertigkeit. Sextus stellt dem *Grundriss der pyrrhonischen Skepsis* folgende, programmatische Passage voran: «[Ich] möchte […] bemerken, dass ich von keinem der Dinge, die ich sagen werde, mit Sicherheit behaupte, dass es sich in jedem Fall so verhalte, wie ich sage, sondern dass ich über jedes einzelne nur nach dem, was mir jetzt erscheint, erzählend (*historikōs*) berichte.»[18] Mit diesem Rückzug entgeht er dem Vorwurf der Selbstwidersprüchlichkeit oder Inkonsequenz seiner Haltung. Zwar sprechen neben psychologischen Tatsachen, wie gezeigt, auch theoretische Gründe gegen die Annahme, dass ein Mensch bezüglich *jeder* These den Eindruck der Gleichwertigkeit haben kann. Man sollte den Pyrrhoniker aber nicht als jemanden verstehen, der alles gleich plausibel findet. Das kann mit «Isosthenie» offensichtlich nicht gemeint sein.[19]

18 PH I, § 4.

19 Wer unter «Isosthenie» versteht, dass sich Gegenpositionen für den Pyrrhoniker jeweils gleich plausibel anfühlen, der muss erklären, wie der Pyrrhoniker angesichts seiner vollkommen ausgeglichenen doxastischen Präferenzen handeln und entscheiden kann. Eine Strategie besteht darin, alltägliche Überzeugungen aus dem Bereich skeptischer Zweifel auszuschliessen und zu sagen, die Isosthenie beziehe sich nur auf umstrittene Fragen über Nichtoffensichtliches (*adēla*). Eine Stelle bei Sextus legt dieses Verständnis besonders nahe: «Wir beziehen unsere skeptischen Schlagworte nicht schlechthin auf alle Dinge, sondern nur auf die verborgenen (*peri tōn adēlōn*) und die von den Dogmatikern untersuchten» (PH I, § 208). Diese Lesart scheint mir allerdings nachteilig gegenüber einer kontextualistischen Lesart. Vgl. Michael Williams: *Scepticism* (Dartmouth: Dartmouth Publishing Company, 1993) S. 57, 82. Siehe auch Markus Gabriel: *Antike und moderne Skepsis* (Hamburg: Junius, 2008) S. 66-77.

Der pyrrhonische Skeptiker kennt das Phänomen sehr wohl, dass ihm eine These plausibler scheint als ihre Negation. Auch er neigt zu bestimmten Meinungen und findet so manches plausibler als das Gegenteil, etwa, dass Menschen nicht auf Bäumen wachsen oder dass etwas, das man aufhebt und loslässt, in der Regel zu Boden fällt. Er glaubt jedoch nicht, dass dieser intellektuelle Druck, den er verspürt und der ihn in eine bestimmte Richtung drängt, begründet ist. Nun legt sich allerdings ein Einwand nahe: Muss der Pyrrhoniker, um dies behaupten zu können, nicht zwischen der *objektiven* Gleichwertigkeit von Gründen und seiner *subjektiven*, letztlich unbegründeten, ungleichen Gewichtung dieser Gründe unterscheiden? Hat er sich mit dem Verweis auf Objektivität aber nicht bereits von seinem Projekt verabschiedet, das darin bestand, nur von seinen momentanen Eindrücken zu berichten, wie ein Historiker des eigenen Innenlebens? Ich denke nicht. Zwar scheint mir der Einwand stichhaltig, der besagt, ein wohlwollender Interpret der pyrrhonischen Skepsis müsse – ebenso wie der Pyrrhoniker selbst – unterscheiden zwischen dem *Gefühl der Gleichwertigkeit* und der *tatsächlichen Gleichwertigkeit* der Gründe *pro* und *contra*. Man sollte dabei allerdings klären, was mit «tatsächlicher Gleichwertigkeit» gemeint ist, schließlich kann damit nicht das gemeint sein, was damit gesagt ist, denn dann müsste man dem Pyrrhoniker Behauptungen über objektive Tatsachen unterstellen. Mir scheint, das Problem lässt sich eleganter lösen: Der Pyrrhoniker muss lediglich unterscheiden zwischen dem subjektiven Gefühl der Gleichwertigkeit und dem Eindruck, den er bezüglich der tatsächlichen, objektiven Gleichwertigkeit hat.[20] Diese Lesart hat einige Vorteile: (1) Der Skeptiker berichtet nur über seine Eindrücke. (2) Er kann zu bestimmten Meinungen neigen, was er aufgrund der unhintergehbaren Parteilichkeit *de facto* auch tut, ohne deshalb den Eindruck der tatsächlichen Gleichwertigkeit aufgeben zu müssen. (3) Er kann sich an der tatsächlichen Gleichwertigkeit der Gründe orientieren, um sich dem Ideal der gefühlten Gleichwertigkeit anzunähern. «Tatsächliche Gleichwertigkeit» meint dabei nur, dass anhand der Kriterien für eine gute Begründung, auf die sich der Skeptiker und der Dogmatiker geeinigt haben (weder zirkulär noch unend-

[20] Myles Burnyeat vertritt die These, «Isosthenie» meine, dass gleich gute Gründe für wie gegen eine These sprechen, während Michael Williams für eine psychologistische Lesart von «Isosthenie» votiert, nach der sich die Gründe für eine These ebenso plausibel und glaubhaft *anfühlen* wie die Gründe dagegen (Myles Burnyeat [Hg.]: *The Skeptical Tradition* [Berkeley: University of California Press, 1983] S. 138; Williams, op. cit. [Fn. 19] S. 49).

lich noch arbiträr dogmatisch), keine Entscheidung zwischen Gegenpositionen gerechtfertigt werden kann. «Tatsächlich» oder «objektiv» heißt hier also nur, «worauf man sich geeinigt hat».

«Aber es gibt doch bessere und schlechtere Begründungen», wird der Dogmatiker einwenden. Mag sein. Aber woran bemisst sich die Validität einer Begründung? Kriterien wie Einfachheit, Sparsamkeit, Plausibilität und Konservativität sind allesamt unzuverlässig, unklar oder relativ. Einfachheit und Sparsamkeit sprechen nicht für die Wahrheit einer Theorie, sind also, abgesehen von deren Interpretationsbedürftigkeit, auch keine Gründe für die Bevorzugung der Theorie. Plausibilität und Konservativität können ebenfalls keine guten Gründe sein, eine Auffassung vor anderen zu bevorzugen, da sie davon abhängen, *für wen* und *unter welchen Umständen* etwas als plausibel oder anschlussfähig gilt.

Der Pyrrhoniker lehrt uns, dass wir die Dinge immer schon aus einer bestimmten Perspektive sehen und daher immer parteiisch urteilen. Für die eigene Perspektive können wir nur zirkulär argumentieren. Wir haben letztlich keine Gründe dafür, dass wir einige Positionen plausibler finden und bestimmte Gründe höher gewichten als andere, wir tun es einfach. Wir haben unsere Perspektive nicht rational gewählt. Sie ist, um mit Wittgenstein zu sprechen, «der überkommene Hintergrund»,[21] auf dem wir zwischen wahr und falsch unterscheiden. Auch der Skeptiker findet aufgrund seines soziohistorischen Umfeldes und seiner Erfahrungen manches plausibler als anderes. Dies sind für ihn jedoch keine Gründe, sondern allenfalls Ursachen, sich entsprechend zu verhalten.

Wie kann der Pyrrhoniker gegen einen Dogmatiker argumentieren, wenn er doch keine Behauptungen aufstellt? Ist er nicht, wie Aristoteles sagen würde, gezwungen, zu schweigen wie eine Pflanze?[22] Wenn der Pyrrhoniker im Gespräch mit dem Dogmatiker Thesen vorbringt, so erhebt er damit weder den Anspruch auf Begründbarkeit noch verleiht er eigenen, antidogmatischen Überzeugungen Ausdruck. Vielmehr bringt er in distanzierter Haltung Gegenargumente vor, von welchen er sich *erhofft*, sie würden den Opponenten zur Kenntnisnahme der Gleichwertigkeit (*isostheneia*) und schließlich zur

21 Ludwig Wittgenstein: *Über Gewissheit*, Werkausgabe in 8 Bänden, Bd. IIX (Frankfurt a.M.: Suhrkamp, 1989) § 94.

22 Aristoteles schreibt im Rahmen der Frage, wie gegen jemand argumentiert werden kann, der das Prinzip des Nicht-Widerspruchs leugnet, es sei «lächerlich, eine Begründung gegen den zu suchen, der ja für nichts eine Begründung hat, insofern er nämlich keine hat; denn ein solcher ist als solcher gleich einer Pflanze» (Met. 1006a).

Zurückhaltung (*epochē*) führen. In einem ersten Schritt versucht der Pyrrhoniker dabei nicht, Gegenpositionen stark zu machen, sondern, die gegnerische Position zu schwächen. Dies tut er, indem er unbegründete Prämissen, unhaltbare oder skeptische Konsequenzen impliziter Annahmen, Inkonsistenzen und ungültige Argumente der dogmatischen Position aufweist. Er zeigt also, dass gewisse Behauptungen selbst vor dem Hintergrund dogmatischer Voraussetzungen unbegründet, unverständlich oder gar widersprüchlich sind.[23] Gelingt ihm keine immanente Kritik, so kontrastiert er die dogmatische Auffassung mit einer konkurrierenden Position und versucht diese ebenso stark zu machen, wie die Position des Gegners. Diese konkurrierende Position ist entweder eine andere dogmatische Position oder eine skeptische Position. Für einen Pyrrhoniker ist eine erkenntnisskeptische Position dabei prinzipiell nicht mehr wert als eine dogmatische. Es kommt immer darauf an, wogegen der Skeptiker argumentiert: Wenn jemand behauptet, es gäbe Bewegung in der Welt, so führt der Pyrrhoniker Zenons Paradoxa an. Falls jemand behauptet, er wisse, dass es eine Welt außerhalb unserer Vorstellungen gibt, so führt er Argumente von Kyrenaikern oder Kynikern an, die zeigen, dass wir es nur immer mit unseren Vorstellungen zu tun haben und keinen Zugang zu den extramentalen Ursachen dieser Vorstellungen haben.[24] Michael Williams schreibt in seinem Aufsatz *Scepticism without Theory*:

[23] Vgl. Barry Stroud, *Contemporary Pyrrhonism*, in *Pyrrhonian Skepticism*, hg. von Walter Sinnott-Armstrong (New York: Oxford University Press, 2004) S. 176.

[24] Sextus Empiricus: *Gegen die Dogmatiker, Adversus mathematicos libri 7-11*, übers. von Hansueli Flückiger (Sankt Augustin: Academia Verlag) I, § 191-192: «Die Kyrenaiker sagen nun, die Widerfahrnisse (*pathē*) seien Kriterien, nur sie würden erfasst und seien untrüglich, von den Dingen aber, welche die Widerfahrnisse herbeiführen, sei keines erfassbar und untrüglich. Denn dass uns etwas Weisses widerfährt, sagen sie, und etwas Süsses, vermag man untrüglich, wahrheitsgemäss, sicher und unwiderlegbar zu sagen. Dass aber das, was das Widerfahrnis bewirkt, weiss ist oder süss ist, ist nicht möglich nachzuweisen. Denn es ist wahrscheinlich, dass jemand auch von Nichtweissem in die Verfassung, dass ihm Weisses widerfährt, gebracht wird, und dass jemandem von Nichtsüssem Süsses widerfährt»; Die Kyniker Anaxarch und Monimos hätten das Wahrheitskriterium ganz aufgehoben, «weil sie das Seiende mit einer bemalten Kulisse (*skēnographia*) verglichen und annahmen, es gleiche dem, was im Schlaf oder im Wahnsinn aufgenommen wird» (GD, 1, 88). Markus Gabriel hat m. E. recht, wenn er gegen Burnyeat und Williams behauptet, dass die Philosophen der griechischen Antike das sogenannte Aussenweltproblem sehr wohl gekannt haben, das wir gemeinhin mit dem Namen «Descartes» verbinden (Gabriel, op. cit. [Fn. 19] S. 56-58).

> So in arguing for sceptical conclusions, the Pyrrhonian is neither grounding his method nor providing an alternative shortcut route to *epochē*: rather he is *applying* his method, *extending* his *epochē* to epistemological disputes. This is perhaps hard for us to see because we are so used to thinking of scepticism as a thesis *within* epistemology, whereas, for Sextus, all epistemological positions, including theoretical scepticism, are just further things to be sceptical about.[25]

Oft verwendet der pyrrhonische Skeptiker Phrasen wie «alles ist unbestimmt», «alles ist unerkennbar» und «jedem Argument steht ein gleichwertiges entgegen (*pantō logō logos isos antikeitai*)». Mit diesen skeptischen Floskeln (*phōnai*)[26] erhebt der Pyrrhoniker aber keine Wahrheitsansprüche, vielmehr sind auch sie als Äußerungen eines gegenwärtigen und subjektiven Eindrucks oder als Resümees der skeptischen Diskurspraxis zu verstehen. Ihnen müsste korrekterweise jeweils ein «es scheint mir gegenwärtig so zu sein, dass (*phainetai moi nyn*)» vorangestellt werden. Solche Phainetai-Aussagen bringen lediglich zum Ausdruck, welchen Eindruck der Sprecher gegenwärtig bezüglich eines bestimmten Sachverhalts hat. Durch diesen permanenten Rückzug auf dasjenige, was ihm gegenwärtig plausibel zu sein scheint, grenzt sich der pyrrhonische Skeptiker von den Skeptikern der Platonischen Akademie ab, den sogenannten «akademischen Skeptikern». Diese vertreten eine *agnostische* Position und firmieren die These, nichts ließe sich erkennen – außer eben der Tatsache, dass sich nichts erkennen lässt.[27] Eine solche Position ist aufgrund ihrer Inkonsequenz letztlich unhaltbar. Denn wer begründen will, weshalb sich nichts erkennen lässt, der muss dabei von Annahmen ausgehen, die er auch erkannt zu haben glaubt. Damit widerspricht er jedoch der Behauptung, er habe nur *etwas* erkannt, nämlich, dass nichts erkennbar sei (außer gerade dieser Erkenntnis). Der pyrrhonische Skeptiker dagegen würde, wie sich gezeigt hat, weder behaupten, nichts sei erkennbar, noch würde er das Gegenteil behaupten. Alles, was er als Praktiker und Therapeut möchte, ist, in seinem Gesprächspartner den Eindruck erzeugen, dass ebenso wenig dafür wie dagegen spricht, dass wir nichts erkennen können.

Wird der Pyrrhoniker aufgefordert, Gründe anzuführen für seine Präferenz der These (G) «Es spricht ebenso wenig für wie gegen *p*», dann kann er entweder (1) auf die scheinbaren *Ursachen* seines Bewusstseinszustandes verweisen oder (2) weitere Thesen anführen, zu denen er neigt und die für (G) zu sprechen scheinen. Die erste Alternative käme einem Begründungs-

25 Williams, op. cit. (Fn. 19) S. 74.
26 PH I (Fn. 1), § 187-188.
27 PH I, § 226.

abbruch gleich, da Ursachen keine Gründe sind. Er würde auf sein autobiographisches Umfeld rekurrieren oder etwas über seinen Charakter und seine Interessen erzählen. Kurz, er würde uns eine Genealogie liefern, aus der klar wird, wie es dazu kam, dass er momentan (G) plausibel findet. Die zweite Alternative würde darin bestehen, das zu tun, was Sextus in seinen Schriften tut, also etwas, das dem ähnelt, was der Dogmatiker macht, wenn er seine Position zu begründen versucht: Der Pyrrhoniker würde etliche Gründe für und gegen *p* anführen, er würde zeigen, dass alle Begründungsversuche am Begründungstrilemma scheitern, dass jede Bewertung der Gründe voreingenommen ist und nichts unbezweifelbar feststeht. Klar gerät dabei auch er in einen infiniten Regress oder einen Begründungszirkel, falls der Dogmatiker nachhakt. Das bringt den Pyrrhoniker im Unterschied zum Dogmatiker jedoch nicht in Verlegenheit, da er weder Begründungs- noch Wissensansprüche erhebt, sondern lediglich berichtet, was ihm momentan plausibel zu sein scheint.

Kann der Skeptiker, von seinen Erlebnissen und Eindrücken abgesehen, wirklich *alles* in Zweifel ziehen? Verwickelt er sich dabei nicht in semantische oder pragmatische Widersprüche? Fest steht, dass er nicht alles *zugleich* bezweifeln kann. Wenn ich etwas Bestimmtes bezweifle, etwa, dass die Welt so ist, wie ich sie sehe, dann gehe ich von bestimmten unhinterfragten Annahmen aus, etwa der Annahme, dass ich gerade keinen Drogenrausch durchlebe und dass die Überlegungen, die ich anstelle, nicht völlig wirr und unschlüssig sind. Zu einem anderen Zeitpunkt kann ich jedoch sehr wohl daran zweifeln, dass mein momentanes Denken normal funktioniert. Angenommen, mein Freund und ich befänden uns momentan in einem Drogenrausch, wüssten es aber nicht. Vielmehr schiene es uns so, als würden wir uns gerade über den Skeptizismus unterhalten und ich würde den Skeptiker verteidigen. Meine Gedanken, die ich äußere, wären zwar ohne jeglichen Zusammenhang, trotzdem schiene es uns so, als ob ich gültige Argumente vorbringen und mein Gegenüber mich verstehen würde. Was ich in einer solchen Situation tatsächlich äußere, würde ein Außenstehender zwar nicht als «Zweifel» bezeichnen. Dies ist jedoch kein Argument gegen den globalen Skeptizismus. Für den radikalen Skeptiker reicht es, dass es *so scheint*, *als ob* er gerade verständliche skeptische Überlegungen anstellen würde. Diese Erklärung kann er auch jemandem geben, der behauptet, man könne nicht bezweifeln, dass die Sätze, die man gerade äußere, Bedeutung haben. Denn, so die Argumentation des Dogmatikers, hätten diese Sätze keine Bedeutung, so würde damit nichts gesagt und also auch nichts bezweifelt. Zweifle man aber *wirklich*, dann würde mit den Sätzen zwar auch etwas gesagt, sie hätten

also eine bestimmte Bedeutung; damit wäre aber auch bestätigt, was der Skeptiker bezweifeln möchte. Das vermeintliche Dilemma ist allerdings kein wirkliches Dilemma, schließlich kann sich der Skeptiker damit begnügen, dass es den *Anschein* hat, als würden seine Sätze etwas bedeuten.[28] Letzteres scheint der Dogmatiker nicht ernsthaft leugnen zu wollen.

1.2 Orientierungsgrundlage

Wie kann ein Pyrrhoniker überhaupt handeln? Woran orientiert er sich, wenn aus seiner Sicht keine bestimmte Handlungsoption gegenüber anderen gerechtfertigt werden kann? Die Frage ist berechtigt, da für den Skeptiker nicht nur alle theoretischen Überlegungen, sondern auch alle praktischen Entscheidungen, alle Lebensformen und sämtliche Handlungen unter rationalen Gesichtspunkten gleichwertig und daher beliebig zu sein scheinen. Mit diesem *Apraxie-Einwand* sahen sich die Pyrrhoniker seit jeher konfrontiert. Die Antwort von Sextus ist verblüffend einfach: Die Orientierungsgrundlagen des Pyrrhonikers sind seine natürlichen Bedürfnisse, die Gewohnheiten (*ethē*), seine Lebensform (*agogē*), seine Erlebnisse (*pathē*) und insbesondere das, was ihm erscheint (*phainomena*).[29] Er tut, was er nicht lassen kann und was er automatisiert hat. Der Pyrrhoniker verlässt sich auf seine Gewohnheiten und orientiert sich an dem, was ihm momentan plausibel und nützlich zu sein scheint, obwohl er gleichzeitig sieht, dass seine derzeitigen Vermutungen und Präferenzen wahrscheinlich das Produkt kontingenter Umstände sind. Insofern gleicht der pyrrhonische Skeptiker der zu einer historistischen und nominalistischen Position neigenden «Ironikerin», wie der Neopragmatist Richard Rorty sie in *Kontingenz, Ironie und Solidarität* charakterisiert:

> «Ironikerin» werde ich eine Person nennen, die drei Bedingungen erfüllt: (1) sie hegt radikale und unaufhörliche Zweifel an dem abschließenden Vokabular,[30] das

[28] Sextus stellt sich diesem Einwand und entschärft ihn durch die Unterscheidung zwischen anzeigenden (*epideiktikon*) und kommemorativen (*hypomnēstikon*) Zeichen. Er behauptet, der Pyrrhoniker beziehe sich mit seinen Worten nicht auf Verborgenes, sondern auf seine eigenen Eindrücke, deren Existenz unhinterfragt sei (PH I, § 130); vgl. die ausführlichere Diskussion in Sextus Empiricus: *Against the Logicians* (PL) (Cambridge, MA: Harvard University Press) II, § 281-297.

[29] PH I, § 23-24.

[30] Ein abschliessendes (*final*) Vokabular zeichnet sich dadurch aus, dass «dem Nutzer keine Zuflucht zu nicht-zirkulären Argumenten mehr bleibt, wenn der

> sie gerade benutzt, weil sie schon durch andere Vokabulare beeindruckt war, die Menschen oder Bücher, denen sie begegnet war, für endgültig nahmen; (2) sie erkennt, dass Argumente in ihrem augenblicklichen Vokabular diese Zweifel weder bestätigen noch ausräumen können; (3) wenn sie philosophische Überlegungen zu ihrer Lage anstellt, meint sie nicht, ihr Vokabular sei der Realität näher als andere oder habe Kontakt zu einer Macht außerhalb ihrer selbst.[31]

Eine ironische Person lebt «immer im Bewusstsein der Kontingenz und Hinfälligkeit ihrer abschließenden Vokabulare»[32] und «meint, nichts habe eine immanente Natur oder reale Essenz».[33] Dieser ironische 'Lifestyle' gleicht der Lebensführung (*agogē*) des Pyrrhonikers insofern, als sich beide relativistisches Gedankengut zu eigen machen und ihr Selbstbild wesentlich dadurch bestimmt ist, dass sie ihre vertrautesten Ansichten als nur zirkulär zu rechtfertigende und in rationaler Hinsicht gleichwertige Positionen eines unauflösbar scheinenden Dissenses begreifen.[34]

Ein pyrrhonischer Skeptiker handelt nach Sextus im Alltag in derselben Weise wie seine nichtskeptischen Mitmenschen. Selbst religiöses Verhalten ist ihm nicht fremd. Er spielt das gesellschaftlich vertraute Spiel mit, da abweichendes Verhalten Aufsehen erregen würde und ihn seine Mitmenschen anschließend zur Rechtfertigung aufforderten. Das Mitläufertum des Pyrrhonikers ist nicht in allen Situationen ein probates Mittel zum Zweck.

Wert seiner Wörter angezweifelt wird» (Richard Rorty: *Kontingenz, Ironie und Solidarität* [Frankfurt a.M.: Suhrkamp, 1991] S. 127).

31 Ibid. S. 127.

32 Ibid. S. 128.

33 Ibid. S. 129.

34 In einem Gespräch mit Helmut Mayer und Wolfgang Ullrich antwortet Rorty auf die Frage, inwieweit der pyrrhonische Skeptiker mit der liberalen Ironikerin verglichen werden kann: «Der Unterschied zwischen den Alten und uns besteht darin, dass wir, abgesehen von denen, die sich zum Zen-Buddhismus hingezogen fühlen, nicht auf Ruhe und Ataraxie aus sind. Uns geht es um Selbsterschaffung. Die Alten waren der Meinung, die Männer und Frauen der Zukunft würden sich von denen, die sie kannten, nicht wesentlich unterscheiden. Wir hingegen hoffen, dass sich unsere fernen Nachkommen sehr stark von uns unterscheiden. Hans Blumenberg hat meiner Meinung nach vollkommen Recht, wenn er sagt, dass wir Modernen eine Haltung zur Zukunft einnehmen, die in früheren Zeiten unmöglich gewesen wäre. Man kann sich die Pyrrhonischen Skeptiker als Proto-Pragmatisten denken: Philosophen, die erklärten, alles müsse als ein Mittel zum Glück der Menschen betrachtet werden. […] Allerdings haben wir Visionen von möglichen Formen menschlichen Glücks, die sie nicht hatten» (Richard Rorty: *Philosophie und die Zukunft. Essays* [Frankfurt a.M.: Fischer Taschenbuch Verlag, 2000] S. 180).

Ist das Verhalten der Menge mit seiner persönlichen Haltung, seinen Präferenzen und seinen doxastischen Neigungen nicht verträglich, wird er seinen eigenen Weg gehen. Zu einer Rechtfertigung seiner Handlungen sieht er sich allerdings weder als Mitläufer noch als Abweichler in der Lage. Er wird bestenfalls eine kausale Erklärung seines Verhaltens geben, d.h. er rekurriert auf seine doxastischen Präferenzen und auf seine Bedürfnisse, die er als psychologische *Ursachen* seines Tuns betrachtet. Er gibt somit keine Rechtfertigung seiner Handlungen und Äußerungen, sondern liefert eine narrative Genealogie derselben: er erzählt uns eine aus seiner Sicht plausible Geschichte, wie es dazu kam, dass er eine bestimmte Handlung ausführte, welche Präferenz ihn dabei leitete und warum ihm zum jeweiligen Zeitpunkt bestimmte Thesen und Argumente plausibler schienen als andere.

2. Rationalität

Meines Erachtens lassen sich in dem bunten Ensemble verschiedener Verwendungsweisen des Ausdrucks «rational» zwei Verständnisse von Rationalität ausmachen, die gleichsam die entgegengesetzten Pole eines Kontinuums bilden: einerseits (1) ein weites, eher formales und weitgehend voraussetzungsloses Verständnis und andererseits (2) ein engeres, gehaltvolleres und voraussetzungsreiches Verständnis von Rationalität. Für die erste Auffassung soll der Begriff *Rationalität*$_1$ stehen, für die zweite der Begriff *Rationalität*$_2$.

2.1 Rationalität$_1$ *als Verstehbarkeit*

Rational$_1$ ist, wer sich um (1) *Konsistenz* und (2) *Kohärenz* bemüht.

(1) Derjenige, dessen Überzeugungssystem vollständig *konsistent* ist, hat keine in sich widersprüchlichen Überzeugungen, noch bestehen zwischen seinen Überzeugungen Widersprüche. Auch bestehen zwischen den logischen und materialen Konsequenzen unterschiedlicher Überzeugungen keine Widersprüche. Nun bleiben uns Menschen aber nicht selten Widersprüche zwischen den begriffslogischen Konsequenzen unserer Überzeugungen verborgen. Die Sokratischen Dialoge Platons führen uns dieses Phänomen deutlich vor Augen. Sokrates verwickelt seine Gesprächspartner in Widersprüche, indem er ihnen bewusst macht, dass die Konsequenzen ihrer Überzeugungen unhaltbar sind oder sich gar widersprechen. Den meisten von uns würde es mit Sicherheit auch so gehen. Wir sind deswegen jedoch nicht irrational. Um die Konsis-

tenzanforderung zu erfüllen, reicht es also, sich in der Regel erfolgreich um Konsistenz zu *bemühen*. Dasselbe gilt für die Kohärenz.

(2) Mit «Kohärenz» meine ich an dieser Stelle die Stimmigkeit und den Zusammenhalt zwischen Überzeugungen, Interessen und Verhalten. Das Verhalten einer rationalen Person soll *grosso modo* Ausdruck ihrer Präferenzen und Überzeugungen sein. Interessen und Überzeugungen sollen ihrerseits aufeinander abgestimmt sein. Zwei Wünsche können zwar konfligieren, eine rationale Person sollte aber in der Lage sein, diese gegeneinander abzuwägen und sich gegebenenfalls von einem der konfligierenden Wünsche zu distanzieren.

Mit der Kohärenzforderung kommen auch die für die Rationalität wichtigen *Gründe* ins Spiel. Rational$_1$ sein heißt, in der Regel etwas Bestimmtes deswegen glauben, wünschen oder tun, *weil* man Gründe dafür zu haben *glaubt*. Eine Handlung kann für uns als Betrachter also unverständlich und unüberlegt scheinen und trotzdem rational$_1$ sein; nämlich dann, wenn eine (verständliche) Überlegung dahintersteckt. Ein Verhalten ist auch dann rational$_1$, wenn die vorangegangene Überlegung aus unserer Sicht auf falschen Annahmen beruht. Sobald wir nämlich wissen, was die vermeintlich unüberlegt handelnde Person glaubt und wünscht, wird ihr Verhalten *verständlich*, d.h., wir können *verstehen*, weshalb sie so handelt.

Neben *intra*individueller oder interner Kohärenz kann auch eine *inter*individuelle oder externe Kohärenz ein Indiz für die Rationalität$_1$ einer Person sein: gemeint ist die ungefähre Übereinstimmung mit der jeweiligen Sprachgemeinschaft und das Bemühen, zwischen unterschiedlichen Theorien oder Überzeugungs-Bereichen ein zusammenhängendes, widerspruchsfreies Überzeugungsgeflecht herzustellen. Wer unbegründet und signifikant vom jeweils herrschenden Konsens abweicht, steht ebenso unter dem Verdacht der Irrationalität$_1$ wie jemand, den es nicht stört, dass er in einem Bereich p, in einem anderen jedoch $\neg p$ glaubt.

Es bleibt noch eine Ergänzung: Bisher wurde implizit davon ausgegangen, dass eine rationale$_1$ Person sich darum bemüht, *zu einem bestimmten Zeitpunkt* konsistent und kohärent zu sein. Wie steht es aber mit transtemporaler Konsistenz und Kohärenz, im Unterschied zu simultaner? Klar scheint zu sein, dass eine Person nicht rational$_1$ ist, wenn sich ihre Überzeugungen und Präferenzen von Tag zu Tag oder gar von Stunde zu Stunde umfassend verändern. Ein gewisses Maß an transtemporaler Beständigkeit scheint also eine weitere notwendige Bedingung der Rationalität$_1$ zu sein. Wie umfassend und wie schnell sich bei rationalen Personen das Überzeugungssystem ändern darf (und unter welchen Umständen dies der Fall sein darf), ist damit allerdings noch nicht geklärt.

2.2 Rationalität$_2$ als Vernünftigkeit

Kriterium für die Rationalität$_2$ im Sinne des engeren und daher voraussetzungsreicheren Rationalitätsbegriffs ist der Grad der *Ähnlichkeit* mit unserer Kultur, die *Vertrautheit* bestimmter Handlungen, Wünsche und Überlegungen. Wenn zwischen zwei Lebensformen, mitsamt deren Welt- und Selbstbildern, nur wenige Gemeinsamkeiten feststellbar sind und die Anzahl geteilter Hintergrundannahmen vergleichsweise gering ist, so neigt man dazu, das abweichende Verhalten der jeweils anderen Kultur als «irrationales Verhalten» zu bezeichnen. Wenn für bestimmte Überzeugungen einer fremden Kultur aus unserer Sicht 'so gut wie nichts spricht', dann sind wir schnell bereit, ihren Überlegungen den Stempel der Irrationalität aufzudrücken. Selbst wenn man erfährt, welche Gründe die Handelnden für das uns fremd anmutende Verhalten anführen, versteht man nicht, wie man *solche* abwegigen und aus unserer Sicht falschen Überzeugungen haben und derart skurrile Lebensformen pflegen kann. Was sie als Grund für ein bestimmtes Verhalten anführen, ist für uns kein wirklicher Grund, entweder weil wir anderer Meinung sind oder weil wir nicht sehen, wie das als Grund Angeführte mit dem zu begründenden Verhalten zusammenhängen soll. Angenommen, die Überlegungen, die eine Person anstellt, basieren nicht nur auf Prämissen, die man für falsch hält, sondern selbst die Art und Weise, *wie* sie schließt, bleibt unverständlich und unnachvollziehbar, so ist man geneigt, ihr nicht nur die Rationalität$_2$ im Sinne der Vernünftigkeit, sondern auch die Rationalität$_1$ im Sinne der Verstehbarkeit abzusprechen. Hier legt sich aber der Verdacht nahe, dass dasjenige, was uns als verständliches Schließen und Begründen gilt, von unseren sonstigen Überzeugungen über die Welt und unseren Interessen nicht gänzlich unabhängig ist. Unser Weltbild gehört, wie Wittgenstein in *Über Gewissheit* schreibt, «zum Wesen dessen, was wir ein Argument nennen. Das System ist nicht so sehr der Ausgangspunkt als das Lebenselement der Argumente.»[35] Die Grenze zwischen Rationalität$_1$ und Rationalität$_2$ ist vermutlich – wie die meisten begrifflichen Abgrenzungen – nicht so eindeutig zu ziehen, wie man hätte denken können.

Die beiden Konzeptionen von Rationalität hängen eng mit der Frage zusammen, was Gründe sind und was Gründe zu guten Gründen macht. Zugespitzt gesagt, heißt rational$_1$ sein nämlich, sich im Denken und Handeln an etwas auszurichten, von dem man *glaubt*, es seien gute Gründe – egal, ob es wirklich gute Gründe sind. Rational$_2$ sein dagegen meint, dass man

35 Wittgenstein, op. cit. (Fn. 21) § 105.

sich im Denken und Handeln *wirklich* an guten Gründen orientiert. Der pyrrhonische Skeptiker hat, da er schlicht *keine* Auffassung vertritt, weder eine eigene Vorstellung davon, was Rationalität ist, noch davon, was gute Gründe sind. Wenn jemand behauptet, pyrrhonische Skeptiker seien irrational, da sie sich nicht an Gründen orientieren, so wird der Pyrrhoniker, ganz im Sinne des maieutisch verfahrenden Sokrates,[36] ausgehend von den Annahmen seines Gesprächspartners gegen dessen Position argumentieren, ohne dabei eigene Thesen vorzubringen. Er wird ihm die Schwächen seines Rationalitätsbegriffs vor Augen führen, seine Konzeption guter Gründe kritisieren und versuchen, ihm andere Auffassungen über Rationalität und Rechtfertigung schmackhaft zu machen.[37] Dem Pyrrhoniker liegt es dabei nicht am Herzen, zu zeigen, dass er rational denkt und handelt. Er möchte nur zeigen, dass ebenso wenig für wie gegen seine Rationalität zu sprechen scheint.

3. Vier Irrationalitätsvorwürfe

Nachdem wir erläutert haben, was ein pyrrhonischer Skeptiker ist und was mit «Rationalität» gemeint sein könnte, sollten wir uns der Titelfrage des Aufsatzes zuwenden: Sind pyrrhonische Skeptiker irrational?

Im Folgenden sollen fünf Irrationalitätsvorwürfe benannt und diskutiert werden:[38]

36 Vgl. Platon: *Charmides*, 155e-157c, und *Theaitetos*, 149a ff.

37 Der Pyrrhoniker wird fragen, ob gute Gründe zu haben heisst, seine epistemischen Pflichten zu erfüllen, oder ob damit nicht vielmehr gemeint ist, dass die Überzeugung durch wahrheitszuträgliche (verlässliche) Methoden gewonnen und gerechtfertigt wurde. Er wird die epistemologische Position des Dogmatikers ins Wanken zu bringen versuchen, indem er durch Beispiele an Intuitionen appelliert, die gegen dessen Konzeption guter Gründe sprechen. Vielleicht wird er ihn fragen, ob die falschen, aber sorgfältig geprüften Meinungen eines Träumenden oder eines kognitiv Zurückgebliebenen gerechtfertigt sind. Ziel ist es, einen *clash of intuitions* zu erzeugen.

38 Die drei ersten der folgenden vier Vorwürfe sind in der einen oder anderen Form im Laufe der Philosophiegeschichte von unterschiedlichen Seiten gegen den pyrrhonischen Skeptiker gerichtet worden. Meine Fragestellung soll an dieser Stelle jedoch eine systematische sein, ich möchte also nicht rekonstruieren, was dem Pyrrhoniker von unterschiedlicher Seite *de facto* unterstellt wurde, sondern, welche Irrationalitätsvorwürfe mit guten Gründen an ihn gerichtet werden *können*.

3.1 Der Inkonsistenzvorwurf

Der Einwand, den ich «Inkonsistenzvorwurf» nennen möchte, unterstellt dem pyrrhonischen Skeptiker, er rede und handle, wie es ihm gerade passt, mitunter *widerspreche* er sich sogar. Er behaupte etwa, alles sei relativ, nehme diese Aussage aber sogleich wieder zurück, da er ja zugebe, dass sich das damit Ausgesagte selbst relativiere. Oder: Er behaupte mit Sicherheit, alles sei unsicher, wobei er sich gleich danach bereit zeige einzuräumen, dass er die Aussage «Alles ist unsicher» selbst für unsicher hält. Auch verteidige er mal diese, mal jene Position, je nachdem, mit wem er es gerade zu tun hat. Er orientiere sich also in keiner Weise an normativen Vorgaben wie Widerspruchsfreiheit und Stimmigkeit. Konsistenz sei allerdings – so der Einwand – eine notwendige Bedingung rationaler Überlegungen. Wer sich nicht um Konsistenz bemühe, sondern vergnügt mal das eine, mal das andere behaupte, der untergrabe damit die Möglichkeitsbedingungen eines rationalen Diskurses.

Zur Verteidigung des Skeptikers muss viererlei gesagt werden:

1. Wenn der Skeptiker aus widersprüchlichen Positionen heraus argumentiert, heißt das nicht, dass er widersprüchliche Überzeugungen hat. Im Gegenteil, er ist von *keiner* Position überzeugt; sie sind lediglich Werkzeuge, um – bei sich selbst, ebenso wie bei seinem jeweiligen Opponenten – den Eindruck eines Gleichgewichts der Argumente (*isostheneia*) herbeizuführen.

2. Der Skeptiker wird es natürlich vermeiden, inkonsistent zu reden, wenn er seine an Konsistenz orientierten Gesprächspartner von einer bestimmten Position abbringen möchte.

3. Der Pyrrhoniker hat theoretische und pragmatische *Gründe* für seine Inkonsistenzen: Die jeweils geltende Logik oder Syllogistik – mitsamt den ihr zugrunde liegenden, unbewiesenen Axiomen – scheint ihm nur *eine* mangelhafte Theorie unter vielen zu sein.[39] Er selbst sieht sich in erster Linie als Praktiker, dem eine Theorie für die eigene Praxis dienlich sein soll. Wenn ihm das Bemühen um Widerspruchsfreiheit zu anstrengend werden sollte, verzichtet er aus pragmatischen Gründen auf Konsistenz.

4. Da Kohärenz und Konsistenz für den Pyrrhoniker – also für jemanden, der die Theorie der Praxis unterordnet – keine unbedingte Geltung beanspruchenden *Normen*, sondern eher psychisch und sozial bedingte *Normalitäten* sind, kann es sein, dass bei ihm bezüglich dessen, was ihm jeweils plausibel scheint, tatsächlich stärkere Schwankungen auftreten als bei Nichtskeptikern.

[39] PH I (Fn. 1), § 157-159.

Aus dieser vergleichsweise stärker ausgeprägten Unbeständigkeit auf die Irrationalität$_1$ des Pyrrhonikers zu schließen, wäre allerdings verfehlt. Er ist nämlich in der Lage, rückblickend eine auf kontingente Ursachen rekurrierende kohärente Geschichte seiner Plausibilitätsschwankungen zu erzählen. Er kann uns also eine verständliche Geschichte darüber erzählen, wie es zu seinen Plausibilitätsschwankungen gekommen ist und welche Überlegungen dabei eine Rolle gespielt haben.

3.2 Der Vorwurf grundloser doxastischer Präferenzen

Der Pyrrhoniker findet manches plausibler als anderes. Er hat zwar keine wirklichen Meinungen (*doxai*), neigt jedoch zu bestimmten Auffassungen mehr als zu anderen. So hat er etwa den Eindruck, dass jeweils gleich schlechte Gründe für wie gegen eine Position sprechen. Auch findet er das Begründungstrilemma plausibler als Letztbegründungsversuche. Er glaubt jedoch nicht, dass gute Gründe für seine intellektuellen Präferenzen sprächen. Er neigt also zu bestimmten Überzeugungen, ohne zu glauben, dafür gute Gründe zu haben. Die doxastischen Präferenzen des Pyrrhonikers richten sich nicht an Gründen aus. Denn diese scheinen ihm schließlich gleichwertig zu sein. Der Pyrrhoniker hat allerdings das unbegründete subjektive Gefühl, die Gründe, die gegen Dogmatiker sprechen, seien höher zu gewichten. Ist jemand aber nicht irrational, wenn er nicht glaubt, für das, was ihm plausibel scheint, gute Gründe geben zu können? Kann man jemanden irrational nennen, der seinem Gefühl folgt, weil dieses, im Unterschied zur Vernunft, ihm erlaubt, sich zu entscheiden? Wäre es für Buridans Esel, der sich in der Mitte zwischen zwei gleich großen Heuhaufen befindet, paradoxerweise nicht vernünftiger gewesen, sich *gegen* die Vernunft und für einen der beiden Heuhaufen zu entscheiden?

Es scheint noch eine weitere Verteidigungsstrategie möglich zu sein: Für die Rationalität des Pyrrhonikers spricht, dass er Gründe geben kann, die dafür sprechen, dass niemand von uns jemals gute Gründe vorbringen kann. Sollte man Nicht-Skeptiker «rational» nennen, weil sie die Gründe des Skeptikers ignorieren, die nahelegen, dass das, was sie für gute Gründe halten, keine guten Gründe sind? Angenommen, alle Menschen würden aufgrund skeptischer Überlegungen zu der Überzeugung gelangen, sie hätten keine guten Gründe für ihre Meinungen, würden diese aber trotzdem aufrechterhalten. Wären dann alle Menschen irrational? Wohl kaum. Man würde geneigt sein, entweder den Ausdruck «gute Gründe» umzudefinieren

oder unter «Rationalität» etwas anderes zu verstehen, um weiterhin die gewohnten Züge im Sprachspiel machen zu können, d.h. psychisch Kranke, Irre, Wunschdenker und Willensschwache als «irrational» zu bezeichnen.

Das stärkste 'Argument' für die Rationalität des Pyrrhonikers scheint an dieser Stelle der Verweis auf unsere Intuition zu sein, dass ein derart scharfsinniger, reflekierter und toleranter[40] Diskussionspartner, wie es der Pyrrhoniker ist, unmöglich irrational sein kann.

3.3 Der Apraxievorwurf

Der Pyrrhoniker kann, so der nächste Einwand, nicht nur der Inkonsistenz, sondern auch der Inkohärenz bezichtigt werden: Es herrsche bei ihm eine Unstimmigkeit zwischen Theorie und Praxis. Seine Praxis stimme nämlich nicht mit seiner angeblichen Meinungslosigkeit überein. Die Inkohärenz besteht gemäß diesem Einwand darin, dass der Pyrrhoniker so handelt, als habe er feste Überzeugungen, gleichzeitig aber behauptet, er würde nichts für wahr halten und habe keine festen Überzeugungen. Obwohl er sich genauso verhält wie seine Mitmenschen, behauptet er, dabei die Meinungen seiner Mitbürger nicht zu teilen, da er schließlich ohne jegliche Meinungen (*adoxastōs*) lebe. Der Apraxie-Vorwurf besagt nun, entweder habe der Pyrrhoniker Meinungen, dann könne er zwar auch handeln wie andere Menschen, sei aber kein meinungsloser pyrrhonischer Skeptiker mehr, oder, er habe keine Meinungen, dann sei er zwar ein echter Skeptiker, könne aber

40 Nach Rorty sollten wir Rationalität eng mit Toleranz verknüpfen, «also mit der Fähigkeit, sich durch Abweichungen vom eigenen Massstab nicht allzu sehr aus der Fassung bringen zu lassen und auf solche Unterschiede nicht aggressiv zu reagieren. Diese Fähigkeit geht einher mit der Bereitschaft, die eigenen Gewohnheiten zu ändern und nicht nur in höherem Masse bekommen zu wollen, was man schon vorher zu haben wünschte, sondern sich selbst umzugestalten und zu einer anderen Person zu werden, die auch andere Dinge erreichen möchte als bisher. Sie geht ebenfalls einher mit der Neigung, sich lieber auf Überredung als Gewalt zu verlassen und lieber zu reden als mit anderen zu kämpfen, sie zu verbrennen oder zu verbannen. Diese Fähigkeit ist eine Tugend, die es Individuen und Gemeinschaften ermöglicht, in friedlicher Koexistenz mit anderen Individuen und Gemeinschaften zusammenzuleben – zu leben und leben zu lassen – und neue, synkretistisch und durch Kompromisse gestaltete Lebensformen zusammenzubasteln» (Richard Rorty: *Rationalität und kulturelle Verschiedenheit*, in *Wahrheit und Fortschritt* [Frankfurt a.M.: Suhrkamp, 2003] S. 269-270).

weder handeln noch sich entscheiden – also könne er sich auch nicht entscheiden, *nicht* zu handeln oder sich *nicht* zu entscheiden.

Der Skeptiker wird darauf antworten, er richte sich im Handeln an den eingespielten und internalisierten Gepflogenheiten seiner Lebensgemeinschaft aus und orientiere sich an den Phänomenen (*phainomena*), also an seinen Sinneseindrücken, seinen Empfindungen und an dem, was ihm derzeit plausibel scheint. Um so zu handeln wie seine Mitmenschen, brauche er keine Meinungen darüber zu haben, wie die Wirklichkeit beschaffen ist. Er brauche einfach seinen natürlichen Neigungen, Gewohnheiten und Vermutungen Folge zu leisten, ohne dabei den Meinungen zuzustimmen, aufgrund welcher seine Mitmenschen das tun, was auch er tut. Den Phänomenen (*phainomena*) ebenso wie seinen Vorstellungen (*phantasma*), Neigungen und Affekten (*pathē*) stimmt er weder aktiv zu noch kann er ihr Vorhandensein leugnen, da sie allesamt unter die Kategorie des Erleidens (*paschein*) fallen. Im Gegensatz dazu inhäriert jeder aktiven Stellungnahme ein libertäres Moment. Der Pyrrhoniker stimmt also nur zu, wo keine Zurückhaltung möglich ist. Da in diesen Fällen ein deliberativer Entscheidungsprozess psychisch unmöglich ist, kann man dem Skeptiker nicht vorwerfen, er hätte sich zurückhalten sollen. «Sollen» scheint, zumindest in vorliegendem Fall, «können» zu implizieren. Dass der Skeptiker zugibt, sein Urteil nicht immer und überall zurückhalten zu können, ist nichts weiter als die Anerkennung der *conditio humana*.

Mit Blick auf dasjenige, was «zwangsweise» und unfreiwillig (*aboulētōs*) zur Zustimmung führt, *kann* sich der Skeptiker gar nicht zurückhalten.[41] Der Skeptiker behauptet damit nichts, sondern gibt lediglich Kunde von seinen Eindrücken. Meist sind dies Urteile über die eigenen Befindlichkeiten, also Urteile über das, was in der modernen Philosophie des Geistes als qualitative und kognitive Bewusstseinszustände bezeichnet wird. So scheint es unmöglich, Schmerzen zu haben, ohne zu meinen, Schmerzen zu haben. Unhinterfragt (*azētētos*) ist nach Sextus auch die Annahme, jemand könne sich darüber täuschen, was ihm gerade der Fall zu sein *scheine*: «Deshalb wird niemand vielleicht zweifeln, ob der zugrunde liegende Gegenstand so oder so

41 David Hume schreibt: «Die Natur nötigt uns mit absoluter und unabwendbarer Notwendigkeit, Urteile zu fällen, ebenso wie sie uns nötigt zu atmen und zu empfinden» (*Ein Traktat über die menschliche Natur*. Bd. 1 [Hamburg: Felix Meiner, 1978] S. 245; «Die Natur hat uns eben in dieser Hinsicht keine Wahl gelassen» [250]). Hume glaubt fälschlicherweise, damit die Praktikabilität des pyrrhonischen Skeptizismus widerlegt zu haben.

erscheint. Ob er dagegen so ist, wie er erscheint, wird infrage gestellt».[42] Sextus beschränkt die Phänomene (*phainomena*) nicht nur auf Sinneseindrücke und Empfindungen, sondern rechnet auch kognitive Bewusstseinsinhalte wie Vermutungen und doxastische Präferenzen unter diese Kategorie:[43] In diesem Sinne *scheint* es dem pyrrhonischen Skeptiker etwa auch, dass Argumente und Gegenargumente jeweils gleich stark seien.

Man könnte sagen, dass der Pyrrhoniker insofern ohne feste Meinungen (*adoxastōs*) durchs Leben kommt, als er lediglich Vermutungen anstellt (*eudokein*). Vor Äußerungen wie «Es scheint mir gegenwärtig so zu sein, dass *p*» schreckt er also nicht zurück, da aus deren Wahrheit weder folgt, (1) er hält es für wahr, dass *p*, noch (2) dass *p*. Sextus schreibt an einer einschlägigen Stelle:

> Wenn wir sagen, der Skeptiker dogmatisiere nicht, dann meinen wir nicht jene Bedeutung von Dogma, in der einige «Dogma» ganz allgemein die Billigung (*to eudokein*) irgendeiner Sache nennen (denn den als Eindruck [*kata phantasian*] aufgezwungenen Erlebnissen [*katēnankasmenois pathesi*][44] stimmt der Skeptiker zu [*synkatatithetai*]).[45]

Auch das Verb «glauben» (*peithesthai*) habe, so Sextus, zwei Bedeutungen:

> einmal das Nichtwiderstreben (*mē antiteinein*), sondern einfache Folgeleisten ohne starke Neigung und Teilnahme, so wie man sagt, dass der Knabe dem Er-

42 PH I (Fn. 1), § 22.

43 Vgl. Andreas Graeser: *Hauptwerke der Philosophie. Antike* (Stuttgart: Reclam, 1992) S. 209: «Nun eignet dem Wort ‹erscheint›, wie auch dem griechischen ‹phainesthai› eine Ambiguität. Es schwankt zwischen einem phänomenologischen Sinn ‹sich zeigen› und einem urteilshaften bzw. epistemischen Sinn ‹meinen›». Der Ausdruck «X scheint F zu sein» kann also nicht immer als «Die Erscheinung von x ist F» wiedergegeben werden – so etwa bei dem Satz «Die Gerechtigkeit scheint gesiegt zu haben». Vgl. auch Hegel und Schelling: «Die Erscheinung ist ihm [dem Pyrrhoniker] aber nicht ein sinnliches *Ding*, hinter welchem von dem Dogmatismus und der Philosophie noch andere Dinge, nämlich die übersinnlichen behauptet werden sollten. Da er sich überhaupt zurückhält, eine Gewissheit und ein Sein auszusprechen, so hat er schon für sich kein Ding, kein Bedingtes, von dem er wüsste, und er hat nicht nötig, der Philosophie weder dieses gewisse Ding noch ein anderes, das hinter diesem wäre, in die Schuhe zu schieben, um sie fallen zu machen» (G. W. F. Hegel: Werke in zwanzig Bänden, Bd. 2: Jenaer Schriften 1801-1807 [Frankfurt a.M.: Suhrkamp, 1983] S. 248).

44 Vgl. «*aboulētō pathei*» (PH I [Fn. 1], § 19).

45 PH I, § 13. Vgl. PH I, § 193: «[…] denn den Dingen, die uns erlebnishaft (*pathētikōs*) bewegen und erzwungenermassen (*anankastikōs*) in die Zustimmung führen, geben wir nach (*eikomen*)».

> zieher glaube; das andere Mal aber das Zustimmen zu etwas mit Entschiedenheit und gleichsam einem Miterleben aufgrund eines starken Wollens.[46]

Der pyrrhonische Skeptiker hat also Meinungen in einem schwachen, höherstufigen Sinne, nämlich solche über das Vorliegen bestimmter Bewusstseinszustände, etwa über das Haben von Vermutungen und sich aufdrängenden Empfindungen. In Fällen, in denen er jedoch aktiv Stellung nehmen kann, d.h. wann immer ein freier deliberativer Entscheidungsprozess zugunsten einer bestimmten These möglich ist, hält er sich zurück.

3.4 Der Unverständlichkeitsvorwurf

Ein weiterer Irrationaliätsverdacht legt sich vor dem Hintergrund des oben skizzierten engeren Rationalitätsbegriffs nahe und behauptet, der pyrrhonische Skeptiker denke und rede zwar konsistent, verhalte sich damit jedoch nicht rational$_2$. Nach diesem Vorwurf sind die Überlegungen eines Pyrrhonikers nämlich derart *fremd und abwegig*, dass Nichtskeptiker oft nicht verstehen, wie jemand *solche* abstrusen Zweifel hegen kann. Der Skeptiker gilt aus dieser Sicht als irrational, weil man nicht sieht, welche Gründe es gibt zu zweifeln, und man daher nicht versteht, weshalb die eigenen Wissensansprüche ungerechtfertigt sein sollen. Der Skeptiker sei irrational, weil er etwas für einen Zweifelsgrund hält, das aus Sicht der meisten kein Grund zum Zweifeln ist. So bezweifelt er, dass es gute Gründe gibt, etwas Bestimmtes zu glauben, obwohl die meisten finden, die Sache sei doch völlig klar – es gäbe eine Wirklichkeit, die in etwa so ist, wie wir meinen; die Sonne gehe auch morgen wieder auf und andere seien keine Zombies.

Gegen diesen Einwand spricht, dass die dogmatischen Gesprächspartner viele der skeptischen Intuitionen des Pyrrhonikers teilen und seine Anforderungen an Wissen und Rechtfertigung zu einem gewissen Grad implizit akzeptieren. Der Pyrrhoniker kann seine Vorbehalte und Bedenken mittels skeptischer Argumente, der Schilderung skeptischer Szenarien und durch den Verweis auf *de facto* existierende alternative Lebensformen, Weltbilder und Auffassungen durchaus verständlich machen.

Ich bezweifle nicht, dass der Skeptiker vom gängigen Sprachgebrauch abweicht. Was wir im Alltag als Fälle von Wissen bezeichnen, würde er nicht «Wissen» nennen. Es ist allerdings nicht klar, ob er einen anderen

[46] PH I, § 230.

Wissensbegriff hat und wir aneinander vorbeireden, oder ob er eine andere Auffassung von Wissen hat als wir und man mit ihm darüber diskutieren kann, welches Verständnis angemessener ist. Jemand, der andere Auffassungen vertritt, redet nicht *ipso facto* falsches oder sinnloses Zeug. Insbesondere dann nicht, wenn er uns auf seine Seite bringen und uns zeigen kann, dass wir unter Rechtfertigung und Wissen *eigentlich* auch das verstehen, was er damit meint, im Alltag aber aus pragmatischen Gründen vieles ausblenden. Wir können nicht in jedem Kontext alles bezweifeln. Einen Historiker wird es schlicht nicht interessieren, zu hören, wir könnten nicht ausschließen, dass die Welt erst vor fünf Minuten erschaffen wurde. Das spricht jedoch nicht gegen den Skeptiker.

Auch wenn jemand unverrückbar an seiner Meinung festhalten sollte, dass für die Zweifel des Skeptikers so gut wie nichts spricht, so muss der *Gehalt* skeptischer Zweifel nicht schon deswegen sinnlos und unverständlich sein, weil der Äußerungs*akt* vielleicht unpassend und unmotiviert ist. Man kann verstehen, *was* eine Person sagt, auch wenn man letztlich nicht versteht, *warum* sie das sagt. Man sollte jemanden jedoch nicht «irrational» nennen, nur weil er gewisse Fragen interessanter und wichtiger findet als der Common Sense.

3.5 Der Selbsttäuschungsvorwurf

Abschließend möchte ich einen letzten Einwand vorbringen, der den Skeptiker unter Verdacht auf *Selbsttäuschung* stellt. Angenommen, Fälle von Selbsttäuschung seien Fälle von Irrationalität und der Pyrrhoniker stehe tatsächlich unter Selbsttäuschungsverdacht, dann steht er damit auch unter Verdacht auf Irrationalität. Worin der Selbsttäuschungsvorwurf besteht und ob er berechtigt ist, soll im Folgenden klar werden.

Als Folge der Instrumentalisierung der Theorie zugunsten einer gelingenden Praxis steht der Pyrrhoniker unter Verdacht, sich Argumente und Überlegungen nach Bedarf zusammenzuschustern und über argumentative Ungleichgewichte hinwegzuschauen. Er würde die Überlegenheit eines Arguments ignorieren, so der Selbsttäuschungvorwurf, nur um den gewünschten und der Gemütsruhe zuträglichen Eindruck der Gleichwertigkeit widerstreitender Thesen und Argumente herbeizuführen. Die Überlegungen und Recherchen des Pyrrhonikers seien motivational verzerrt. Seine theoretische Praxis sei stark interessegeleitet und somit in keiner Weise unvoreingenommen. – Dessen sei sich der Skeptiker *als* Skeptiker aber wohl bewusst. Man

könnte ihm also unterstellen, er müsse sich gleichsam *selbst täuschen*, um die These der Gleichwertigkeit aller Argumente plausibel finden zu können, denn schließlich ahnt er, dass er sich seine Überlegungen zum Zwecke einer gelungenen Praxis zurechtmacht. Die Gründe für die Gleichwertigkeit verlieren aber ihre Überzeugungskraft, sobald man weiß, dass sie auf die eigenen lebenskünstlerischen Interessen zugeschnitten sind. Wer bemerkt, dass er einen Gedankengang plausibel findet, *weil er ihn plausibel finden möchte*, dem geht damit die Plausibilität des Gedankens verloren. Wer vermutet, dass sein eigenes Denken letztlich *Wunschdenken* ist, wird anders anfangen zu denken, da er sich als rationales Wesen in seinem Denken an Gründen und nicht an Wünschen ausrichten möchte.

Der Skeptiker wird auf den Selbsttäuschungsvorwurf entgegnen, er sei in erster Linie ein *Therapeut* und kein Theoretiker. Er versuche sich also nur dann an Gründen auszurichten, wenn er sich davon einen lebenspraktischen Nutzen erhofft. In bestimmten Fällen aber liege ihm daran, Argumente nach Bedarf zu konstruieren, um ein doxastisches Gleichgewicht herzustellen, auf das die epistemische Zurückhaltung und – mit etwas Glück – die erhoffte Gemütsruhe folgt. Bekennt sich der Pyrrhoniker angesichts dieses Vorwurfs also zu einer lebensdienlichen Selbsttäuschung und damit zur lebensdienlichen Irrationalität?

Man könnte einwenden, absichtliche Selbsttäuschung sei ein Ding der Unmöglichkeit. Rationale Wesen seien mitunter dadurch ausgezeichnet, dass die für sie geltenden psychischen Gesetzmäßigkeiten der Meinungsbildung ein solches Wunschdenken nicht zuließen. Ist das aber wirklich eine so ausgemachte Sache? Kennen wir selbst das Phänomen des interessegeleiteten Forschens nicht auch? An dieser Stelle kann und möchte ich kein Argument liefern, das meinen Verdacht stützen würde; ich versuche es mit einer Geschichte: Wir treffen im Antiquariat zufällig auf ein bestimmtes philosophisches Buch, machen uns die Mühe und lesen uns ein, besuchen Seminare, schreiben Arbeiten und geraten allmählich in ein soziales Netzwerk, innerhalb dessen jeder den Autor des Buches zu schätzen gelernt hat und seine Thesen plausibel findet. Wir lernen zu argumentieren wie der Autor, finden andere Positionen allmählich immer unplausibler, werden angefragt, etwas zu diesem Autor zu publizieren, und haben schließlich immer weniger Zeit, Energie und Interesse, konkurrierende Positionen zu würdigen. Ist das nicht interessegeleitetes Forschen? Und liegt nicht der Verdacht nahe, dass wir mit den Überlegungen und Forschungen, die wir anstellen und die für unsere Meinungsbildung relevant sind, Ziele verfolgen, die mit Wahrheit nicht viel zu tun haben? Finden wir aber aufgrund der Einsicht, dass die Prozesse

der Überzeugungsbildung interessegeleitet waren, unsere Überzeugungen unplausibel? Geben wir unsere intellektuellen Gewohnheiten und Überzeugungen deswegen auf? Wohl kaum. Der Pyrrhoniker ist in diesem Punkt also vielleicht gar nicht so verschieden von rationalen Nicht-Pyrrhonikern.

4. Resümee

Sind pyrrhonische Skeptiker irrational? Fest steht, dass die Überlegungen des Pyrrhonikers weder inkohärent (1) noch unverständlich (4) sind. Seine Praxis steht weder in Widerspruch zu seinen Überlegungen noch ist sie völlig unabhängig von ihnen (3). Der Pyrrhoniker glaubt zwar nicht, aus guten Gründen zu handeln. Dies spricht allerdings nicht gegen seine Rationalität, da er Gründe geben kann, die dafür sprechen, dass niemand gute Gründe für seine Entscheidungen und Überzeugungen hat (2). Seine Überlegungen sind jedoch anfälliger für Interessen als die Überlegungen eines Nichtskeptikers, da die Rationalität für einen therapeutischen Skeptiker keinen Selbstzweck, keine absolute Norm darstellt. Das heißt aber auch, dass er nicht nur damit rechnet, dass er interessegeleitet forscht und überlegt, sondern sich gegebenenfalls auch affirmativ dazu verhält. Der Pyrrhoniker unterscheidet sich von uns mitunter darin, dass er die Selbsttäuschung nicht nur ahnt, sondern sie auch zulässt – soweit die psychologischen Gesetze der Überzeugungsbildung das erlauben.

Wenn Selbsttäuschung zu einer gelingenden Lebenspraxis beiträgt, dann sträubt sich der pyrrhonische Skeptiker nicht dagegen. In Fällen, in denen ihn die eigene Rationalität an seinem Glück hindert, zieht er die Irrationalität vor. Es bleibt jedoch – wie ich zu zeigen versucht habe – fraglich, ob sich rationale Personen in diesem Punkt hinreichend stark von einem pyrrhonischen Skeptiker unterscheiden. Und falls sie das tun, bliebe immer noch die Frage, ob sie damit auf dem richtigen Weg sind.

Studia philosophica 68/2009

Marcello Ostinelli

I dilemmi morali e il significato del rincrescimento

The paper argues for the plausibility of the thesis that there are no genuine moral dilemmas. In particular, I examine Bernard Williams' objection that agent-regret in a situation of conflict among contradictory moral obligations would prove the existence of such dilemmas.

Contrary to Williams' claim, agent-regret in these situations is not sufficient to establish the existence of such moral dilemmas. Williams' argument cannot undermine the logical consistency of moral theories, nor their capacity to solve moral conflicts.

I also consider Monika Betzler's proposal that agent-regret can be considered 'rational in a looser sense' without implying an inconsistency in moral theory. In order to discuss the plausibility of this proposal, I focus on the interpretation of agent-regret as an emotional response which also restores offended moral integrity.

I claim, in conclusion, that it is reasonable to assume this proposal, insofar as moral theory can determine some criteria of judgement which allow a distinction between authentic and false modalities of integration.

Accade talvolta a qualcuno di affrontare situazioni che impongono la scelta tragica di sacrificare vite umane innocenti. Pagine famose della letteratura mondiale ci permettono di rielaborare sul piano dell'immaginazione queste tragiche esperienze e di approfondirne la comprensione. Ne sono illustri esempi il conflitto tra Antigone e Creonte, tra gli obblighi personali dell'una e quelli impersonali dell'altro; il dramma di Agamennone che per far tornare a spirare i venti giusti sulla flotta greca deve sacrificare la figlia Ifigenia; e quello altrettanto tragico di Abramo cui Dio ordinò il sacrificio di Isacco. In tempi a noi più vicini William Styron in una pagina che è ormai divenuta un luogo classico su questo tema racconta la scelta di Sophie per evitare che almeno uno dei due figli non sia messo sul convoglio destinato al campo di Birkenau.[1] Una situazione simile è immaginata da Bernard Williams per mettere in evidenza la differenza tra una teoria etica che considera la condotta umana in una prospettiva alla terza persona; e, d'altro lato, una teoria che adotta la prospettiva di chi compie l'azione: il caso ipotetico è quello di Jim

1 William Styron: *Sophie's Choice* (London: Jonathan Cape, 1979) chap. XV, trad. it. *La scelta di Sophie* (Milano: Leonardo, 1990) cap. XV, pp. 583-587.

che è costretto a scegliere tra lasciare che accada che venti persone innocenti vengano fucilate oppure essere l'autore dell'uccisione di una di loro per salvare la vita delle altre.[2] È a queste scelte tragiche, in particolare a quelle situazioni che comportano inevitabilmente la perdita di qualche vita umana innocente, come quelle descritte da William Styron e da Bernard Williams, che rivolgerò la mia attenzione.

La questione non è soltanto pratica; è rilevante pure dal punto di vista teorico, come è dimostrato dal dibattito filosofico contemporaneo. Riguarda infatti lo statuto della teoria etica: se essa possa ragionevolmente contribuire alla riuscita della vita morale o, quanto meno, ad evitare il suo fallimento; oppure se di essa si possa fare a meno; se essa debba fornire principi che consentano di risolvere i conflitti morali e di prescrivere ciò che il soggetto morale deve fare oppure se essa debba identificare di fronte a scelte tragiche come quelle di Sophie e di Jim anche le risorse emotive da cui attingere una risposta riparatrice della propria integrità morale offesa.

Risale a Bentham la concezione di teoria etica, sostenuta con vigore dall'utilitarismo contemporaneo, che ritiene che in qualunque situazione, anche in quelle di conflitto morale estremo, esiste sempre l'azione che tutto considerato è migliore di altre e compierla è nostro dovere. Soddisfano questo criterio le teorie etiche monistiche. Per il monismo etico non si danno dilemmi morali genuini, vale a dire situazioni in cui due doveri non soverchiabili obbligano ad azioni incompatibili. Ciò che l'agente considera un dilemma morale è semplicemente il risultato della sua difficoltà epistemica: egli non sa cosa fare perché è incapace di giudicare correttamente la situazione. Non si nega l'esistenza di conflitti morali; però ogni buona teoria etica deve essere in grado di risolverli. Una teoria etica che non sia in grado di fugare la perplessità morale di fronte ad una situazione problematica e di decidere ciò che si deve fare, per quanto difficile sia il giudizio sulla situazione e per quanto tragica possa risultare la scelta, va considerata incoerente e indeterminata o, nel migliore dei casi, superficiale. Al più la teoria può ammettere l'esistenza di un dilemma se risulta da un precedente sbaglio morale. E' questo il caso di ciò che Tommaso d'Aquino denominava dilemma *secundum quid*, quando il soggetto morale a seguito di un precedente sbaglio morale deve onorare

[2] Bernard Williams: *A Critique of Utilitarianism*, in John J. C. Smart, Bernard Williams: *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973), trad. it, *Utilitarismo: un confronto* (Napoli: Bibliopolis, 1973) qui pp. 123-124 (= *U*).

due obblighi incompatibili.[3] In questi casi la perplessità morale dipende esclusivamente da una precedente scelta sbagliata del soggetto ma non deriva da un'insufficienza della teoria a prescrivere l'azione giusta. Una teoria etica consistente esclude la possibilità che il soggetto morale possa trovarsi in uno stato di perplessità *simpliciter*, ovvero che possa esistere un conflitto vero tra due obblighi morali se antecedentemente alla situazione in cui si trova egli non ha compiuto sbagli. Per Tommaso la possibilità che chi ha compiuto uno sbaglio sia perplesso su quel che deve fare non è un argomento contro la consistenza della teoria: «non est inconveniens, ut aliquo supposito, home peccatum vitare non possit».[4] Tuttavia, anche nel caso di un dilemma *secundum quid*, è questione controversa che una teoria etica non sia in grado di prescrivere la condotta morale giusta a chi abbia compiuto lo sbaglio.

Forse l'espressione più conseguente della tesi monistica che una buona teoria etica deve essere in grado di fugare qualsiasi perplessità morale si trova in *Moral Thinking* di Richard M. Hare. Egli distingue tra due livelli del pensiero morale. È al livello intuitivo dei principi *prima facie* che riguarda le situazioni ordinarie della vita quotidiana che un conflitto di doveri può sembrare insolubile. In effetti, quando il livello intuitivo del pensiero morale è confrontato con una situazione eccezionale può trovarsi in difficoltà. In questi casi lo soccorre il livello critico che sarà in grado di dire cosa si deve fare. A questo livello il pensiero morale potrà dimostrare quanto avrebbe scritto un pastore dello Yorkshire fuori della sua chiesa, che «Se avete dei doveri contrastanti, uno di essi non è il vostro dovere». Per Hare infatti la perplessità del filosofo è la prova che la sua comprensione del conflitto di doveri è rimasta superficiale: «Le diverse concezioni dei filosofi morali nei riguardi del conflitto di doveri costituiscono un sintomo estre-

[3] Questa interpretazione venne sostenuta in particolare da Alan Donagan in diversi importanti contributi: *The Theory of Morality* (Chicago: University of Chicago Press, 1977); *Consistency in Rationalist Moral Systems*, in *The Journal of Philosophy* 81 (1985) pp. 291-309; *Moral Dilemmas, Genuine and Spurious: A Comparative Anatomy*, in *Moral Dilemmas and Moral Theory,* ed. by H. E. Mason (Oxford: Oxford University Press, 1996) pp. 11-22. Nella sua interpretazione Donagan si è basato su *Summa Theologiae*, Iª-IIae q. 19 a. 6 arg. 3 I-II; IIª-IIae q. 62 a. 2 arg. 2; IIIª q. 64 a. 6 ad 3; e a *Quaestiones disputatae de veritate*, q. 17 a. 4 ad 8. Più recentemente M. V. Dougherty ha suffragato l'interpretazione di Donagan con ulteriori prove in: *Perplexity Simpliciter and Perplexity Secundum Quid: A Look at Some Contemporary Appeals to Thomas Aquinas*, in *International Philosophical Quarterly* 41 (2001) pp. 469-480.

[4] Tommaso d'Aquino: *Quaestiones disputatae de veritate*, q. 17 a. 4 ad 8.

mamente chiaro del livello di comprensione e di penetrazione raggiunto dal loro pensiero intorno alla moralità; forse è proprio nell'esame di ciò che viene detto su questo argomento, che la superficialità si rivela nel modo più immediato».[5]

È noto che la tesi che non esistono dilemmi morali intrattabili è stata oggetto di numerose critiche. Un'obiezione importante contro il monismo riguarda l'esistenza di obblighi morali che sono giustificati da valori incommensurabili i quali rendono gli obblighi incomparabili. Qui non considero questo argomento, che com'è noto, originariamente fu proposto da Thomas Nagel.[6] Vi è poi l'argomento che sostiene che l'esistenza di dilemmi morali genuini sia provata per il fatto che quando in una situazione di conflitto tra doveri morali contraddittori (potendo cioè fare ciò che è richiesto dall'uno ma non ciò che è richiesto dall'altro, o perché non è possibile fare entrambe le cose o perché fare l'una preclude che si possa fare l'altra) il soggetto che sceglie a favore dell'uno proverà rincrescimento per aver negletto l'altro. L'esistenza di questo sentimento proverebbe che il conflitto di obblighi costituiva un dilemma morale autentico. È del significato di questo argomento che intendo occuparmi in questa sede, nella scia di un dibattito ormai molto esteso e complesso.

In un articolo pubblicato nel 1965 Bernard Williams sosteneva che, a differenza di quanto ritengono i fautori del monismo etico, risolvere un dilemma morale non è un'operazione mentale simile a eliminare un errore. Se così fosse, sarebbe come abbandonare una credenza sbagliata. La constatazione che due credenze sono tra loro in conflitto costringe a indebolirne una e poi ad abbandonarla. Se risolvere un dilemma morale assomigliasse all'eliminazione di un errore, allora «si tratta solo di stabilire quale dei due asserti conflittuali indicanti un dovere sia quello vero, e poiché quegli asserti non possono essere entrambi veri, decidere correttamente per uno di essi non

5 Richard M. Hare: *Moral Thinking* (New York: Oxford University Press 1981) cap. II, § 1, trad. it. *Il pensiero morale* (Bologna: Il Mulino, 1989) p. 58.

6 Thomas Nagel: *The Fragmentation of Value*, in T. N.: *Mortal Questions* (Cambridge: Cambridge University Press, 1979) pp. 128-141, trad. it.: *La frammentazione del valore*, in *Questioni mortali* (Milano: Il Saggiatore, 1986) pp. 127-139. L'argomento di Nagel si può così riassumere: (1) taluni valori sono tra di loro incommensurabili; (2) se ci sono valori incommensurabili, allora vi saranno anche obblighi incommensurabili; (3) obblighi incommensurabili generano conflitti insolubili, cioè dei dilemmi morali.

può che voler dire sbarazzarsi dell'errore contenuto nell'altro».[7] In questo modo si esclude l'esistenza di dilemmi genuini e li si considera alla stregua di conflitti morali che in linea di principio la teoria è in grado di risolvere.

Secondo Williams tuttavia, i conflitti tra doveri «assomigliano di più ai conflitti tra desideri che ai conflitti tra credenze».[8] La prova è che in entrambi i casi l'alternativa a cui non è possibile dare seguito non può essere eliminata, come accade invece ad una credenza falsa. In un conflitto di obblighi «non penso in termini di eliminazione dell'errore»[9]; al contrario riconosco la presenza di entrambi gli obblighi e anche quello che non è stato adempiuto, alla stregua di un desiderio respinto, può persistere e ricomparire nella forma di un «residuo emotivo» che esprime il rincrescimento per ciò che è perduto. In generale, il rincrescimento comporta «il desiderio che le cose siano andate diversamente»;[10] qui però si tratta di «rincrescimento dell'agente» *(agent-regret)*, cioè di «quel rincrescimento che una persona può avvertire solo nei confronti delle sue azioni passate (o delle azioni alle quali per lo meno ritiene di aver contribuito)».[11] Il rincrescimento dell'agente si distingue perciò dal rincrescimento dello spettatore, che in via di principio chiunque potrebbe provare se venisse a conoscenza di quel particolare stato di cose. Il rincrescimento dell'agente è pure diverso dal rimorso. Mentre si prova rimorso soltanto per azioni volontarie, «noi abbiamo sentimenti di rincrescimento come agenti anche nei confronti di azioni involontarie».[12]

A Williams l'esistenza di un rincrescimento dell'agente sembrava sufficiente per una «critica fondamentale» degli assunti delle teorie etiche monistiche perché, se l'obiezione è pertinente, sarebbe provato che non rendono giustizia alla peculiare fenomenologia che accompagna l'esperienza personale del conflitto tra doveri. Infatti, nonostante che le conseguenze di cui il soggetto morale si rincresce non siano il risultato di proprie azioni volontarie, il rincrescimento che egli prova (allo stesso modo del rimorso ma a differenza

7 Bernard Williams: *Ethical Consistency*, in B. W.: *Problems of the Self. Philosophical Papers 1956-1972* (Cambridge University Press: Cambridge, 1973), trad. it. *Coerenza etica*, in *Problemi dell'io* (Il Saggiatore: Milano, 1990) p. 213 (= *CE*).

8 *CE*, p. 209.

9 *CE*, p. 210.

10 Bernard Williams: *Moral Luck*, in B. W.: *Moral Luck. Philosophical Papers 1973-1980* (Cambridge: Cambridge University Press, 1981), trad. it. *Sorte morale*, in *Sorte morale* (Milano: Il Saggiatore, 1987) p. 46 (= *SM*).

11 *SM*, p. 42.

12 *SM*, p. 45.

del rincrescimento dello spettatore) «comporta da parte dell'agente stesso il desiderio di non aver fatto ciò che ha fatto».[13]

Secondo Williams le teorie etiche monistiche ci porterebbero a concludere che è irrazionale[14] provare rincrescimento per il dovere che è stato tralasciato, se la scelta che ho compiuto è la migliore tutto considerato: «se sono convinto di aver agito per il meglio e se non ho motivo di rimproverarmi di essermi messo da solo in una situazione conflittuale, allora è semplicemente irrazionale provare rincrescimento».[15] Tuttavia, a parere di Williams, questo giudizio contraddice il modo comune di rappresentarsi il carattere appropriato di un agente morale. In una situazione di conflitto di obblighi morali come quella che Sophie fu costretta ad affrontare senza propria precedente colpa, qualunque essere umano decente proverebbe i suoi stessi sentimenti. Se la scelta compiuta non avesse lasciato nell'agente alcun rincrescimento, vi sarebbe ragione di interrogarsi sulle qualità morali di chi l'ha compiuta. Così Sophie potrebbe sentirsi colpevole, ritenendo di aver sbagliato, qualunque cosa ella avesse fatto.

Considerando la statuto della teoria la conseguenza generale è che anche le distinzioni etiche più elementari vacillano, al punto che per Bernard Williams, se ciò non è sufficiente per disfarsi interamente dell'idea di moralità, non può non modificarne il concetto. Quello che ci resterà sarà «certamente un concetto meno importante di quanto non sembri essere il nostro»,[16] che sembra mettere in dubbio la possibilità di una teoria etica.

Secondo Williams, poiché per le teorie etiche monistiche è irrazionale provare rincrescimento per quanto accaduto e desiderare di aver agito diversamente, in esse c'è qualcosa che non va. È valida la critica di Williams? Che cosa prova la sua obiezione?

Se con rincrescimento Williams intende il sentimento di colpa di aver compiuto uno sbaglio, allora è irrazionale provarlo, se chi compie l'azione ha scelto ciò che era giusto fare in quella situazione. Qui non c'è ragione di sentirsi in colpa (come invece succede nel dilemma *secundum quid*) perché non c'è stato alcuno sbaglio. Questo è ciò che sostiene una teoria etica monistica. Secondo Williams, invece, «dev'essere un errore supporre che qui abbiamo a che fare con un caso di incoerenza logica».[17]

13 *SM*, p. 46.
14 *CE* (n. 7), p. 210.
15 Ibid.
16 *SM* (n. 10), p. 57.
17 Bernard Williams: *Conflicts of Values,* in B. W.: *Moral Luck. Philosophical*

Come già osservò Philippa Foot[18] la pertinenza dell'obiezione di Williams è dubbia. Se l'argomento riguardasse il nostro modo di rappresentarci la situazione (come dice Williams),[19] allora esso proverebbe ben poco.

Una cosa è affermare che, se ho sbagliato nella mia condotta, è verosimile che io provi un sentimento di colpa per lo sbaglio compiuto, se sono una persona moralmente decente. Cosa ben diversa invece è credere che l'esistenza di qualcosa che sembra assomigliare a un sentimento di colpa sia sufficiente per provare che io ho compiuto uno sbaglio morale. Infatti se mi sento in colpa per uno sbaglio che però non ho realmente compiuto, dal momento che ho scelto l'azione che era giusto fare, non ho ragione di provare una colpa anche se ve ne sono di buone per provare rincrescimento per quanto è accaduto. In effetti, l'esistenza di un «residuo emotivo» non prova che la scelta compiuta fosse sbagliata; né prova che, qualunque cosa avessi fatto, avrei sbagliato; né prova tanto meno che il conflitto di obblighi fosse un dilemma morale genuino.

Contrariamente a quel che suppone Williams, se è vero che Sophie non può salvare sia Eva che Jan ma soltanto uno dei due, allora ha l'obbligo di salvare o Eva o Jan, per quanto tragica sia questa scelta. In effetti, se non è in suo potere salvare sia Eva che Jan, sarebbe un errore «pensare che erano entrambe le cose che avev[a] il dovere di fare»,[20] e conseguentemente è irrazionale che Sophie si senta colpevole per non aver salvato uno dei suoi due figli. Ciò che Sophie deve fare in realtà è salvarne uno, o Eva o Jan, anche se vorrebbe salvarli entrambi (e probabilmente sarebbe disposta a dare anche la sua vita per questo). Mentre in situazioni normali l'obbligo di fare il bene di Eva e quello di fare il bene di Jan valgono congiuntamente, nel caso della situazione tragica in cui si trova Sophie gli obblighi devono essere disgiunti in quanto ella può salvare soltanto uno dei suoi figli. Pertanto, contrariamente a ciò che pensava Williams, il fatto che Sophie non deve considerarsi moralmente responsabile di non aver salvato Eva (piuttosto che Jan) non costituisce una falsificazione del pensiero morale.[21]

Papers 1973-1980 (Cambridge: Cambridge University Press, 1981), trad. it. *Conflitti tra valori,* in *SM*, p. 101 (= *CV*).

18 Philippa Foot: *Moral Dilemmas Revisited*, in *Modality, Morality, and Belief*, ed. by Walter Sinnott-Armstrong (Cambridge: Cambridge University Press 1995) pp. 117-128, ristampato in Philippa Foot: *Moral Dilemmas and Other Topics* (Oxford: Clarendon Press 2002) pp. 175-188, qui pp. 184-185.

19 *CV* (n. 17), p. 101.

20 *CE* (n. 7), p. 223.

21 Ibid.

Tuttavia, ciò non nega ogni valore e qualsiasi significato alle emozioni che prova chi senza propria colpa è costretto a compiere una scelta tragica. È sufficiente non scambiare il rincrescimento per quanto è accaduto malgrado le intenzioni del soggetto morale, con il rimorso che egli prova per le conseguenze negative di una scelta di cui a giusta ragione dev'essere invece ritenuto responsabile. Mentre il rincrescimento è compatibile con la coscienza di aver compiuto il giusto, date le circostanze in cui il soggetto morale si è trovato a scegliere e provarlo non è irrazionale; il rimorso, invece, è il sentimento morale appropriato di chi non ha compiuto il proprio dovere. Pertanto, nel caso di Sophie, che è quello di chi ha agito come era giusto fare, sarebbe irrazionale provare rimorso perché nessuno sbaglio è stato compiuto.

Vi è dunque una differenza tra queste emozioni morali che ha rilevanza per la teoria etica: la distinzione tra rimorso e rincrescimento, che già Richard Hare aveva evidenziato,[22] inficia l'argomento di Williams, se esso intendeva provare l'esistenza di un dilemma morale genuino attraverso l'esperienza del rincrescimento dell'agente. È sufficiente avere la cura di non confondere l'uno con l'altro per evitare questo errore.

D'altra parte l'attitudine di una teoria etica nei confronti dei «residui emotivi» non è univoca. Monika Betzler ha chiarito opportunamente la differenza tra due attitudini opposte delle teorie etiche che escludono l'esistenza di dilemmi morali genuini: tra l'attitudine di quelle teorie per le quali, una volta che sia stata scelta l'azione migliore, considerare il rincrescimento non è rilevante per una comprensione più profonda del punto di vista dell'agente; e l'attitudine di una teoria per la quale il rincrescimento dell'agente può essere considerato «razionale in un senso più vago» («rational in a looser sense»), senza che ciò comporti un'incongruenza per la teoria.[23]

Data questa possibilità di considerare il rincrescimento come un «residuo emotivo» non semplicemente irrazionale ma «razionale in un senso più vago», si tratta di determinarne il significato. Che cosa c'è di razionale nel rincrescimento?

Recentemente Carla Bagnoli ha proposto di intendere il rincrescimento come una modalità di integrazione della propria coscienza morale e di ricostruzione della propria integrità, quando, come nella scelta tragica di Sophie,

22 Hare, op. cit. (n. 5) p. 61.

23 Monika Betzler: *Sources of Practical Conflicts and Reasons for Regret*, in *Practical Conflicts. New Philosophical Essays*, ed. by Peter Baumann, Monika Betzler (Cambridge: Cambridge University Press, 2004) pp. 197-222, qui pp. 199-200.

«le nostre azioni non sono espressive della nostra identità pratica».[24] In condizioni nelle quali la scelta è tragica, gli atteggiamenti con cui l'agente affronta il conflitto e sopporta le conseguenze della sua azione «sono più espressivi e anzi costitutivi dell'integrità dell'agente che non l'azione».[25] Per esprimere la propria identità valgono di più gli stati emotivi (in una gamma che va dal rincrescimento alla disperazione) che l'azione che, pur essendo quella che è giusto fare, nuoce alla propria integrità morale, comportando la perdita della vita di persone innocenti come sua conseguenza prevedibile ma non intenzionale. Per la teoria etica, quanto meno per una teoria etica che considera la condotta umana nella prospettiva dell'integrità di chi compie l'azione, non è sufficiente la valutazione dell'azione che scelgo di compiere per risolvere il conflitto morale; dev'essere considerata anche la capacità di dare una risposta emotiva appropriata che sia «espressiva delle persone che siamo».[26]

In effetti sembra ragionevole che la teoria etica consideri seriamente questa fenomenologia delle emozioni morali se si ritiene che essa non esaurisca il suo compito con la prescrizione dell'azione giusta. La considerazione del significato delle risposte emotive riparatrici della propria integrità morale, in particolare del rincrescimento dell'agente, non è incoerente con l'idea di una teoria etica monistica. Infatti, poiché ciò non comporta l'allentamento del vincolo di consistenza della teoria, ovvero l'affermazione che esistano dilemmi morali genuini, essa non apre la strada alla demolizione della teoria etica così come usualmente è intesa.

In precedenza, nell'esame dell'argomento di Bernard Williams, ho sostenuto, nella scia della critica di Philippa Foot, che l'esperienza personale del rincrescimento, per quanto importante e significativa sia per chi la prova, non basta a dimostrare che il conflitto esperito sia un dilemma morale genuino. Nonostante che a Jim non può essere rimproverata alcuna colpa (se dovesse rifiutare di piegarsi al ricatto di uccidere una persona innocente onde salvarne altre diciannove poiché «ognuno è responsabile di ciò che egli stesso fa, piuttosto che di quello che fanno gli altri»)[27] lo smarrimento morale che egli prova quando assistesse alla fucilazione di venti persone innocenti non può essere considerato fuori luogo. Attraverso questo sentimento a Jim diventa manifesta la «percezione di un fallimento». Di quale fallimento?

24 Carla Bagnoli: *Dilemmi morali* (Genova: De Ferrari, 2006) p. 16.
25 Ibid. p. 95.
26 Ibid. p. 149.
27 *U* (n. 2) p. 124.

Non della propria scelta, ovviamente, perché tanto basterebbe a negare che quella scelta fosse quella giusta. Infatti, mentre nel caso del rimorso il soggetto manifesta attraverso questo sentimento una specie di insoddisfazione riguardo al proprio carattere e alle proprie azioni;[28] nel caso della perplessità morale dopo una scelta che ragionevolmente non può essere considerata uno sbaglio si tratta della manifestazione del sentimento di impotenza di fronte alla cattiva sorte, agli elementi dell'esistenza umana che sfuggono al controllo dell'agente ma che egli è costretto ad accettare come costitutivi della propria esistenza. In situazioni come queste risulta evidente ciò che già Williams aveva notato, che «Il pensiero che sta alla base del rincrescimento in generale è all'incirca questo: come sarebbe meglio se fosse andata diversamente», senza che però ciò significhi necessariamente che se io avessi agito diversamente, ora le cose starebbero meglio. Lo smarrimento di Jim, la sua perplessità morale sono dunque l'espressione appropriata della propria tragica inadeguatezza di fronte alle prove più difficili della vita.

Qual è il difetto di questa interpretazione? Essa può fornire un alibi a un atteggiamento autoindulgente; potrebbe prestarsi infatti a legittimare una pratica di ricostruzione dell'integrità di sé anche in presenza di scelte discutibili dal punto di vista morale. Se provare rimorso è irrazionale quando la scelta compiuta è quella che era giusto fare; provare rincrescimento potrebbe essere una forma di mistificazione e di autoinganno quando l'azione che l'agente ha compiuto non era quella che era doveroso fare; per esempio, se Jim non ha fatto tutto ciò che era in suo potere per evitare la fucilazione di persone innocenti senza tuttavia prestarsi a diventare l'autore dell'uccisione di una di esse per evitare che altri le uccida tutte. Ciò che dunque si richiede è una teoria etica di queste pratiche di ricostruzione dell'integrità di sé che sia in grado di discriminare tra integrazioni moralmente legittime della nostra identità pratica e integrazioni autoindulgenti, mistificatorie e autoingannevoli. È dunque interessante che la teoria etica esplori le diverse forme dell'integrazione personale a cui risponde il rincrescimento del soggetto. L'analisi fenomenologica non è autosufficiente; essa richiede che la teoria etica stabilisca dei criteri di giudizio che sappiano distinguere tra modalità autentiche e modalità fasulle di integrazione.

28 Bagnoli, op. cit. (n. 24) p. 35.

Studia philosophica 68/2009

Johannes D. Balle

Emotionaler Logos. Werterfahrung und Deliberation in einer Theorie emotionaler Kultivierung

There are two interesting concepts in the theory of reason: deliberation and value-experience. A cognitivist theory stresses the objectivity of values and claims that there are real values in the world which can be represented in practical propositions. On the other hand, the subjectivist denies the reality-dependency of values and focuses on forms of practical reasoning. Furthermore there is a strong tradition in philosophy of dividing the explanation of action into two: conative and cognitive contents. I shall argue for a moderate Aristotelian conception, claiming an intersubjective constitution of values which could be perceived through emotional content. Emotions are a third, irreducible category in explaining human behaviour. But there is one important feature about emotions: they could have a cultivated status or not. If they are cultivated, we should speak of qualitative thoughts or practical habits. There are three claims I would like to defend: (1) A Frankfurt-style conception of deliberation and freedom of will cannot explain the normativity of practical reasons. We need a critical and value-sensitive notion of emotional reason. (2) We explain value-sensibility by regarding some emotions as value-perceivers. This in turn has something to do with our social and moral education, what Aristotle called hexis. (3) We learn about the importance of the intersubjective forming of our emotions as habits from Aristotle. But, there must also be a kind of non-instrumental form of free practical reasoning that enables us to constitute personal importance. Emotional logos is the ability on the basis of our emotional value-experience (a) to get clear about things that really matter for us context-independently and context-dependently; and (b) to get the right answer what to do and how to do this in a specific situation by also considering dynamic personal values. No doubt, forming this practical sense is a life-long task.

1. Werterfahrung und Deliberation

Was für uns zählt, ist eine Frage, die auf zweierlei Weisen beantwortet werden kann: Durch Verweis auf Werte, die wir passiv wahrnehmen, etwa in der Folge unserer Erziehung und der Interaktion mit der sozialen Umwelt. Andererseits, indem man auf das blickt, was wir aktiv im Zuge praktischer Überlegungen als das Gute beurteilen.

Man kann zunächst objektivistische bzw. kognitivistische Konzepte über Gründe und Werte von subjektivistischen oder non-kognitivistischen Modellen unterscheiden.[1] Ein Kognitivist scheint etwa folgende Ansicht zu vertreten: Personen sind in der Lage, evaluative Eigenschaften, die für eine Situation entscheidend sind, passiv wahrzunehmen, weshalb von wahren Überzeugungen gesprochen werden kann. Ein Nonkognitivist hingegen vertritt eine projektionische Theorie, bei der nicht evaluativen Tatsachen, sondern subjektiven Reaktionsmustern eine systematische Bedeutung zukommt.

Ich halte beide Positionen in ihrer extremen Form für einseitig und will zu zeigen versuchen, weshalb die These, dass Personen das Gute wollen, nicht auf einseitige Weise erörtert werden sollte. Ich denke, wir sollten hierzu das kognitivistische Modell über Wertwahrnehmungen durch eine autonomietheoretische Konzeption praktischen Überlegens erweitern. Dass es sich dabei um keinen Widerspruch handeln muss, sondern vielmehr um einen Prozess der wechselseitigen Abhängigkeit von Werterfahrungen und Deliberationen, das soll deutlich werden, indem die Rolle der Emotionen im Zusammenhang der Konstitution von Werten und praktischen Bedeutsamkeiten skizziert wird. Ich werde in diesem Zusammenhang die Konzeption eines emotionalen Logos[2] verteidigen, der sich als gründegebende Fähigkeit auf der Grundlage emotionaler Lernerfahrungen beschreiben lässt, wodurch die herkömmliche Dichotomie von kognitiven und konativen Erklärungsbeständen praktischer Rationalität ebenso überwunden werden soll wie die einseitige Explikationsweise von Wertkonstitution durch subjektive Projektionen oder objektive Repräsentationen.

1 Hierzu zählen etwa Martha C. Nussbaum: *Upheavals of Thought* (Cambridge: Cambridge University Press, 2001) oder Robert C. Solomon: *On Emotions as Judgements*, in *Philosophical Quarterly* 25 (1988) S. 183-191. Klassischerweise könnte man David Hume als Vertreter eines Subjektivismus bezeichnen sowie seine Nachfolger, etwa Bernard Williams: *Problems of the Self*, in B. W.: *Philosophical Papers 1956-1972* (Cambridge: Cambridge University Press, 1973, 1995).

2 Der Begriff «Logos» betont einen Unterschied zum herkömmlichen Vernunftbegriff in folgender Hinsicht: Gemeint ist nicht nur die Rationalität im Sinne eines Grundvermögens, sondern eine eingeübte Fähigkeit wie etwa gutes Reden, Schreiben oder Argumentieren. Der gründegebende Charakter des «Logos» impliziert begriffliche und qualitative Aspekte, womit eine Verstehens- und Deutungsfähigkeit bezeichnet wird, die als begriffliche Phänomenalität aufzufassen ist und im Rahmen einer «kognitiven Phänomenologie» erforscht wird.

Das zweistämmige Wunsch-Überzeugungs-Modell erschwert die Darstellung rationaler Kontrolle bzw. Kultivierung dessen, was uns in Abhängigkeit von intersubjektiven Rationalitätsstandards am Herzen liegt. Andererseits bietet ein klassisches kognitivistisches Modell zu wenig Spielraum für die Berücksichtigung des deliberativen Selbstverstehens, auf das etwa Charles Taylor großen Wert legt.[3] So wie eine humesche Konzeption zu wenig Spielraum lässt für die kognitivistische Intuition von evaluativen Aussagen mit Wahrheitsbedingungen, so sollte andererseits die kantische Konzeption abgelehnt werden aufgrund ihres allzu rigiden Verallgemeinerungsprinzips in Form des Sittengesetzes. Dieser generalisierte Handlungsmaßstab legt eine ent-individualisierte und dekontexualisierte Wollenskonzeption nahe, die anthropologisch unplausibel ist.[4] Die Konzeption des emotionalen Logos demgegenüber wäre als aristotelisches Modell aufzufassen, dessen Grundidee wie folgt lautet: Durch praktische Deliberation formen wir in einem lebenslangen Lernprozess unsere Emotionen ebenso wie es unseren Sensibilitäten vorbehalten ist, unsere Urteile zu kultivieren.[5]

2. Kognitivismus und Nonkognitivismus

Es ist die alte platonische Frage: Erstreben wir unsere Ziele aufgrund ihrer Güte oder sprechen wir einem Ziel gerade deshalb das Güteprädikat zu, weil wir es erstreben? Fürchten wir jemanden, weil er furchterregend ist, oder ist er furchterregend, weil wir ihn fürchten? Allgemein gesagt: Sehen wir etwas als gut, weil es tatsächlich gut ist, oder ist es für uns deshalb gut, weil wir es als etwas Gutes empfinden?

Was also bedeutet es, das Gute zu wollen? Ist unser Wollen ein zweckunabhängiges «gutes» Wollen oder ist es ein «gutes Wollen» nur insofern, als

3 Vgl. Charles Taylor: *Self-Interpreting Animals*, in C. T.: *Human Agency and Language. Philosophical Papers Vol. I* (Cambridge: Cambridge University Press, 1985) S. 45-76.

4 Vgl. die lehrreiche Darstellung handlungstheoretischer Positionen von Kirsten B. Endres. Die Autorin plädiert freilich entgegen der hier favorisierten Linie für eine systematische Trennung von Gründen und Werten: Kirsten B. Endres: *Praktische Gründe. Ein Vergleich dreier paradigmatischer Theorien* (Frankfurt a.M: Ontos Verlag, 2003).

5 Zum Begriff der «Persönlichkeit» verweise ich auf Michael Quantes lehrreiche Ausführungen, denen ich wertvolle Hinweise verdanke, Michael Quante: *Person* (Berlin: Walter de Gruyter, 2007).

wir ein gutes Ziel erstreben? In einem grundlegenden Sinn wäre es zunächst unplausibel zu behaupten, es läge im Interesse von Akteuren, gegen ihre personale Signifikanz[6] zu handeln. Es leuchtet ein, dass praktische Gründe immer die guten Gründe des Akteurs sein sollten. Man kann unterstellen, dass Personen grundsätzlich interessiert daran sind, ihre Handlungsgründe so zu wählen, dass sie in ein System von evaluativen Selbst-Aussagen passen, die von ihnen im Falle einer Befragung ausdrücklich bestätigt würden. Also fällt die Antwort auf die Frage, was Personen erstreben, leichter: Personen wollen das Gute als den Handlungszweck, den sie selbst im Lichte gewisser Werte und Überzeugungen als guten betrachten. Gute Gründe hängen damit von personalen Werten bzw. von «praktischer Signifikanz» ab.

Wenn es also im Interesse von Akteuren liegt, gute Gründe zu entwickeln, die von einem System personaler Signifikanz abhängen, dann gerät die Entstehung praktischer Signifikanz in den Blickpunkt des Interesses. Sind es nun objektive Tatsachen oder subjektive Projektionen, die darüber entscheiden, was für uns Wert und Bedeutung besitzt?

Ich möchte dafür argumentieren, dass zum Verständnis evaluativer Einstellungen wesentlich ihre emotionale Erfahrungsqualität gehört. Entscheidend ist der Zusammenhang zwischen emotionaler Erfahrung, evaluativen Eigenschaften und deliberativen Urteilen. Die entscheidende Frage lautet, wie unsere Emotionen über Prozesse des Überlegens zu emotional geformten selbstkritischen und profilkohärenten Haltungen werden können. Meine Antwort lautet in einem Wort: Verantwortlich hierfür ist das dynamische (kontextbezogene und kontextübergreifende) Zusammenspiel von praktischer Überlegung und emotionaler Werterfahrung. Dadurch kann zweierlei gelingen: Eine Klärung dessen, was uns am Herzen liegt, sowie eine Entscheidung darüber, wie dies nach den Anforderungen der jeweiligen Situation zu realisieren ist.

In welchem Abhängigkeitsverhältnis stehen nun aber evaluative Eigenschaften und evaluative Einstellungen? David Wiggins vertritt die Ansicht, dass Werteigenschaften und entsprechende Einstellungen infolge eines historisch-kulturellen Prozesses aufeinander abgestimmt seien.[7] Tatsächlich herrscht in den meisten Fällen eine gewisse Harmonie zwischen evaluativen

6 Ich verwende zur Bezeichnung von Bedeutsamkeiten den technischen Begriff «praktische Signifikanz», der sowohl personale als auch moralische Werte umfasst.

7 Vgl. David Wiggins: *A Sensible Subjectivism?*, in D. W.: *Needs, Values, Truth* (Oxford: Blackwell, 1987) S. 185-214.

Eigenschaften und korrelierenden Einstellungen, eine Stabilität, die wir als individuelle Lernerfahrung allgemeiner Werte innerhalb sozialer Kontexte deuten werden. Wir wollen also Wiggins zustimmen, dass die Konstitution praktischer Signifikanz als intersubjektiver Prozess zu betrachten ist, wobei «Werteigenschaften» und «Einstellungen» reziprok voneinander abhängen.

Daher sollten wir nicht länger von einer einseitigen Betonung entweder subjektiver Projektionen oder objektiver Repräsentationen sprechen. Wir werden in den folgenden Abschnitten genauer sehen, wie man die zugrunde liegende Aktivität der Entstehung praktischer Signifikanz fassen könnte. An dieser Stelle sei nur so viel gesagt: (1) Es handelt sich um Erfahrungsgehalte innerhalb spezieller kultureller Kontexte, weshalb wir es mit einem sozialen Lernen zu tun haben. (2) Praktische Überlegungen besitzen immer auch nicht-instrumentalistischen Charakter. Die Tatsache, dass ich allgemein gelernt habe, dass es falsch ist zu lügen oder zu betrügen, gibt mir nicht in jeder heiklen Situation eine Antwort auf die Frage, was ich tun soll. Aber auch grundsätzlich erwäge ich nicht nur Mittel, sondern auch Ziele. (3) Auch passiv aufgenommene Werte müssen sich einer kritischen Prüfungen unterziehen lassen; deren Grundlage sind emotionale Selbsterfahrungen mit Blick auf intersubjektive Werte innerhalb gemeinsamer kultureller Kontexte. Wir werden sehen, dass die involvierten Überlegungen weder instrumentellen noch logischen, sondern hermeneutischen Charakter besitzen.[8]

3. Eine defizitäre Theorie rationaler Kontrolle

Wie könnte die humesche Idee einer Bifurkation von konativen und kognitiven Aspekten überwunden werden? Ein erster Schritt wurde bereits durch die Betonung der zentralen Rolle emotionaler Erfahrungen skizziert. Der intersubjektive Bestimmungsgrund erfolgt dabei im Rahmen einer Bewertungspraxis: Praktische Signifikanz wird in Abhängigkeit nicht nur deliberativer Urteile, sondern indirekt, als Fundament praktischer Propositionen, in Abhängigkeit emotionaler Werterfahrungen beschrieben. Von Interesse ist dabei, wie Emotionen und Urteile Signifikanzkonstitution ermöglichen und sich dabei gegenseitig beeinflussen. Wenden wir uns hierzu dem Modell der rationalen Kontrolle unserer Wünsche zu, das Harry Frankfurt vorgelegt hat,

8 Vgl. Taylor, op. cit. (Fn. 3).

um sogleich auch seine Grenzen zu beleuchten. Wir werden sehen, dass bei Frankfurt die Idee fehlt, das Gute als das kritisch bewertete Gute zu wollen.[9]

Zunächst erweist sich Frankfurts Modell[10] als lehrreiches Beispiel. Seine Grundidee lautet: Personen sind Wesen, die in der Lage sind, emphatisch wollen zu können, was sie wollen. Sie besitzen die Fähigkeit, nur diejenigen Motive als tatsächlich aktivierende zu akzeptieren, die sie durch Identifikation als solche betrachten wollen. Doch Frankfurts «rückhaltlose Identifikation» mit bestimmten Wünschen kann kein hinreichendes Kriterium sein für Selbstbestimmung mit Blick auf das, was uns am Herzen liegt.

Betrachten wir hierzu Manipulationsfälle. Mit Frankfurt könnte man auch hier von «vorbehaltloser Identifikation» sprechen. Der Anhänger des totalitären Regimes mit seiner menschenverachtenden Ideologie etwa könnte ohne jede Wertekritik darauf beharren, sich nun einmal mit ganzem Herzen und also vorbehaltlos mit den Zielen des Unrechtstaates zu identifizieren. Ich denke, man kann einer solchen Person nicht trockenen Auges einen aufgeklärten Willen zusprechen, einfach deshalb, weil wir nicht behaupten sollten, dass eine vorbehaltlose Identifikation mit irgendwelchen Gründen ein Kriterium richtigen Handelns darstellt. Hier wäre eine darüber hinausgehende rationale Kontrolle in Abhängigkeit von Werten erforderlich, wobei gute Handlungsgründe vom drittpersonalen Standpunkt aus nachvollziehbar sein sollten. So wäre praktische Signifikanz als das Ergebnis einer potentiell selbstkritischen Reflexion aufzufassen. Die von Frankfurt veranschlagte interne Kohärenz mit Blick auf einen bestimmten Zeitpunkt kann daher nicht ausreichen, es müssen weitere Kriterien hinzutreten. Daher sollte sich eine personale Identifikation auch anhand kontextübergreifender (besser: dynamischer) und normativer Maßstäbe bewerten lassen. Eine subjektive situationsbezogene Identifikation könnte sich tatsächlich als inkohärent erweisen innerhalb eines situationsüberschreitenden Identitätsprofils eines Akteurs. Mit Blick auf das Manipulationsargument jedenfalls erweist sich Frankfurts Modell als defizitär.

Wir sehen, dass die Möglichkeit einer Überprüfung der eigenen Werte und Maßstäbe im Lichte potentiell nachvollziehbarer Argumente vom drittpersonalen Standpunkt bei Frankfurt ebenso fehlt wie das Kriterium einer

9 Vgl. Bennett Helm: *Emotional Reason. Deliberation, Motivation, and the Nature of Value* (Cambridge: Cambridge University Press, 2001) S. 161-199.

10 Vgl. Harry Frankfurt: *Freedom of the Will and the Concept of a Person*, in *Journal of Philosophy 68* (1971) S. 5-20.

kontextübergreifenden selbstkritischen Kohärenzprüfung.[11] Halten wir im Gegenzug fest, was wir von Frankfurt lernen: Personen zeichnen sich dadurch aus, dass sie den Unterschied kennen zwischen dem, was sie in ihrem Innersten wollen, und dem, was sie sich ab und zu oder auch fälschlicherweise oder im Modus der Sucht usw. wünschen. Dies aber, und das scheint mir der Kern des Defizits zu sein, hängt nicht nur an passiven Erfahrungen, sondern ist das Ergebnis irgendeiner Form selbst-kritischer Aktivität, die deutlich zu unterscheiden ist vom passiven Charakter erlebter Motive. Wir lernen bei Frankfurt also die Notwendigkeit der Unterscheidung zwischen passiver Motiverfahrung und aktiver Motivaneignung.

Sofern es uns um eine Konzeption der rationalen Kontrolle unserer Motive geht, stehen wir nun vor folgenden Fragen: (1) Wie bilden sich die Maßstäbe der «kritischen Selbstrevision»?[12] (2) Was leistet die Rede vom intersubjektiven Charakter von Gründen und Werten? (3) Welche Rolle spielt dabei die Überlegung? (4) In welchem Verhältnis stehen Überlegung und Werterfahrung?

Ich werde im folgenden Abschnitt erläutern, inwiefern Emotionen als passive Wertwahrnehmungen zu verstehen sind. Hierbei soll an einige aristotelische Intuitionen über den Zusammenhang von Werten und Haltungen erinnert werden. Die Vorstellung einer emotionalen Werterfahrung, die durch intersubjektive Habitualisierung ermöglicht wird, ist eine originär aristotelische Idee. Wie aber ist diese Konzeption mit der Idee einer selbstkritischen Willensbildung zu verbinden? Wir sollten diese Frage systematisch diskutieren, indem wir die Möglichkeit nicht-substantialistischer Formen praktischen Überlegens erörtern, die auch bei Aristoteles zu finden sind.[13] Oft wird den Verteidigern von Wertwahrnehmungen vorgeworfen, es handle sich hier um unkritische Habitualisierungen im Rahmen eines tugendethischen Theorierahmens. Das Konzept des emotionalen Logos ist als kritische Fortentwicklung der aristotelischen Sensibilitätstheorie aufzufassen, das Elemente autonomer Selbstkritik berücksichtigt.

11 Eine Diskussion der Defizite von Frankfurts Konzeption lege ich vor in: Johannes D. Balle: *Gründe wollen. Praktische Abduktion, Emotionale Erfahrung und die Intentionalität des Wollens*, in *Salzburger Jahrbuch für Philosophie* 53 (2008) S. 35-51.

12 Vgl. Marcus Willascheks lehrreichen Aufsatz: M. W.: *Was will ich wirklich? – Zum Zusammenhang zwischen Freiheit, Rationalität und praktischer Identität*, Frankfurt 2006 (http://www.trl-frankfurt.de/index.php?article_id=191&clang=0.

13 Vgl. Holmer Steinfath: *Orientierung am Guten* (Frankfurt a.M.: Suhrkamp, 2001).

4. Emotionale Wertwahrnehmung

Anders als bei «humeschen» Konzeptionen sollten wir Emotionen als eigenständiges Kriterium für Handlungserklärungen betrachten, das weder auf Wünsche noch auf Überzeugungen reduzierbar ist.[14] Emotionen sind der Motor der individuellen Existenz, doch bleibt festzuhalten, dass sie in ganz unterschiedlicher *Kultivierung* auftreten. Ihre Formung geschieht durch Reflexion, weshalb im Mittelpunkt unseres Interesses die Idee der wechselseitigen Abhängigkeit von Emotionen und Urteilen steht. Hierzu soll der zweifache «Gründecharakter» von Emotionen erörtert werden, als Direktrechtfertigung oder als Basen praktischer Abduktionen.

Praktische Urteile liefern uns unmittelbare Handlungsgründe, sofern sie durch Emotionen rationalisierbar sind. Dies geschieht oft *nicht-inferentiell*, indem in emotionalen Erfahrungen Werte direkt wahrgenommen werden. In anderen Fällen erfolgt die Rationalisierung durch Überlegungen. Dass umständliche Überlegungen nicht ständig praktikabel sind, dürfte einleuchten. Potentiell jedoch sollte ein Akteur seine Handlung mit guten Gründen rechtfertigen können.[15]

Achten wir zunächst auf die These, dass Emotionen oft systematisch relevant für situationsbezogene Urteile sind. Man könnte mit Aristoteles behaupten, dass praktisches Wissen stets erfahrungsabhängig ist und eingeübt oder gelernt werden will. Im Rahmen einer Werttheorie könnte man contra Kant und pace Aristoteles von der Erfahrungsabhängigkeit praktischer Propositionen sprechen. Im Gegensatz dazu könnte ein Kantianer behaupten, moralisches Wissen sei erfahrungs-unabhängig, weshalb er auf Prinzipien zurückgreife, deren Rechtfertigungsgehalt keine Erfahrungen impliziere. Ich plädiere für die Erfahrungsabhängigkeit praktischer Deliberation aufgrund der eingangs beschriebenen dynamischen Grundsituation menschlicher Pra-

14 Vgl. Sabine A. Döring, C. Peacocke: *Handlungen, Gründe und Emotionen*, in *Deutsche Zeitschrift für Philosophie* 4 (2002) S. 81-103.

15 Beide Rechtfertigungstypen greifen lebenspraktisch ineinander. Denken wir an die Aristotelische Idee einer klugen und tugendhaften Person: Das moralische Gutsein ihrer Handlung ist auf der Grundlage ihrer charakterlichen Erziehung zu begreifen. Eine tugendhafte Charakterbildung impliziert, in spezifischen Situationen Emotionen zu erleben, die als Mittel der Erkenntnis der moralischen Richtigkeit einer bestimmten Handlung dienen. Die praktische Erkenntnis erfolgt oftmals nicht-inferentiell, bisweilen wäre selbst ein «Phronimos» auf kritische Überlegungen angewiesen. McDowell sieht dies anders: Vgl. auch John McDowell: *Virtue and Reason*, in *The Monist* 62 (1979) S. 331-350.

xis und favorisiere eine Auffassung, welche die lebensweltliche Behauptung stützt, dass auch kontextbezogene Erfahrungen eine unverzichtbare Basis praktischer Deliberation darstellen. Worin liegt nun genau die Bedeutung von Emotionen für Werterfahrungen?

Humesche Konzeptionen ziehen es vor, Handlungen unter Zuhilfenahme des Zweikomponentenmodells zu erläutern, das sich aufgrund seiner Eleganz großer Beliebtheit erfreut. Indem Emotionen auf Wünsche reduziert werden, verlieren sie den Status einer eigenständigen Kategorie für Handlungserklärungen. Dabei dient die spezifische Passensrichtung als Kriterium ihres motivationalen Charakters. Eine Handlung wird vollzogen, weil die handelnde Person einen bestimmten Wunsch hegt, der wiederum mit einer Überzeugung in Zusammenhang steht, die darüber aufklärt, mit welchen Mitteln der Wunsch in Anbetracht der Weltsituation zu verwirklichen ist. Damit gilt, dass im Zusammenhang mit einer Überzeugung jeweils ein Wunsch den Grund bildet, wobei dieser Wunsch nicht nur den Grund, sondern auch die Ursache der Handlung darstellt.

Hiergegen wollen wir behaupten, dass Emotionen eine eigenständige rationalisierende Funktion besitzen, die durch das Passensmodell nicht dargestellt wird. Stellen wir uns die Schulangst eines Kindes vor. Das Gefühl basiert auf realen Erfahrungen. Es wäre unangemessen, diesem Erfahrungsgehalt schon deshalb eine Rolle in Handlungserklärungen abzusprechen, weil wir ihn für subjektiv halten. In Wahrheit besitzt der emotionale Gehalt des Kindes *repräsentationalen Wert*, unabhängig von Wünschen und Überzeugungen. Das Kind hält diese Emotion und das, was sie repräsentiert, für wahr, weil sie auf Tatsachen fundiert ist.[16]

In einer Hinsicht sind Emotionen mit *Sinneswahrnehmungen* vergleichbar aufgrund ihres intentionalen Gehaltes.[17] Wie Wünsche richten sich Emotionen auf Gegenstände und besitzen intentionale Objekte. Infolge ihres intentionalen Gehaltes rechtfertigen sie wie Sinneswahrnehmungen in *nicht-inferentieller Weise*. Die durch emotionale Erfahrungen angezeigten Gründe rationalisieren Handlungen und *bewegen* zu praktischen Urteilen. Doch sind Emotionen nicht rein *dispositional* zu analysieren, da sie auch *bewertende Elemente* bezüglich des involvierten Gegenstandes beinhalten. Meine Scham vor Nacktheit in der Öffentlichkeit impliziert meine *Überzeugung*, dass es

16 Vgl. Sabine Döring: *Kann Willensschwäche rational sein?*, in *Gehirne und Personen*, hg. von Martina Fürst, Wolfgang Gombocz, Christian Hiebaum (Heusenstamm: Ontos, 2008) S. 55-71.

17 Döring/Peacocke, op. cit. (Fn. 14) S. 81-103.

für mich unangemessen ist, in dieser Weise aufzutreten. Es gibt Gründe dafür. Daher sind emotionale Erfahrungen durch dispositionale Analysen nicht erschöpfend darzustellen. Die zugeschriebenen Eigenschaften besitzen impliziten Charakter, so dass es nicht erforderlich wird, dem intentionalen Objekt der Emotion alle Merkmale aktual und bewusst zuzuschreiben.[18]

Andererseits ist der emotionale Gehalt kein *Werturteil*, denn offenbar ändert sich meine emotionale Situation nicht zwingend durch *bessere Einsicht*. Ein starker Kognitivismus wie etwa bei Nussbaum oder Salomon wäre hier abzulehnen.[19] Die entsprechende Veränderung bedarf stattdessen der Übung und Gewöhnung, was an die Aristotelische Idee einer klugen Lebenspraxis im Sinne des *«Habitus»*[20] erinnert und ins Zentrum unserer Überlegungen weist: Emotionen als Werterfahrungen sind keine Überzeugungen, die man einfach ändern könnte, sondern es bedarf ihrer allmählichen *Formung* unter Leitung eines emotionalen Logos. Darin gleichen sie Wahrnehmungsgewohnheiten, die man nur schrittweise im Ausgang doxastischer Inhalte verändern kann. Sofern man Emotionen daher im Zusammenhang mit Gründen und Werten diskutiert, entdecken wir im Kern des «emotionalen Logos» das Prinzip des *Lernens aus Selbst-Erfahrungen*, das sich in *praktischen Haltungen* niederschlägt.

Das formale Objekt[21] der Emotion wird vom Akteur in einer gewissen Weise *bewertet*, was wiederum der Zuschreibung einer Reihe von Eigenschaften entspricht, die den intentionalen Gehalt der emotionalen Erfahrung bestimmen. Emotionen dienen als rationalisierende Gehalte und sind nichtinferentielle Mittel für Handlungserklärungen. Ihre motivationale Kraft indessen verdanken sie ihrer *Phänomenalität*. Emotionen sind in einer gewissen Hinsicht, strukturell betrachtet, *phänomenale Perzeptionen und besitzen eine phänomenale Intentionalität:*[22] Emotionen repräsentieren Welt-

18 Vgl. Johannes D. Balle: *Gefühlte Gründe. Emotionale Erfahrung in einer Theorie praktischer Propositionen* (Essen: Deutscher Kongress für Philosophie XXI, Sektionsbeitrag).

19 Vgl. Nussbaum, op. cit. (Fn. 1) und Solomon, op. cit. (Fn. 1).

20 Aristoteles' Erfahrungsverständnis leitet sich aus den grundlegenden Ethismos-Erfahrungen ab; die Gewohnheit im besten aus-gebildeten(!) Sinne bildet den Gehalt der «Hexis», wobei vor allem der Aspekt der Einübung betont wird. Vgl. Aristoteles: *Nikomachische Ethik*, VIII 11, 1152a 31 ff.

21 Vgl. Kenny, Anthony: *Action, Emotion and Will* (London: Routledge, 1963).

22 Attraktiv ist die phänomenologische Idee, evaluative Eigenschaften als «Resultanzeigenschaften» (vgl. hierzu Jonathan Dancy: *Moral Reasons* [Oxford: Blackwell, 1993] S. 74, 76, 78) im Sinne der Gestaltwahrnehmung aufzufassen. Dieses

zustände. Sie besitzen *intentionalen Gehalt*, der die Welt in wahrer oder falscher Weise beschreiben kann. Das formale Objekt der Emotion indessen ist nicht notwendigerweise ein Bestandteil ihres repräsentationalen Gehaltes.[23] Emotionen besitzen keine implizit *doxastische Struktur* wie Werturteile. Sie weisen gleichermaßen phänomenologische und intentionale Aspekte auf.[24]

Sofern Emotionen als evaluative Wertwahrnehmungen aufgefasst werden, besitzen auch evaluative Eigenschaften phänomenalen Charakter. McDowell ist der Ansicht, es handle sich bei evaluativen Tatsachen um «sekundäre Eigenschaften».[25] Diese Redeweise impliziert eine andere wichtige These: Durch unsere Emotionen lernen wir die Welt *tatsächlich* kennen. Sie fühlen sich nicht nur in irgendeiner Weise an, sondern besitzen tatsächlich informativen Gehalt und das eine ist nicht ohne das andere zu haben. Wenn uns durch Emotionen Eigenschaften der Welt zugänglich werden, dann stellen sie unabdingbare Quellen unserer Welterkenntnis dar. Sie repräsentieren die Welt als Kulturwelt.[26] Bedeutsamkeiten als Werterfahrungen sind dann Bestandteile einer gemeinsamen Kulturwelt. Werterfahrungen hängen demnach nicht einseitig vom Subjekt und dessen Projektionen ab, sondern sind das Ergebnis der intersubjektiven Praxis des Gründegebens.[27]

Damit wird aber auch deutlich, dass es sich hier um eine moderate kognitivistische Variante handelt: Als Kulturprodukte repräsentieren Werterfahrungen evaluative Tatsachen, die keineswegs einseitig subjektabhängig zu konzipieren sind. Gegen projektionistische Theorien wäre diese Variante insofern im Vorteil, als die Möglichkeit eines systematischen Irrtums durch das Instrument der rationalen Überprüfung durch Dritte verhindert würde. Hier kommt das Element der praktischen Überlegung ins Spiel, das nachfolgend dargestellt wird. Sofern gefühlsbasierte Wertungen eines Akteurs einer

Modell ist mit einem moderaten Kognitivismus kompatibel. Doch in welchem Sinne implizieren Emotionen Begriffe? Ich favorisiere eine gestalttheoretische Konzeption «begrifflicher Qualia» als kognitive Phänomenologie.

23 Vgl. Ronald de Sousa: *The Rationality of Emotion* (Cambridge, Mass.: Harvard University Press, 1987) S. 122-123.

24 Vgl. Peter Goldie: *The Emotions. A Philosophical Exploration* (Oxford: Clarendon Press, 2000).

25 Vgl. John McDowell: *Values and Secondary Qualities*, 1985, in J. M.: *Mind, Value, and Reality* (Cambridge, Mass.: Harvard University Press, 1998) S. 131-150.

26 Vgl. ibid. S. 134.

27 Vgl. McDowell über «Bildung» in einer aristotelischen Konzeption: John McDowell: *Mind and World* (Cambridge, Mass.: Harvard University Press, 1994) S. 87-88, 123-124.

intersubjektiven Prüfung nicht standhalten, wären wir im Recht, wenn wir die Angemessenheit der evaluativen Einstellung bezweifelten. Diese Idee einer sozialen Konstitution evaluativer Begriffe durch die Wertgemeinschaft spielt nicht nur in zeitgenössischen Theorien eine Rolle,[28] sondern verdichtet sich in der aristotelischen Habituslehre.[29]

5. Praktische Haltungen

Die Unterscheidung von rationalen und motivierenden Elementen wirft im Zusammenhang mit emotionalen Erfahrungen Fragen auf. Nachdem wir sahen, welchen Stellenwert Emotionen für Werterfahrungen besitzen, und auf diesem Wege die klassische Unterscheidung von kognitiven und konativen Bestandteilen modifizierten, suchen wir nach einer alternativen Konzeption. Ich werde hierzu die aristotelische «Hexis» als «praktische Haltung»[30] deuten, deren Funktion die «logosartige» Formung von Emotionen darstellt. Damit nähern wir uns der für unsere Zwecke zentralen Frage nach dem Zusammenspiel von Werterfahrungen und Deliberation. Da wir bereits den Emotionen sowohl motivierenden als auch rationalisierenden Charakter zusprachen, wird eine Erklärung notwendig, weshalb hier ein weiteres rationalisierendes Element in Form praktischer Überlegungen berücksichtigt werden soll.

Bei Aristoteles finden wir zunächst eine Antwort auf die Frage, wie die Normativität praktischer Signifikanz zu verstehen ist. Aristoteles verweist auf das anerzogene Ethos einer Gemeinschaft. Durch die entsprechende Erziehung gewinnen Akteure praktische Haltungen, deren Charakter essentiell sozial ist. Aufgrund ihres gemeinschaftlichen Charakters erwei-

28 Vgl. Goldie, op. cit. (Fn. 24).

29 Ein ähnliches Projekt verfolgen Helm, op. cit. (Fn. 9), und Jan Slaby: *Gefühl und Weltbezug* (Paderborn: Mentis, 2008). Anders als bei Helm verfolge ich eine intersubjektivistische Konzeption; anders als bei Slaby steht die Idee einer emotionalen Kultivierung im Zusammenhang einer kritischen Sensibilitätstheorie im Zentrum des Interesses, wobei das Herzstück dieser emotionalen Autonomiekonzeption eine kognitive Phänomenologie der Bewertung ist, die Antwort auf die Frage gibt: Was bedeutet es, das Gute zu sehen?

30 Man kann den «habitus» als «praktische Haltung» bezeichnen, wobei dieser Begriff in meiner Verwendungsweise lebensweltliche praktische Orientierungsfähigkeiten umfasst, die im Umkreis der Phänomenologie Edmund Husserls und Martin Heideggers berücksichtigt wurden.

sen sich Haltungen als träge oder passiv, wenn auch als revisionsfähig: Ein «Phronimos» besitzt eine für ihn typische Klugheit («Phronesis»), die als Verstandestugend die Fähigkeit darstellt, gut begründete Entscheidungen zu treffen, die über die vorherige Charakterformung insofern hinausgehen, als es sich dabei um ein essentiell kontextbezogenes Urteil handelt. Bei Aristoteles handelt es sich um eine Praxis des Gründegebens auf Grundlage eines gemeinschaftlichen Logos. Die Basis der Bewertung stellt eine kulturelle Praxis dar, deren Regeln und Formen zunächst nicht von einem externen Standpunkt aus zu bewerten sind. Um sie zu beurteilen, sollte man sie verstehen, was wiederum erfordert, ihre Praxis aktiv zu beherrschen und emotional habitualisiert zu haben. Das entsprechende Kriterium ist also bei Aristoteles ein internes und der Praxis inhärentes, weshalb die Bewertungsmaßstäbe notwendigerweise Bestandteile derselben darstellen. Kritik an einer Entscheidung bezieht sich immer auf Vorstellungen, die innerhalb der betreffenden Praxis verständlich werden. Dabei besitzen die Bewertungskriterien emotionalen und begrifflichen Charakter, die einzelnen Formen sind erst im holistischen Zusammenspiel des ganzen Ethossystems zu verstehen. So wird klar, dass Aristoteles als Vater einer intersubjektiven Konstitutionstheorie zu betrachten ist. Die Standards praktischer Signifikanz bemessen sich an der kulturellen Praxis. Es ist diese aristotelische Idee, der sich die eingangs getroffene Behauptung schuldet, es handle sich bei der hier favorisierten Version weder um einen Subjektivismus noch um einen Objektivismus.

Personen erfahren Werte, die sie durch Erziehung, Sozialisation und Bildung kennen und schätzen lernten. Wir sahen, dass emotionale Werterfahrungen motivationalen und rationalen Charakter besitzen. Andererseits sollten wir auf die Idee einer Kritik der eigenen Gründe nicht verzichten, denn die Gefahr einer Irrationalität lauert in zwei Richtungen: Folgen wir blind der ersten Quelle, so führt das bisweilen zu erheblichen Unstimmigkeiten, etwa wenn wir in heiklen Fällen ungeprüft den Maximen unserer Erziehung folgen. Setzen wir indessen auf die zweite Quelle des praktischen Überlegens, so würde eine dermaßen übertheoretisierte Konzeption zu Recht irritieren. Eine kritische Sensibilitätstheorie sollte beide Elemente zusammendenken: Das kritische Potential einer Prüfung der habitualisierten Direktrechtfertigung bildet ein notwendiges, wenn auch gewiss nicht alltägliches Korrektiv zur Ermittlung praktischer Gründe. Dabei hängen Urteile von emotionalen Erfahrungen ab, die ihnen ihrerseits eine individuelle Form verleihen. Der emotionale Logos ist die treibende Kraft einer moderaten aristotelischen Konzeption: Personen sind Wesen, die sich darum bemühen, praktische

Bedeutung, die ihrerseits Gründe fundiert, in Form von Deliberation auf Grundlage emotionaler Erfahrungen zu prüfen.[31]

Dieses modifizierte «aristotelische» Modell könnte als Alternative zu zeitgenössischen Modellen rationaler Kontrolle betrachtet werden, weil hierbei die Idee einer intersubjektiven Konstitution durch ein Modell individueller kritischer Selbstrevision erweitert wird. Anders als bei Frankfurt gibt es sowohl bei Kant als auch bei Aristoteles diesen Freiraum selbstkritischer, normativer und kontextbezogener Überlegungen.[32]

Nun betont freilich auch Frankfurt den Aspekt der Aneignung von Motiven. Doch Aristoteles, Kant und neuerdings etwa Michael Bratman mit dessen Betonung einer vernünftigen Stabilität[33] setzen auf den kritischen Maßstab, der nicht einfach willkürlich vorgegeben ist und wieder verworfen werden kann. Anders als bei Kant jedoch, der die Gleichsetzung des normativen Maßstabes mit dem allgemeinen Sittengesetz fordert, finden wir bei Aristoteles die Ansätze einer Konzeption des emotionalen Logos, die auf der Grundlage kulturspezifischer Sensibilitäten eine kluge Prüfung der Situation sowie der eigenen Emotionen und Einstellungen zulässt. Aristoteles spricht von der kritischen Verantwortung des Individuums für seinen Habitus und

31 Das «Sehen» des Guten bietet folgende Aspekte: Das Sehen guter Gründe in Abhängigkeit von Erfahrung und Erziehung sowie das prospektive Sehen der Situation ihrer Möglichkeit nach – ein zugleich affektives und rationales Sehen. Peter Goldie spricht von einem «personal point of view» und meint damit ebenso die Einheit des Empfindungsvermögens und begrifflicher Kompetenzen: Personen sind Wesen, die über die richtigen Begriffe als auch über affektive Dispositionen verfügen, um so die Anerkennung als Mitglied einer Kulturgemeinschaft zu verdienen. Sie werden in Evaluationskollektive hineinsozialisiert und ihre kultivierte Befähigung zur Orientierung am Guten zeigt sich im emotionalen Logos, der Sensibilität und Urteilskraft impliziert. Der Aristotelische «Phronimos» sieht das Gute, denn er verfügt über ethische Erziehung sowie dianoetische Fähigkeit, wobei die moralische Erziehung im Resultat eine emotionale Sensibilität darstellt, die situativ in kluger Weise zum Tragen kommt. Aristoteles beschränkt Überlegungen dabei nicht auf instrumentalistische Formen, sondern berücksichtigt ebenso zweckegenerierende Rationalität. Der «Phronimos» vermag eigenständige Überlegungen anzustellen, doch gleicht es einer Karikatur zu behaupten, der Akteur müsse in jeder Situation neu deliberieren.

32 Vgl. Ralf Elm: *Erfahrung und Klugheit bei Aristoteles* (Paderborn, München, Wien, Zürich: Schöningh, 1996) S. 263-272.

33 Vgl. Bratmans instruktive Kritik an Frankfurt: Michael Bratman: *Eine überlegte und vernünftige Stabilität*, in: Harry Frankfurt: *Sich selbst ernst nehmen* (Frankfurt a.M.: Suhrkamp, 2007) S. 109-124.

erweitert damit das Spektrum praktischer Rationalität auf substantielle Zielüberlegungen.[34] Dabei sind die Darstellungen des Aristoteles anthropologisch angemessener als Kants rigider Rückgriff auf ein formales allgemeines Sittengesetz: Aristoteles sieht lediglich ein kluges situationsbezogenes Beratschlagen geformter Emotionen und Einstellungen vor. Dass der «Phronimos» dabei einem Idealziel entspricht, passt gut zu unserer Intuition, dass gute Gründe nicht einfach zu finden sind, sondern einer geübten deliberativen Befähigung bedürfen. Der normative Maßstab, der contra Kant[35] nicht

[34] «Prohairesis» und «Bouleusis» gehen «pros to telos» und werden üblicherweise instrumentalistisch dargestellt. In der *Nikomachischen Ethik* VI und VII sowie in *De Anima* III 7 ist jedoch auch von Zielüberlegung die Rede. In der *Nikomachischen Ethik* 1114 a 7-13 wird die Verantwortung für die eigene Haltung betont, die über instrumentelle Vernunftgründe hinausgeht. Vgl. David Wiggins: *Deliberation and Practical Reason*, in: D. W.: *Needs, Values, Truth* (Oxford: Clarendon Press, ^{3}1998) S. 215-237.

[35] Man sollte in anderem Zusammenhang Stärken und Schwächen der kantischen Konzeption diskutieren. Sofern wir hier die Frage der emotionalen Kultivierung verhandeln, nur so viel: Aristoteles ist optimistischer als Kant bezüglich der Frage, ob der Mensch dem moralisch Guten gerne nachkommt. Offenbar lassen sich das individuelle Glück und das der Gemeinschaft bei Aristoteles nicht trennen. Ich schließe mich ausdrücklich Andreas Trampota an, der gegen Nancy Sherman (*Kantian Virtue. Priggish or Passional?*, in A. Reath, B. Hermann, C. M. Korsgaard: *Reclaiming the History of Ethics. Essays for John Rawls*, [Cambridge: Cambridge University Press, 1997]) bei Kant von einer «Überformung» statt von «Kultivierung» der Gefühle spricht und letzteres bei Aristoteles erkennt. Nach Sherman betont Kant den aktiven Wahlcharakter moralischer Emotionen. Fraglich ist, ob Kant auch behauptet, dass uns ausgebildete Gefühle Werte und Problemzusammenhänge zugänglich machen und so die scharfe kantische Trennlinie zwischen Verstand und Emotionen zurückgenommen wird. Diese These von Sherman lehne ich mit Andreas Trampota ab: «Kant beschreibt die Empfindsamkeit als ein rationales Vermögen, dessen Aufgabe es ist, das Quantum an passivem, non-kognitivem Affiziertwerden im Dienste der Moral zu steuern. Mit ihr sind keinerlei Gefühle rationaler Natur verbunden, die einen Einfluss auf die Qualität unserer Wahrnehmung der Wirklichkeit haben. Die Rede vom Kultivieren von Gefühlen impliziert also bei Kant nicht, dass Gefühle an der Vernunft partizipieren und deshalb in dem Sinne kultiviert werden können, dass sie uns beim Unterscheiden der sittlich relevanten Eigenschaften unterstützen, indem sie uns auf spontan-unreflektierte Weise auf Güter aufmerksam machen, die in die Erwägungen der Vernunft einbezogen werden müssen. Das Kultivieren von Emotionen im Dienst der Moralität meint bei Kant nur, dass die für die Emotionen kennzeichnende passive Empfänglichkeit, das Affiziertwerden-können durch die Gefühle von Lust und Unlust, gezielt im Dienst der

starr und unveränderlich vorgegeben ist und stattdessen erfahrungsoffenen Charakter besitzt und an attraktivem vorbildhaften Verhalten ausgebildet wird, ist dabei contra Hume nicht identisch mit Wünschen, sondern bildet ein kritisch angeeignetes Kriterium. Die Schemata der Wunschbewertung sind also meine, sofern sie in der jetzigen Form nicht nur das Ergebnis von Erziehung und Umwelt sind, sondern, wie Willaschek zu Recht fordert,[36] von selbstkritischen Revisionen abhängen, die ich im Lichte früherer Erfahrungen und Überlegungen vornahm. Eine Konzeption der kritischen Beurteilung der Maßstäbe auf Grundlage habitualisierter Sensibilitäten bietet das Modell der praktischen Abduktion.

6. Praktische Abduktion

In praktischen Überlegungen erörtern Personen nicht nur, auf welche Weise vorgegebene Zwecke zu realisieren sind, sondern sie dienen darüber hinaus auch der Ausbildung ihrer Zwecke. Der instrumentalistischen Sichtweise wäre eine substantialistische hinzuzufügen, wie dies in Ansätzen auch bei Aristoteles zu finden ist: Personen kalkulieren nicht nur die Mittel zu vorgegebenen Zwecken, sondern bilden selbst Ziele, Zwecke, Maßstäbe, Absichten aus.[37]

Welche Rolle spielen Emotionen im Rahmen solcher Reflexionen? Nach den bisherigen Ausführungen kann man sagen, Emotionen dienten als Indikatoren für Bewertungen. Deliberation ist auf Gefühle angewiesen, da diese *zeigen*, welchen Wert eine Sache *für die Person* besitzt. Ohne Gefühle wüsste der Akteur in vielen Fällen nicht, was zu tun ist. Das bedeutet, Emotionen geben minimal eine *Bewertungsrichtung* ab und verbinden individuelle Ansprüche mit normativen Erfordernissen.

Moralität eingesetzt wird. Deshalb ziehe ich es vor, im Zusammenhang mit Kant nicht von einem Kultivieren von Emotionen zu sprechen, sondern beschränke den Gebrauch dieser Redeweise auf eine Theorie der Emotionen aristotelischen Typs, welche die Möglichkeit einer Vervollkommnung unseres Gefühlslebens qua seiner impliziten Rationalität kennt.» Andreas Trampota: *Autonome Vernunft oder moralische Sehkraft? Das epistemische Fundament der Ethik bei Immanuel Kant und Iris Murdoch* (Stuttgart: Verlag W. Kohlhammer, 2003) S. 79 (F. 297).

36 Vgl. Willaschek, op. cit. (Fn. 13).

37 Vgl. Johannes D. Balle: *Indexikalität, kognitive Dynamik und praktisches Überlegen*, in *Beiträge der Österreichischen Ludwig Wittgenstein Gesellschaft* 12 (Kirchberg 2004) S. 25-27.

Praktisches Entscheidungsvermögen setzt voraus, dass die Person das Urteil, das durch den intentionalen Gehalt der emotionalen Erfahrung gerechtfertigt wird, erstens für wahr hält und zweitens als *passende Handlung* mit Blick auf eine Situation sowie über diese hinaus fällt. Dabei haben praktische Propositionen unterschiedliche Erfüllungsgrade. Nicht in jedem Fall ist eine Handlung, welche durch emotionsbasierte Urteile nahegelegt wird, auch ratsam und wird von der Person als insgesamt passend empfunden. In solchen Fällen kommt man nicht umhin, über eine Revision der bisherigen Zwecke oder gar Maßstäbe nachzudenken. Wenn etwa jemand infolge seiner Kinderliebe das erste Mal ein Gefühl der Abneigung gegen einen menschenverachtenden Staat verspürt und diese Emotion für angemessen hält, so hat er damit einen veritablen Grund, seine Kinder aktiv vor diesem Staat zu schützen. Emotionen zeigen hier neue Bewertungsmaßstäbe auf. Eine erste Bedingung der praktischen Indikatorfunktion von Emotionen mit Blick auf Überlegungen könnte daher lauten: Akteure sollten *potentiell* in der Lage sein, ein entsprechendes Urteil auf der Grundlage ihrer Emotion mit Blick auf die Entscheidungssituation zu formulieren.

Das Akzeptanzkriterium praktischer Gründe fordert, dass Urteile nicht nur in unkritischer Weise auf Grundlage von Emotionen getroffen werden. *Situationsbezogenheit* und *kontextübergreifende Aspekte* fließen gleichermaßen mit ein, weshalb emotionale Erfahrungen, Pläne, Situationswahrnehmungen und die praktische Identität der Person als Bedingungen praktischen Urteilens zu nennen sind. Es sind gerade diese Aspekte, über die ein Phronimos gut zu beratschlagen weiß, um eine situations-, profil- und ethosadäquate Entscheidung zu treffen. Praktischen Urteile dieser komplexen Art können nicht einfach deduziert oder induziert werden, sie sind *hypothetische Erweiterungsschlüsse aufgrund komplexer Erfahrungsgehalte*. Dies gilt insbesondere für Handlungssituationen, die neuartige Herausforderungen darstellen. Dass der Akteur einerseits in der Lage ist, seinem Gefühl «guten Gewissens» zu vertrauen, hängt, wie wir sahen, mit dem kulturellen Status der Wertwahrnehmung innerhalb einer Gemeinschaft zusammen. Das entsprechende praktische Urteil jedoch, welches ein Gefühl im Falle von Divergenzen und Widersprüchen als einen Handlungsgrund autorisiert, geht darüber hinaus und ist kein logischer Schluss, sondern eine Sinnhypothese über die individuelle Bedeutung einer Erfahrung in einer Situation. Man könnte dies eine «praktische Abduktion» nennen.

Die abduktive Überlegung ist auf emotionalen Erfahrungen fundiert. Personen erfahren Gründe, doch der Modus ihres Gründe-Erkennens ist nicht zu verwechseln mit ihrem Inhalt. Personen nehmen in dieser Weise

echte Gründe wahr. Oft stellt sich im Nachhinein heraus, dass es *für die betreffende Person* richtig war, so zu handeln. Wenn Gefühle uns auch (wie Sinneswahrnehmungen) bisweilen über den wahren Zusammenhang der Dinge täuschen, so ist doch ebenso wahr, dass wir der Ansicht zuneigen, dass es oft gute Gründe für emotionsbasiertes Handeln gibt. Gefühle leisten also trotz ihres opaken Charakters einen wertvollen Beitrag in praktischen Überlegungen. Eine systematische Beschreibung praktischer Rationalität im Rückgriff auf *gefühlte Gründe* könnte sich daher als ein erfolgversprechender Weg herausstellen. Ein erster Schritt in diese Richtung lautet: Intelligente bzw. kultivierte Gefühle repräsentieren in bestimmten Kontexten echte Gründe, die vom Akteur im Modus des abduktiven Urteils mit Blick auf Erfordernisse seiner praktischen Identität über Kontexte hinweg zu entwickeln sind. Es geht also um ein Zweifaches: Erstens die Herausbildung praktischer Signifikanz als rationale Formung dessen, was uns am Herzen liegt; zweitens die entsprechende prospektive Beurteilung einer Situation vor diesem Hintergrund mit dem Ziel der passenden Handlungswahl. Im Kern handelt es sich hierbei m. E. um begriffliche Erlebnisse, weshalb eine Konzeption des emotionalen Logos als Teil einer kognitiven Phänomenologie der Werte zu betrachten ist

Wir sehen, dass Akteure durch Emotionen rechtfertigende Gehalte zugleich als motivierende erfahren. Die Maßstäbe des Handelns leuchten unmittelbar ein, etwa infolge von Erziehungsstandards, die «in Fleisch und Blut» übergegangen sind. Praktische Abduktionen sind Formen *potentieller kritischer Selbstrevision*. Sie sind auf emotionale Erfahrungen und praktische Sensibilitäten fundiert. Man könnte sagen, dass sie uns das Auffinden passender Bewertungsmuster ermöglichen, diese Muster sollten zur praktischen Identität passen, weshalb sie vom Individuum nicht einfach frei erfunden werden. Sie sind hermeneutische Schlüsse auf die beste Erklärung in Anbetracht der eigenen Sensibilitäten sowie den Erfordernissen einer spezifischen Situation. Es leuchtet ein, dass sich eine solche Musterwahrnehmung nicht nur an internen Kohärenzkriterien bemisst, die entsprechenden Maßstäbe sollten darüber hinaus auch einer kritischen Prüfung vom *drittpersonalen Standpunkt* standhalten. Gefühlsbasierte Abduktionen ermöglichen es hier, die beiden Intuitionen zu verbinden, von denen wir anfangs sprachen: Wertwahrnehmung und Deliberation. Ein Modell der rationalen Wahl, wie wir es etwa bei Frankfurt finden, weist zwar die richtige Richtung, wird aber erst durch das Konzept des emotionalen Logos vollständig: Kritische Intersubjektivität und Dynamik auf der Grundlage emotionaler Erfahrungen erweitern die Idee einer rationalen Formung

unserer Affekte und Wünsche, indem hier die Ansätze einer *Konzeption emotionaler Autonomie* zum Tragen kommen.

7. Praktischer Sinn

Praktische Urteile erfolgen in Abhängigkeit von intelligenten Emotionen, denen Werterfahrungen zugrunde liegen. Sofern wir mit Aristoteles der Ansicht sind, dass Personen diejenigen Wesen sind, die den Logos als gemeinschaftliche Fähigkeit des «Gründegebens» verstehen, öffnet sich ein Blick auf die Quellen des «praktischen Sinns»: Wertsensibilität, Selbstverstehen und Deliberation.[38] Dabei orientiert sich der emotionale Logos nicht einseitig an kontextübergreifenden Maßstäben, sondern vermittelt diese «dynamisch» mit der klugen Situationsbeurteilung. Gerade der Kluge weiß um die Freiheiten, die sich nicht nur auf die Verwirklichung von Zwecken beschränken, sondern auch auf die «Herzensfreiheit»[39] als Kultivierung emotionaler Bedeutsamkeiten. Wir sahen: Wenn Personen in praktischen Abduktionen neue Maßstäbe entwickeln und alte prüfen, hängt dies auch von der gemeinschaftlichen Praxis ab. Die sinnerschließende Dimension «gefühlter Gründe» zeigt sich, wenn Altes überdacht und in Anbetracht neuer Konstellation neu beurteilt wird, wobei es sich weder um induktive oder deduktive noch um instrumentelle Überlegungen handelt, sondern um die Ausübung hermeneutischer Fähigkeiten.[40] Man sollte den «praktischen Sinn» als eingeübte Fähigkeit des emotionalen Logos verstehen, Absichten, Werte, Selbstverständnis und Kontextwahrnehmung miteinander zu vermitteln, damit gute Gründe ermittelt sowie Wege ihrer Verwirklichung aufgezeigt werden.

Damit gibt es eine Antwort auf die eingangs gestellte Frage nach den Gründen unseres Strebens: Sofern es sich um ein kultiviertes Wollen handelt, lernt man die Dinge so zu sehen, dass das eigene Wollen möglichst ein kritisch-kultiviertes ist. Es stellt eine Lernaufgabe dar, praktische Gründe in Abhängigkeit derjenigen Werte zu sehen, die mein Selbstverständnis strukturieren. Doch diese guten Gründe stehen in motivationaler Abhängigkeit von jenem Guten, das dem gemeinschaftlichen Wohl dient.

38 Vgl. Elm, op. cit. (Fn. 32) S. 252.
39 Vgl. Helm, op. cit. (Fn. 9) S. 161-199.
40 Vgl. Friederike Rese: *Praxis und Logos bei Aristoteles* (Tübingen: J.C.B. Mohr, 2003) S. 103-130.

Sind wir damit am Ende weitergekommen oder bleibt die platonische Frage ungelöst? Man könnte sich auf den Standpunkt stellen, dass wir unter einer Bedingung einen Fortschritt erzielten: Sofern Personen ihren *Möglichkeiten* nach beurteilt werden, greifen sie potentiell nach jenem Guten aus, das ihr egoistisches Wollen transzendiert. Im Mittelpunkt einer solchen *normativen Konzeption* steht die Idee der moralischen Entwicklung und emotionalen Kultivierung. Dafür bedarf es der Bildung eines emotionalen Logos, der die Kulturform unseres Wollens darstellt und der Logik des Vorbildes folgt, das uns sehen lässt, was gut ist. In diesem Sinne könnte Gutes zu tun gleichbedeutend sein mit einem Leben in Eudaimonia und Autarkia.

Studia philosophica 68/2009

Christine Clavien

Jugements moraux et motivation à la lumière des données empiriques*

This paper contains an 'affective picture': a story, extensively supported by empirical data, about the way I take people to judge and behave morally; a picture in which the respective roles of reflective and affective processes are explained. According to this picture, different sorts of judgements have to be distinguished, some being cognitively more complex than others. 'Sophisticated judgements' are displayed at the level of rational considerations and allow for moral thinking, whereas 'basic value judgements' are a primitive and non-reflective way of assessing the world and are motivating. As we shall see, this affective picture has some consequences for the traditional internalism-externalism debate in philosophy; it highlights the fact that motivation is primarily linked to 'basic value judgements' and that the norms and judgements we openly defend do not have a particular effect on our actions, unless we are inclined to have an emotional attitude that conforms to them.

Key words: affect, emotion, emotional reaction, externalism, internalism, justification, moral judgement, motivation, norm, value.

1. Le problème de la motivation

Raymond croit en l'obligation morale d'aider toute personne en détresse. En se promenant un dimanche matin, il voit un homme étendu de tout son long sur le trottoir. Il est clair que l'homme a été battu et qu'il souffre horriblement. Alors que Raymond aurait le temps et les moyens de se porter en aide, il décide sans raison particulière de passer son chemin. Ce faisant, il est parfaitement conscient du fait qu'il transgresse une norme morale à laquelle il souscrit ; en d'autres termes, il agit consciemment à l'encontre de ce qu'il pense devoir faire.

* Mes remerciements vont à Otto Bruun, Julien Deonna, Ronald de Sousa, Julien Dutant, Curzio Chiesa, Chloe FitzGerald, Jean Gayon, Peter Goldie, Daniel Schulthess, et Fabrice Teroni pour avoir pris le temps de critiquer et discuter les idées présentées dans cet article.

Cet exemple est assez troublant. Quoique parfaitement crédible, il contredit l'idée souvent défendue que le fait d'accepter une norme et les jugements qui en découlent, est un facteur motivant à l'action. C'est du moins l'avis des défenseurs de la position internaliste selon laquelle les considérations morales – en particulier le fait de produire un jugement moral – motivent les sujets à agir conformément à ces considérations. Un internaliste[1] défend l'idée qu'il existe une connexion interne ou nécessaire entre nos jugements moraux et la motivation à produire les actions prescrites par ces jugements. Cette position est compatible, soit avec le point de vue selon lequel les jugements moraux motivent par eux-mêmes, ou avec le point de vue selon lequel ils sont nécessairement mais indirectement connectés à la motivation.

Le comportement de Raymond entre en contradiction directe avec l'internalisme et semble confirmer la position externaliste selon laquelle, même si l'on peut établir une relation entre le jugement et la motivation, il s'agit d'une connexion externe et contingente ; en d'autres termes, les jugements moraux ne possèdent aucune force motivante en eux-mêmes pas plus qu'ils ne sont nécessairement connectés à la motivation.

Un internaliste pourrait objecter que l'exemple de Raymond n'est pas réaliste ; ou alors, il pourrait être tenté d'ajouter des détails à l'histoire – il pourrait ajouter par exemple que Raymond était en réalité incliné à aider l'homme blessé mais d'autres motifs égoïstes l'en ont empêché. Ces tenta-

1 Tel qu'il est décrit ici, l'internalisme s'approche des positions auxquelles s'intéressaient David Brink et Jonathan Dancy. « [internalism states that] moral considerations necessarily motivate » (David Owen Brink : *Moral realism and the foundations of ethics* [Cambridge, New York : Cambridge University Press, 1989] p. 42). « [It is] the conception according to which it is impossible for an agent to make a sincere cognitive moral judgement and not to be motivated accordingly » (Jonathan Dancy : *Practical reality* [Oxford, New York : Oxford University Press, 2000] p. 21). Certains auteurs ont raffiné la thèse internaliste en y intégrant la condition supplémentaire de la rationalité pratique : la connexion internaliste est supposée fonctionner uniquement pour des agents doués de raison pratique (Michael Smith : *The moral problem* [Oxford, Cambridge : Blackwell, 1995]). Cette position ne sera pas discutée dans cet article parce qu'elle s'éloigne à mon avis de la question originale, d'autant plus qu'il n'est pas évident de saisir ce qu'il faut entendre par « agent doué de raison pratique ». En revanche il vaut la peine d'ajouter une clause d'exception en faveur de l'internalisme qui exclut de la discussion toute situation où l'agent souffre de désordres motivationnels susceptibles de l'affecter durablement – par exemple la dépression (Sigrún Svavarsdóttir : *Moral cognitivism and motivation*, in *The Philosophical Review* 108 [1999] pp. 161-219).

tives de décrédibiliser ou de redécrire les exemples dérangeants sont évidemment rejetées par les externalistes ; selon eux, de telles pratiques reviennent à un pur et simple refus de traiter la question.

Le débat entre l'internalisme et l'externalisme a une longue histoire et ce n'est pas mon intention de répéter ici des arguments déjà connus. Mon propos consistera à soutenir l'idée qu'il est possible de prendre au sérieux des exemples comme celui de Raymond tout en résolvant le débat sur la motivation morale au moyen d'une simple distinction entre deux sortes de jugements. Je tâcherai de montrer que l'internalisme est vrai pour les « jugements spontanés » et l'externalisme pour les « jugements sophistiqués ». Plus précisément, je soutiendrai que la motivation ne provient pas de considérations d'ordre conceptuel – ce qui est le cas des jugements moraux sophistiqués – mais plutôt d'évaluations rapides et non réflexives, lesquelles sont directement responsables de nos jugements évaluatifs spontanés. Cette thèse sera défendue en deux étapes.

Il s'agira d'abord de présenter et développer un « tableau affectif » qui décrit la manière dont les gens évaluent des situations et forment leurs propres valeurs et normes. Sans chercher de preuves philosophiques pour soutenir les détails de ce tableau, je tâcherai d'en consolider les traits en recourant à des données et explications empiriques issues de diverses sciences comme la psychologie, l'anthropologie, l'économie empirique ou les neurosciences.

Dans un second temps, je montrerai que mon tableau affectif pose un nouveau cadre pour réfléchir à la question de la motivation ; nous verrons que lorsque qu'il est combiné à une hypothèse humienne sur la motivation, ce cadre permet de remettre sérieusement en question la position internaliste et, dans une moindre mesure, l'externalisme. Cela nous permettra de comprendre pourquoi la décision de Raymond est à la fois possible et déroutante.

2. *Les grandes lignes du tableau affectif*

Afin de résoudre le cas de Raymond, je vais élaborer un tableau affectif qui consiste à mettre en évidence les ingrédients et mécanismes psychologiques sous-jacents à l'activité morale.[2] Il s'agira de proposer dans un premier

[2] Pour les besoins de cet article, je ferai usage d'une conception large de la notion de moralité, qui comprend toutes les réactions, réflexions et décisions relatives aux devoirs liés au bien-être et aux besoins des êtres humains. De même, la notion de jugement sera utilisée en un sens large : une simple expression de réaction

temps, une description de ce qui se passe dans nos esprits lorsque nous jugeons une situation comme devant être réalisée ou évitée, et dans un second temps, de déceler ce qui nous motive à agir. La notion de « tableau affectif » réfère au fait que l'analyse se situe au niveau purement descriptif – il s'agit de dépeindre de manière systématique la pensée et l'activité morales – et accorde une place particulière à l'aspect affectif.

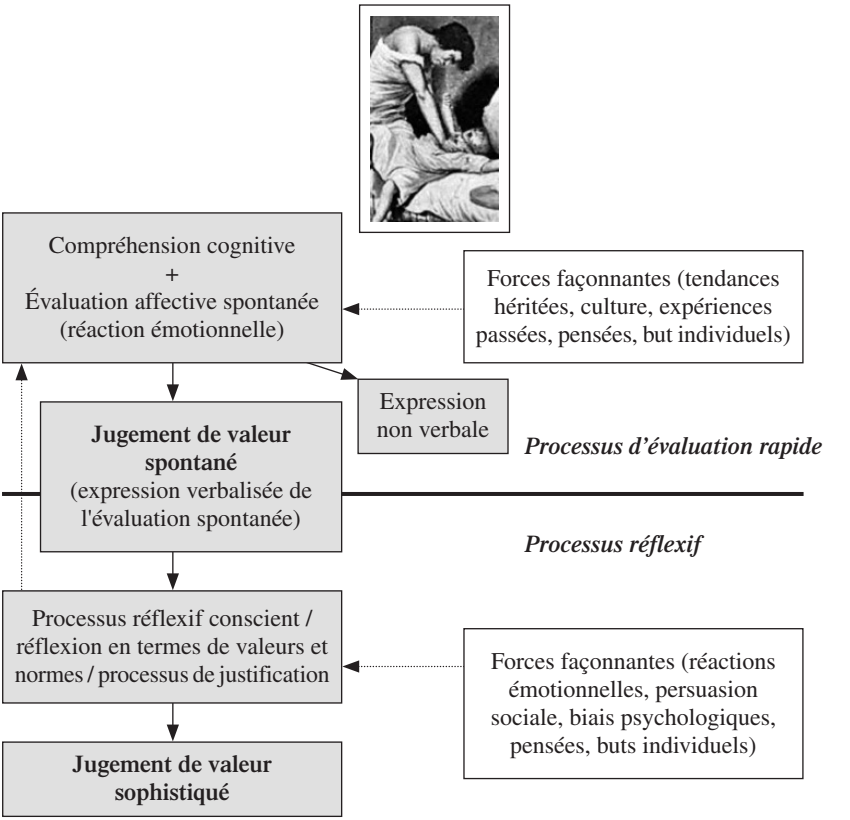
Le tableau affectif n'entre pas sous la coupe de deux conceptions classiques qui accordent une large place à l'affect : l'émotivisme pour qui les jugements moraux sont des expressions de réactions émotionnelles,[3] et l'expressivisme qui considère les jugements moraux comme des expressions d'états d'esprit au sujet d'émotions.[4] Il me semble que ces deux manières de concevoir les jugements moraux minimisent excessivement notre capacité de mener des raisonnements complexes et, par le moyen de processus d'inférence et sans l'aide des émotions, de tirer des conclusions au sujet de ce qu'il faut faire. Nous verrons que l'analyse du jugement moral proposée ici est plus complexe. Voici (à la page 183) un schéma qui sera explicité dans les sections qui suivent.

En deux mots, les grandes lignes de mon tableau affectif sont les suivantes : il y a deux processus dans l'activité morale – ou plus généralement dans l'activité évaluative. Le premier, que j'appellerai « processus d'évaluation rapide », est émotionnel et automatique alors que le second, que j'appellerai « processus réflexif », plus raffiné, inclut les réflexions conscientes et raisonnements d'inférence. Dans cet article, j'avancerai l'idée que seul le premier processus est capable de nous motiver à agir alors que seuls les éléments du processus réflexif peuvent être justifiés. En présentant ce double modèle, je montrerai qu'il existe différents types de jugements et que les plus sophistiqués n'ont pas le pouvoir de motiver à l'action.

émotionnelle peut être considérée comme un jugement. Ce choix n'est pas problématique dans la mesure où il existe une abondante littérature en philosophie morale qui conçoit les jugements moraux de cette manière – par exemple le courant émotiviste.

3 Alfred J. Ayer : *Language, truth, and logic* (London : V. Gollancz, 1946).

4 Selon la théorie d'Allan Gibbard : *Wise choices, apt feelings : A theory of normative judgment* (Cambridge : Harvard University Press, 1990) par exemple, les jugements moraux sont des expressions d'états d'esprit complexes qui consistent dans le fait de penser qu'il est justifié de ressentir une certaine émotion face à un certain type de situation.



Les flèches représentent des influences causales, soit directes (flèches pleines), soit diffuses (flèches en pointillé). Les boîtes grises représentent des activités de l'agent, et les boîtes blanches représentent les influences extérieures.

3. Données empiriques sur les jugements moraux

Cette section est consacrée à la présentation d'une collection de données issues de la psychologie et des neurosciences, y compris certaines expériences en imagerie cérébrale. Ces données contribuent à éclairer le schéma présenté ci-dessus et les relations qu'il établit entre nos réactions émotionnelles et nos choix réflexifs sophistiqués. Nous verrons que les études empiriques nous aident à comprendre la relation entre les jugements moraux et les émotions.

Elles montrent notamment que les jugements moraux ne résultent généralement pas de processus d'inférence au cours desquels nous appliquons consciemment une norme ou une valeur à une situation ; il semblerait plutôt que la plupart des évaluations soient intuitives, émotionnelles, largement automatisées.

Pour commencer, un ensemble d'études brise le vieux mythe de la rationalité de nos esprits moraux et soutient la distinction proposée entre le processus d'évaluation rapide et le processus réflexif.

L'étude la plus connue a été faite par Jonathan Haidt[5] et montre que la plupart des gens condamnent de manière impulsive les pratiques incestueuses ou d'autres violations de tabous et soutiennent ce verdict même au terme d'une discussion où ils sont forcés d'admettre qu'ils ne disposent d'aucune bonne raison pour fonder leur jugement. Selon Haidt, ces résultats mettent en doute la rationalité de la moralité en montrant qu'une part non négligeable de notre activité morale ne peut pas être influencée par un processus réflexif. En d'autres termes, un groupe important de jugements moraux semble se fonder sur de simples réactions intuitives et non sur des raisons morales.

Il me semble que la manière la plus éclairante d'interpréter ces résultats est de comprendre nos évaluations intuitives comme des réactions émotionnelles ; on pourrait les comprendre comme des évaluations affectives spontanées. Nous verrons à la section suivante que cette idée de faire des émotions un constituant essentiel à l'activité morale n'est pas neuve. Mais pour le moment, considérons d'autres données empiriques susceptibles de nous apporter des informations intéressantes au sujet de la manière intuitive et émotionnelle dont nous évaluons les états de faits.

Thalia Wheatley et Jonathan Haidt[6] ont mené une expérience sur des sujets hautement hypnotisables. Sous condition d'hypnose, ils ont expliqué aux sujets qu'ils éprouveraient du dégoût chaque fois qu'ils liraient un mot arbitraire – par exemple le mot *donc*. Une fois réveillés, ils ont demandé aux sujets de lire et juger moralement une série de petites histoires assez communes – dont certaines n'étaient même pas particulièrement pertinentes au plan moral – qui contenaient ou non le mot lié au dégoût hypnotique. Les résultats sont impressionnants : systématiquement, les sujets condamnent moralement les histoires contenant les mots qui causent en eux du dégoût.

5 Jonathan Haidt : *The emotional dog and its rational tail : A social intuitionist approach to moral judgment*, in *Psychological Review* 108 (2001) pp. 814-834.

6 Thalia Wheatley, Jonathan Haidt : *Hypnotically induced disgust makes moral judgments more severe*, in *Psychological Science* 16 (2005) pp. 780-784.

Cette étude montre que des émotions causées de manière artificielle modulent nos jugements et parfois même génèrent des réactions morales.

Dans une expérience congruente menée par Greene et collègues[7], les sujets doivent mener des expériences de pensée sur différentes variantes de situations problématiques comme celle du trolleybus où il s'agit de choisir de tuer une personne afin d'éviter la mort de cinq autres personnes.[8] Au cours de l'expérience, on demande aux sujets quelle action, parmi deux options données, ils considèrent comme moralement adéquate. Le cerveau des sujets est scanné à l'aide de la technique d'imagerie cérébrale afin de détecter les zones du cerveau qui s'activent. Les résultats montrent que l'engagement émotionnel – révélé par une activité accrue des zones du cerveau liées aux émotions – influence largement les jugements moraux : s'imaginer devoir pousser une personne sous un trolleybus roulant à pleine vitesse afin de le stopper, et par là, sauver cinq passagers, est émotionnellement plus saillant que s'imaginer pousser une manette qui va orienter la trajectoire du trolleybus sur une voie où se trouve un homme – dans cette deuxième option, la vie de l'homme sera également sacrifiée pour stopper le trolleybus. Cette différence d'engagement émotionnel induit les sujets à condamner la première action et à juger la seconde comme moralement permissible, alors même que dans les deux cas, la vie d'une personne est sacrifiée pour sauver celle de cinq autres.

La manière la plus simple d'interpréter ces données est de dire que l'engagement émotionnel influence fortement nos jugements moraux. Toutefois, il est important de remarquer que ces données ne permettent pas de prouver que les réactions émotionnelles exercent un effet causal sur les jugements moraux ; on pourrait imaginer qu'elles soient des effets secondaires, eux-mêmes causés par les jugements. Cette position est défendue par une série de chercheurs qui ont mené des études similaires sur le dilemme du trolleybus ou des dilemmes similaires. Voici un résumé de leurs résultats et de la manière dont ils les interprètent.

Dans différentes études, des psychologues ont testé la manière dont les gens justifient leurs jugements moraux.[9] On présente aux sujets des

7 Joshua D. Greene et al. : *An fMRI investigation of emotional engagement in moral judgment*, in *Science* 293 (2001) pp. 2105-2108.

8 Le premier modèle de cette expérience de pensée est dû à la philosophe Philippa Foot.

9 John Mikhail : *Aspects of the theory of moral cognition : Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect*, in *Social Science Research Network* (2005) pp. 1-129, http://papers.ssrn.

problèmes moraux comme celui du trolleybus. Puis, on leur demande dans un premier temps quelles actions ils considèrent comme moralement acceptables ou moralement condamnables et dans un deuxième temps comment ils justifient leur jugement. L'analyse des réponses données par les sujets suggère que certaines règles intuitives profondément ancrées dans nos esprits guident les jugements des gens même si elles ne font pas surface dans leurs raisonnements conscients ; lorsqu'il s'agit de justifier leurs évaluations spontanées, les sujets n'invoquent pas toujours les règles qui les semblent avoir poussés à formuler ces jugements. Parmi ces règles intuitives et parfois même inconscientes, il y a notamment celle selon laquelle un tort causé de manière intentionnelle en tant que moyen pour atteindre une fin est moralement plus condamnable que le même tort s'il est conçu comme l'effet secondaire d'un but que l'on cherche à atteindre.[10] Comme autre exemple, les gens conçoivent un tort causé directement avec contact physique comme moralement plus condamnable que le même tort causé sans contact physique.[11]

Des auteurs comme John Mikail pensent que ces résultats sont incompatibles avec l'idée que l'engagement émotionnel influence le jugement moral. Marc Hauser considère que ces données sur les jugements intuitifs supportent la position selon laquelle les émotions ne précèdent pas mais découlent des nos jugements moraux. Selon lui, un système gouverné par des règles déclenche une évaluation et cette évaluation déclenche parfois – mais pas systématiquement – une réponse émotionnelle. Ce système gouverné par des règles serait un sens du bien et du mal qui a évolué au cours de millions d'années et serait partagé par l'ensemble des êtres humains ; il précède à la fois les jugements et les émotions.

Toutefois, il convient de noter que les données présentées par ces auteurs ne fournissent pas des informations permettant de prouver une telle inter-

com/sol3/DisplayAbstractSearch.cfm (15.12.2007) ; Marc D. Hauser : *Moral minds : How nature designed our universal sense of right and wrong* (New York : Ecco, 2006) ; Fiery Cushman, Liane Young, Marc Hauser : *The role of conscious reasoning and intuition in moral judgment : testing three principles of harm*, in *Psychological Science* 17 (2006) pp. 1082-1089.

10 Il s'agit en fait de la doctrine du double effet qui avait déjà été introduite par Thomas d'Aquin, *Summa theologica*, Qu. 64, Art. 7.

11 Pour des résultats similaires, voir aussi les travaux de Jonathan Baron : *Judgment misguided : Intuition and error in public decision making* (New York : Oxford University Press, 1998) et de Robert J. Nisbett, Timothy D. Wilson : *Telling more than we can know : Verbal reports on mental processes*, in *Psychological Review* 84 (1977) pp. 231-259.

prétation. Il me semble que ce débat repose sur une mécompréhension. En observant de plus près les données et les interprétations présentées dans les deux groupes de recherche, on constate qu'elles sont parfaitement compatibles. Pour peu que l'on considère les réactions émotionnelles comme des actes d'évaluation, les deux interprétations peuvent être combinées. Les résultats de toutes les expériences sont parfaitement compatibles avec l'idée que les jugements intuitifs sont des activations de programmes affectifs, c'est-à-dire des réactions émotionnelles. Une classe particulière d'états mentaux – par exemple la prise de conscience qu'une souffrance est causée de manière intentionnelle plutôt que de manière non intentionnelle – déclenche une réaction émotionnelle typique qui, à son tour, cause une évaluation morale consciente.[12] De plus, en cas de réactions émotionnelles conflictuelles, certaines réactions peuvent être systématiquement plus prégnantes que d'autres. Par exemple, le fait de causer du tort à autrui par le biais d'un contact physique peut déclencher une réaction émotionnelle plus forte que le fait de causer le même tort de manière indirecte. Si un tort est ressenti comme plus important que l'autre, cela influencera les jugements moraux consciemment défendus. Une régularité dans la force de nos réactions émotionnelles face à certaines classes de situations peut donner l'impression que les jugements évaluatifs sont gouvernés par des règles intuitives. Mais cette illusion est due à une tendance à focaliser notre attention sur les évaluations exprimées par les gens ; c'est exactement la manière dont Mikail, Hauser et collègues procèdent. Mais si l'on pense plutôt en termes de réactions émotionnelles différenciées, on peut expliquer ces règles apparentes sans devoir postuler un sens moral spécialisé et gouverné par des règles intuitives.

Une telle analyse qui accorde un rôle important aux émotions se trouve confirmée par des expériences récentes menées sur des sujets normaux et des patients souffrant de lésions du cortex frontal et ventromédian – c'est-à-dire des patients présentant des déficiences au niveau affectif.[13] Au quotidien, ces patients présentent des comportements moraux anormaux et un manque notoire d'intérêt pour les règles morales ; en revanche, leurs capacités intellectuelles sont préservées. Les résultats d'expériences basées sur

12 Pour une analyse similaire, voir Jonathan Haidt, Craig Joseph : *Special Issue on Human Nature*, in *Daedalus* (2004) pp. 55-66.

13 Elisa Ciaramelli et al. : *Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex*, in *Social Cognitive and Affective Neuroscience* 2 (2007) pp. 84-92; Michael Koenigs et al.: *Damage to the prefrontal Cortex Increases Utilitarian Moral Judgements,* in *Nature* 446 (2007) pp. 908-1011.

le dilemme du trolleybus ont montré que même si ces patients se montrent parfaitement capables de distinguer entre les situations moralement et non-moralement pertinentes,[14] leurs jugements au sujet des situations morales sont biaisés par rapport aux sujets normaux. Probablement à cause de leurs déficiences émotionnelles, les patients considèrent les deux scénarios du trolleybus mentionnés plus haut – pousser une manette ou précipiter une personne sur les voies – comme également acceptables. Ces études semblent montrer que les réactions émotionnelles sont des ingrédients essentiels au moins pour certains jugements.

4. Le processus d'évaluation rapide

Dans ce qui suit, je présente en détail mon tableau affectif à la lumière des résultats et analyses de la section précédente. Je m'inspire également d'écrits philosophiques sur les émotions. Commençons avec le premier volet du tableau : le processus évaluatif affectif.

Lorsque nous sommes témoins de certaines situations nous formons des « évaluations affectives spontanées ». On pourrait les considérer comme des jugements primitifs.[15] Plus précisément ces évaluations sont des sortes de réactions émotionnelles composées de deux parties intrinsèquement liées l'une à l'autre : une évaluation rapide non inférentielle[16] d'un état de fait, et un état affectif particulier lié à une sensation particulière. En utilisant le terme de « sensation », je voudrais exprimer l'idée que les réactions émotionnelles possèdent une certaine phénoménologie dans notre expérience personnelle qui est intimement liée à l'évaluation.[17]

14 Cela est sans doute dû au fait qu'au cours de leur vie, avant que leur maladie ne se déclare, ces patients ont appris à faire ce genre de distinctions.

15 Ronald de Sousa : *The rationality of emotion* (Cambridge : MIT Press, 1990) ; Sabine A. Döring : *Seeing what to do : Affective perception and rational motivation*, in *Dialectica* 61 (2007) pp. 363-394.

16 Pour une analyse du caractère non inférentiel de l'évaluation émotionnelle, voir Döring, ibid.

17 En philosophie des émotions, il y actuellement un grand débat autour de la question de savoir si l'aspect affectif de la réaction émotionnelle est causé par l'aspect évaluatif (Jenefer Robinson : *Deeper than reason : Emotion and its role in literature, music, and art* [Oxford : Oxford University Press, 2005]) ou si les deux aspects se confondent en un tout indissociable (Peter Goldie : *The emotions : A philosophical exploration* [Oxford, New York : Oxford University Press, 2000] ;

Par son aspect évaluatif, une réaction émotionnelle possède un contenu intentionel – au sens où elle est dirigée vers un objet.[18] Elle est une manière de prêter une attention sélective à certaines caractéristiques d'une situation et de les percevoir d'une certaine manière.[19] Dans un contexte moral, une évaluation affective spontanée est une manière primitive et non réflexive d'interpréter une situation comme « devant être réalisée » ou « devant être évitée ».[20] Cette évaluation s'exprime par un état affectif qui implique des sensations négatives ou positives – comme celles liées au fait de se sentir moralement dégoûté ou de se sentir bienveillant.

Cette manière d'interpréter le monde dépend de différents facteurs : certaines tendances génétiquement innées, nos expériences passées et l'environnement social dans lequel nous nous trouvons. Voyons dans le détail comment cela fonctionne. Dans une certaine mesure, la manière dont nous évaluons les faits moralement pertinents dépend de systèmes simples et rapides dont les résultats, après réflexion, nous satisfont généralement assez bien – quoique ce ne soit pas toujours le cas.[21] Grâce à ces raccourcis

Döring, op. cit. [n. 15]). Pour mon propos, il ne me paraît cependant pas indispensable de prendre position dans ce débat. Il me paraît suffisant de dire qu'une réaction émotionnelle est à la fois affective et évaluative et que les deux aspects sont intimement liés l'un à l'autre.

18 Dans la mesure où les réactions émotionnelles ont un contenu intentionel, on peut dire qu'elles sont cognitives. Toutefois, cela n'implique pas que les réactions émotionnelles se réduisent à des jugements propositionnels, c'est-à-dire à des sortes de croyances (voir Döring, op. cit. [n. 15]). Cela n'implique pas non plus – comme semblent le penser certains (Döring, op. cit. ; Goldie, op. cit. (n. 17) ; Christine Tappolet : *Emotions et valeurs* [Paris : Presses universitaires de France, 2000]) – que les émotions nous fournissent un accès épistémique à des valeurs existantes. En fait, cette idée est assez problématique. Comme le souligne Kevin Mulligan (*Intentionality, knowledge and formal objects*, in *Disputatio* 23 [2007] pp. 1-17, http://www.fil.lu.se/HommageaWlodek/ [15.12.2007]), la connaissance n'est normalement pas conçue comme une réaction. C'est la raison pour laquelle il est difficile de concevoir les émotions ou les affects comme des producteurs de connaissance. Je ne poursuivrai pas cette discussion ici puisque mon projet n'est pas de déterminer ce qui est bon et mal, pas plus que d'interroger la fiabilité ou la justesse des réactions émotionnelles.

19 Voir de Sousa, op. cit. (n. 15).

20 Notons que cette interprétation du monde via l'émotion n'implique pas que l'objet conçu d'une certaine manière possède en lui-même la propriété qu'on lui attribue de manière intuitive (à ce propos, voir aussi la note 18).

21 Richard S. Lazarus : *Emotion and adaptation* (New York : Oxford University Press, 1991) ; Paul Ekman : *Basic Emotions*, in *Handbook of cognition and*

mentaux, nous avons plus ou moins les mêmes réactions émotionnelles face à des situations similaires. Par exemple, nous considérons généralement que les actions qui causent un tort à autrui sont moralement plus condamnables que les omissions qui causent un tort à autrui ;[22] ou de manière plus générale, nous blâmons systématiquement les actions qui causent de la souffrance chez autrui[23] de même que les comportements non loyaux ou opportunistes. Il semblerait donc qu'il y a un ensemble de mécanismes ou de capacités spécialement conçues pour générer des évaluations rapides au sujet de ce qui est acceptable ou non.

Au-delà de ces aspects rigides, nos facultés émotionnelles sont plastiques et ouvertes, au sens où elles se développent en fonction de nouvelles expériences. Cela signifie qu'elles sont influencées par nos expériences passées et par le contexte dans lequel nous nous trouvons.[24] En d'autres termes, toute réaction émotionnelle est particulière, car imbriquée dans la vie d'un sujet avec toutes ses contingences culturelles et personnelles. Considérons un exemple. Martial est grondé plusieurs fois par ses parents pour avoir caché les jouets de sa petite sœur. Disons que ces reproches l'ont mis mal à l'aise. Suite à cela, Martial associe de manière inconsciente le fait de subir des reproches avec le fait de se sentir mal à l'aise. Et puisque cette sensation est déplaisante, son cerveau marquera – au sens de la théorie des marqueurs somatiques de Damasio[25] – les reproches ainsi que les situations susceptibles de les déclencher comme étant « à éviter ». Dans ce contexte, on peut parler de mécanismes appris par opposition aux mécanismes purement innés qui déclenchent une sensation émotionnelle qui à son tour guide les futures évaluations et comportements. Dans notre exemple, chaque fois que Martial éprouve le désir d'une action susceptible

emotion, éd. T. Dalgleish, M. J. Power (Chichester, New York : Wiley, 1999) pp. 45-60.

22 Il vaut la peine de remarquer que ce principe fonctionne même dans les cas où l'action causerait moins de dommages que l'omission (à ce propos, voir Baron, op. cit. (n. 11) ; Cass R. Sunstein : *Moral heuristics*, in *Behavioral and Brain Sciences* 28 [2005] pp. 531-542).

23 Shaun Nichols : *Sentimental rules : On the natural foundations of moral judgment* (Oxford : Oxford University Press, 2004).

24 La proportion dans laquelle les émotions sont encodées génétiquement ou le résultat de l'éducation ou de la culture est une question très débattue quoique mal posée. Il n'est pas possible d'y apporter une réponse satisfaisante car chaque cas est différent.

25 Antonio R. Damasio : *L'erreur de Descartes : la raison des émotions* (Paris : Odile Jacob, 2001/1995).

de déclencher le reproche de ses parents – par exemple, cacher les jouets de sa sœur –, son cerveau reconstituera le marqueur somatique du malaise – même si c'est de manière moins vive que dans les circonstances réelles – et cela le motivera à se restreindre d'agir de la sorte.[26]

Nos évaluations affectives spontanées dépendent également des divers motifs que nous avons par ailleurs. Par exemple, des recherches en psychologie sociale ont montré que les gens sont profondément influencés par le désir de maintenir des relations sociales plaisantes. Ce motif guide leurs attitudes évaluatives et la manière dont ils traitent l'information.[27] De manière plus générale, beaucoup d'études ont également été menées sur le phénomène de la contagion émotionnelle, la tendance à ressentir des émotions similaires à celles d'autrui.[28] En résumé, notre environnement social dirige nos réactions émotionnelles dans le sens de la consistance avec celles de nos voisins.

Avant de conclure cette section, il me paraît important de préciser encore deux points. Premièrement, même si nous sommes passifs dans nos réactions émotionnelles – au sens où elles s'imposent à notre conscience de manière soudaine, incontrôlée et sans effort –, cela n'empêche pas que ces réactions émotionnelles puissent découler de pensées et croyances complexes ; elles peuvent résulter de la prise de conscience de spécificités très fines d'une situation ; elles peuvent même être produites au cours d'une réflexion rationnelle. De plus, leur grain particulier dépendra en bonne partie de nos capacités intellectuelles et du contexte culturel dans lequel nous avons grandi car les émotions complexes sont des choses que nous apprenons et qui sont largement forgées par la culture.

Deuxièmement, même si les réactions émotionnelles sont des évaluations, on peut se demander s'il est pertinent de les considérer comme des jugements. Les réactions émotionnelles nous fournissent des inputs pour le raisonnement et la formulation de jugements moraux conscients. En un sens, on peut dire que ce sont des dispositions à croire. Toutefois, si l'on est enclin

26 De plus, comme nous le verrons plus loin, au cours du processus réflexif, Martial attribuera probablement une valeur négative à ce type d'actions.

27 Serena Chen, David Shechter, Shelly Chaiken : *Getting at the truth or getting along : accuracy versus impression motivated heuristic and systematic processing*, in *Journal of Personality and Social Psychology* 71 (1996) pp. 262-275.

28 Marvin L. Simner : *Newborn's response to the cry of another infant*, in *Developmental Psychology* 5 (1971) pp. 136-150 ; Elaine Hatfield, John T. Cacioppo, Richard L. Rapson : *Emotional contagion* (Cambridge, New York, Paris : Cambridge University Press, 1994).

à considérer les jugements comme des expressions verbales, les réactions émotionnelles ne pourraient pas être inclues dans cette catégorie.

5. Les jugements de valeur spontanés

Une réaction émotionnelle est généralement exprimée ; elle ne reste pas tacite. Il y a deux manières de le faire. La première est de l'exprimer sous forme non verbale. C'est le cas par exemple lorsque notre faciès se contracte et exprime la colère, ou lorsqu'on crie « Ah ! ». Cela correspond plus ou moins à la manière dont les émotivistes conçoivent les jugements moraux.[29] Mais cette manière d'exprimer une évaluation affective spontanée reste encore dans le domaine de l'incontrôlable ; l'expression n'est que l'extension automatique et directe de la réaction émotionnelle interne et s'inscrit dans une polarité positive – approbation – ou négative – désapprobation.

La deuxième manière d'exprimer une évaluation affective spontanée comporte une expression verbale. C'est-à-dire que l'on peut conceptualiser les réactions émotionnelles en termes de « c'est horrible ! » ou « c'est bien ! ». Il ne s'agit pas uniquement d'une expression d'approbation ou de désapprobation – comme l'expression non verbale – mais d'un énoncé conscient et réflexif par le moyen duquel on attribue une valeur à quelque chose. Il semblerait donc qu'il s'agisse d'une sorte de reconstruction cognitive de la réaction émotionnelle. Cet apport cognitif, qui s'ajoute à la forme primitive du jugement de valeur, permet en fait la production d'un nouveau jugement légèrement plus complexe : ce que je vais appeler un « jugement de valeur spontané ». L'idée est qu'en mettant des mots sur une évaluation affective spontanée, nous ne nous contentons pas de l'exprimer mais produisons en fait un nouveau jugement. De cette manière, nous faisons un premier pas en direction d'un processus réflexif dans le cadre duquel il s'agit de trouver de bonnes raisons pour nos réactions émotionnelles et pour les jugements, normes et valeurs que nous sommes enclins à accepter.

Toutefois, même s'ils sont légèrement plus raffinés – car nous ne nous contentons pas d'exprimer une attitude mais attribuons une valeur à une situation –, ces jugements spontanés demeurent relativement confus et s'effectuent également de manière largement automatique. En produisant un jugement spontané, nous affirmons sans plus de précision que quelque

29 Ayer, op. cit. (n. 3).

chose est bon ou mauvais parce que nous le ressentons de cette manière.[30] Sur une échelle de complexité, les jugements spontanés se situent donc entre deux extrêmes : la simple expression d'approbation ou de désapprobation et l'assertion d'un énoncé normatif à caractère objectif.

Il semblerait que ce soit ce genre de jugements que les psychologues comme Haidt ou Hauser testent dans leurs expériences lorsqu'ils focalisent l'attention sur les jugements rapides et intuitifs. Il vaut la peine de remarquer ici que les philosophes de la morale s'intéressent généralement aux jugements sophistiqués – ceux que l'on peut justifier – plutôt qu'aux jugements spontanés décrits dans cette section. Les psychologues de la morale en revanche, considèrent les deux sortes de jugements sans réellement les distinguer et les traitent plus ou moins de la même manière. Je doute de la pertinence d'une telle approche et favoriserais au contraire un modèle explicatif qui différencie clairement les deux sortes de jugements. De ce point de vue, je pense que mon tableau affectif aide à réconcilier les résultats des études empiriques avec les approches philosophiques, précisément parce qu'il introduit une distinction entre les jugements spontanés et sophistiqués. Ces derniers font l'objet du second volet de mon tableau affectif auquel est consacrée la prochaine section.

6. Le processus réflexif

Cette section rend compte du versant hautement cognitif de l'activité morale ; c'est ici qu'entrent en jeu les « jugements de valeur sophistiqués ».

Nous nous trouvons souvent confrontés à des désaccords entre nos jugements évaluatifs spontanés et ceux produits par autrui, ou entre nos jugements et le comportement d'autrui. Parce que notre survie dépend largement de notre capacité de mener une vie cohérente et coordonnée à celle de nos voisins, nous éprouvons un double besoin : d'un côté, nous voulons nous prouver à nous-mêmes et à autrui la pertinence de nos réactions émotionnelles, de l'autre côté nous désirons qu'autrui partage nos réactions émotionnelles. Ces deux objectifs nous incitent à nous engager dans une activité complexe au cours de laquelle nous réfléchissons sur les raisons qui justifient nos réactions émotionnelles et jugements spontanés – c'est ce

[30] A ce niveau de mon explication, je ne voudrais pas défendre une position volontariste qui accorderait un rôle trop important à la volonté dans la production des jugements spontanés.

que j'appelle le « processus réflexif » – et mettons en pratique de manière plus ou moins consciente diverses méthodes pour influencer les réactions et convictions de nos voisins.

Dans le cadre du processus réflexif, nous cherchons à justifier nos jugements spontanés. Nous pouvons le faire de manière minimale en nous demandant simplement s'ils sont appropriés à la situation ; il s'agit alors de savoir si tous les aspects pertinents ont été pris en compte, si notre engagement personnel a altéré notre jugement, etc. Nous pouvons également – et c'est ce qui est le plus intéressant en matière de morale – chercher à leur donner une justification forte au sens de fondement ; dans ce cas, nous faisons reposer nos jugements sur des normes et des valeurs. Dès lors, pour les besoins de notre tableau, il serait intéressant de disposer d'une explication de la manière dont nous définissons et choisissons les normes et valeurs auxquelles nous adhérons. Je vais argumenter en faveur d'un modèle tripartite selon lequel il y a trois déterminants pour nos normes et valeurs : les affects – et en particulier les émotions –, les considérations rationnelles, et la persuasion sociale.

Il semblerait que nous choisissions nos normes et valeurs en grande partie en fonction des sensations qui sont causées en nous dans des situations concrètes. Si un état de choses cause en nous une forte sensation désagréable, dès que nous y réfléchissons, nous aurons tendance à attribuer une valeur négative à cet état de choses – et inversement pour les sensations positives.[31] Ce lien peut s'établir de deux façons : la première met en jeu les sensations qui font partie d'une réaction émotionnelle alors que la seconde met en jeu des sensations simples.

Les réactions émotionnelles, c'est-à-dire les évaluations affectives spontanées, peuvent révéler ce que nous valorisons inconsciemment. En effet, il semble qu'une partie de ce que nous valorisons nous est rendu épistémologiquement accessible précisément par la médiation des réactions émotionnelles.[32] Par exemple, si nous sommes horrifiés à la vue de notre voisin qui bat son enfant pour le plaisir – et par là produisons une évaluation affective spontanée –, nous aurons tendance à attribuer une valeur négative à ce type de comportement et à établir une norme correspondante qui interdit ce genre d'action. Selon cette explication, notre esprit est déjà imprégné de manière

31 De Sousa, op. cit. (n. 15); Christopher Hookway : *Emotions and epistemic evaluations*, in *The cognitive basis of science*, éd. P. Carruthers et al. (Cambridge, New York : Cambridge University Press, 2002) pp. 251-262.

32 Voir Goldie, op. cit. (n. 17) pp. 48-49.

inconsciente par un certain nombre de valeurs et de règles et les réactions émotionnelles servent à nous en faire prendre conscience. Toutefois, ce n'est certainement pas la seule manière de définir nos valeurs. Si c'était le cas, nous naîtrions avec un dispositif de valeurs prédéterminé et notre seul horizon serait de les découvrir.

Le choix d'une valeur peut aussi résulter d'une simple sensation associée à une représentation conceptuelle de la cause de cette sensation. Souvenons-nous de Martial ; après s'être fait gronder par ses parents, il éprouve un certain malaise chaque fois qu'il conçoit l'idée de cacher les jouets de sa sœur. Cette sensation de malaise associée à un certain type de situation l'incitera à valoriser négativement la situation en question ; il pensera que c'est mal de cacher les affaires de sa sœur. En bref, nous avons tendance à attribuer une valeur négative à ce qui cause en nous des sensations désagréables – et inversement pour les valeurs positives.[33] De plus, les valeurs et normes que nous choisissons sous l'influence de simples sensations peuvent même supplanter des valeurs et normes induites par des réactions émotionnelles préalablement ancrées en nous. Prenons un exemple parlant : Prosper prône la fidélité à tel point qu'il est dégoûté à l'idée même d'un individu qui trompe son partenaire. Mais un jour il rencontre Célestine et craque … il trompe sa femme avec elle. Cette expérience s'avère si plaisante qu'il cesse d'éprouver de l'aversion envers les partenaires infidèles.[34] L'exemple de Prosper et Célestine illustre également un autre principe : la puissance des sensations joue un rôle dans la manière dont nous choisissions nos normes et valeurs ;

33 Sabine Döring (op. cit. [n. 15]) défend un point de vue similaire mais basé sur une compréhension des émotions comme perceptions. Selon elle, « en faisant l'expérience d'une émotion, le monde apparaît au sujet comme s'il était tel que l'émotion le lui représente ». Elle ajoute que « les émotions jouent un rôle dans le raisonnement avant que le raisonnement ne joue son rôle dans la rationalisation d'une action » – cela dit, je m'éloigne de la position de Döring lorsqu'elle prétend que les émotions permettent de percevoir des valeurs extérieures (à ce propos, voir n. 18 et 20).

34 Je ne voudrais pas affirmer par là que tous les individus confrontés à une situation similaire réviseraient leurs valeurs après avoir rencontré une Célestine. Certains pourraient admettre que leur comportement est moralement condamnable et qu'ils souffrent de faiblesse de la volonté. D'autres pourraient maintenir leur aversion morale face à l'infidélité mais dans le même temps se trouver des raisons spécifiques pour justifier leur comportement. Mon intention avec cet exemple est plutôt d'illustrer le fait que nos émotions exercent une influence importante sur nos choix normatifs, même si un effort de réflexion rationnelle permet d'atténuer leurs effets.

plus une sensation est forte, plus nous avons tendance à attribuer une valeur à ce qui en est la cause.

Cette analyse est confirmée par des données empiriques. Les études mentionnées plus haut sur les patients souffrant de lésions au cortex préfrontal et ventromédian[35] montrent que les jugements de valeur dépendent de notre expérience affective et de notre compréhension des émotions d'autrui. L'étude a révélé que les patients présentant des déficiences au niveau émotionnel considèrent comme moralement acceptable de précipiter un homme sur les voies du trolleybus.[36]

Voilà pour une première explication de la manière, probablement la plus courante, de choisir nos normes et valeurs. Il existe cependant d'autres manières de les définir. Nous pouvons nous lancer dans des raisonnements d'inférence plus complexes et, partant de normes et valeurs préalablement acceptées, en déduire de nouvelles. C'est à ce stade que la pensée rationnelle prend sa place dans l'activité morale. Toutefois, sans parler du fait que tout processus d'inférence part de prémisses données, il faudrait se garder d'y accorder trop d'importance.[37] Comme le font bien remarquer Joshua Greene et Jonathan Haidt,[38] il est vrai que les gens s'engagent souvent dans des débats moraux réels ou fictionnels mais la plupart du temps, ces efforts sont dirigés vers le renforcement ou la transmission de jugements, valeurs ou normes sur lesquels les sujets ont déjà fixé leur choix par avance.

Le troisième déterminant des normes et valeurs est la persuasion sociale. Dans les contextes sociaux, les gens tentent de s'influencer mutuellement et d'assurer un consensus avec leurs amis ou alliés. Il existe même des mécanismes psychologiques – auxquels est consacrée toute une littérature empirique – qui régissent et assurent l'efficacité de cette influence mutuelle. Selon les anthropologues évolutionnistes Joseph Henrich, Robert Boyd et collègues, les êtres humains sont largement influencés dans leurs jugements

35 Ciaramelli et al., op. cit. ; Koenigs et al., op. cit. (n. 13).

36 Pour des données similaires, voir Adina Roskies : *Are ethical judgments intrinsically motivational ?*, in *Philosophical Psychology* 16 (2003) pp. 51-66 ; *Patients with ventromedial frontal damage have moral beliefs,* in *Philosophical Psychology* 19 (2006) pp. 617-627. Cette chercheuse a pu montrer que les patients ventromédians – ceux qui souffrent de déficits au niveau des réactions émotionnelles – peuvent produire des jugements moraux même en l'absence de réactions émotionnelles.

37 Haidt, op. cit. (n. 5).

38 Joshua D. Greene, Jonathan Haidt : *How (and where) does moral judgment work ?*, in *Trends in Cognitive Sciences* 6 (2002) pp. 517-523.

et pratiques par un certain nombre de biais psychologiques. Les biais sont des sortes de règles d'apprentissage social du type ' Copie qui a le plus de succès ! ', ou 'Copie la majorité ! ' Un des biais les plus influents est celui du conformisme : les êtres humains ont tendance à reproduire les comportements les plus fréquents de la population dans laquelle ils évoluent.[39] Un autre biais est celui du prestige : les êtres humains tendent à prendre pour modèle des individus qui paraissent avoir du succès ou qui semblent posséder des qualités ou des connaissances supérieures.[40] En ayant recours aux techniques de la théorie des jeux ou de la simulation par ordinateur, ces chercheurs tentent de déterminer les conditions d'adaptation évolutionnaire de ces mécanismes psychologiques et émettent l'hypothèse qu'ils ont évolué parce qu'ils permettent aux individus de bénéficier à peu de frais des avantages liés à l'adoption d'un comportement, d'une norme, d'une valeur ou d'une coutume. Chandra Sripada et Stephen Stich[41] renforcent à l'aide d'exemples concrets cette hypothèse de l'existence de biais psychologiques.[42] Faisant référence à différentes études empiriques, ils montrent que les êtres humains sont souvent incapables de juger correctement les avantages et désavantages de variantes culturelles – par exemple dans le domaine des innovations ou des tabous liés à la nourriture – et prennent leurs décisions sous l'influence des biais du conformisme et du prestige.[43] De plus il a été montré dans diffé-

39 Joseph Henrich, Robert Boyd : *The evolution of conformist transmission and the emergence of between-group differences*, in *Evolution and Human Behavior* 19 (1998) pp. 215-241.

40 Joseph Henrich, Francisco J. Gil-White : *The evolution of prestige : Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission*, in *Evolution and Human Behavior* 22 (2001) pp. 165-196.

41 Chandra Sripada, Stephen P. Stich : *Evolution, culture and the irrationality of the emotions*, in *Emotion, evolution, and rationality*, éd. D. Evans, P. Cruse (Oxford, New York : Oxford University Press, 2004) pp. 133-158 ; Daniel M. Fessler, Carlos David Navarrete : *Meat is good to taboo : Dietary proscriptions as a product of the interaction of psychological mechanisms and social processes*, in *Journal of Cognition and Culture* 3 (2003) pp. 1-40.

42 Allan Gibbard (op. cit. [n. 4]), un des rares philosophes de la morale qui prennent les données empiriques réellement au sérieux, est un précurseur d'un bon nombre d'idées développées dans cet article. Entre autres choses, il intègre dans son explication de l'activité morale, des mécanismes psychologiques comme l'« influence normative » (normative influence) – similaire au biais du prestige – ou l'exigence de cohérence (demand for consistency) – similaire à l'émotion épistémique qui sera introduite plus loin dans cette section.

43 Ces auteurs soulignent également le fait que malgré leur fonction de révélateurs d'adaptations, ces biais peuvent mener à la propagation d'innovations et pratiques

rentes expériences psychologiques menées par Jody Davis et Caryl Rusbult[44] que nous sommes directement influencés dans nos jugements par ceux de nos amis, alliés ou proches parents. Cela confirme selon eux l'existence du phénomène d'« alignement d'attitude » (*attitude alignment*), une tendance, chez les partenaires d'interaction, à modifier leurs attitudes respectives de manière à ce qu'elles convergent.[45]

En bref, nous choisissions systématiquement des valeurs et normes congruentes avec nos émotions, lesquelles sont partiellement innées et partiellement acquises au moyen de l'apprentissage individuel au grès des interactions avec notre environnement social. Cela n'empêche en rien l'émergence de conflits à l'intérieur même de nos productions morales – c'est-à-dire nos actions, les valeurs ou normes auxquelles nous souscrivons, les jugements de valeur spontanés ou sophistiqués que nous produisons – ou entre nos productions morales et celles de nos voisins. C'est ici que l'activité rationnelle refait surface dans le tableau affectif : c'est elle qui attire notre attention sur ce genre d'incohérences. Cette prise de conscience déclenche en nous une émotion épistémique[46] que l'on peut appeler « demande de cohérence » et se caractérise par un sentiment d'inconfort.[47] Ce sentiment nous inclinera à tenter de rétablir la cohérence.[48] Nous ne pourrons pas nous empêcher de procéder à des raisonnements inférentiels, à ouvrir notre cœur

hautement maladaptives si elles sont prises isolément ; c'est par exemple souvent le cas en matière de tabous alimentaires.

44 Jody L Davis, Caryl E. Rusbult : *Attitude alignment in close relationships*, in *Journal of Personality and Social Psychology* 81 (2001) pp. 65-84.

45 Pour plus de données, voir Haidt : *The emotional dog*, op. cit. (n. 5). Précisons également que les mécanismes psychologiques mentionnés sont étroitement liés aux émotions. Par exemple, le biais du prestige se manifeste par une certaine sensibilité émotionnelle face à des personnes que l'on considère comme prestigieuses ; il induit la tendance à copier les pratiques, et adopter leurs croyances et valeurs.

46 Certains préfèreront parler de mécanisme psychologique (Gibbard, op. cit. [n. 4]).

47 En effet, promouvoir à la fois p et –p, nous met dans une situation analogue à un dilemme pratique où il est impossible de réaliser tous les buts que nous nous sommes fixés.

48 Pour des études empiriques supportant cette idée, voir Haidt : *The emotional dog*, op. cit. (n. 5) ; Gordon B. Moskowitz, Ian Skurnik, Adam D. Galinsky : *The history of dual process notions, and the future of pre-conscious control*, in *Dual-process theories in social psychology*, éd. S. Chaiken, Y. Trope (New York : Guilford Press, 1999) pp. 12-36 ; Paul Thagard : *Conceptual revolutions* (Princeton : Princeton University Press, 1992).

à nos sentiments les plus profonds et à chercher l'avis de notre entourage jusqu'à ce que l'harmonie soit rétablie. Enfin, l'activité rationnelle signifie également la prise en considération des questions d'ordre pratique et l'orientation du choix de nos normes en fonction de leur réalisabilité et des autres buts que nous nous sommes fixés.

Les grandes lignes de mon tableau affectif sont maintenant tracées. Ce tableau mériterait évidemment d'être développé davantage mais la présentation qui en a été faite est suffisante pour permettre de prendre position dans le débat entre l'internalisme et l'externalisme. Dans ce qui suit, je commencerai par défendre l'idée que la motivation morale est toujours déclenchée par les réponses affectives avant de prendre position dans le vieux débat philosophique.

7. L'hypothèse humienne affective

L'approche humienne de la motivation est une conception philosophique qui a déjà une longue histoire. Selon cette théorie, la motivation morale dépend toujours d'un état conatif préexistant, lequel est habituellement conçu en termes de désir.[49] Je pense qu'il y a de bonnes raisons d'accepter une telle analyse de la motivation à la différence qu'au lieu des désirs, je suggèrerai de mettre la focale sur les états affectifs – dans leur forme simple ou en tant qu'ils font partie intégrante de phénomènes plus complexes comme les épisodes émotionnels par exemple ; au fond, les désirs motivent précisément parce qu'ils sont intrinsèquement liés à des états affectifs. Appelons cette position révisée une « hypothèse humienne affective ». Cette dernière semble d'ailleurs plus proche de la position défendue par Hume lui-même lorsqu'il écrit : « Il paraît évident que les fins ultimes des actions humaines ne peuvent jamais, en aucun cas, être expliquées par la raison, mais qu'elles se recommandent entièrement aux sentiments et aux affections des hommes, sans dépendre aucunement des facultés intellectuelles. ».[50] Selon Hume, les considérations rationnelles ne peuvent

[49] Notons que l'approche humienne n'impose aucune prise de position particulière au niveau du débat entre l'internalisme et l'externalisme. Même si beaucoup de philosophes d'obédience humienne défendent une position externaliste, cela ne doit pas nécessairement être le cas. Michael Smith, op. cit. (n. 1) par exemple est un humien internaliste.

[50] David Hume : *Enquête sur les principes de la morale* (Paris : Flammarion, 1991 [1751]) Appendix I, V.

en elles-mêmes générer ni des affects ni des actions. Dans cette section, j'argumenterai en faveur de cette idée.

L'hypothèse humienne affective semble la meilleure explication possible pour une série de données empiriques issues de la psychologie. Jennifer Beer et collègues[51] ont mené des études comparatives entre des personnes en bonne santé et des patients souffrant de dommages cérébraux au niveau du cortex orbifrontal, la partie du cerveau connue pour être liée aux processus émotionnels. En faisant jouer un certain nombre de paramètres, les expérimentateurs ont pu mettre en évidence une corrélation nette entre les troubles de régulation du comportement social et l'absence d'émotions liées à la conscience de soi (*self-conscious emotions*) comme la honte, l'embarras ou la fierté. Il semblerait donc que la capacité d'avoir des émotions liées à la conscience de soi dans des circonstances appropriées est un facteur essentiel à la régulation du comportement social. Dans la même veine, différentes études sur des psychopathes ou patients avec des lésions des parties ventromédiane et préfrontale du cerveau[52] révèlent que ces patients présentent un dysfonctionnement de leur système affectif alors même qu'ils ne montrent aucun déficit dans leurs raisonnements pratiques. D'un côté, ils sont parfaitement capables de penser en termes de normes, d'un autre côté en revanche, ils montrent peu d'intérêt pour les normes sociales et morales et se comportent de manière antisociale – à un degré nettement plus dramatique dans le cas des psychopathes.

Ces données empiriques révèlent l'existence d'une corrélation forte qui relève probablement d'un lien essentiel entre les processus affectifs et l'action morale. De ce point de vue ces données semblent plus compatibles avec une analyse humienne de la motivation qu'avec une approche rationaliste selon laquelle les croyances morales suffisent à nous motiver à agir.[53] Plus

51 Jennifer S. Beer et al. : *The regulatory function of self-conscious emotion : Insights from patients with orbitofrontal damage*, in *Journal of Personality and Social Psychology* 85 (2003) pp. 594-604 ; Jennifer S. Beer et al. : *Orbitofrontal cortex and social behavior : Integrating self-monitoring and emotion-cognition interactions*, in *Journal of Cognitive Neuroscience* 18 (2006) pp. 871-879.

52 Damasio, op. cit. (n. 25) ; James Blair, Derek Robert Mitchell, Karina Blair : *The psychopath : Emotion and the brain* (Malden : Blackwell, 2005) ; Antoine Bechara et al. : *Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex*, in *Cerebral Cortex* 6 (1996) pp. 215-225 ; Nicola S. Gray et al. : Forensic psychology : *Violence viewed by psychopathic murderers*, in *Nature* 423 (2003) pp. 497-498.

53 Thomas Nagel : *The possibility of altruism* (Oxford : Clarendon Press, 1970) ; Dancy, op. cit. (n. 7).

précisément, ces données soutiennent l'hypothèse humienne affective parce que l'explication la plus simple et directe pour les importantes corrélations observées est de considérer les processus affectifs comme des ingrédients nécessaires à la prise de décision. Ce n'est bien sûr pas un argument décisif mais voici une raison supplémentaire de favoriser cette approche. Le philosophe Matteo Mameli[54] a analysé de manière détaillée les résultats des expériences de Damasio sur des patients présentant des lésions préfrontales. Il constate qu'en plus d'être incapables d'agir de manière sociale, ces patients ont également beaucoup de peine à prendre des décisions pratiques de tous les jours, malgré le fait qu'ils soient tout à fait capables de mener des raisonnements logiques.[55] Si, comme le suggèrent ces données, les considérations rationnelles ne sont pas suffisantes pour motiver à prendre des décisions, on peut se demander quel est l'ingrédient manquant. Mameli pense que c'est précisément parce que les patients ont des déficiences au niveau du système affectif qu'ils éprouvent des difficultés à prendre des décisions. Il en déduit que nos capacités générales à prendre des décisions pratiques, y compris les décisions morales, dépendent de manière cruciale du bon fonctionnement de notre système affectif ; en d'autres termes, nos décisions pratiques sont le produit de nos émotions. Ainsi il écrit : « Chez les êtres humains, le choix entre différentes actions […] est toujours déterminé par les réactions émotionnelles [*emotional feelings*] causées par la prise en considération des résultats probables des différentes actions possibles, et non par une froide analyse des coûts et bénéfices ».[56] Un corrélat important de cette explication est que même si les énoncés moraux résultent de processus d'inférence conscients, le choix des actions n'est pas dû à ce processus. En effet, un processus d'inférence est indépendant de l'affect ; ce n'est rien de plus qu'une succession de croyances qui suivent certaines règles. Dans les termes de Mameli : « les croyances au sujet de ce qui est socialement appro-

54 Matteo Mameli : *The role of emotions in ecological and practical rationality*, in *Emotion, evolution, and rationality*, éd. D. Evans, P. Cruse (Oxford, New York : Oxford University Press, 2004) pp. 159-178.

55 Il semblerait qu'en plus de leurs déficits affectifs, les psychopathes présentent des difficultés de langage – ils peinent à catégoriser les notions sémantiques abstraites – d'attention ou d'orientation (Kent A. Kiehl : *A cognitive neuroscience perspective on psychopathy : Evidence for paralimbic system dysfunction*, in *Psychiatry Research* 142 [2006] pp. 107-128). Toutefois, les psychopathes sont habituellement capables de raisonnements inférentiels. C'est tout ce dont nous avons besoin pour donner du crédit à l'analyse de Mameli, op. cit. (n. 54).

56 Mameli, op. cit. (n. 54) p. 171, ma traduction.

prié dans certaines circonstances n'exercent aucune force motivationnelle sur nos prises de décision, sauf si elles peuvent déclencher des réactions émotionnelles [*emotional feelings*] qui nous motivent à agir en accord avec le contenu de nos croyances ».[57] Précisons que Mameli n'affirme pas l'impossibilité d'être motivé à produire des actions dont on sait qu'elles auront pour conséquence de déclencher des sensations désagréables. Ces cas sont possibles à condition que l'on prédise pour les mêmes actions, des résultats plaisants d'un autre point de vue et que ces aspects plaisants l'emportent sur les aspects négatifs.[58]

Considérons maintenant deux modèles concurrents à l'hypothèse humienne affective et demandons-nous s'ils sont capables de la surpasser. On pourrait défendre l'idée que les réactions morales affectives ne sont rien de plus que des épiphénomènes d'une cognition morale intuitive ou tout simplement de nos croyances morales. Selon cette approche, nous comprenons – ou du moins pensons comprendre – ce qui est la bonne chose à faire ; cette connaissance ou simple croyance nous motive à agir et déclenche habituellement une réaction émotionnelle. Le problème d'une telle interprétation réside dans son coût théorique : elle soulève plus de questions qu'elle n'en résout. En effet, emprunter cette voie argumentative implique de devoir trouver un nouvel éclairage pour d'autres aspects du phénomène à expliquer. Si la cognition ou croyance morale est suffisante pour motiver, pourquoi nous sentons-nous souvent outrés ou coupables ? Si ces émotions se réduisent à des épiphénomènes, pourquoi sont-elles si largement partagées dans l'espèce humaine ? Quel rôle jouent-elles si ce n'est pas un rôle motivant ? Pourquoi certains types d'émotions – par exemple, la peur des serpents – sont clairement responsables de nos actions alors que d'autres types d'émotions – celles qui sont supposées épiphénoménales – sont dénuées d'une telle efficacité causale ?

Une autre alternative – sans doute plus intéressante – serait d'affirmer que les croyances morales sont motivantes parce qu'elles sont capables, en elles-mêmes – sans autres médiateurs –, de générer des désirs[59] ou des

57 Ibid.

58 Mameli, op. cit. (n. 54) p. 172. Dans la même veine, Joshua Greene (*Cognitive Neuroscience and the Structure of the Moral Mind* in *The Innate Mind: Structure and Contents,* ed. P. Carruthers et al. [Oxford, NewYork : Oxford Unversity Press, 2005] pp. 338-352) passe en revue une série de recherches qui montrent que l'efficacité des normes sociales dépend d'émotions suffisamment fortes pour motiver à l'action.

59 Dancy, op. cit. (n. 7).

réactions émotionnelles qui, à leur tour, génèrent des actions. Mais cette stratégie explicative soulève également de nouvelles questions. Selon cette approche, il devrait exister un processus par le biais duquel les croyances morales causent l'activation de désirs ou de mécanismes affectifs qui, ultimement, nous incitent à agir. Mais cette connexion causale n'apparaît pas dans tous les cas de figure ; les croyances au sujet de la beauté d'un objet ou de la validité d'un argument par exemple ne génèrent aucun désir particulier. On peut donc légitimement se demander pourquoi précisément les croyances morales génèrent des désirs ou des réactions émotionnelles.

De manière plus problématique encore, les deux modèles alternatifs décrits ci-dessus se trouvent confrontés à la difficulté d'expliquer pourquoi les patients souffrant de lésions du cortex ventromédian sont apparemment capables de penser qu'ils devraient agir d'une certaine manière sans pour autant être motivés à le faire.[60]

En fin de compte, le principe de parcimonie s'impose de lui-même et incite à adopter l'hypothèse humienne affective : une réponse affective est une cause nécessaire à l'action, c'est le seul ingrédient capable de disposer les êtres humains à agir.

8. Retour au débat sur l'internalisme versus externalisme

L'hypothèse humienne affective, combinée avec mon tableau affectif, permet de tirer une série de conclusions intéressantes dans le contexte du débat sur l'internalisme et l'externalisme. Comme nous allons le voir, la balance penchera finalement du côté d'une position hybride entre l'internalisme et l'externalisme.

60 Adina Roskies : *Are ethical judgments intrinsically motivational ?*, op. cit. (n. 36); Antonio R. Damasio, Daniel Tranel, Hanna Damasio : *Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli*, in *Behavioural Brain Research* 41 (1990) pp. 81-94. On pourrait objecter que les patients souffrant de lésions du cortex ventromédian ne sont pas capables de reporter correctement leurs propres états mentaux ; soit ils se trompent au sujet de leurs véritables jugements ou alors ils n'endossent pas sincèrement leurs assertions morales (Michael Cholbi : *Belief attribution and the falsification of motive internalism*, in *Philosophical Psychology* 19 [2006] pp. 607-616) ; mais ces interprétations ont été réfutées de manière convaincante par Adina Roskies : *Patients with ventromedial frontal damage*, op. cit. (n. 36).

L'application de l'hypothèse humienne au tableau affectif met en valeur deux caractéristiques importantes des jugements de valeurs. Premièrement les jugements spontanés sont intrinsèquement motivants puisque ce sont des expressions de réactions émotionnelles. Plus précisément, en raison du fait que les jugements spontanés sont intrinsèquement liés à des états affectifs – les réactions émotionnelles –, ils sont un facteur clef pour l'action morale. Deuxièmement les produits du processus réflexif – les valeurs et normes acceptées ainsi que les jugements sophistiqués – ne sont pas motivants en eux-mêmes puisqu'ils ne possèdent aucune dimension émotionnelle. La motivation doit provenir d'une source externe ; soit d'une réaction émotionnelle corrélée au jugement sophistiqué, soit d'un désir d'obéir au jugement en question – par exemple parce que nous voulons éviter la punition ou soigner nos relations.

Cette analyse fournit une explication simple des cas de conflit entre nos jugements moraux et nos actions. Elle permet de considérer la possibilité de juger qu'une chose devrait être faite sans pour autant être motivé à le faire. Une telle situation ne peut cependant surgir qu'en cas d'élaboration froide d'un jugement sophistiqué qui ne correspond pas à l'évaluation émotionnelle de la situation. Souvenons-nous de l'exemple de Raymond qui refuse son aide à une personne en difficulté. Il se peut que Raymond ne ressente rien de particulier dans cette situation, même s'il est conscient du fait que son attitude est moralement condamnable au vu des principes moraux auxquels il adhère. Une telle situation peut par exemple être due au fait que Raymond est extrêmement fatigué ce jour-là ou au fait que son esprit est occupé par la triste perspective de son divorce agendé au lendemain …

L'analyse proposée ici permet de remettre en question la thèse philosophique internaliste. Si mon analyse est correcte, les jugements moraux et les considérations morales en général ne motivent pas forcément à l'action ; à l'exemple de Raymond, il est possible pour un agent de produire sincèrement un jugement moral sophistiqué sans pour autant être incliné à le suivre. D'un autre côté, l'externalisme est également affaibli au niveau des jugements de valeur spontanés puisque l'hypothèse humienne combinée au tableau affectif nous présentent ces jugements comme nécessairement motivants.

Au premier abord, cette position hybride peut s'avérer assez dérangeante puisqu'elle met en évidence l'inefficacité causale des jugements moraux paradigmatiques, ceux qui peuvent être fondés sur des normes et des valeurs. Dès lors, on pourrait penser que l'action moralement vertueuse est toujours produite par hasard, si bien que l'on ne peut pas vraiment parler d'action morale. Mais il s'agit là d'une conclusion trop pessimiste. D'une part, les actions

vertueuses sont intrinsèquement liées à des jugements évaluatifs spontanés. D'autre part, les liens entre nos réactions automatiques et notre activité réflexive sont extrêmement forts. De fait, la plupart de nos normes, valeurs et jugements sophistiqués sont intiment corrélés à nos réactions émotionnelles parce que ces dernières influencent largement nos évaluations réflexives.[61] En outre, l'effet inverse n'est pas négligeable : les produits du processus réflexif peuvent exercer un impact indirect sur nos réactions émotionnelles. Par exemple, en affinant notre compréhension cognitive de la situation jugée, nous pouvons déceler de nouvelles caractéristiques pertinentes qui vont modifier nos réactions émotionnelles ou en déclencher de nouvelles. De manière encore plus intéressante, je pense qu'il faut prendre au sérieux la possibilité de manipuler les mécanismes émotionnels ancrés dans nos esprits, ceux qui causent nos réactions émotionnelles – à l'exemple de l'expérience du dégoût hypnotique mentionnée plus haut. Par exemple, si nous voulons qu'une de nos normes ou valeurs ait une force motivante – ou du moins soit corrélée à nos motivations concrètes –, nous pouvons tenter de forger un mécanisme émotionnel correspondant. Dans une certaine mesure nous manipulons déjà nos émotions sans même avoir à nous fixer consciemment ce but. Malheureusement, ces questions sont extrêmement complexes et délicates, si bien que je ne pourrai pas les développer plus avant dans le cadre de cet article. Quoiqu'il en soit, en prenant conscience de l'enchevêtrement complexe des relations entre les processus affectifs et réflexifs, on comprend qu'il n'y a aucune raison de déprécier les actions qui découlent des premiers.

9. Conclusion

Dans cet article, j'ai détaillé les deux volets d'un tableau affectif représentant la manière dont les gens jugent et se comportent moralement. Ce tableau est largement inspiré d'études empiriques. Je n'ai pas cherché à adopter une attitude philosophique de confrontation vis-à-vis des ces études. Plutôt que de chercher des raisons de douter de la vérité ou de la pertinence de leurs conclusions, je suis partie du présupposé que ces analyses empiriques sont

[61] C'est d'ailleurs la raison pour laquelle l'internalisme motivationnel est convaincant au premier abord : les jugements sophistiqués semblent motivants parce qu'ils sont habituellement corrélés avec des réactions émotionnelles congruentes – mais en réalité, seules ces dernières sont les vecteurs de nos actions morales.

en mesure de nous dire quelque chose d'intéressant au sujet de la manière dont nous pensons et agissons moralement et peuvent être intégrées dans mon tableau affectif. Une tâche future serait bien sûr de renforcer ce tableau à l'aide de données empiriques supplémentaires et davantage d'arguments de nature philosophique. En attendant, j'espère avoir pu relever de manière convaincante deux effets positifs des processus affectifs dans le contexte de l'activité morale. Premièrement, ils ne dénaturent pas nos jugements mais les façonnent de manière directe – dans le cas des jugements spontanés – et indirecte – dans le cas des jugements sophistiqués. Deuxièmement, ils sont responsables de notre comportement moral ; les jugements de valeur spontanés sont intrinsèquement motivants parce qu'ils résultent directement de nos réactions émotionnelles. De plus, le tableau de l'activité morale proposé dans cet article permet d'expliquer pourquoi les normes auxquelles nous souscrivons et les jugements sophistiqués qui leurs sont associés ne prescrivent pas toujours ce que nous choisissons en fin de compte de faire ; ils échouent à nous motiver lorsqu'ils ne sont pas soutenus par une évaluation affective spontanée.

Studia philosophica 68/2009

Daniel Schulthess

La psychologie politique d'Adam Smith : Biais cognitifs et différences sociales dans la *Théorie des sentiments moraux* (1759)

In his Theory of Moral Sentiments *(1759), Adam Smith does not deal only with interpersonal moral issues. He also addresses some economic and political consequences that tie with his analysis of 'sympathy'. Interestingly, these socially relevant outcomes do not feature as products of sympathy proper, but rather as by-products of certain 'irregularities' or biases which affect the way sympathy actually works. The stability of a political society through a system of 'ranks' which are spontaneously granted a share of authority thus gets projected beyond the reach of a rational approach, e.g. of a contractarian character. Although the idea of a spontaneous order certainly attracts Adam Smith here as elsewhere, his approach of the economic and political sphere in TMS is nevertheless tinged with a tone of moral criticism which must be taken seriously.*

1. Introduction

Dans son œuvre de 1759, Adam Smith explore avec un sens heuristique aigu et une exactitude méticuleuse nombre de facettes inattendues de l'action intentionnelle ordinaire. Celle-ci a des contours imposés : elle procède de motivations clairement repérables chez l'agent ainsi que de croyances en réseau permettant à ce dernier d'agir en vue de réaliser ses buts. Ensemble les motivations et les croyances fixent les grandes lignes de l'action intentionnelle de façon pratiquement transparente pour l'agent lui-même. Tel est le tableau de l'action qu'il est convenu d'appeler aussi rationnelle.

Mais au-delà de cette auto-présentation transparente de l'action, l'agent n'est en rien omniscient et il peut ignorer de multiples à-côtés de ce qu'il fait : tout en contribuant à réaliser ses propres buts, son action possède alors des facettes inattendues et à ce titre non intentionnelles. Smith s'intéresse de près à ces facettes inattendues – tout particulièrement quand elles possèdent une pertinence morale, économique et sociale plus large, et bien entendu lorsqu'elles ne procèdent pas du pur hasard. Son raisonnement de la *Richesse*

des nations[1] sur la « main invisible » est bien connu. Il consiste à contraster, en ce qui concerne l'activité économique et sur fond de division du travail, une composante intentionnelle et une composante non intentionnelle de l'action. La première composante, c'est la poursuite délibérée et anticipatrice de l'intérêt privé ; l'autre composante, l'effet non délibéré et non anticipé que comporte cette poursuite, à savoir la promotion de l'intérêt du public. A ce dernier, la « main invisible » procure l'abondance de biens communément recherchés.[2] Par la mise en évidence de cette structure, Smith figure parmi les théoriciens les plus influents de l'« ordre spontané » cher à la théorie économique libérale. Et comme on sait cette structure définit une approche spécifique dans les sciences sociales, une façon de déterminer leur objet.

Cependant Smith est loin de propager des vues unilatérales au sujet de la sphère économique, comme on s'en aperçoit dans la *Théorie des sentiments moraux* ;[3] par le même art du contraste entre les composantes intentionnelle et non intentionnelle de l'action, il assortit sa mise en place de la sphère économique dans la *TMS* d'une sorte d'analyse morale désabusée particulièrement intéressante. Dans son ouvrage, Smith ne cesse de s'intéresser aux « irrégularités » (*irregularities*) qui se font jour dans les mécanismes psychologiques qu'il entreprend d'étudier. Dès la première partie de *TMS*, il pose les mécanismes les plus basiques de la « sympathie », qui ressemble à l'« empathie » de la psychologie cognitive d'aujourd'hui. Si les bases sont simples, les répercussions en sont profondes et constitutives de la société humaine. Quelquefois, justement, les répercussions de la sympathie ne sont pas directes mais obliques. Les « irrégularités » venant affecter la sympathie sont souvent profondes et stables, et Smith ne laisse pas de leur attribuer un impact théorique important, notamment en matière politique. Il effectue ainsi une sorte de critique indissociablement psychologique et sociale, où ce qui est psychologiquement discutable devient politiquement indispensable. Dans notre exposé, nous traiterons de deux moments de la *TMS* particulièrement intéressants à cet égard. Nous nous efforcerons d'identifier aussi clairement que possible le mécanisme proposé et de mettre en évidence à son sujet la réflexion généralisante de Smith. Nous noterons aussi la curieuse préemption de questions normatives par des analyses factuelles, qui constitue une des marques de fabrique du courant Hume-Smith dans le cadre des Lumières écossaises.

1 1776 ; ci-après *RN*.
2 Voir *RN*, IV.ii, § 9 ; I.ii, § 1-2
3 Ci-après *TMS*.

2. *D'un penchant exagéré pour les moyens*

Dans la 4e partie de *TMS*, Smith traite du sentiment de l'approbation portant sur l'action d'autrui. L'action d'autrui est vue ici dans un sens large qui inclut les effets extérieurs de l'action. L'intérêt de Smith se porte donc aussi sur la mise en place d'enchaînements de moyens et de fins. Dans ce cadre, le thème de l'utilité des moyens se présente sous le jour suivant : notre approbation se renforce à mesure de l'utilité dont nous sommes les spectateurs, c'est-à-dire de la mise en place la mieux appropriée de moyens réalisant des fins, par exemple lorsqu'il s'agit d'organiser les habitations des hommes.[4] Reprenant un exemple de Hume,[5] Smith localise d'abord dans le maître de maison l'approbation avec laquelle ce dernier considère son propre habitat. Ensuite, par sympathie, le spectateur entre dans les vues du maître de maison.

A un moment donné, toutefois, une distorsion se fait jour : c'est que l'approbation des moyens vient à se rendre autonome et à prendre une dynamique propre. Là où la cohérence voudrait que l'approbation des moyens ne se laissât calibrer, en toute transparence, que sur la valeur des buts que ces moyens permettent de réaliser, les choses se dérèglent. Par une irrégularité sui generis, dont Smith revendique la découverte, l'utilité des moyens jouit d'une plus haute approbation que les buts visés en tant que tels (§ 3). Elle devient alors elle-même une finalité dérivative et fantasmatique quoique hautement approuvée.

Une dimension esthétique s'impose aussi dans le même contexte : nous rattachons une sorte de beauté aux arrangements coordonnés de moyens visant certaines fins. Smith continue en ajoutant un profil social à ses analyses : les personnes puissantes et riches se signalent par leur disposition à multiplier et à embellir les moyens en vue de finalités qui leur sont propres. Par l'irrégularité esquissée plus haut, le spectateur de telles personnes admire non pas tellement la véritable satisfaction et le bonheur que celles-ci peuvent atteindre, mais surtout la vaste panoplie des moyens du bonheur qu'elles mettent en place (§ 8). Par un effet d'imagination, en général imperméable à la raison, nous nous enthousiasmons devant la belle mise en œuvre des systèmes utiles, nous fermant par contrecoup à l'appréciation sérieuse et réfléchie de la valeur des fins que de tels systèmes permettent réellement d'atteindre. Si les fins des personnes riches et puissantes, dûment examinées pour elles-mêmes, s'avèrent accessibles à presque tous (nous reprendrons ce

4 *TMS*, IV.1.

5 *Treatise of Human Nature,* II.ii.5.

point ci-dessous), personne ne semble s'en apercevoir. En réalité, l'action suit alors ces effets d'imagination : aveugles à la médiocrité des fins poursuivies, les hommes sont prêts à s'embarquer pour elles dans les entreprises les plus périlleuses et les plus fatigantes.

Smith rattache alors à sa critique morale plutôt désenchantée une facette économique et sociale. La recherche même de systèmes utiles accompagnée d'une cécité systématique à la valeur des buts réellement atteignables par ces systèmes, devient un moteur primordial de l'économie humaine. Smith met en valeur spécifiquement son impact sur l'agriculture : les grands propriétaires terriens mettent en exploitation de grands domaines et produisent beaucoup de ressources. Ils pensent produire à leur seul profit et accumuler les moyens qui rendent la vie humaine réussie et agréable. En fait, cependant, indépendamment même de leurs propres intentions, peut-être même sans qu'ils s'en avisent :

> Ils sont conduits par une main invisible à faire presque la même distribution des produits nécessaires à la vie, qui aurait été faite, si la terre eût été divisée en portions égales entre ses habitants, et ainsi sans le vouloir, sans le savoir, ils servent l'intérêt de la société et ils procurent des moyens pour la multiplication de l'espèce.[6]

L'analyse morale n'est donc pas sans déboucher sur un volet politico-économique. En gros, la distorsion qu'elle identifie suffit à mettre en place un cadre à peu près adéquat pour le bon fonctionnement économique de la société :

> Il est heureux que la nature nous trompe de cette façon [= en nous présentant en imagination des biens tels que la perspective de leur acquisition justifie pour nous la mise en œuvre d'un immense labeur, n. d. l'A.]. C'est cette illusion qui suscite et entretient le mouvement perpétuel de l'industrie du genre humain.[7]

Quoi qu'on puisse penser de cette économie politique mélancoliquement animée par l'illusion, Smith tend à confier à de tels mécanismes moraux basiques la charge d'instaurer des cadres politico-sociaux pour ainsi dire auto-entretenus et à ce titre solides et permanents.

6 *TMS*, IV.i, § 10.
7 Ibid.

3. Les effets d'une sympathie asymétrique

Dans la première partie de la *TMS*, nous trouvons une autre analyse d'irrégularité possédant un impact politico-économique puissant. La troisième section du chapitre a pour titre : « Effets de la prospérité et de l'adversité sur le jugement des hommes concernant la convenance (*propriety*) de l'action ; et pourquoi il est plus facile d'obtenir leur approbation dans le premier état que dans le second ». Ici encore, Smith prolonge l'élaboration des mécanismes de la sympathie dans le sens d'une psychologie politique. Le philosophe écossais travaille avec trois variations indépendantes, et dont la composition produit d'intrigants effets.

3.1 Première variation : les biens plafonnés, les maux non

Le premier plan auquel nous devons nous intéresser concerne les biens et les maux attachés à la vie humaine. Le point que Smith souligne d'abord dans un esprit quelque peu épicurien est que les biens de cette vie sont facilement atteignables : être en bonne santé, échapper à l'endettement, avoir la conscience tranquille, voilà un véritable bonheur facile à trouver, à portée de presque tous et peu susceptible d'être jamais dépassé. Mais ensuite – et là on est loin d'Epicure –, l'étendue des maux de l'homme est sans limites.[8] Quel que soit le malheur qui frappe un homme, il semble qu'il y en ait toujours qui peuvent venir péjorer une situation déjà mauvaise. Les biens et les maux de la vie sont donc fortement asymétriques.

3.2 Deuxième variation : la sympathie inégale

Sur un deuxième plan se situe maintenant la possibilité de sympathiser avec l'homme heureux ou malheureux. Ici, Smith s'explique longuement sur le fait contre-intuitif que « la sympathie est plus forte avec la joie qu'avec le chagrin ». Intrinsèquement, nous venons de le voir, la variation des biens et des maux s'avance peu dans le positif, et indéfiniment dans le négatif. La sympathie comportant toujours une certaine inertie, il est déjà plus facile de la pratiquer dans le premier que dans le second. A cela s'ajoute que le niveau de corrélation des sentiments du sympathisant avec les sentiments

8 *TMS*, I.iii.i, § 7-8.

du sympathisé est très différent dans un cas et dans l'autre : la corrélation des sentiments qui s'effectue par la sympathie est très supérieure pour la joie que pour la peine. Le sujet prête à controverse, car ne pense-t-on pas spontanément que la sympathie s'adresse d'abord aux souffrances des hommes ? Aussi Smith appuie-t-il son point sur une multiplicité d'observations. Il note qu'en tant que sympathisants, nous tendons souvent à supprimer les sentiments (eux-mêmes pénibles !) acquis par sympathie pour la peine ; cette tentative de suppression elle-même peut d'ailleurs nous rendre attentifs à la peine des autres. Dans le cas de la joie, cependant, nous n'avons pas de tendance similaire (§ 4). Certes, nous pouvons être sujets à la jalousie. Mais s'il y a jalousie, il n'y a pas du tout de joie par sympathie, et donc pas de tendance à supprimer les sentiments acquis par sympathie, ceux-ci n'existant pas du tout. Cependant l'existence de la jalousie a pour conséquence que souvent nous prétendons sympathiser avec la joie alors que nous ne le faisons pas. Un indice du fait que « la sympathie est plus forte avec la joie qu'avec le chagrin », c'est encore que dans l'expression de la peine, nous tendons à pratiquer une retenue, dont l'expression de la joie n'est pas assortie. Cela s'explique de la façon suivante : dans le cas de l'expression de la peine, nous craignons pour cette conduite le défaut de sympathie, toujours menaçant ; mais nous ne le craignons pas dans le cas de l'expression de la joie (§ 9).

Dans la même ligne de raisonnement, Smith observe encore que nous résistons aux pleurs visibles en public, alors que nous résistons beaucoup moins aux rires ostensibles en public. Dans le cas des pleurs visibles en public, nous craignons le défaut de sympathie ; mais nous ne le craignons pas dans le cas des rires ostensibles en public (§ 10-11). C'est ainsi que la nature, souligne le philosophe, a mis en place une sympathie limitée avec la peine (§ 12).

3.3 Attirer l'attention du monde

Les faits mis en évidence sous les deux rubriques précédentes ont diverses conséquences directes et indirectes :

> C'est parce que l'humanité est disposée à sympathiser plus complètement avec notre joie qu'avec notre peine, que nous faisons montre de nos richesses, et que nous cachons notre pauvreté.[9]

9 *TMS*, I.iii.2, § 1.

Et plus profondément, l'asymétrie de la sympathie est l'origine de l'ambition, comme l'annonce le titre du deuxième chapitre de cette section : par l'ambition, les hommes font effort pour éviter la pauvreté et pour accumuler des richesses, et cela bien au-delà des nécessités de la nature. Nous assistons à cette étape à un fort couplage des dimensions psychologique et économique. C'est pour attirer « l'attention du monde » que nous nous lançons dans de multiples entreprises (§ 1). Désormais, nous recherchons des plaisirs et évitons des peines en fonction de l'attention d'autrui (Smith n'a pas manqué de lire Rousseau). En particulier, nous recherchons les plaisirs de la sympathie et fuyons les peines de l'indifférence et de la désapprobation : mais indirectement, puisque nous contrôlons ce qui en donne l'occasion. L'ambition apparaît, en tant que celle-ci nous porte à rechercher des plaisirs et à éviter des peines sur un plan autre que celui des « nécessités de la nature ». Comme l'a souligné Albert Hirschman, Smith rompt par cette analyse de l'ambition économique avec l'argumentation libérale antérieure, qui trouvait dans l'intérêt économique un antidote approprié aux passions destructrices des hommes : ici l'ambition n'est plus que la fusion des diverses dispositions qui la précèdent.[10]

3.4 La troisième variation : les joies imaginaires de la prospérité

Au-delà des points discutés jusqu'ici, Smith souligne que l'imagination nous fait exagérer bien à tort la joie qu'amène la prospérité. Lorsque nous envisageons la vie des gens puissants et riches, nous la remplissons de joie de façon telle que nous ne pourrions pas, pour nous-mêmes et au vu de tous nos désirs, nous fixer des objets plus achevés, plus élevés. Par « les préjugés de l'imagination », nous sommes donc affectés d'un biais cognitif systématique sur la structure de la joie :

> Dans tous nos songes éveillés et dans toutes nos rêveries oisives, c'est cet état même que nous avions esquissé comme l'ultime objet de nos désirs.[11]

Dans notre imagination, la joie de la prospérité répond pour ainsi dire exactement au principe symétrique de celui qui prévaut en fait pour les maux : quelle que soit la joie qui puisse être atteinte, il nous semble qu'il y en a toujours une plus grande qui peut découler d'une prospérité accrue.

10 Voir *The Passions and the Interests*, II.ii.2.
11 *TMS*, I.iii.2, § 2.

Pour l'étude de cette « bulle eudémonique », l'analyse proposée dans notre première section peut aussi être sollicitée : il existe en effet une intersection du thème de la joie supposée des riches avec le thème des instruments du bonheur qu'ils sont portés à mettre en place.[12]

3.5 La disposition à servir les puissants

Exagérant la joie présumée que procure la prospérité, admirant ceux qui sont réputés y parvenir, participant par sympathie à cette joie imaginée, et tirant donc quelque chose de celle-ci, les hommes en viennent à vouloir maintenir les puissants dans cette joie, à vouloir les aider.[13] Sur ce fond biaisé se construit donc un désir de servir qui répercute sur un plan politique les conséquences de l'admiration. Par manière de contraste, l'adversité que rencontrent les hommes peine à susciter la sympathie et entraîne plutôt le mépris. Sur ces bases psychologiques, Adam Smith construit une politique de style naturaliste qui comporte une forte ambiguïté : elle paraît à la fois irrationnelle et inévitable : « La Nature nous apprend à nous soumettre aux rois par égard pour eux ».[14] Smith se prononce donc ici, comme dans ses *Cours sur le droit (Lectures on jurisprudence)*, contre les théories du contrat social, fondées sur la notion d'intérêt mutuel et dans lesquelles la conformité des actions aux obligations encourues découle du contrat lui-même. Pour lui les sources de l'obéissance aux autorités politiques sont complètement différentes des sources rationnelles qu'alléguaient les contractualistes.

Sur la foi de ces analyses, l'effet non délibéré et non anticipé des mécanismes de la sympathie – qui eux-mêmes ne sont l'objet que d'une volonté diffuse quoique systématique de par le plaisir qu'on y prend[15] – sera de maintenir les « ordres » dans la société (*ranks*). A travers cette question des « ordres » se joue en fait la question de l'autorité politique. Smith souligne à plusieurs reprises dans son œuvre l'ancrage psychologique – lointain mais puissant – de l'autorité politique.[16] Pour Smith, le bon fonctionnement de la

12 Cf. *TMS*, IV.i, §§ 8-10.
13 *TMS*, I.iii.2, § 3 ; cf. III.iii, § 10.
14 *TMS*, I.iii.2, § 3.
15 *TMS*, I.i.3.
16 Cf. *TMS*, VI.ii.1, § 20.

société dépend de l'équilibre des ordres qui lui sont propres.[17] Les sentiments moraux, dans le cheminement décrit ci-dessus, travaillent à le maintenir.

4. Conclusion

Les analyses qui nous ont retenus répondent au schéma général suivant : elles veulent produire par des mécanismes psychologiques les structures intersubjectives dont on attendait avant Smith qu'elles fussent produites par l'ordonnancement de la société conformément à des règles inspirées par le droit naturel. Certains biais systématiques – dont la critique morale est modérée mais non complètement absente – sont ainsi mis en rapport avec le plan politique, où d'ailleurs ils prennent finalement une portée plus positive. L'invocation d'une nature semi-providentielle n'est pas rare dans ce contexte.[18]

L'aperçu que nous avons donné dans ces pages montre que la *TMS* est une œuvre complexe et ambitieuse, animée par une visée analytique très articulée. Certes, elle se situe initialement sur le terrain psychologique des sentiments qu'elle veut baliser, notamment en se fondant sur les mécanismes de la sympathie. Mais ce n'est pas sa seule ambition. En effet, pour aller à l'autre extrême, elle veut rattacher aux sentiments moraux basiques des aspects sociétaux et politiques complexes, mettant en jeu tout ce qui relève de la pensée politique moderne. C'est de ce programme ambitieux que nous avons voulu donner ici un aperçu, en soulignant que les chapitres de la *TMS* que nous avons approchés sont parmi les plus représentatifs de Smith pour le découvrir.

17 Voir *TMS*, VI.ii.2, §§ 7-10.
18 Cf. aussi *TMS*, II.iii.3.

Willensschwäche und Selbsttäuschung
Acrasie et mauvaise foi

Studia philosophica 68/2009

Christophe Calame

Mais quoi, ce sont des fous !

Descartes 1642

Dans la première de ses *Méditations métaphysiques*, Descartes écrit :

> Et comment est-ce que je pourrais nier que ces mains et ce corps-ci soient à moi ? Si ce n'est peut-être que je me compare à ces insensés, de qui le cerveau est tellement troublé et offusqué par les noires vapeurs de la bile, qu'ils assurent constamment qu'ils sont des rois, lorsqu'ils sont très pauvres ; qu'ils sont vêtus d'or et de pourpre, lorsqu'ils sont tout nus ; ou s'imaginent être des cruches, ou avoir un corps de terre. Mais quoi ? ce sont des fous (sed amentes sunt isti) et je ne serais pas moins extravagant si je me réglais sur leur exemple.[1]

Foucault 1961

Dans son *Histoire de la folie à l'âge classique*,[2] Michel Foucault, qui avait jusque-là semblé suivre une voie phénoménologique dans ses réflexions philosophiques à propos des concepts de la psychiatrie,[3] allant même jusqu'à préfacer la traduction de certains essais de Ludwig Binswanger en français,[4] va au contraire dramatiser, durcir et *dédialectiser* le rapport entre raison et folie, en montrant comment l'exclusion *pratique* de la folie du champ social la constitue dans son altérité même. Travaillant sur la grande collection de documents recueillie à la bibliothèque de l'Université d'Uppsala, Foucault découvre dans les archives ce que nulle présentation de malade ni expérience clinique personnelle n'aurait pu lui offrir : le recul philosophique qui permet

1 René Descartes : *Méditations métaphysiques*, tome II (Paris : éd. Alquié, Classiques Garnier, 1967) pp. 405-406.

2 Michel Foucault : *Folie et déraison, histoire de la folie à l'âge classique* (première édition : Paris : Plon, 1961 ; Paris : Gallimard, 1972) (ci-après HF).

3 Michel Foucault : *Maladie mentale et personnalité* (Paris : PUF, 1954). Rééd. *Maladie mentale et psychologie* (Paris : PUF, 1961 ; Paris : PUF, « Quadrige », 1997).

4 Ludwig Binswanger : *Le rêve et l'existence* (Paris : Desclée de Brouwer, 1954).

de mettre en perspective large non tant le problème de la folie que le partage même entre raison et folie.

A cette enquête d'un nouveau type, qui se présente sous le masque de l'histoire pour mieux se soustraire rhétoriquement à toute problématisation philosophique, s'ajoute une affirmation très clairement *généalogique*, au sens de Nietzsche : l'origine du grand partage entre raison et folie est *obscure*, et ses dimensions pratiques et politiques *occultées* par l'évidence médicale, clinique, asilaire. En fait, dans le grand partage entre les vivants et les morts, le fou a pris la place du lépreux en Occident. Alors que le Moyen Age, comme l'Antiquité, reconnaissait la transcendance de la folie, et que la raison pouvait encore parfois tenter d'engager un dialogue oraculaire avec la folie,[5] la Renaissance embarque la folie dans la dérision de l'embarquement, ou la récupère dans le discours amusant de son éloge humaniste[6].

Mais, grâce aux archives d'Uppsala, Foucault a pu montrer comment et à quel point nous n'en sommes plus là : depuis l'Age classique, la folie entendue, dans son sens le plus large, comme l'ensemble des comportements bannis par la société d'ordre, fait l'objet d'une proscription non plus *centrifuge* (nef des fous), mais bien *centripète* (enfermement). Pour ce faire, Foucault rappelle *l'oubli* des rapports complexes, en *miroir*, qui existaient au Moyen Age et à la Renaissance entre raison et folie, au profit du *Grand Renfermement*, l'institution de l'Hôpital général (nullement destiné à des *malades* à l'origine). Michel Foucault établit bien que le projet de l'Age classique portait non seulement sur les fous délirants, mais aussi sur tous les sujets en difficultés dans la conduite de leur vie : vagabonds, chômeurs, prostituées, libertins, prodigues, suicidaires qui tentent de se « défaire ».

Et l'oubli lui-même du Renfermement est l'œuvre de la psychiatrie qui, à partir du mythe de la « libération » des fous par Pinel sous la Révolution française, organise l'asile psychiatrique comme une structure fondamentalement médicale, même si elle hérite largement de la structure carcérale sur laquelle elle s'est édifiée et à laquelle elle a succédé socialement. Dans le rapport de la raison à la folie, Foucault veut établir que l'exclusion n'est pas une constante historique, mais bien un trait commun entre ordre classique et psychiatrie bourgeoise. Foucault apporte les résultats de son enquête à charge

5 La ruse de Tristan ne se comprend qu'à l'horizon d'une telle entente de la folie : la preuve qu'il est fou est qu'il sait tout. Le savoir qui désole Iseult fait pourtant rire le roi Marc.

6 Erasme, *Éloge de la folie*.

dans le procès de la psychiatrie, et surtout dans le grand débat sur la nature constitutionnelle ou sociale de la folie.

L'Histoire de la folie a suscité une abondante littérature de protestation, particulièrement de la part des psychiatres, mais aussi des opposants au structuralisme dans les années 80.[7] Il faut rappeler que sa parution est contemporaine de la naissance de l'antipsychiatrie, du développement d'institutions alternatives à l'asile, et bien sûr de la fermeture progressive des grandes Bastilles de la folie. Nous ne pouvons entrer ici dans le détail de cette discussion (qui fut souvent très agressive). Il nous suffira de rappeler quelques points de méthode qui caractérise l'approche de Foucault, et qui sont indispensable à la compréhension de ce qui va suivre :

1. La méthode de Foucault n'est historique que par le recours aux archives (Uppsala). Son objet est le *discours* bien plus que l'événement. Et dans « l'ordre du discours »,[8] Foucault s'attache aux *discontinuités* bien plus qu'aux continuités, dans le but de faire apparaître le *refoulé*, non tant au sens psychanalytique qu'au sens « archéologique » : le discours du passé est comme *enfoui* parmi les strates des archives.
2. Le dégagement « archéologique » des discours du passé produit un effet de *stupeur* (celui que Freud appréciait dans sa collection de statues ?). La culture générale reprend soudain la mesure d'un passé méconnu. L'effet de cette stupeur est toujours critique, il introduit une relativisation du présent, un sentiment d'étrangeté au présent immédiat (aux notions psychiatriques).
3. Foucault ne se borne pas à citer d'obscurs documents ou des essais justement oubliés (ce qu'on lui a parfois reproché). Il ponctue tous ses livres par l'évocation de très grandes figures de la littérature et de l'art, qui se succèdent comme des cariatides ou des atlantes sur une façade bourgeoise (pour redonner une apparence plus organique à la géométrie des murs ?). Ainsi Jérôme Bosch, Don Quichotte, Descartes, Le neveu de Rameau, Sade, Nerval, Roussel, Artaud. Cet usage rhétorique de la culture dans l'argumentation ne signifie nullement que l'ordre du discours est déterminé par les figures littéraires. Au contraire, c'est lui qui porte leur intelligibilité ultime.

7 Cf. Luc Ferry, Alain Renaut : *La pensée 68, essai sur l'anti-humanisme contemporain* (Paris : Gallimard, 1985).

8 Michel Foucault : *L'ordre du discours, leçon inaugurale au Collège de France* (Paris : Gallimard, 1971).

Et Descartes dans tout cela ?

Michel Foucault présente le texte des *Méditations* cité en ouverture comme le « symptôme » philosophique du Grand Renfermement, parce que Descartes y distinguerait clairement deux traitements opposés de l'irrationalité : *l'erreur* qui peut-être dissipée par la recherche méthodique des évidences, – et la *déraison*, condamnée, elle, sans appel (être sujet, *c'est d'abord ne pas être fou*). Sans insister ici sur le rapport entre philosophie pure et pratique sociale du « renfermement », rapport aucunement problématisé par Foucault encore une fois, il convient d'abord de remettre en scène la confrontation entre Renaissance et Age classique, à travers les figures de Montaigne et de Descartes.

Montaigne, dans le chapitre XXVII[9] du Premier livre des Essais, « C'est folie que de rapporter le vray et le faux à nostre suffisance », se refuse à déclarer que « telle chose est fauce et impossible » parce que « c'est se donner l'advantage d'avoir dans la teste les bornes et limites de la volonté de Dieu et de la puissance de nostre mère nature » (il parle là « des esprits qui reviennent, ou du prognostique des choses futures, des enchantemens, des sorcelleries » etc., bref de l'occulte). Le chapitre de Montaigne cité par Foucault tourne en fait autour de la défense des miracles des saints par les catholiques dans les Guerres de Religion françaises, et non de la folie. Et Montaigne de citer des événements historiques dont la prise de connaissance défie la raison, et d'invoquer sans la pouvoir nommer la télépathie, ou transmission de pensée : le Comte de Foix connaît la défaite du Roi de Castille le lendemain de la bataille, le pape Honorius organise les funérailles de Philippe Auguste le jour même, la défaite d'Antonius en Allemagne est proclamée à Rome à la fin de la journée, etc. La suspension du jugement requise par le scepticisme de Montaigne porte sur des phénomènes qui ne sont pas du même ordre que le rêve ou l'erreur.

Néanmoins, Foucault va utiliser une paraphrase poétique de Montaigne : « On n'est pas toujours sûr de ne pas rêver, jamais certain de n'être pas fou ».[10] Peu importe ici si cette paraphrase de Montaigne correspond ou non au mouvement propre du texte cité, il s'agit pour Foucault de l'opposer aux méditations cartésiennes de la manière la plus claire, à propos des arguments du rêve et de l'erreur :

9 Et non pas XXVI comme l'indiquent toutes les versions de HF (n. 2).
10 HF, p. 58.

1. *La vérité ne glissera pas tout entière dans la nuit*. En effet, le rêve ne pourrait nous tromper s'il ne présentait un *résidu de vérité* (ce papier, ce feu, cette robe de chambre) et l'imagination élaborer ses créations fantastiques qu'à partir de *natures plus simples* qu'elle ne peut elle-même produire. Ce qui n'est pas le cas du pauvre fou tout nu qui se prend pour un roi vêtu de pourpre, ou une cruche au fragile corps de verre ? Si le délire manifeste si évidemment son impossibilité, et se révèle incapable de tromper celui qui n'est pas fou, c'est que la folie est définitivement écartée de l'horizon de la raison.
2. Moi qui pense, je ne peux pas être fou :

> On ne peut, en revanche, même par la pensée, supposer qu'on est fou, car la folie justement est condition d'impossibilité de la pensée.[11] Désormais, la folie est exilée […]. Une ligne est ainsi tracée qui va bientôt rendre impossible l'expérience si familière à la pensée d'une Raison déraisonnable, d'une raisonnable Déraison.[12]

Derrida 1963

Dès la parution de *L'Histoire de la folie*, Jacques Derrida prend pour cible ce passage et objecte à sa lecture par Michel Foucault, dans une conférence du 4 mars 1963 au Collège philosophique d'abord,[13] conférence reprise ensuite dans la *Revue de Métaphysique et de Morale* (1964, 3/4), et enfin dans le recueil intitulé *L'écriture et la différence*.[14] Bien sûr, la richesse et le souffle de cette conférence exceptionnelle vont être ici réduits aux seuls arguments critiques qui portent sur le texte de Descartes.

Tout d'abord, Derrida s'interroge sur le projet même de Foucault : comment peut-on dire la folie sans la sousmettre à la raison ? Lorsque Foucault affirme par exemple que *le logos grec n'avait pas de contraire*,[15] cela signifie-t-il, ce qui serait fort inquiétant, qu'il contenait la folie en son sein ? D'autre part, Derrida se demande comment comprendre *l'historicité* de Descartes et comment Foucault peut bien penser le statut de l'intervention carté-

11 HF, p. 57.
12 HF, p. 58.
13 Selon Elisabeth Roudinesco (*Philosophes dans la tourmente* [Paris : Fayard, 2005]), Foucault assista à la séance sans prendre la parole.
14 Jacques Derrida : *L'écriture et la différence* (Paris: Seuil, « Tel Quel », 1967) (ci-après ED).
15 HF (n. 2).

sienne dans le rapport de la raison à la folie. Et comment comprendre celle même de Foucault, faisant strictement appel au seul langage de l'exclusion et de l'interdiction pour décrire le partage cartésien incriminé : « Faire l'histoire de l'historicité ? Faire l'histoire de l'origine de l'histoire ? ».[16] Et là où Foucault renvoie aux pratiques sociales du Grand Siècle, Derrida répond par l'inclusion originaire de la folie dans la raison et la *dissension* progressive : « Comme toujours, la dissension est interne. Le dehors (est) le dedans, s'y entame et le divise selon la déhiscence de l'Entzweiung hégélienne ».[17]

Ensuite, et c'est le point le plus important de sa lecture, Derrida relativise le passage sur la folie incriminée par Foucault dans les *Méditations*, en faisant remarquer qu'on pourrait bien l'attribuer à l'interlocuteur du dialogue indirect, c'est-à-dire au « non-philosophe imaginaire »,[18] et par conséquent être largement inclus et dépassé par l'affrontement vertigineux avec le Malin Génie qui va suivre. A l'appui de cette lecture, Derrida cite le *Sed forte* qui introduit l'argument et que le duc de Luynes n'a pas traduit en français. On pourrait donc reconstituer ainsi l'échange philosophique :

1. Le philosophe : Nous devons mettre en doute la réalité sensible, car les sens nous mentent parfois, à travers les méprises de la perception et les créations fantastiques de l'imagination.
2. Le lecteur, objectant : Mais vous ne pouvez pas nier que vous êtes dans ce fauteuil, vêtu d'une robe de chambre, sans ressembler aux fous qui se prennent pour des rois alors qu'ils sont très pauvres, ou se croient vêtus royalement alors qu'ils sont tout nus.
3. Le philosophe : Il existe pourtant une situation limite, et pourtant familière, dans laquelle les sens nous trompent bien plus gravement que dans la folie, c'est l'expérience du rêve.

Une telle lecture du passage, que Derrida présente d'ailleurs comme *préalable* à toute question concernant l'historicité ou le rôle historique de Descartes, amène donc l'inclusion de la folie dans le rêve, et non l'exclusion de la folie par le rêve. Inclusion incomplète cependant puisque subsiste dans le rêve même des éléments intelligibles qui lui confèrent précisément ce pouvoir de séduction que l'« extravagance » seule ne lui donnerait pas.

16 ED (n. 14), p. 68.
17 ED, p. 62.
18 ED, p. 78.

En revanche, dans l'affrontement avec l'argument du Malin Génie,[19] les termes mêmes de la folie semblent repris par Descartes, là même où ce qui faisait la vérité du rêve est mise en question. Selon Derrida, l'expérience relative de la folie qu'on enferme ou qu'on tourne en dérision devient alors l'expérience philosophique par excellence. Lorsque Descartes décrit la suspension hyperbolique du jugement, il écrit en effet : « Je me considérerai moi-même comme n'ayant point de mains, point d'yeux, point de chair, point de sang … », ne se trouve-t-il pas au plus près de l'expérience de ceux qui « s'imaginent être des cruches ou avoir un corps de terre » ? « L'acte du Cogito vaut même si je suis fou, même si ma pensée est folle de part en part »[20] et par conséquent il constate que la philosophie n'écarte nullement la folie, mais en fait l'expérience à son point le plus élevé.

Enfin, Derrida conclut son évocation affolante du Cogito en se demandant si, par sa lecture restrictive de Descartes, Foucault n'a pas procédé à « un puissant geste de protection et de renfermement. Un geste cartésien pour le XX^e^ siècle ».[21]

Foucault 1972

Foucault ne laissera pas cette attaque sans réponse, et publiera une longue argumentation en faveur de sa lecture de Descartes en 1972, dans la réédition de *l'Histoire de la folie*.[22] Sur le plan herméneutique, on assiste donc à l'opposition exemplaire de deux lectures, l'une en extériorité et l'autre en intériorité. L'enjeu pour l'histoire des idées contemporaines en est important : l'irrationalité est-elle une production *sociale* et historique ou une *expérience* proprement philosophique qui renvoie à l'origine de la pensée ?

Pour affronter Derrida sur son propre terrain,[23] Foucault commence par revenir sur les termes mêmes dont se sert Descartes. Les fous sont appelés dans le texte latin *insani*, terme aussi bien médical que familier: « Etre in-

19 C'est-à-dire avec Dieu lui-même et sa volonté absolue, comme l'indiquait précisément le Discours de la Méthode, où le Malin Génie n'intervient pas encore.

20 ED (n. 14), p. 85.

21 Ibid.

22 Michel Foucault : *Histoire de la folie* (Paris : Gallimard, « Bibliothèque des Histoires », 1972).

23 Et lui faire la leçon ! Le thème du maître et du disciple fait intrinsèquement partie de cet échange philosophique, sans qu'on puisse en mesurer l'ironie, à mon avis plus apparente que réelle.

sanus c'est se prendre pour ce qu'on n'est pas, c'est croire à des chimères, c'est être victime d'illusions ».[24] Or si je prenais exemple[25] sur les fous, dit Descartes, je serais *demens*, terme juridique qui indique la restriction des droits et de la responsabilité de certains individus. Et l'expression même par laquelle le philosophe stigmatise les fous, *sed amentes sunt isti*, montre pour Foucault la forte rupture qui est présente dans le texte de Descartes entre rêve et folie. « L'extravagance du rêve garantit son caractère démonstratif comme exemple : sa fréquence assure son caractère accessible comme exercice ».[26] L'évocation du rêve ne fait pas perdre au philosophe ses droits à la conduite et à la poursuite de la méditation, tandis que celle de la folie – Foucault ne relève pas la proposition de lecture de Derrida à propos du dialogisme latent dans le texte de Descartes : est-ce là une concession tacite ? – rend toute démonstration impossible.

Et la conclusion de Foucault, très violente :

> Je dirai que c'est une petite pédagogie, historiquement bien déterminée qui, de manière très visible, se manifeste. Pédagogie qui enseigne à l'élève qu'il n'y a rien hors du texte […] Pédagogie qui inversement donne à la voix du maître cette souveraineté sans limites qui lui permet indéfiniment de redire le texte.[27]

On ne reviendra pas ici sur les rapports « pédagogiques » noués par Foucault et Derrida dans le cadre de L'Ecole Normale supérieure, mais bien sur le geste d'exclusion philosophique : le maître n'est pas un maître ou n'est qu'un maître, qui redit indéfiniment le texte philosophique, dans le cadre de sa « petite pédagogie ».

Ferry et Renaut 1985

Dans l'ouvrage bien trop décrit qui a fait le procès de toute une époque, *La pensée 68*,[28] Luc Ferry et Alain Renaut affirment que le différend entre Foucault et Derrida peut être « facilement désamorcé » si on considère que Descartes ne peut éviter le soupçon de ressembler à un fou en mettant en cause la perception au premier stade de son parcours, mais que le jugement

24 HF (n. 2), p. 590.
25 Tout le rapport aux fous est marqué par le registre de la comparaison, tandis que celui du rêve obéit au registre de la mémoire, donc de l'intériorité (HF, p. 589).
26 HF, p. 585.
27 HF, p. 602.
28 Ferry/Renaut, op. cit. (n. 7).

philosophique peut prouver ultérieurement sa validité et donc légitimer la suspension de la croyance aux vérités sensibles. Cette lecture kantienne de Descartes, articulée autour de la théorie du jugement, ne peut faire l'objet d'une discussion ici, mais elle introduit l'idée que le « désaccord n'engage pas l'essentiel ».[29] Renvoyant ainsi dos à dos « nietzschéisme français » et « heideggerisme français », leurs deux bêtes noires, les auteurs nous donnent néanmoins une idée juste de l'unité de principe des deux lectures opposées : « Le texte ne se laisse comprendre qu'à partir d'autre chose que lui-même». Foucault était décédé au moment de la parution de *La pensée 68*, mais aurait-il répondu alors aux imputations de surinterprétation des textes en parlant d'une « très petite pédagogie » ?

Revenant sur le détail de l'argumentation des deux auteurs à propos de Descartes, mon exposé n'en veut pas moins montrer à quel point, dès ses premiers débuts, au temps de sa fortune littéraire et philosophique à peine naissante, le « structuralisme » ou le « postmodernisme » n'a été qu'un regroupement artificiel de positions fort opposées, même si une certaine convergence se dessine dans le procès des philosophies de la conscience, et donc de la « modernité »[30] dans son partage, quel qu'il soit, toujours bien trop court et contraignant entre le rationnel et l'irrationnel. *L'Histoire de la folie*, dans ce sens, malgré toute la force de sa démonstration historique, nous pousserait à la nostalgie de l'humanisme pré-kantien, et enfin au retour à la tolérance modeste de Montaigne.

29 Op. cit. p. 126.
30 Ibid.

Studia philosophica 68/2009

Thomas Sturm

Selbsttäuschung: Wer ist hier (ir)rational und warum?*

I argue that both psychological and philosophical studies of self-deception suffer from serious weaknesses, albeit different ones. On the one hand, psychologists often use varying and unreflective conceptions of self-deception in their research. On the other hand, philosophers either ignore the necessity of paying attention to psychological research – or, if they do, they use empirical studies of human cognition and reasoning without realizing that theories and data are loaded with highly problematic assumptions. These weaknesses become centrally important in discussions about whether self-deception is a rational or irrational phenomenon. Both parties have tried to advance their views without clearly stating which normative theory of rationality they are committed to, and without explaining how this theory can be used to study or assess self-deception. More thorough interdisciplinary work is required to overcome naive conceptions and one-sided methodologies in the study of self-deception.

1. Einleitung

Während sowohl Philosophen als auch Psychologen am Phänomen der Selbsttäuschung interessiert sind, kommt es zu wenig Austausch; und wo er stattfindet, sprechen die Vertreter beider Disziplinen nicht selten aneinander vorbei. Viele derzeitige Analysen in der Philosophie sind zwar scharfsinnig, aber verlieren sich zu oft in fingierten Beispielen, von denen offen ist, ob ihnen reale Fälle von Selbsttäuschung entsprechen, und die auch zu weit von der empirischen Forschung über Voraussetzungen und Konsequenzen von Selbsttäuschung entfernt sind. Psychologen wiederum stellen häufig

* Dieser Beitrag baut auf einer erheblich überarbeiteten Übersetzung meines Beitrags *Self-Deception, Rationality, and the Self* (in *Teorema* 26 [2007] S. 73-95). Insbesondere habe ich das dort behandelte Thema der Rolle des Selbst in der Selbsttäuschung entfernt und beschäftige mich hier ausführlicher mit der Rationalität oder Irrationalität der Selbsttäuschung. Ich danke Gerd Gigerenzer für Diskussionen besonders zu «overconfidence»-Studien und Anton Hügli für verschiedene hilfreiche Fragen.

empirische Hypothesen über Selbsttäuschung auf, ohne die erforderliche Begriffsanalyse hinreichend durchgeführt zu haben. Manchmal werden so verschiedene Konzeptualisierungen von Selbsttäuschung miteinander vermengt und daher problematische Thesen darüber formuliert, was die empirischen Daten tatsächlich zeigen.[1] Dabei sollten Vertreter beider Disziplinen mehr zusammenarbeiten. Wie ich im Folgenden zeigen möchte, erfordert insbesondere die umstrittene Frage, ob Selbsttäuschung rational oder irrational ist, dringend mehr interdisziplinäre Kooperation.

Dazu möchte ich zunächst verdeutlichen, wie sich philosophische und psychologische Studien zur Selbsttäuschung in ihren Methoden und Hauptfragen unterscheiden (2). Dann werde ich argumentieren, dass die vertrauten philosophischen Methoden der Begriffsanalyse nicht ausreichen, um ein adäquates Verständnis von Selbsttäuschung zu erreichen (3). Die naheliegende Hoffnung ist, dass Psychologen hier weiterhelfen. Am Beispiel von einflussreichen Arbeiten zu der grundlegenden Frage, ob Selbsttäuschung überhaupt empirisch nachweisbar ist, wird jedoch deutlich werden, dass psychologische Studien nicht zuletzt gründlicherer begrifflicher Analysen bedürfen (4). Den gleichsam umgekehrten Punkt mache ich, indem ich philosophische Arbeiten diskutiere, die sich auf scheinbar stabile psychologische Forschungsergebnisse stützen. Ich denke hier besonders an den gründlichsten derartigen Ansatz: Alfred Meles nicht-intentionalistische oder deflationäre Konzeption der Selbsttäuschung. Mele zufolge sollte die Selbsttäuschung in den gewöhnlichen Fällen als eine Art Vorurteil oder verzerrte Meinung (*biased belief*) verstanden werden, in dem die Voreingenommenheit – und damit die behauptete Irrationalität der Selbsttäuschung – *motiviert* ist, aber keine Absicht im vollen Wortsinn erfordert. Meles Nicht-Intentionalismus vertraut wesentlich auf das «heuristics and biases»-Programm in der aktuellen psychologischen Forschung. Doch er unterschätzt dabei innerpsychologische Debatten über Theorien menschlichen Urteilens und Schließens dramatisch (5-7). Dass Selbsttäuschung irrational ist, darf nicht einfach vorausgesetzt werden. Denn zum einen liegt es nahe anzunehmen, dass sich praktisch alle Beispiele von Selbsttäuschung auch rationalisieren lassen. Zum anderen haben derartige Versuche jedoch zu wenig

[1] Das ist kein auf das Phänomen der Selbsttäuschung beschränktes Problem: Die Psychologie ist voll von nichtempirischen Annahmen, die oft auch notwendig sind, um den Gegenstand der Forschung genau zu bestimmen; doch häufig werden diese Annahmen nicht ausreichend geklärt. Vgl. Jochen Brandtstädter, Thomas Sturm: *Apriorität, Erfahrung und das Projekt der Psychologie*, in *Zeitschrift für Sozialpsychologie* 35 (2004) S. 15-32.

überdacht, was die Annahme von Rationalität erfordert, und zu oft unklare Begriffe von Selbsttäuschung verwendet. Zudem erfordert die Charakterisierung von Selbsttäuschung als (ir)rational substanzielle und umstrittene Rationalitätsnormen (8). Philosophen wie Psychologen werden sich offenbaren müssen: Entweder sie explizieren und begründen ihre Rationalitätsstandards – oder ihre Beispiele von Selbsttäuschung, ja sogar die verbreitete Annahme der Existenz von Selbsttäuschung, werden dubios bleiben. Selbst wenn wir am Ende bei der Auffassung bleiben sollten, dass Selbsttäuschung irrational ist, kann dies in seriöser Weise nur behauptet werden, wenn die Rationalitätsstandards reflektiert werden (9). Philosophen wie Psychologen haben sich also bislang ziemlich irrational verhalten, wenn sie meinten, sie könnten die jeweils andere Disziplin ignorieren.

2. *Ein philosophisches Paradoxon und ein Problem der Psychologie*

Plato hielt die Selbsttäuschung für eines der größten Übel, da der Betrüger in der Seele selbst sei, keinen Schritt weit entfernt vom Betrogenen (*Kratylos* 428d). Bischof Butler, Adam Smith, Kant, Sartre und andere haben darüber geschrieben. Seit der letzten Hälfte des 20. Jahrhunderts jedoch ist die Literatur hinsichtlich des Themas geradezu explodiert. Dabei lassen sich zwei neue Entwicklungen feststellen: eine philosophische und eine psychologische.

In früheren Jahrhunderten haben die Philosophen die Selbsttäuschung meist als ein ethisches Problem behandelt. In den letzten Jahrzehnten jedoch haben sie sich auf das theoretische Problem des «Paradoxons der Selbsttäuschung» konzentriert. Ein theoretisches Rätsel tritt nämlich auf, wenn man die Selbsttäuschung anhand des Vorbilds der absichtlichen Täuschung einer anderen Person analysiert. Eine solche Täuschung liegt etwa vor, wenn Cheney beabsichtigt, Powell vom Gegenteil dessen zu überzeugen, was seine (Cheneys) eigene beste Überzeugung ist, und dies mit verschiedenen vorsätzlichen Maßnahmen durchführt. Wenn man Selbsttäuschung analog konzeptualisiert, dann muss eine Person sich absichtlich dazu bringen, an etwas zu glauben, das sie im selben Moment nicht glaubt. Ist so etwas überhaupt möglich? Diskussionen über dieses Problem dominieren die philosophische Debatte über Selbsttäuschung seit der zweiten Hälfte des 20. Jahrhunderts.[2]

[2] Inzwischen werden sogar zwei Formen des Paradoxons unterschieden, ein «statisches» (Wie kann man gleichzeitig p und ¬p glauben?) und ein «dynamisches» (Wie kann man sich absichtlich dazu bringen, p zu glauben, während und

Psychologen wiederum haben ‘Selbsttäuschung’ und verwandte Begriffe ebenfalls schon länger verwendet. Neu in den letzten Jahrzehnten ist, dass sie sich für die Fragen interessieren, die sie mit empirischen Mitteln beantworten wollen. Kann man beispielsweise überhaupt nachweisen, dass es Fälle von Selbsttäuschung gibt? Was sind die Mechanismen und Funktionsweisen von Selbsttäuschung? Wie passt Selbsttäuschung zu empirischen Theorien über das Denken und Schließen des Menschen?

3. Ansätze zur Auflösung des Paradoxons

Zur begrifflichen Auflösung des Paradoxons der Selbsttäuschung gibt es verschiedene Ansätze. Einige Autoren wie etwa Raphael Demos in einem klassischen Artikel nehmen an, dass Selbsttäuschung möglich ist, weil die widersprüchlichen Meinungen auf verschiedenen Bewusstseinsebenen auftreten: Während die bevorzugte Meinung transparent ist und eine damit inkompatible Meinung nach außen hin auch abgelehnt wird, existieren sichere Indizien dafür, dass letztere Meinung unbewusst vorhanden ist.[3] Herbert Fingarette dagegen hat gefordert, die Rede von Meinungen und unbewussten Zuständen aufzugeben. Vielmehr sollten wir von verschiedenen «Verpflichtungen» (*commitments*) sprechen, die wir im Falle von Selbsttäuschung nicht fähig sind auszusprechen.[4] Robert Audi wiederum macht ein moderateres Angebot. Er stellt fest, dass eine der zwei inkompatiblen Propositionen nicht in Form einer Meinung akzeptiert wird, sondern als «aufrichtige Bekundung».[5] Donald Davidson schließlich hat behauptet, dass die Selbsttäuschung durch die Aufteilung des Geistes einer Person in unhabhängige Gruppen von Zuständen ermöglicht wird – unabhängig in dem

gerade weil man ¬p glaubt?) Vgl. Alfred Mele: *Recent Work on Self-Deception*, in *American Philosophical Quarterly* 24 (1987) S. 1-17; A. M.: *Self-Deception Unmasked* (Princeton: Princeton UP, 2001). Diese Unterscheidung benötige ich im Folgenden nicht.

3 Raphael Demos: *Lying to Oneself*, in *Journal of Philosophy* 57 (1960) S. 588-595; eine bis in die jüngste Zeit immer wieder vertretene Position: vgl. Baljinder Sahdra, Paul Thagard: *Self-Deception and Emotional Coherence*, in *Minds and Machines* 13 (2003) S. 213-231. Demos vertritt dabei nicht etwa ein starkes freudianisches Konzept des Unbewussten.

4 Herbert Fingarette: *Self-Deception* (London: Routledge & Kegan Paul, 1969).

5 Robert Audi: *Self-Deception, Action, and Will*, in *Erkenntnis* 18 (1982) S. 133-158.

Sinn, dass die gewöhnlichen logischen und epistemologischen Beziehungen zwischen ihnen zerrissen sind, obwohl die Zustände kausal miteinander verbunden bleiben, so dass die Annahme eines gespaltenen Selbst vermieden werden kann.[6]

Alle diese Explikationsversuche haben Probleme, die ich hier nicht diskutieren möchte. Wichtig ist vielmehr, den ihnen gemeinsamen Maßstab zu erfassen. Allen gemein ist nämlich die Suche nach der Lösung des Paradoxons der Selbsttäuschung. Und eine Explikation löst das Paradoxon so gut wie eine andere. Dabei habe ich nur eine begrenzte Anzahl der begrifflichen Optionen umrissen.[7] Die schiere Vielfalt von Möglichkeiten verdeutlicht, dass es mehr als fraglich ist, ob man durch begriffliche Auflösung des Paradoxons allein das Phänomen der Selbsttäuschung angemessen erfassen kann.

4. Existenzielle Erörterungen über Selbsttäuschung mittels eines Experiments – und seine Probleme

Gibt es nicht einen offensichtlichen Lösungsweg? Ist es nicht eine natürliche Aufgabe der Psychologie, dieses Phänomen angemessen zu beschreiben und zu erklären, die diese Disziplin auch besser erledigen kann als irgendeine andere? Leider wirkt ein näherer Blick in die entsprechende Literatur ernüchternd; Psychologen geraten hinsichtlich der Selbsttäuschung schnell in Schwierigkeiten.

Um dies zu zeigen, möchte ich mit einer vielleicht überraschenden Frage beginnen: Kommt Selbsttäuschung überhaupt jemals vor? Auch Psychologen und andere empirische Wissenschaftler sind oft fraglos von der Existenz dieses Phänomens ausgegangen. So ist vermutet worden, dass falsche Selbst-

6 Donald Davidson: *Deception and Division*, in *The Multiple Self*, hg. von John Elster (Cambridge: Cambridge University Press, 1986) S. 79-82. Die stärkere Annahme wird Davidson etwa von Alexander Bird (*Rationality and the Structure of Self-Deception*, in *European Review of Philosophy* 1 [1994] S. 19-38) zugeschrieben; tatsächlich behauptet wird sie z.B. von Amelie O. Rorty (*Akratic Believers*, in *American Philosophical Quarterly* 20 [1983] S. 175-183).

7 Auch etwa die in der Literatur oft mitbehandelte Frage, wie man Selbsttäuschung von Wunschdenken unterscheiden kann, schränkt die Suche nach dem besten Konzept nur geringfügig ein. Weitere Optionen s. unten in Abschnitten 5-6; umfassende Bibliographien finden sich im Internet: http://consc.net/mindpapers/5.1l.5.13, http://philpapers.org/browse/self-deception/ und http://plato.stanford.edu/entries/self-deception/.

darstellungen in Persönlichkeitstests eher infolge von Selbsttäuschung als Fremdtäuschung auftreten.[8] Auch wurde mithilfe von Selbsttäuschung zu erklären versucht, warum Menschen an Hypothesen festhalten, selbst wenn deren Nichtbestätigung bereits erfolgt ist.[9] Selbsttäuschung wird auch als ein im Alltagsleben allseits präsenter Zustand angenommen. So wird sie im Verneinen von Krankheit, im fahrlässigen Verhalten von Autofahrern und im zuversichtlichen Optimismus von Arbeitslosen oder Soldaten im Kampfeinsatz gesehen.[10] Die Beispiele lassen sich vermehren.

Doch in keiner dieser Studien ist nachgewiesen worden, dass Selbsttäuschung überhaupt jemals auftritt. Haben die genannten Hypothesen also eine Basis? Auch angesichts des *prima facie* paradoxen Charakters des Phänomens sollte die Annahme ihrer Existenz nicht einfach hingenommen werden, so beliebt und alltäglich sie auch sein mag. Vielleicht werden Zuschreibungen von Selbsttäuschung nur vorgenommen, um vermeintlichen Selbstbetrügern irrationales oder gar unmoralisches Verhalten vorzuwerfen? Ist Selbsttäuschung lediglich ein Konstrukt unserer Kultur, von dem wir uns besser befreien sollten?

Vor diesem Hintergrund ist es keine bloß eitle Aufgabe, wenn Psychologen versuchen, die Existenz von Selbsttäuschung experimentell zu demonstrieren. Im frühesten und zugleich wohl auch letzten Ansatz einer derartigen Untersuchung sind Gur und Sackeim von der Idee Demos' ausgegangen, dass die widersprüchlichen Meinungen auf «verschiedenen Ebenen des Bewusstseins» gehalten werden. Genauer behaupten sie folgende Bedingungen als notwendig und hinreichend für Selbsttäuschung:

1. Das Individuum hat zwei einander ausschließende Meinungen (p und $\neg p$).
2. Diese zwei Meinungen werden gleichzeitig gehalten.
3. Das Individuum ist sich nur bei einer der Meinungen bewusst, dass es sie hat.
4. Der Akt, der bestimmt, welche Meinung unbewusst ist, ist ein motivationaler Akt.[11]

8 Paul E. Meehl, Starke R. Hathaway: *The K Factor As a Suppressor Variable in the Minnesota Multiphasic Personality Inventory*, in *Journal of Applied Psychology* 30 (1946) S. 525-564.

9 Peter C. Wason, Philip N. Johnson-Laird: *Psychology of Reasoning* (Cambridge, MA: Harvard University Press, 1972).

10 Vgl. Sahdra/Thagard, op. cit. (Fn. 3) S. 213.

11 Ruben C. Gur, Harold A. Sackeim: *Self-Deception: A Concept in Search of a Phenomenon*, in *Journal of Personality and Social Psychology* 37 (1979) S. 147-169, hier S. 149 (meine Übersetzung). – Vgl. Harold A. Sackeim, Ruben

Zum Nachweis von Selbsttäuschung werden dann Stimmenerkennungsexperimente eingesetzt. In einem typischen Experiment werden Einzelpersonen gefragt, ob sie eine auf Tonband aufgenommene Stimme für ihre eigene oder die einer anderen Person halten. Während die Probanden darüber berichten, werden Hautreflexe gemessen. Das soll Auskunft darüber geben, ob die Probanden dann, wenn sie ihre eigene Stimme nicht zu erkennen vorgeben, dies unterschwellig doch tun. Gur und Sackeim argumentieren zudem mit Hilfe von Persönlichkeitstests sowie einem Fragebogen speziell zur Selbsttäuschung, dass Menschen mit negativem Selbstwertgefühl oder mit diskrepanten Meinungen darüber, was sie selbst sind und was sie sein sollten, Konfrontation mit sich selbst – etwa mit der eigenen Stimme – unangenehm finden. Entsprechend sei die geäußerte «Nichterkennung» der eigenen Stimme motiviert. Umgekehrt finden Personen mit positivem Selbstwertgefühl an der Konfrontation mit der eigenen Stimme sogar Gefallen. So sei Selbsttäuschung tatsächlich nachweisbar.

Diese Studie ist in der psychologischen Forschung durchaus einflussreich geworden. Sie hat den Weg für weitere Untersuchungen von Mechanismen und Funktionen der Selbsttäuschung geöffnet, etwa für psychologische Erklärungen, Wirkungen oder den möglichen Nutzen dieses Phänomens.[12] Auf solche Studien komme ich später zurück.

Gegen Gurs und Sackeims Ansatz lassen sich zumindest drei Einwände vorbringen. Erstens ist argumentiert worden, dass die Aufgabe, die eigene

C. Gur: *Self-Deception, Self-Confrontation, and Consciousness*, in *Consciousness and Self-Regulation*, hg. von G. E. S. D. Shapiro (New York: Plenum, 1978) S. 139-197; G. E. S. D. S.: *Self-Deception, Other-Deception and Self-Reported Psychopathology*, in *Journal of Consulting and Clinical Psychology* 47 (1979) S. 213-215. Sie folgen Demos' Begriffsexplikation, weil sie Selbsttäuschung in Analogie zum Phänomen der «perceptual defense» sehen. Menschen neigen manchmal dazu, unangenehme Wahrnehmungen zu meiden, aber damit das überhaupt möglich ist, muss man einen bestimmten Reiz erst einmal erfassen. Das Konzept des Wahrnehmens eines Reizes bedeutet daher zum einen so viel wie 'vom sensorischen System erfasst werden' und zum anderen so viel wie 'bewusst erfahren werden'.

12 Vgl. etwa George A. Quattrone, Amos Tversky: *Self-Deception and the Voter's Illusion*, in *The Multiple Self*, hg. von John Elster (Cambridge: Cambridge University Press, 1986) S. 35-58; Joan S. Lockhard, Delroy L. Paulhus (Hg.): *Self-Deception: An Adaptive Mechanism?* (Englewood Cliffs, NJ: Prentice-Hall, 1988); Delroy L. Paulhus, Douglas B. Reid: *Enhancement and Denial in Socially Desirable Responding*, in *Journal of Personality and Social Psychology* 60 (1991) S. 307-317.

Stimme zu erkennen, problematisch ist, da ähnliche Ergebnisse erzielt werden, wenn die Testperson Stimmen von Bekannten erkennen soll. Dieser Einwand ist freilich schwach: Er zeigt nicht, dass Fehler bei der Erkennung anderer Stimmen nicht auch auf Selbsttäuschung beruhen können (weil sie etwa auch motiviert sind). Ein zweiter Einwand jedoch zieht stärker:[13] Wegen der gewählten Explikation des Begriffs von Selbsttäuschung, demzufolge es sich um eine Spannung zwischen gleichzeitig, aber auf verschiedenen Bewusstseinsebenen gehaltenen *Meinungen* handelt (Bedingungen 1-3 bei Gur und Sackeim), müssen die im Experiment gemessenen Hautreflexe Indikatoren von (unbewussten) Meinungen sein. Diese Annahme ist jedoch alles andere als zwingend. Beispielsweise mögen die Hautreflexe ja nur Unsicherheit darüber ausdrücken, was die Probanden öffentlich über die gehörten Stimmen vorgeben, oder auch eine Sensibilität für die gehörten Stimmen.[14] Drittens ist zweifelhaft, ob Gur und Sackeim wirklich Demos' Begriff der Selbsttäuschung verwendet haben. Ihre Beispiele zeigen nicht, dass aufseiten des Selbstbetrügers auch eine Intention vorliegen muss, sich selbst zu täuschen, wie Demos dies gefordert hat.[15] Bloß ein Motiv zu haben (Bedingung 4 bei Gur und Sackeim) ist schwächer: Es erfordert beispielsweise keine praktischen Überlegungen über Zwecke und Mittel, die zu einer gefestigten Absicht führen. Man sollte sich noch einmal die Parallele zur intentionalen Fremdtäuschung vor Augen halten. Wir würden nicht von einer Lüge sprechen, wenn z.B. Hans (der glaubt, dass p) unbeabsichtigt verursacht, dass Franz glaubt, dass $\neg p$. Wir würden dies selbst in dem Fall nicht tun, in dem Hans zwar auch wünscht oder ein Motiv dafür hat, dass Franz glaubt, dass $\neg p$, aber Hans etwa aus moralischen Gründen der Versuchung zum Lügen widersteht, während Franz aus irgendeinem anderen Grund heraus zu der Meinung gelangt, dass $\neg p$. Bei intentionaler

13 Der Einwand stammt von William Douglas, Keith Gibbins: *Inadequacy of Voice Recognition as a Demonstration of Self-Deception*, in *Journal of Personality and Social Psychology* 44 (1983) S. 589-592. Zur Erwiderung: Harold A. Sackeim, Ruben C. Gur: *Voice Recognition and the Ontological Status of Self-deception,* in *Journal of Personality and Social Psychology* 48 (1985) S. 1365-1368.

14 Vgl. Alfred Mele: *Recent Work on Self-Deception*, in *American Philosophical Quarterly* 24 (1987) S. 1-17. Später hat Sackeim dies eingeräumt (Harold A. Sackeim: *Self-deception: A Synthesis*, in Lockhard/Paulhus, op. cit. [Fn. 12] S. 146-165). Für weitere Kritik an vermeintlichen empirischen Demonstrationen des Haltens widersprüchlicher Meinungen vgl. Mele: *Self-Deception Unmasked*, op. cit. (Fn. 2) S. 76-93.

15 Demos, op. cit. (Fn. 3) S. 588.

Fremd- wie Selbsttäuschung muss der Wunsch auch bei höherstufiger Überlegung gewollt sein.

Die Probleme dieses experimentellen Existenzbeweises von Selbsttäuschung sind zum Teil methodologischer oder empirischer Natur (wie beim ersten vorgeführten Einwand), zum Teil begrifflicher Natur (wie bei den beiden letzteren). Daher ist es nicht leicht zu sagen, was zuerst zu tun wäre, um voranzukommen. Sollte man eine andere Begriffsexplikation wählen? Aber welche? Oder lieber das Experiment ändern? Oder gleich beides? Vermutlich würden ähnliche Probleme beim Gebrauch anderer Konzeptualisierungen auftreten. Wenn man etwa annimmt, dass die Selbsttäuschung z.B. eine noch tiefere Aufspaltung des Selbst erfordert als eine Verschiedenheit von Bewusstseinsebenen, dann wäre es womöglich erneut unnütz, den Existenzbeweis mittels Hautreaktionen durchzuführen.

Manche Autoren behaupten nun, empirische Studien zur Selbsttäuschung seien ohnehin überflüssig, da es sich nur um ein soziales oder kulturelles Konstrukt handele, wie der zuvor angesprochene skeptische Einwand schon angedeutet hatte.[16] Jedoch wäre es verfrüht, eine empirische Demonstration von Selbsttäuschung völlig auszuschließen, auf deren Basis man den Begriff dieses Phänomens gebrauchen könnte, um weitere Untersuchungen und Theorien zu entwickeln. Wegen der vielfältigen Möglichkeiten der Begriffsexplikation ist dies eine schwierige Aufgabe; nicht weniger, aber eben auch nicht unbedingt mehr. Diese brauche ich hier nicht zu verfolgen – das wäre schließlich nur in Zusammenarbeit mit empirischen Untersuchungen sinnvoll. Ziel war zunächst nur zu verdeutlichen, dass Psychologen sich nicht immer ausreichend um die notwendige konzeptuelle Analyse bemühen.

5. Sparsam sein? Selbsttäuschung als «biased belief»

Nun möchte ich gleichsam die umgekehrte Schwäche auf philosophischer Seite illustrieren: den Versuch, den Begriff der Selbsttäuschung mittels beliebter empirischer Theorien über Mechanismen und Funktionen menschlichen Denkens und Urteilens zu analysieren. So ein Ansatz entspricht schließ-

[16] Kenneth J. Gergen: *The Ethnopsychology of Self-Deception*, in *Self-Deception and Self-Understanding*, hg. von Mike W. Martin (Lawrence, Kansas: Kansas University Press, 1985) S. 228-243; Brian L. Lewis: *Self-Deception: A Postmodern Reflection*, in *Journal of Theoretical and Philosophical Psychology* 16 (1996) S. 49-66.

lich der Hoffnung, nicht nur (aber auch) das genannte Paradoxon aufzulösen, sondern darüber hinaus ein sachlich adäquates Konzept von Selbsttäuschung zu gewinnen. Der bislang gründlichste derartige Ansatz findet sich zweifellos in den Publikationen Alfred Meles.

Mele behauptet, dass das Modell der intentionalen interpersonellen Täuschung unnötig ist, um Selbsttäuschung korrekt zu beschreiben. Vielmehr sollten wir Selbsttäuschung als eine Art von voreingenommener oder von aufgrund bestimmter kognitiver Mechanismen verzerrter Meinung (*biased belief*) auffassen. Damit bezieht Mele sich auf Studien des als «heuristics and biases» bezeichneten, höchst ernst zu nehmenden psychologischen Forschungsprogramms (dessen Hauptvertreter, Daniel Kahneman und Amos Tversky, dafür immerhin einen Nobelpreis für Ökonomie erhalten haben).[17] Diesem Programm zufolge ist unser Denken und Urteilen von kognitiven «Heuristiken» bestimmt, die passable Faustregeln darstellen, jedoch universell angewandt zu systematischen Fehlern führen. Die Fehler werden identifiziert, indem man gewisse Normen als Standards ansetzt – Regeln der Logik, der Wahrscheinlichkeitstheorie oder der Statistik etwa. Letztere Normen lassen sich auch als das «Standardmodell» von Rationalität bezeichnen.[18] Ich werde das Programm in den Abschnitten 7-8 noch näher diskutieren.

Meles Position zufolge sind nicht alle voreingenommenen oder verzerrten Meinungen Selbsttäuschungen. Für letztere ist es spezifisch, dass die Verzerrung oder Voreingenommenheit *motiviert* ist.[19] Es bedarf aber keiner vollen Absicht, keiner intentionalen Handlung und auch keines Festhaltens an sich widersprechenden Meinungen.[20] Meles Konzept ist daher sparsam oder deflationär. Er leugnet dabei nicht, dass intentionale Selbsttäuschungen möglich sind. Die von ihm formulierten Bedingungen sollen hinreichend, aber nicht notwendig sein. Allerdings nimmt er an, dass die gewöhnlichen Fälle von Selbsttäuschung eher seiner Analyse als den intentionalistischen

17 Alfred Mele: *Real Self-Deception,* in *Behavioral and Brain Sciences* 20 (1997) S. 91-102; Mele: *Self-Deception Unmasked*, op. cit. (Fn. 2). Für andere Ansätze, die dieses Forschungsprogramm aufgreifen, vgl. Ariela Lazar: *Deceiving Oneself or Self-Deceived?*, in *Mind* 108 (1999) S. 263-290; David Patten: *How Do We Deceive Ourselves?*, in *Philosophical Psychology* 16 (2003) S. 229-246.

18 Vgl. Edward Stein: *Without Good Reason* (Oxford: Clarendon Press, 1996).

19 Vgl. Ziva Kunda: *The Case for Motivated Reasoning*, in *Psychological Bulletin* 108 (1990) S. 480-498.

20 Der letzte Punkt wurde schon von Frederick A. Siegler (*Demos On Lying to Oneself*, in *Journal of Philosophy* 59 [1962] S. 469-475) gegen Demos, op. cit. (Fn. 3) vorgebracht.

Begriffsexplikationen entsprechen. Man bemerke auch einen wichtigen Vorteil seines Ansatzes: Mit einem derartigen moderaten Konzept von Selbsttäuschung, bei dem sich kein Paradoxon mehr stellt, und das zudem in eine Theorie eingebettet ist, die umfassend Verzerrungen oder Voreingenommenheiten in unseren Urteilen und Entscheidungen behandelt, stellt sich die Frage der Existenz der Selbsttäuschung nicht, oder jedenfalls nicht in der dramatischen Weise wie beim intentionalistischen Begriffsverständnis. Wenn wir häufig unter verzerrten Meinungen leiden und wenn Selbsttäuschungen oft ein Spezialfall hiervon sind, braucht man die Existenz von Selbsttäuschung kaum zu bezweifeln.

Nach dieser Auffassung ist Selbsttäuschung nach wie vor ein irrationales Phänomen, allerdings nicht mehr aus Gründen der Meinungsinkonsistenz oder wegen der Absurdität des Vorhabens, sich selbst absichtlich in die Irre zu führen – sondern vielmehr, weil sie irgendwelche Normen der Logik, der Wahrscheinlichkeitstheorie oder der Statistik verletzt. Man betrachte etwa den von Mele aufgegriffenen Befund, dass 94% der Universitätsprofessoren von sich glauben, besser in ihrem Beruf zu sein als ihre durchschnittlichen Kollegen.[21] Dies kann keinesfalls für alle diese Professoren gelten. Doch obwohl sie irren, müssen sie keine widersprüchlichen Meinungen besitzen oder sich selbst absichtlich getäuscht haben. Vielleicht haben sie ihre Ansicht aus den Feedbacks von Studenten gefolgert, die wiederum aufgrund des gewöhnlichen Wissensvorteils zwischen Lehrern und Schülern denken mögen, ihre Dozenten seien schlicht brillant. Die Professoren mögen daher Fakten über statistische Verteilungen übersehen. Wenn der Befund akzeptabel ist, dann erliegen die Professoren der Täuschung aus verständlichen Motiven und Prozessen – und täuschen sich daher in Meles Sinn selbst.

6. Illusionen des Nicht-Intentionalismus

Doch ist Meles Position überzeugend? Zwei Einwände von eher zwiespältiger Überzeugungskraft seien in diesem und ein weiterer und deutlich stärkerer in den nächsten zwei Abschnitten vorgeführt.

Erstens könnte man vermuten, dass in manchen Fällen von Selbsttäuschung kein Motiv vorliegt. So würde Meles zentrale, das Phänomen kennzeichnende Bedingung verfehlt sein. Dieser Einwand ist zumindest

21 Entnommen aus Thomas Gilovich: *How We Know What Isn't So* (New York: Free Press, 1991) S. 77.

aufgeworfen worden, um zu betonen, dass Meles Annahme, Selbsttäuschung sei stets motiviert, lediglich ein unbegründetes Postulat darstellen mag oder vielleicht nur aus einer Überverallgemeinerung von als typisch angesehenen Fällen herrührt.[22] Allerdings ist dieser Einwand nicht sonderlich zwingend. Man muss ja bei Begriffsexplikationen auch immer von Standardfällen ausgehen, und ohne die genannte Bedingung würde Selbsttäuschung sich überhaupt nicht mehr von anderen verzerrten Meinungen, Vorurteilen oder Täuschungen unterscheiden.

Eine zweite, plausiblere Frage geht in die umgekehrte Richtung: Können nicht einige von unseren Vorurteilen motiviert sein, ohne schon Selbsttäuschungen darzustellen? Wenn dies so wäre, wären Meles Bedingungen nicht nur nicht notwendig, sondern auch nicht hinreichend. In Richtung solcher Überlegungen hat Holton auf das Beispiel eines Rassisten verwiesen, der aus erkennbaren Motiven heraus bei seinen Überzeugungen bleibt, auch wenn die Argumente gegen den Rassismus deutlich vorgeführt worden sind.[23] Holton führt diesen Fall an, weil er denkt, dass Selbsttäuschung weniger durch die Person selbst *verursacht* ist als vielmehr eine Täuschung darstellt, die schlicht die Person selbst *betrifft*. Man irrt sich, anders gesagt, nicht selbst (schon gar nicht intentional nach Holtons Position), sondern *über* sich selbst. Der Rassist hingegen hat keine motivierte falsche Meinung über seine eigene Person, sondern über andere. Doch es ist überhaupt nicht einleuchtend, diese Einschränkung vorzunehmen. Man mag sich selbst über die Handlungen der eigenen Partner täuschen oder über die Talente und den Charakter der eigenen Kinder oder dass die eigene Regierung die Beteiligung an einem Krieg nicht aus den besten moralischen Gründen verweigert. Holton will mit seiner Explikation die ganzen Paradoxien der Selbsttäuschung erledigen. Doch auch wenn man gegenüber der übermäßigen philosophischen Vorliebe für die Beschäftigung mit Begriffsrätseln skeptisch sein darf, ist das zu handstreichartig.

Es gibt bessere Beispiele dafür, dass verzerrte Meinungen motiviert sein können, ohne schon Selbsttäuschungen darzustellen. Bereits Demos hat bestritten, dass der seine Fähigkeiten überschätzende Wissenschaftler sich selbst täuscht, da eben keine Intention für die Täuschung vorgelegen hat.[24] Ein Beispiel dieser Art wird von Mele freilich andersherum interpretiert, mit

22 Vgl. Patten, op. cit. (Fn. 17).

23 Richard Holton: *What is the Role of the Self in Self-Deception?*, in *Proceedings of the Aristotelian Society* 101 (2000) S. 53-69, hier S. 59.

24 Demos, op. cit. (Fn. 3) S. 588.

Verweis auf den oben erwähnten statistischen Befund.[25] So scheinen beide Positionen ihre Begriffsvorstellungen nur fruchtlos hin- und herzuschieben. Ich werde jedoch zeigen, dass Mele mit seiner Position nicht durchkommt (unten, Abschnitt 7).

Zudem kann der Intentionalist eine andere Route einschlagen und die Bedingung bestreiten, dass die beiden inkompatiblen Meinungen *gleichzeitig* gehalten werden müssen – also statt einer synchronen Trennung eine diachrone Trennung der relevanten geistigen Zustände verlangen. Dies kann die Position stärken, der zufolge Motivation allein nicht hinreichend ist. Beispiele hierfür sind leicht denkbar. Ein Mädchen, dessen Häschen gestorben ist und dessen boshafter Bruder es dauernd an den Trauerfall erinnern will, mag bei jedem Versuch des Bruders die Hände auf die Ohren legen, schreien und aus dem Zimmer laufen, um die schmerzhafte Information zu vermeiden. Durch seine intentionalen Handlungen bewahrt es geflissentlich seine Meinung, der Hase lebe noch.[26] Ich hörte vor einer Weile von einem philosophischen Vortrag, in dem der Redner energisch für Ansichten argumentiert hatte, die auf gar keinen Fall für wahr gehalten werden können, weil sie viel zu verworren sind (der Vortrag handelte über Hegel, den «Mythos des Gegebenen» und dergleichen). Ein klar denkender und uneitler Philosoph verließ den Saal und flüsterte zu einigen Kollegen: «Ich glaube, dies war eine sehr, sehr subtile Art von Selbsttäuschung.» Wie könnte man diese Bemerkung verstehen? Vielleicht hatte der Vortragende erfasst, dass ein Philosoph, der über die erwähnten Themen und Autoren redet, es zu Bekanntheit und zum Ruf eines tiefen Denkers bringen kann, und wollte dies auch erreichen. Er mag sich dazu einseitig auf entsprechende Artikel und Bücher gestürzt haben und irgendwann eben zu den fraglichen Meinungen gelangt sein, da ihm niemand klipp und klar gesagt hatte, dass das keine Garantie für die Vermeidung von Unsinn ist. So bildete er die verzerrten Meinungen und verdrängte immer wieder den aufkeimenden Gedanken, dass diese Meinungen gar nicht glaubwürdig sind. Ich will hier offenlassen, ob derartige Fälle typisch sind. Dass die Selbsttäuschung als «sehr, sehr subtil» charakterisiert wurde, zeigt vielleicht, dass sie eher untypisch ist. Allerdings gibt es Autoren, die das anders sehen.[27]

25 Mele: *Self-Deception Unmasked*, op. cit. (Fn. 2) S. 3, 11.

26 Das Beispiel stammt von Christian Perring (*Direct, Fully Intentional Self-Deception Is Also Real*, in *Behavioral and Brain Sciences* 20 [1997] S. 123-124).

27 So José L. Bermudez: *Self-Deception, Intentions, and Contradictory Beliefs*, in *Analysis* 60 (2000) S. 309-319; etwas anders Anna Nicholson: *Cognitive Bias, Intentionality, and Self-Deception*, in *Teorema* 26 (2007) S. 45-58.

Sicher würde Mele wieder antworten, dass solche Fälle eine schlechte Explikation des Konzeptes der Selbsttäuschung darstellen. Er verweist zum Beispiel darauf, dass man bei den vorgeführten intentionalen Handlungen zwischen solchen unterscheiden sollte, die aus irgendwelchen Absichten heraus geschehen, und solchen, die direkt aus einer Absicht der Selbsttäuschung heraus erfolgen. Doch selbst wenn das Mädchen oder der Philosoph intentionale Handlungen vollzogen, die zur Selbsttäuschung führten, ohne ursprünglich dieses Ziel gehabt zu haben, kann Mele nicht ausschließen, dass solche Fälle vorkommen. Damit jedoch stehen wir vor einer Sackgasse: Anhänger des Nicht-Intentionalismus und des Intentionalismus sprechen offenbar von verschiedenen Phänomenen. Ihre Konzeptualisierungen bezwecken, das Paradox der Selbsttäuschung zu lösen, und beide sind darin erfolgreich. Doch wenn man von diesem oder ähnlichen begrifflichen Problemen besessen ist, wird man über die Phänomene keine wirkliche Einsicht erlangen. Insofern ist Meles Ansatz inkonsequent durchgeführt: Obwohl er sich stärker als andere Philosophen auf psychologische Forschungen stützen will, bleibt er vorwiegend im üblichen philosophischen Verfahren des Gebrauchs von ausgedachten Beispielen und Begriffsanalysen befangen. Er bietet zwar einige empirische Argumente für die Existenz der «gewöhnlichen» Fälle von Selbsttäuschung, aber keinen Beleg dafür, dass diese den Großteil von Selbsttäuschung ausmachen. Nun möchte ich zeigen, dass auch seine empirischen Argumente auf schwachen Füßen stehen.

7. Das «heuristics and biases»-Programm und seine Grenzen

Ich will weitere in der Literatur erhobene Einwände nicht erörtern, da sie an vergleichbaren Mängeln leiden wie die eben vorgeführten. Es gibt ohnehin einen dritten und destruktiveren Einwand, der noch nicht klar genug betrachtet worden ist. Er betrifft das von Mele genutzte psychologische Forschungsprogramm von «heuristics and biases» (HB-Programm).

Dem HB-Programm zufolge ist unser Denken und Urteilen von kognitiven «Heuristiken» bestimmt, die passable Faustregeln darstellen, jedoch universell angewandt zu systematischen Fehlern führen.[28] Das Programm

[28] Vgl. zur Übersicht etwa Amos Tversky, Daniel Kahneman: *Judgment Under Uncertainty: Heuristics and Biases*, in *Science* 185 (1974) S. 1124-1131; Richard E. Nisbett, Lee Ross: *Human Inference* (Englewood Cliffs, NJ: Prentice Hall, 1980). Thomas Gilovich, Dale W. Griffin, Daniel Kahneman (Hg.): *Heuristics and Bi-*

benötigt für jeden empirischen Versuch eine zuvor bestimmte Norm, an der dann das Verhalten des Probanden gemessen wird – wie in Abschnitt 5 angesprochen, sind dies in aller Regel Normen des Standardmodells von Rationalität. Hier sind vier Beispiele für das HB-Programm:

(1) Mittels der Wason-Kartenwahlaufgabe hat man geprüft, ob Versuchspersonen die Wahrheitsbedingungen der *materialen Implikation* der Logik befolgen. Wenn die Probanden aufgefordert werden, ihnen vorgelegte Karten (mit Buchstaben und Zahlen oder anderen Inhalten) als Instanzen der Verknüpfung $p \rightarrow q$ zu prüfen, so übersehen sie dabei regelmäßig die für die Falsifizierung wichtigen $\neg q$-Karten. Dies ist oft als ein übermäßiger Hang zu Bestätigungen («confirmation bias») angesehen worden – so, als ob wir für gewöhnlich Anti-Popperianer wären. [29]

(2) Im «Linda problem» wird den Probanden eine Person beschrieben: Linda sei eine aufgeweckte, intelligente und an politischen Themen und Aktivitäten interessierte ehemalige Philosophiestudentin. Dann ist gefragt worden, ob es wahrscheinlicher sei, dass Linda *(A)* eine Bankangestellte ist oder *(A&B)* eine Bankangestellte und Feministin ist. In Kahnemans und Tverskys Studien hierzu haben die Versuchspersonen ganz überwiegend Antworten gegeben, wonach *(A&B)* wahrscheinlicher sei als *(A)*. Das widerspricht jedoch der *Konjunktionsregel* der Wahrscheinlichkeitstheorie – der Regel, der zufolge ein Ereignis *A* niemals weniger wahrscheinlich sein kann als die Konjunktion der (unabhängigen) Ereignisse *A* und *B* (formal: $p(A) \geq p(A\&B)$). Erklärt wurde dies so, dass die Probanden ihr Urteil aufgrund der «Repräsentativität» der Informationen über Linda bilden («representativeness heuristics»).[30]

(3) *Basisraten* sind ein wichtiger Bestandteil zur Bildung von Urteilen über Wahrscheinlichkeiten, werden jedoch häufig ignoriert («base rate neglect»). Probanden wurden in Tests etwa nach dem wahrscheinlichen Beruf einer Person gefragt. Dabei erhielten sie neben der Basisrate (es seien z.B. 30 Anwälte und 70 Ingenieure in einer Population von 100) nur Informationen über die Person, die in keine Richtung Hinweise enthielten. Die Probanden urteilten zumeist, dass es gleich wahrscheinlich

ases: The Psychology of Intuitive Judgment (Cambridge: Cambridge University Press, 2002).

29 Peter Wason: *Reasoning About a Rule*, in *Quarterly Journal of Experimental Psychology* 20 (1966) S. 273-281.

30 Amos Tversky, Daniel Kahneman: *Extensional Versus Intuitive Reasoning: Conjunction Fallacy in Probability Judgment*, in *Psychological Review* 90 (1983) S. 293-315.

sei, ob die Person Anwalt oder Ingenieur sei (50%). Trotz der erhaltenen Basisrateninformation schienen sie also erneut die inhaltliche Beschreibung stärker heranzuziehen.[31] Sogar statistisch ausgebildete Angehörige der medizinischen Fakultät von Harvard (!) neigen dazu, Basisraten zu ignorieren, zumal bei der Schätzung bedingter Wahrscheinlichkeiten – wie bei der Frage, wie wahrscheinlich eine Krankheit ist, gegeben das Vorliegen bestimmter Symptome oder Testergebnisse.[32] Die fehlerhaften Urteile werden zum Beispiel dadurch erklärt, dass die Probanden sich eher auf leicht verfügbare Informationen stützen («availability heuristic»).

(4) Wenn Personen eine Reihe von Sachfragen erhalten, diese beantworten und hinterher befragt werden, für wie wahrscheinlich sie es halten, dass diese oder jene ihrer Antworten korrekt ist, überschätzen sich viele Versuchspersonen («overconfidence bias»).[33]

Die Vertreter des HB-Programms behaupten regelmäßig, dass diese und viele andere Fehler stabil sind. Die Subjekte wurden beispielsweise nach der ersten Versuchsreihe aufgeklärt – sie wurden auf die relevante Norm hingewiesen. Dennoch begingen sie die Fehler danach aufs Neue. Man spricht in dieser Tradition daher von «kognitiven Illusionen», um darauf hinzuweisen, dass weder das Wissen, dass eine bestimmte Meinung falsch ist, noch das Wissen, warum sie es ist, zur Korrektur der Meinung führt. Da viele der Studien in diesem Programm so unerfreulich ausgefallen sind, ist nicht selten geurteilt worden, dass wir Menschen allgemein wohl ziemlich irrational sind. Obwohl Kahneman und Tversky hier stets eher zurückhaltend formuliert haben, haben andere Vertreter des HB-Programms, und zumal manche ihrer Leser die Schlussfolgerung gezogen, die Resultate hätten «bleak implications for human rationality».[34]

31 Tversky/Kahneman, op. cit. (Fn. 28).

32 Das «Harvard medical school problem» (Ward Casscells, Arno Schoenberger, Thomas B. Grayboys: *Interpretation by Physicians of Clinical Laboratory Results*, in *New England Journal of Medicine* 299 [1978] S. 999-1001). Die Berechnung hat gemäß der Bayes'schen Regel zu erfolgen, was auch gewisse Schwierigkeiten auslöst: $p(A/B) = (p(B/A) \times p(A))/p(B)$.

33 Sarah Lichtenstein, Baruch Fischhoff, L. D. Phillips: *Calibration of Probabilities*: *The State of the Art to 1980*, in *Judgment under Uncertainty: Heuristics and Biases*, hg. von Daniel Kahneman, Paul Slovic, Amos Tversky (Cambridge, Cambridge University Press, 1982) S. 306-334.

34 Richard B. Nisbett, Eugene Borgida: *Attribution and the Psychology of Prediction*, in *Journal of Personal and Social Psychology* 32 (1975) S. 932-943, hier S. 935; vgl. u.a. Massimo Piattelli-Palmarini: *Inevitable Illusions: How Mistakes*

Mele bezieht sich beispielsweise auf das Vorziehen von bestätigenden Evidenzen («confirmation bias»), leicht verfügbaren Informationen («availability heuristic») oder auch Fälle des Überschätzens der Korrektheit eigener Meinungen («overconfidence bias»). Diese und andere Mechanismen könnten durch Motive angeregt werden und so Selbsttäuschung erzeugen.[35] Kann Mele sich auf das HB-Programm verlassen, um Selbsttäuschung und ihre Irrationalität zu erklären?

Eine erste Merkwürdigkeit ist, dass Wason und Johnson-Laird die «confirmation bias» beim Testen von Hypothesen dadurch erklärt haben, dass die Versuchspersonen einer Selbsttäuschung anheimfallen. Der problematische Status dieses Konzepts wurde dabei nicht bemerkt.[36] Mele dürfte sich natürlich nicht auf diese Erklärung des Hangs zu Bestätigungen stützen. Wollte man die Erklärung zudem auf andere fehlerhafte Urteile oder Schlussfolgerungen ausdehnen, dann ließe sich das HB-Programm noch weniger zur Erklärung von Selbsttäuschung heranziehen – es sei denn, man hat eine Schwäche für zirkuläre Erklärungen. Konsequenter im Rahmen des Programms ist es, das Phänomen selbst als eine psychologisch zu erklärende Fehlleistung anzusehen. So haben Quattrone und Tversky (mit Übernahme der Begriffsexplikation von Gur und Sackeim) Selbsttäuschung der Verwechslung von kausalen und «diagnostischen» Zusammenhängen zugeordnet, einem Mechanismus, der auch andere Urteile und Entscheidungen in die Irre leite.[37]

Natürlich verfällt Mele nicht dem Fehler von Wason und Johnson-Laird. Doch, was wichtiger ist, sein Ansatz ignoriert die erheblichen Grenzen des HB-Programms. Dieses Programm wird schon seit einer Weile mit guten Gründen angegriffen.

Zum einen ist gezeigt worden, dass viele vom HB-Programm erzeugte Daten experimentelle Artefakte darstellen. So beruht die Missachtung der wahrscheinlichkeitstheoretischen Konjunktionsregel (Beispiel [2] oben) wohl auf Mehrdeutigkeiten der entscheidenden Termini des Materials. Im alltäglichen Sprachgebrauch besitzen ‘und’ und ‘wahrscheinlicher als’ noch andere legitime Bedeutungen als die in Logik und Wahrscheinlichkeitstheorie gebrauchten. Wenn das beachtet wird, kann man die scheinbar

of Reason Rule our Minds (New York: Wiley, 1994); Stephen Stich: *Could Man Be an Irrational Animal?*, in *Synthese* 64 (1980) S. 115-135.

35 Mele: *Self-Deception Unmasked*, op. cit. (Fn. 2) S. 3, 11, 28-31.

36 Wason/Johnson-Laird, op. cit. (Fn. 9).

37 Quattrone/Tversky, op. cit. (Fn. 12). Vgl. dazu unten Fn. 41.

fehlerhaften Antworten auch als in gewisser Hinsicht rational interpretieren. Den Versuchssubjekten wird Linda als aufgeweckte, an politischen Themen interessierte ehemalige Philosophiestudentin beschrieben; und dann wird gefragt, ob es wahrscheinlicher ist, dass Linda *(A)* eine Bankangestellte ist oder *(A&B)* eine Bankangestellte und Feministin ist. Wenn man von einem Experimentator vor eine so scharfe Alternative gestellt wird, mag man die erste Alternative *(A)* leicht so verstehen, dass bei ihr *(B)* ausgeschlossen sein soll – als ob *(A)* so viel bedeutet wie *(A&¬B)*. Das aber macht (*A*) in der Tat zur unplausiblen Antwort; also liegt *(A&B)* nicht so fern. Diese Wahl der Antwort mag noch dadurch verstärkt werden, dass die Probanden annehmen, dass die *Beschreibung* von Linda doch für die Aufgabenlösung relevant sein muss – warum sollte sie sonst angeführt werden? Gegeben die Beschreibung, scheint es doch plausibel, dass Lindas Bankangestelltendasein keineswegs ausschließt, dass sie gleichwohl Feministin ist. Zudem lassen sich durch bestimmte Testvarianten die fehlerhaften Antworten drastisch verringern. Wählt man statt Fragen nach subjektiven Wahrscheinlichkeiten Fragen nach objektiven Häufigkeiten («Von 100 Personen mit Lindas Charakteristika, sind wie viele *(A)* Bankangestellte oder *(A&B)* feministische Bankangestellte?»), verschwinden die Verzerrungen drastisch: Es kann schließlich nicht mehr feministische Bankangestellte als Bankangestellte geben. Hinter dieser Option steckt vermutlich mehr: Auch andere scheinbar unvermeidliche Fehlleistungen lassen sich nämlich beseitigen, wenn man statt nach subjektiven Wahrscheinlichkeiten nach natürlichen Häufigkeiten fragt.[38] Davidsons bekannter Auffassung zufolge sollten wir das *principle of charity* befolgen, also menschliches Verhalten so weit wie möglich als rational interpretieren. Wie diese Überlegungen zeigen, können wir das in sehr konkreter Weise.

Ähnliche Punkte ließen sich auch für die HB-Programm-Beispiele (I) und (III) vorbringen, aber aus Platzgründen verzichte ich hier darauf. Es sei nur noch bemerkt, dass dies keinesfalls Argumente sind, die aus dem

[38] Klaus Fiedler: *The Dependence of the Conjunction Fallacy on Subtle Linguistic Factors*, in *Psychological Research* 50 (1988) S. 123-129; Lola L. Lopes: *The Rhetoric of Irrationality*, in *Theory & Psychology* 1 (1991) S. 65-82; Gerd Gigerenzer: *How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases*, in *European Review of Social Psychology* 2 (1991) S. 83-115; Ralph Hertwig, Gerd Gigerenzer: *The «Conjunction fallacy» Revisited: How Intelligent Inferences Look Like Reasoning Errors*, in *Journal of Behavioral Decision Making* 12 (1999) S. 275-305; Gerd Gigerenzer, Ulrich Hoffrage: *How to Improve Bayesian Reasoning Without Instruction: Frequency Formats*, in *Psychological Review* 102 (1995) S. 684-704.

berüchtigten philosophischen Lehnstuhl gegen die empirische Forschung vorgebracht werden. Viele der Punkte sind von Psychologen gesichtet und getestet worden – freilich teils von sprachphilosophisch informierten Psychologen: Die Reinterpretation des Verhaltens der Probanden im Linda-Test mittels der Annahme, die Informationen über Linda müssten doch relevant für die Aufgabenlösung sein, stützt sich auf Grice' Arbeiten zu Maximen der Konversation.[39]

Ich will lieber den Fall der sich selbst überschätzenden Wissenschaftler, den Mele als Beispiel von Selbsttäuschung ansieht, in ähnlicher Weise angreifen – was teils auf der Kritik am obigen Beispiel (IV) beruhen wird. Drei Punkte seien genannt:

(1) Es müsste auch hier geprüft werden, ob die Testfragen fair und ohne Mehrdeutigkeiten in den entscheidenden Fragestellungen formuliert worden sind. Was soll eine Frage wie «Halten Sie sich für kompetenter als den durchschnittlichen Wissenschaftler?» schon bedeuten? Kompetenter im Bereich der Lehre oder der Forschung? Und in welchen Forschungsbereichen? Welcher Durchschnitt sollte hier gemeint sein? Wenn Wissenschaftler nach ihren Fähigkeiten gefragt werden, mögen sie natürlich zuerst an das denken, worin sie wirklich Leistungen vollbracht haben. Der Blick auf diese eingegrenzten Kompetenzbereiche mag es sein, der sie meinen lässt, dass sie ihre durchschnittlichen Kollegen überragen.

(2) Es ist empirisch gezeigt worden, dass sich der scheinbar starre Hang zur Überschätzung («overconfidence bias») der Korrektheit eigener Meinungen sehr wohl beseitigen lässt. Wenn man die Versuchspersonen nämlich nicht fragt, für wie wahrscheinlich sie es halten, dass diese oder jene ihrer Antworten korrekt ist, sondern danach, wie hoch die Zahl der für korrekt gehaltenen Testantworten insgesamt ist (also wieder nach Häufigkeiten statt nach subjektiven Wahrscheinlichkeiten fragt), verschwinden die Fehlleistungen in dramatischem Umfang.[40]

(3) Sofern man die Urteile der Wissenschaftler als relativ zu ihren herausragenden Leistungen interpretieren kann, muss bei einem solchen Urteil kein Motiv vorliegen. Daher ist es fraglich – und bedürfte einer empirischen

[39] Hertwig/Gigerenzer, *The «Conjunction fallacy» Revisited*, op. cit., beziehen sich etwa auf Paul Grice: *Logic and Conversation*, in *Syntax and Semantics 3: Speech Acts*, hg. von Peter Cole, Jerry L. Morgan (New York: Academic Press, 1975) S. 41-58.

[40] Gerd Gigerenzer, Ulrich Hoffrage, H. Kleinbölting: *Probabilistic Mental Models: A Brunswikian Theory of Confidence*, in *Psychological Review* 98 (1991) S. 506-528.

statt einer begrifflichen Untersuchung –, ob es sich in Meles Sinn um Selbsttäuschung oder «nur» um eine Art von durch kognitive Heuristiken erzeugte verzerrte Meinung handelt.

Zweifellos ist es offen, ob sich alle von Mele angesprochenen Kandidaten für empirisch nachgewiesene Selbsttäuschungen so bestreiten lassen. Ich muss es an dieser Stelle dem Leser überlassen, hier selbst weiter nachzudenken.[41] Aber klar sollte sein, dass Mele keineswegs sicher sein kann, dass

41 Einen von Vertretern des HB-Programms als Selbsttäuschung beschriebenen Fall bestreitet Mele selbst, aber aus ganz anderen Gründen als den oben vorgeführten. Es handelt sich um die Studie von Quattrone/Tversky, op. cit. (Fn. 12): Probanden wurde hier erzählt, sie nähmen an einem Test über Typen von Herzen teil. Ihnen wurde gesagt, dass es Herzen zweierlei Typs gebe: Typ 1 neige zu mehr Krankheiten, während Typ 2 davon frei sei. Die Versuchspersonen wurden dann aufgefordert, ihren Arm so lange wie möglich in eiskaltes Wasser zu halten. Danach wurden sie auf ein Trainingsrad gesetzt, und während einer Hälfte der Gruppe gegenüber behauptet wurde, dass solch ein Training bei Personen mit Typ-2-Herz die Schmerztoleranz erhöhe, erklärte man der anderen Gruppe, das Training verringere die Schmerztoleranz von Typ-2-Herzen. In einer zweiten Versuchsrunde sollten die Probanden wieder den Arm in kaltes Wasser halten: bei fast allen verlängerte bzw. verkürzte sich die Ausdauer. Doch sie leugneten, dass sie versucht hätten, ihre Schmerztoleranz auszudehnen oder zu verringern (Quattrone und Tversky haben Mittel angewendet, um Lügen auszuschließen). Mele (*Self-Deception Unmasked*, op. cit. [Fn. 2] S. 84-93) bestreitet, dass die Probanden widersprüchliche Meinungen haben müssen: Die erste Gruppe muss gar nicht unbewusst glauben, dass sie versucht hätten, ihre Schmerztoleranz auszudehnen. Freilich könnten Quattrone und Tversky hier erwidern, dass es sich gleichwohl um Selbsttäuschung handeln könnte – zwar nicht nach Gurs und Sackeims Konzeption, sondern etwa nach Meles, die ja keine widersprüchlichen Meinungen erfordert. Eine andere Strategie der Kritik an Quattrone und Tversky wäre die, die behauptete Irrationalität der Probanden zu bezweifeln. Quattrone und Tversky sehen das Verhalten der Subjekte als irrational an, weil sich darin eine Verwechslung von kausalen und diagnostischen Zusammenhängen ausdrücke. Das heißt, das HB-Programm unterstellt hier die Richtigkeit der sog. kausalen Entscheidungstheorie: Wir sollen Handlungen nur wählen, um damit gewünschte Ziele zu verursachen, nicht aber, um Evidenzen von Zuständen zu erhalten, auf die wir keinen Einfluss haben (hier etwa: darauf, welches Herz man hat, hat man keinen Einfluss, und daher ist der Versuch, die Schmerztoleranz zu erhöhen (bei Gruppe 1) der Versuch, ein günstigeres Bild der eigenen Gesundheit zu gewinnen). Man mag nun die Grundannahme der kausalen Entscheidungstheorie bestreiten und an der sogenannten evidenzialen Entscheidungstheorie festhalten, die einfach fordert, eine Maximierung des konditionalen erwarteten Nutzenwertes zu verfolgen, ohne Rücksicht auf Kausalzusammenhänge. Dies ist jedoch nicht leicht, schon gar nicht im Fall der Selbsttäuschung. Jean-Pierre

das HB-Programm eine gesicherte Basis für seine Konzeption und für die Existenz von Selbsttäuschungen darstellt. Angesichts des problematischen Charakters von Selbsttäuschung gilt erneut die Maxime: Die Beweispflicht liegt aufseiten dessen, der die Existenz des Phänomens behauptet.

8. Wie (weit) lässt Selbsttäuschung sich rationalisieren?

Die Debatte über das HB-Programm geht noch tiefer. Sie betrifft nämlich sogar die Basis von Rationalitätsstandards selbst. Wann – bei welchen Aufgaben – dürfen wir welche Regeln als normativ gültig einsetzen? Was ist eigentlich das Fundament von Rationalitätsstandards? Dies kann ich hier nicht weiter ausführen, da die Debatte zu komplex und in vielen Hinsichten offen ist.[42] Klar ist allerdings, dass es dubios ist, Selbsttäuschung nach dem HB-Programm als motivierte Art von verzerrten Meinungen zu verstehen und *deshalb* als irrational zu bewerten. Umgekehrt formuliert: Es lässt sich zumindest in Grenzen plausibel machen, dass Selbsttäuschung aus Gründen rational sein kann, die von anderer Art sind als die vom Standardmodell vorgegebenen Prinzipien des Urteilens und Entscheidens.

Eine erste Konkretisierung dieser Idee ist die, dass Selbsttäuschung auf praktischer (im Sinne von instrumenteller) Rationalität beruht: dass

Dupuy (*Rationality and Self-Deception,* in *Self-Deception and Paradoxes of Rationality*, hg. von Jean-Pierre Dupuy [Stanford: CSLI Publications, 1998] S. 113-150) behauptet so eine Rationalisierung der Selbsttäuschung zu liefern, doch er baut letztlich nur darauf auf, mittels des Evidenzialismus Entscheidungen zu rationalisieren. Es bleibt dunkel, ob sich das auf Meinungen oder ähnliche kognitive Einstellungen übertragen lässt. Mir ist keine klare Diskussion von Quattrones und Tverskys Studie in dieser Richtung bekannt.

42 Vgl. etwa Michael Bishop: *Reflections on a Normative Psychology*, in *Philosophie: Grundlagen und Anwendungen*, hg. von Ansgar Beckermann, Holm Tetens, Sven Walter (Paderborn: Mentis, 2008) S. 249-262; L. Jonathan Cohen: *Can Human Irrationality Be Experimentally Demonstrated?*, in *Behavioral and Brain Sciences* 4 (1981) S. 317-331; Alvin Goldman: *Epistemology and Cognition* (Cambridge, MA: Harvard University Press, 1986); A. G.: *Human Rationality: Epistemological and Psychological Perspectives*, in *Philosophie: Grundlagen und Anwendungen*, hg. von Ansgar Beckermann, Holm Tetens, Sven Walter (Paderborn: Mentis, 2008) S. 230-247; Stein, op. cit. (Fn. 18); Thomas Sturm: *What Is the Foundation of Norms of Rationality?*, in *Philosophie: Grundlagen und Anwendungen*, hg. von Ansgar Beckermann, Holm Tetens, Sven Walter (Paderborn: Mentis, 2008) S. 189-201.

sie ein Resultat der Verbindung von subjektiven Zielen mit angemessenen praktischen Überlegungen ist.[43] Dies ist allerdings eine starke Art des Intentionalismus, die allenfalls für einige Fälle geeignet ist. Einen moderateren Ansatz bietet die psychologische Konzeption von Rationalität an, die den Hauptkonkurrenten zum HB-Programm darstellt. Dieser moderate Ansatz lehnt die Annahme des HB-Programms ab, dass Rationalitätsnormen nur die Regeln des «Standardmodells» sein können. Stattdessen sollten wir die Idee einer «beschränkten Rationalität» *(bounded rationality)* zulassen: Regeln sind nur insofern rational oder normativ gültig, als sie eingegrenzten Aufgaben und Kontexten angepasst sind.[44] Dieser Ansatz geht zudem davon aus, dass Normen eine evolutionäre Anpassung aufweisen müssen, um erfolgreiches Räsonnieren zu ermöglichen. Dies konvergiert mit einer Tendenz in empirischen Studien über Selbsttäuschung. Anstatt die Irrationalität der Selbsttäuschung hervorzuheben, wird zunehmend betont, welche Vorteile die Selbsttäuschung haben mag: etwa die Verminderung von Stress oder die Erhaltung von Selbstachtung und Wohlbefinden[45] oder das effektivere Verstecken der wahren Absichten vor anderen Personen: Täuschung hat einen adaptiven Nutzen, und Selbsttäuschung verstärkt die Fähigkeit zur Täuschung.[46]

Doch können alle Fälle von Selbsttäuschung so rationalisiert werden? Obwohl es verführerisch ist, dagegen Gedankenexperimente oder empirische Gegenbeispiele anzuführen – etwa Studien, die die Selbsteinschätzungen von Autofahrern als Selbsttäuschungen beurteilen, die keinesfalls rational sind – werde ich mich solcher Einwände hier enthalten. Ich möchte lieber drei andere Schwierigkeiten hervorheben:

43 So etwa Amelia O. Rorty: *Belief and Self-Deception*, in *Inquiry* 15 (1972) S. 387-410; Davidson, op. cit. (Fn. 6).

44 Vgl. Gerd Gigerenzer: *Adaptive Thinking* (New York: Oxford University Press, 2000).

45 Larry J. Jamner, Gary E. Schwarz: *Self-Deception Predicts Self-Report and Endurance of Pain*, in *Psychosomatic Medic*ine 48 (1986) S. 211-223; James F. Welles: *Self-Deception as a Positive Feedback Mechanism*, in *American Psychologist* 41 (1986) S. 325-326; Lockhard/Paulhus, op. cit. (Fn. 12); Sahdra/Thagard, op. cit. (Fn. 3).

46 Robert Trivers: *Social Evolution* (Menlo Park, CA: Benjamin/Cummings, 1985); R. T.: *The Elements of a Scientific Theory of Self-Deception*, in *Annals of the New York Academy of Sciences* 907 (2000) S. 114-131; Christopher C. Byrne, Jeffrey A. Kurland: *Self-deception in an Evolutionary Game*, in *Journal of Theoretical Biology* 212 (2000) S. 457-480.

(1) Zunächst sollte klar sein, dass es sich um Rationalisierungen handelt, die aus externer oder Beobachterperspektive gebildet werden. Angenommen, Selbsttäuschung erfüllt die erwähnten Funktionen und liefert diese oder andere nützliche Ergebnisse, dann kann sie das – zumindest typischerweise – nur, insofern sie den Subjekten nicht bewusst ist und von ihnen nicht intentional geplant wurde (lassen wir hier die Grenzfälle intentionaler Selbsttäuschung außer acht, bei denen auf die Forderung der Gleichzeitigkeit der inkompatiblen Meinungen verzichtet wird). Wir verlangen aber von Rationalität nicht bloß, dass sie zu vorteilhaften Resultaten führt. Wir verlangen auch, dass sich die Subjekte im Prinzip bewusst sein können, *was sie tun und warum sie es tun*. Wenn eine Gruppe von Affen zufällig korrekte Kalkulationen oder Schlussfolgerungen auf Papier kritzelt, ist das noch kein hinreichendes Zeichen von Rationalität. Man muss auch auf in gewissem Umfang reflektierte Weise zu den Resultaten gelangt sein. Darauf baut schließlich das Lernen, Prüfen und Verbessern von Strategien und Lösungen im Urteilen und Entscheiden auf. Sobald diese Forderung aber auf Selbsttäuschung angewendet wird, ist nicht zu sehen, wie sie im gleichen Sinn eine rationale Strategie sein kann, wie wir es von anderen Problemlösungsmethoden meinen. Das gilt übrigens unabhängig davon, ob man in normativer Hinsicht das Standardmodell oder das adaptive Modell von Rationalität bevorzugt.

(2) Ein methodologisches Problem der angedeuteten Rationalisierungen besteht erneut darin, dass manche Psychologen nur ungenaue Vorstellungen von Selbsttäuschung haben. So untersucht Whittaker-Bleuler den Zusammenhang zwischen Täuschung, Selbsttäuschung und sozialer Dominanz im Tennis.[47] Sie argumentiert wie folgt: Indem ein Tennisspieler sich über seine Leistungsfähigkeit, den allgemeinen Stand des Spiels und ähnliches täuscht, kann er seine Unsicherheit besser gegenüber dem Gegner verbergen. So mag er den Kopf aufrecht halten, statt ihn zaudernd zu schütteln, oder es vermeiden, immer wieder wie zur Übung Trockenübungen mit dem Tennisschläger zu machen – kurz, er mag *cool* bleiben. Der Grad der Selbsttäuschung muss Whittaker-Bleuler zufolge dann besonders hoch sein, wenn der sich selbst täuschende Spieler in einen Punkterückstand gerät und sich dennoch besonders dominant verhält. Egal, wie das Spiel verlief,

47 Sharon Whittaker-Bleuler: *Deception and Self-Deception: A Dominance Strategy in Competitive Sport*, in Lockhard/Paulhus, op. cit. (Fn. 12) S. 212-228; für Schwimmer: Joanna E., Carolina F. Keating: *Self-Deception and Its Relationship to Success in Competition*, in *Basic and Applied Social Psychology* 12 (1991) S. 145-155.

Pete Sampras hat bekanntlich fast nie Unsicherheit oder Schwäche gezeigt. Jedoch macht Whittaker-Bleuler nicht klar, warum man das als Selbsttäuschung verstehen muss. So müssen in diesem Fall ja keine inkompatiblen Meinungen vorliegen. Sampras hat vielleicht die Hinweise auf den aktuellen Punktestand oder auf die letzten Punkte nie so wichtig genommen wie seine Fitness, seine exzellente Technik, seine Fähigkeit, sich vollkommen auf den nächsten Ballwechsel zu konzentrieren. Es mag auch bezweifelt werden, dass jemand, der mehr Grand-Slam-Titel gewonnen hat als Rod Laver, dies wegen (oder auch nur teils wegen) selbsttäuscherischer Strategien erreicht hat. Er mag einfach ein berechtigtes Vertrauen in sich selbst haben. Auch der Versuch, ihm eine wenn auch noch so kleine Überschätzung der eigenen Fähigkeiten zuzuschreiben, wäre absurd, weil man dann auch den Besten Selbstüberschätzung unterstellen würde – was es wiederum unmöglich macht, klare Begriffe von berechtigtem Vertrauen in sich selbst und unberechtigter Selbstüberschätzung zu bilden. Daher kann nicht einfach auf ein anderes Konzept von Rationalitätsstandards zurückgegriffen werden, um kurzerhand zu zeigen, dass Selbsttäuschung rational ist.

Sahdra und Thagard[48] haben in einem Artikel auf empirische Studien verwiesen, in denen Selbsttäuschung zur Erklärung von verbreiteten Verhaltensweisen herangezogen werde: den «positiven Illusionen» von Soldaten im Kampfeinsatz,[49] dem Verneinen von Krankheit,[50] dem zuversichtlichen Optimismus von Selbständigen[51] und dem fahrlässigen Verhalten von Berufskraftfahrern.[52] Doch nur in der letzten dieser vier Studien herrscht begriffliche Klarheit. Auch diese Autoren folgen Sackeim und Gur. In der ersten Studie wird der Begriff nicht genau analysiert und manchmal mit «self-serving bias», manchmal mit dem Konzept von positiven Illusionen gleichgesetzt. Im Grunde könnte dieser Autor (Wrangham) auch einfach von Irrtümern sprechen. In der zweiten und dritten Publikationen kommt der

48 Sahdra/Thagard, op. cit. (Fn. 3).

49 Richard Wrangham: *Is Military Incompetence Adaptive?*, in *Evolution and Human Behavior* 20 (1999) S. 3-17.

50 Rainer Goldbeck: *Denial in Physical Illness*, in *Journal of Psychosomatic Research* 43 (1997) S. 575-593.

51 Gholamreza Arabsheibani et al.: *And a Vision Appeared Unto Them of a Great Profit: Evidence of Self-Deception Among the Self-Employed*, in *Economics Letters* 67 (2000) S. 35-41.

52 Timo Lajunen et al.: *Impression Management and Self-Deception in Traffic Behavior Inventories*, in *Personality and Individual Differences* 22 (1996) S. 341-353.

Begriff der Selbsttäuschung gar nicht oder so gut wie gar nicht vor, weder als explanatorisches Konstrukt noch als etwas, dessen Existenz demonstriert oder erklärt wird. Umso weniger kann in diesen Studien gezeigt worden sein, dass das, was hier jeweils rationalisiert wird, wirklich ein Fall von Selbsttäuschung ist.

(3) Schließlich sollten wir nicht übersehen, dass zahlreiche der oft als «gewöhnlich» bezeichneten Fälle von Selbsttäuschung Beispiele von Lebenssituationen voller Unklarheiten und Ungewissheiten sind. Diese erlauben schwerlich eindeutige Zuschreibungen von Meinungen oder anderen für Selbsttäuschung relevanten propositionalen Einstellungen. Ein Fall wie der des sich selbst täuschenden Ehemanns mag dabei noch lösbar sein. Wenn er seine Frau verfolgen und am Ende das untreue Handeln beobachten kann, so reduziert sich seine Unsicherheit auf Null, und er mag sich wohl sogar selbst eingestehen können, zuvor einer Selbsttäuschung erlegen zu sein. Aber unzählige andere Fälle sind ganz anders. Kann man sich jemals im Klaren darüber sein, die beste Partnerin oder den besten Partner geheiratet zu haben, den optimalen Job gefunden zu haben? Kann man sich sicher sein, wie die Aussichten für die eigenen künftigen Unternehmungen stehen, sei es als Sportler oder Soldat, Geschäftsmann oder Glücksritter? Sofern die Antworten hier negativ ausfallen, steht auch die Zuschreibung von propositionalen Einstellungen, insbesondere von Meinungen und Motiven, auf schwankendem Grund – oft zu schwankend, um fairerweise eine schwerwiegende Zuschreibung wie die der Selbsttäuschung oder der Irrationalität zu erlauben.

9. Was gefordert wäre

Was kann man aus dem Gesagten für die weitere Erforschung der Selbsttäuschung lernen? Zumindest zweierlei: Erstens benötigen empirische Studien und Theorien über die Selbsttäuschung eine klare Analyse des Phänomens, aber genauso eine bewusst gemachte und reflektierte normative Theorie der Rationalität. Zweitens wären empirische Nachweise von Selbsttäuschung die notwendige Basis für jegliches weitere Forschen und Theoretisieren über das Phänomen. Solche Nachweise erfordern nicht nur, dass man im konkreten Fall nachweist, dass die begrifflichen Bedingungen für Selbsttäuschung erfüllt werden. Zusätzlich ist auch folgendes notwendig:

(1) die Wahl einer speziellen Regel;
(2) der Nachweis, dass diese Regel für gegebene Testaufgaben die normativ angemessene ist;

(3) der Nachweis, dass die Regel von den Probanden in der Test-Situation erfasst werden kann; und
(4) der Ausschluss der Möglichkeit, dass angebliche Fälle von Selbsttäuschung wohlwollend als etwas Anderes interpretiert werden können.

Es sollte klar sein, wie anspruchsvoll solche Forderungen sind. Ich behaupte keinesfalls, dass man ihnen nicht nachkommen kann. Wir sollten jedoch philosophische Konzeptualisierungen ablehnen, die ungeprüft gewisse empirische Befunde und Theorien aus der Psychologie übernehmen. Ebenso wenig sollten wir glauben, wir wüssten schon, dass Selbsttäuschung irrational (oder rational) ist. Es bedarf erheblich verbesserter Anstrengungen zu einer Zusammenarbeit von Philosophen und Psychologen, bevor wir bei diesem Thema weiterkommen können.

Studia philosophica 68/2009

JULIUS SCHÄLIKE

Selbstkontrolle. Synchrone contra diachrone Analyse von motivationalem Zwang und Willensschwäche

Someone who suffers from motivational compulsion (e.g. a drug addict) has problems to control her will. How is this deficiency to be analyzed? I argue that a synchronic analysis fails, whereas a diachronic analysis is well suited. In addition, a diachronic analysis is able to appropriately distinguish compulsion from weakness of will.

Was geschieht, wenn ein Subjekt unter *innerem Zwang* handelt? Offenbar etwas der folgenden Art: Das Subjekt bewertet eine Handlung H1 als die beste, identifiziert sich auf diese Weise mit ihr und bildet den entsprechenden autonomen Willen; zugleich wünscht es jedoch, H2 zu vollziehen, eine Handlung, die es als schlecht beurteilt, und bildet den entsprechenden heteronomen Willen. Im Wettstreit unterliegt der autonome Wille dem heteronomen, und zwar nicht nur faktisch, sondern notwendig: der autonome Wille *kann* sich nicht durchsetzen. *Willensschwäche* scheint ein verwandtes Phänomen zu sein: auch hier unterliegt der autonome Wille faktisch; der Unterschied liegt darin, dass er sich hätte durchsetzen *können*. Wie aber ist dieser Wettstreit und die Rede von Durchsetzungsfähigkeit genauer zu verstehen? Es bieten sich zwei unterschiedliche Analysen an, eine synchrone und eine diachrone. Ich werde argumentieren, dass die synchrone unhaltbar, die diachrone hingegen konsistent ist und für den relevanten Phänomenbereich erschöpfende explanatorische Kraft besitzt.

1. Die synchrone Analyse

Die synchrone Analyse unterscheidet zu einem Zeitpunkt zwischen dem autonomen und dem heteronomen Willen des Subjekts und interpretiert motivationalen Zwang dahin, dass dasjenige Wollen, *das das Subjekt situativ als das autonome erachtet*, sich nicht gegen das heteronome Wollen durchsetzen kann. Ein paradigmatischer Vertreter einer synchronen Analyse ist Harry

Frankfurt.[1] Frankfurt unterscheidet zwischen Wünschen unterschiedlicher Reflexionsstufen und identifiziert den autonomen Willen mit demjenigen Willen, der im Lichte bestimmter höchststufiger Wünsche gebilligt wird. Zum einen ist es jedoch unklar, warum gerade die höchststufigen Wünsche das autonome Wollen definieren sollten;[2] zum anderen bestehen Zweifel, ob hier die Rede von Unfähigkeit des Subjekts, sein autonomes Wollen gegen das heteronome durchzusetzen, tatsächlich wörtlich zu verstehen ist. Wäre es dem Subjekt nicht doch möglich gewesen, dem Drang nach der Droge zu widerstehen? Zahlreiche Philosophen, darunter Rogers Albritton, Joel Feinberg, Immanuel Kant, Jay Wallace und Gary Watson, argumentieren, dass man es nicht wörtlich nehmen kann.[3] Warum die synchrone Analyse problematisch ist, zeigt sich, wenn man sich klar macht, wie sich ein Wille bildet.

Grundsätzlich kann dies auf unterschiedliche Weise geschehen. Oftmals handeln wir gewohnheitsmäßig oder spontan, ohne vorher zu überlegen und Entscheidungen zu treffen. In solchen Fällen werden Wünsche direkt handlungskausal wirksam. In anderen Fällen hingegen überlegen wir und formen unseren Willen in einer Entscheidung. Die Entscheidung, hier und jetzt etwas zu tun, ist im Normalfall hinreichend dafür, es zu tun, sofern keine äußeren Hindernisse entgegenstehen – wobei als «außen» alles gelten kann, was außerhalb des willensbildenden Systems liegt, also auch Hindernisse, die mit dem eigenen Körper verbunden sind. Wenn ich entscheide, jetzt meinen Arm zu heben, wird er sich normalerweise heben, falls der Arm nicht gefesselt oder gelähmt oder auf andere Weise gehindert ist. Angenommen also, das Subjekt hat überlegt den autonomen Willen gebildet, den Arm nicht nach

1 Harry G. Frankfurt: *Freedom of the Will and the Concept of a Person* (1971), in *Free Will. Second Edition,* hg. von Gary Watson (Oxford: Oxford University Press, 2003) S. 322-336.

2 Vgl. Gary Watson: *Free Agency* (1975), in *Agency and Answerability. Selected Essays* (Oxford: Oxford University Press, 2004) S. 13-32; Julius Schälike: *Spielräume und Spuren des Willens. Eine Theorie der Freiheit und der moralischen Verantwortung* (Paderborn: Mentis, erscheint voraussichtlich 2010).

3 Vgl. Rogers Albritton: *Freedom of Will and Freedom of Agency* (1985), in *Free Will*, op. cit. (Fn. 1) S. 408-423; Joel Feinberg: *What is so Special About Mental Illness?*, in *Doing and Deserving. Essays in the Theory of Responsibility* (Princeton: Princeton University Press, 1970) S. 272-292; Immanuel Kant: *Kritik der praktischen Vernunft*, Anmerkung zu § 6; R. Jay Wallace: *Addiction as Defect of the Will: Some Philosophical Reflections* (1999), in *Free Will,* op. cit. (Fn. 1) S. 424-452; Gary Watson: *Disordered Appetites: Addiction, Compulsion, and Dependence* (1999), in *Agency and Answerability. Selected Essays* (Oxford: Oxford University Press, 2004) S. 59-87.

der Droge auszustrecken, sie nicht zu greifen und nicht einzunehmen. Wie nun könnte dieser Wille durch den heteronomen Willen, den Arm auszustrecken, die Droge zu greifen und sie einzunehmen, überwältigt werden? Die erste Möglichkeit ist, dass der heteronome Wille sich ebenfalls in einer Überlegung und einer Entscheidung formt. Dann aber müsste das Subjekt gleichzeitig oder kurz hintereinander zwei Mal überlegen und entscheiden. Wie aber kommt es dazu, dass ein heteronomer Wunsch sich der deliberativen und dezisionalen Ressourcen des Subjekts bedienen kann? Offenbar müsste er bereits an dieser Stelle den Widerstand des autonomen Willens brechen. Das erste Symptom des inneren Zwanges bestünde also darin, dass erzwungene Überlegungen durchgeführt werden.

Eine Überlegung stellt einen zeitlich ausgedehnten Prozess dar, der mehrere Schritte enthält. Der heteronome Wunsch müsste, um Kontrolle über diesen Prozess zu erlangen, das Subjekt zwingen, diese Schritte zu vollziehen. Während die Vorstellung, jemand werde durch physischen Zwang dazu gebracht, seinen Arm zu heben, problemlos verständlich ist, ist unklar, wie Zwang im Falle des Vollzugs willensbildender Operationen zu verstehen ist. Ein Schritt jedoch muss offenkundig vom Subjekt aktiv vollzogen werden: die Ausbildung der Intention, des Willens, H2 zu tun, durch das Fällen einer Entscheidung. Dieser Schritt hat Handlungscharakter, und er ist es, der H2 erst zur Handlung des Subjekts macht. An dieser innersten Stelle der Willensbildung gibt es jedoch keinen Platz für zwingende Faktoren; die Schritte müssen vom Subjekt selbst vollzogen werden, sonst verlieren sie den Handlungscharakter, mit der Folge, dass auch die Körperbewegung, etwa H2, diesen Charakter verliert.[4] In der Willensbildung, die sich im Zuge eines Entscheidungsprozesses vollzieht, manifestiert sich das Subjekt als aktiv; dieser Prozess steht unter seiner direkten Kontrolle. Es gehört zum Sinn von Entscheidungen, dass Aktivität und Kontrolle gewährleistet sind. Entschei-

4 Jay Wallace hat diesen Sachverhalt wie folgt beschrieben: «By volition [...] I mean a kind of motivational state that [... is] directly under the control of the agent. Familiar examples of volitional states in this sense are intentions, choices, and decisions. It is distinctive of states of these kinds that we do not think of them to belong to the classes of mere events in our psychological minds, along with sensations, moods, passing thoughts, and such ordinary states of desire as being very attracted to the chocolate cake in front of one at the café. Rather, intentions, decisions, and choices are examples of the phenomenon of agency itself. [...] The difference, I would suggest, marks a line of fundamental importance, the line between the passive and the active in our psychological lives» (Wallace, op. cit. [Fn. 3] S. 437).

dungen können deshalb *in einem bestimmten Sinne* nicht erzwungen sein. Natürlich können sie insofern erzwungen sein, als es für das Subjekt nicht in Frage kommt, bestimmte Entscheidungen zu treffen, etwa weil furchtbare Folgen angedroht werden. Dennoch kann eine Entscheidung nicht in dem Sinne erzwungen werden, dass es dem Subjekt nicht mehr *freisteht*, die Folgen zu tragen. Das Subjekt *kann* wählen, wird jedoch nicht so dumm oder unmoralisch sein, die Option, zu der es genötigt wird, nicht zu ergreifen. Ein Zwang, der *direkt* die Entscheidung herbeiführte, würde die Entscheidung nicht zu einer heteronomen machen, sondern würde ihr den Entscheidungscharakter nehmen, sie zu einem psychischen Ereignis anderer Art machen.

Es ist somit unmöglich, dass ein Wunsch einen Entscheidungsprozess gleichsam «kidnappt», indem er seinen Verlauf erzwingt. Aber könnte ein heteronomer Wille sich bilden und einem autonomen mit unüberwindlicher Macht entgegentreten, indem ein heteronomer Wunsch einen heteronomen Willen direkt, ohne Einschluss eines Überlegungs- und Entscheidungsprozesses, hervorbringt? Grundsätzlich sind Wünsche dazu in der Lage, Handlungen direkt zu verursachen; dies geschieht ja etwa im Falle von Gewohnheitshandlungen. In diesen Fällen jedoch existieren keine gleichzeitig ablaufenden, auf denselben Handlungsbereich bezogenen Überlegungs- und Entscheidungsprozesse. Die automatischen Prozesse treten *an die Stelle* der überlegungskontrollierten Prozesse, nicht aber *in Konkurrenz* zu ihnen. Wäre es anders und die unmittelbaren Prozesse verursachten Körperbewegungen an der Überlegung – den «Ichprozessen», wie Tugendhat sie nennt[5] – vorbei, so wären diese Bewegungen nicht als Handlungen, ihre Ursache nicht als Wille zu verstehen. Sie hätten keinen Willenscharakter, da ein Kausalmechanismus bestimmte Minimalanforderungen an Kontrolle erfüllen muss, um als Wille gelten zu können. Zwar ist es gerade die Pointe der Idee, synchrone Willensfreiheit könne fehlen, dass das Subjekt Kontrolle verliert, doch darf der Kontrollverlust natürlich nicht so groß sein, dass das bewirkte Ereignis keine Handlung mehr darstellt. Die Analyse schießt somit über ihr Ziel hinaus.

Strebungen, die an der Überlegung und Entscheidung des Subjekts vorbei Bewegungen verursachten, täten dies somit gänzlich ohne Beteiligung des Subjekts. Diese Beschreibung wird zwar von Vertretern synchroner Konzepte akzeptiert, wie Harry Frankfurts Charakterisierung eines willentlich unfreien Süchtigen wider Willen zeigt: «[he is] helplessly violated by his own

5 Ernst Tugendhat: *Anthropologie statt Metaphysik* (München: C. H. Beck, 2007) S. 69.

desires.»[6] Stillschweigend wird hier jedoch angenommen, dass die kausale Rolle der Wünsche, die ja Wünsche des Subjekts sind, den Handlungscharakter der Ereignisse sicherstellt. Schließlich sind Wünsche ja grundsätzlich in der Lage, Handlungen direkt zu verursachen – warum nicht auch an der Überlegung vorbei? Doch wie sich nun gezeigt hat, ist dies nicht möglich. Man sagt zwar, dass man einer Versuchung nicht widerstehen kann, meint damit aber normalerweise, wie Albritton bemerkt, nicht, dass etwa ein sexueller Wunsch insofern Zwang ausübt, als er das Subjekt gewaltsam ins Bett wirft. Täte er es, so läge kein willensunfreies Handeln vor, da gar nicht im relevanten Sinne gehandelt worden wäre.[7] Wenn ein Wunsch gemäß dem Modell synchronen Zwangs am praktischen Standpunkt der Person vorbei kausal wirksam würde, so geschähe dies also in einer Weise, die die Bewegungen zu unwillentlichen Ereignissen machte. Gary Watson hält die Rede von «motivationalem Zwang», die er in früheren Arbeiten[8] verteidigt hat, heute deshalb für inadäquat:

> I am now inclined to think that 'being captivated', or in some cases even 'being possessed', are superior images to that of being compelled, which suggests something too external: that *you* aren't really involved, except as a bystander or victim. In states of captivation, it is not that you aren't into it, but that *you* are (temporarily) transformed (not displaced) by a superior power.[9]

Es zeigt sich, dass das Bestreben derer, die ein synchrones Modell vertreten, darzulegen, dass ein Wunsch und ein Wille «fremd» bzw. «extern» sein kann, zum Scheitern verurteilt ist. Die Synchronisten sind mit einem Dilemma konfrontiert: ein Wille, der in der Weise, die für synchronen motivationalen Zwang erforderlich ist, kausal wirksam würde, wäre allzu fremd und extern, um noch ein Wille sein zu können. Wäre der Wille hinreichend intern, um als Wille gelten zu können, so könnte er keinen Zwang auf das Subjekt ausüben, denn dann würde er dessen praktischen Standpunkt konstituieren.

In welcher Weise können Wünsche ein Subjekt zu einem bestimmten Handeln «zwingen»? Wie sich gezeigt hat, kann dies nicht nach der Art von hydraulischen oder mechanischen Kräften geschehen, die von außen auf das Subjekt einwirken und denen es dadurch zu widerstehen versuchen kann, dass es sich ihnen mit aller Kraft entgegenstemmt. Diese Konzep-

6 Frankfurt, op. cit. (Fn. 1) S. 328.
7 Albritton, op. cit. (Fn. 3) S. 420.
8 Watson, op. cit. (Fn. 2).
9 Gary Watson: *Agency and Answerability. Selected Essays* (Oxford: Oxford University Press, 2004) S. 3, Fn. 3.

tion ist, wie Gary Watson feststellt, von zweifelhafter Kohärenz, da sie auf eine «Externalisierung» von Wünschen hinausläuft. Wir stehen zu unseren eigenen Wünschen nicht wie zu Lawinen; wenn sie uns «mitreißen», dann nicht wie übermächtige physikalische Kräfte, also unwillentlich, sondern als Einstellungen, in deren Licht bestimmte Optionen reizvoll erscheinen, so dass sie uns «verführen».[10] Feinbergs Befund, kein Wunsch sei so stark, dass wir unfähig seien, ihm zu widerstehen,[11] ist Watson zufolge nicht so zu verstehen, dass wir mit unbegrenzten Willenskräften ausgestattet sind, sondern so, dass «motivationale Fähigkeit» gänzlich anders zu analysieren ist als «physische Fähigkeit».[12]

Wie aber könnte eine solche Analyse aussehen? Bei den relevanten Phänomenen handelt es sich um Fälle, in denen Wünsche einen problematischen Einfluss auf den Willen erlangen. Bestimmte Wünsche drohen, das Subjekt zu Handlungen zu bewegen, welche von ihm nicht «eigentlich» gewollt werden. Dies drohen sie jedoch nicht in der Weise zu bewerkstelligen, dass sie am Willen des Subjekts vorbei handlungskausal wirksam werden, sondern dadurch, dass sie diesen Willen *umlenken*. In der Möglichkeit einer Differenz zwischen dem «eigentlich» Gewollten und dem faktischen, verführten Wollen liegt die Möglichkeit eines heteronomen Willens. Was genau ist jedoch unter einem «eigentlichen» Wollen zu verstehen, und was heißt in diesem Zusammenhang «Willensstärke» bzw. «motivationale Fähigkeit»?

2. Sucht und Hypnose: die diachrone «Jeckyll&Hyde»-Analyse

Sucht lässt sich charakterisieren als langfristige Disposition, die das Subjekt anfällig dafür macht, bestimmten Handlungsimpulsen nachzugeben, die als *Suchtwünsche* bezeichnet werden können. Solche Wünsche sind durch vier Merkmale gekennzeichnet:[13] (1.) Sie sind kaum durch Überlegungsprozesse bezüglich der Frage, wie vorteilhaft ihre Realisierung ist, zu beeinflussen, sie stellen sich periodisch immer wieder ein. (2.) Sie sind ungewöhnlich intensiv, weshalb man hier von einem *craving*

10 Watson, op. cit. (Fn. 3) S. 64ff.

11 «Strictly speaking no impulse is irresistible; for every case of giving in to a desire […] it will be true that, if the person had tried harder, he would have resisted successfully» (Feinberg, op. cit. [Fn. 3] S. 282).

12 Watson, op. cit. (Fn. 3) S. 66.

13 Vgl. Wallace, op. cit. (Fn. 3) S. 426-427.

spricht, welches die Süchtigen nach der Droge verspüren. (3.) Sie sind mit Dispositionen verbunden, Lust und Leid zu empfinden. Werden sie nicht befriedigt, so bilden sich äußerst unangenehme Entzugserscheinungen aus. (4.) Die Empfänglichkeit für Suchtwünsche hat eine physiologische Grundlage, die mit bestimmten Veränderungen im Belohnungssystem des Gehirns zusammenhängt.

Sucht hindert Subjekte daran, zu tun, wozu sie die besten Gründe haben. Sie verursacht somit irrationales Verhalten. Manchmal stellt die Irrationalität, die mit Sucht einhergeht, eine Form von Willensschwäche dar, manchmal hingegen eine Form von motivationalem Zwang. Willensschwäche unterscheidet sich von motivationalem Zwang nach einer geläufigen Definition dadurch, dass nur im letzteren Fall das Subjekt *unfähig* ist, die richtige Handlung zu vollziehen. Dies wirkt sich auf die *moralische Verantwortung* des Süchtigen aus: bei zwanghaftem Verhalten entschuldigen wir seine Taten, bei Willensschwäche nicht.

Wie aber ist die Unfähigkeit des Zwanghaften genau zu verstehen? Welche relevante Fähigkeit rationaler Kontrolle wird durch Sucht beeinträchtigt? Worin besteht der Unterschied zwischen motivationalem Zwang und Willensschwäche, wenn das Kriterium der Fähigkeit nicht in der Weise greift, wie es die synchrone Analyse behauptet?

Ich werde versuchen, zu zeigen, dass sich ein Teil der Rationalitätsproblematik, die mit Sucht verbunden ist, *diachron* analysieren lässt. Diese Analyse stellt Sucht in die Nähe zu Hypnose. Wer unter dem Einfluss von Hypnose handelt, tut nicht das, was er autonom will. Doch gibt es zum Zeitpunkt des Handelns keinen okkurenten, vom Subjekt als autonom erachteten Willen, der von einem heteronomen überwältigt würde, vielmehr hat die Hypnose den autonomen Willen vorübergehend beseitigt und durch einen neuen Willen *ersetzt*. Der autonome Wille stellt sich erst dann wieder ein, wenn die Hypnose endet. Dann mag das Subjekt konstatieren, dass es nicht im Sinne dessen gehandelt hat, was es eigentlich, autonom will. Das autonome Wollen ist das jetzt wieder okkurente, nicht hypnoseinduzierte Wollen.

Man kann sagen, dass der Hypnotisierte *irrational* handelt: Er tut nicht das, wozu er im Lichte seines autonomen Wollens die besten Gründe hat, wobei das autonome Wollen einfach das Wollen zu einem bestimmten Zeitpunkt darstellt. Er *registriert* seine Irrationalität auch, und er erfährt sich in dem Geschehen als unfrei: er war nicht fähig, etwas gegen die Rationalitätseinbuße zu tun. Wie dieses Freiheitsdefizit genauer zu verstehen ist, wird uns noch beschäftigen.

Zunächst jedoch einige Bemerkungen zur Frage der *moralischen Verantwortung*. Es ist klar, warum es unangemessen wäre, jemanden für das, was er unter Hypnose tat, verantwortlich zu machen: sobald die Hypnose endet, haben diese Handlungen nichts mehr mit dem Subjekt zu tun, es gibt weder einen okkurenten Willen noch Willensdispositionen, die in Bezug zu diesen Handlungen stünden. Die kausalen Quellen der Handlungen existieren nicht mehr, es waren Wünsche, die mit dem Ende der Hypnose zu existieren aufhörten. Der Wille des nicht-hypnotisierten Subjekts ist kausal in diese Handlungen nicht involviert. Der aus der Hypnose Erwachte kann mit vollem Recht sagen: Das war ich nicht! Natürlich war er kausal involviert, durch seinen Körper. Dies ist jedoch keine moralisch relevante Form der Involvierung. Es ist keine Form der Involvierung *als Person* oder *als Akteur*. Als Person oder als Akteur ist man nur dann involviert, wenn man durch seinen Willen oder seine volitiven Dispositionen involviert ist.

Auch Süchtige entschuldigen sich für ihr Verhalten oftmals, indem sie sagen: «Ich war nicht ich selbst.» In der diachronen Analyse erweist sich eine solche Aussage als durchaus angemessen. Sie lässt sich so verstehen, dass unter dem Einfluss von physiologischen Prozessen, die mit der Drogensucht zusammenhängen, die praktische Identität des Subjekts oszilliert. Ich bezeichne eine solche Fluktuation der Präferenzen als ein «Dr. Jekyll und Mr. Hyde-Phänomen». Dr. Jekyll hat eine Substanz eingenommen, welche ihn dazu disponiert, sich periodisch in Mr. Hyde zu verwandeln, eine Person mit gänzlich verschiedenen volitiven Dispositionen.

Der Süchtige verwandelt sich zwar nicht – wie im Falle von Hypnose oder Dr. Jekyll und Mr. Hyde – in eine komplett andere Person, doch ändern sich einige seiner volitiven Dispositionen. Angenommen, diese veränderten Dispositionen werden handlungskausal wirksam, sie verursachen die Handlung H. Insofern diese Dispositionen vom Subjekt hinterher als Fremdkörper betrachtet werden, lässt sich mit Recht sagen, dass es gar nicht in relevanter Hinsicht personal in H involviert war, ähnlich wie Dr. Jekyll nicht als Person in die Taten des Mr. Hyde verwickelt ist.

Was genau macht Sucht zu einem Jekyll&Hyde-Phänomen? Unter dem Einfluss von Sucht können sich die intrinsischen Präferenzen eines Subjekts verändern. Während es zunächst kein Verlangen nach der Droge hat, bilden sich aufgrund der mit der Sucht verbundenen physiologischen Prozesse Suchtwünsche. Hierbei ist zu unterscheiden zwischen Wünschen, die darauf zielen, die unangenehmen Entzugserscheinungen zu beenden, und solchen, die *intrinsisch* auf den Konsum der Droge gerichtet sind. Beide Sorten von

Wünschen sind mit typischen Süchten verbunden. Die zweite Wunschsorte ist es, die zu einer Jekyll&Hyde-Analyse passt. Nennen wir sie *intrinsische Drogenwünsche*.

Bilden sich intrinsische Drogenwünsche, so wird der Süchtige gleichsam vorübergehend ein anderer, seine personale volitive Identität fluktuiert. Dies stellt ein Hindernis für das Subjekt dar, weil der Verlust der Fähigkeit zu volitiver Kontinuität durchaus unerwünscht sein kann: Hinterher bedauert der Akteur, aufgrund eines Willens gehandelt zu haben, mit dem er sich nicht identifiziert. Ähnlich wie Dr. Jekyll, ist er nicht in der Lage, darauf Einfluss zu nehmen, dass und wann diese Verwandlungen erfolgen. Er kann seine ursprünglichen Intentionen nicht aufrechterhalten, weil er sich in seinem volitionalen Kern, seinen Handlungszielen, verändert und seinen Willen im Lichte der neuen Präferenzen umorientiert.

Man könnte bezweifeln, dass die intrinsischen Drogenwünsche tatsächlich nichts mit dem Süchtigen zu tun haben. Sicher sind sie phasenweise nicht okkurent. In einem bestimmten Sinne wurzeln die Drogenwünsche aber durchaus in den volitiven Dispositionen des Subjekts, so dass man, anders als bei dem, der aus der Hypnose erwacht, nicht sagen kann, es gäbe nichts in seinen Dispositionen, woran sich die Zuschreibung der Taten knüpfen könnte. Doch handelt es sich bei diesen Dispositionen um Dispositionen zu Dispositionen, also gewissermaßen um Dispositionen zweiter Stufe: in den Phasen, in denen der Spiegel der Drogensubstanz im Blut oberhalb einer bestimmten Schwelle liegt, ist das Subjekt nicht disponiert, Verbrechen zu verüben, um auch zukünftig diesen Spiegel halten zu können; sobald der Spiegel jedoch sinkt, ändern sich diese volitiven Dispositionen, die Bereitschaft etwa zu Beschaffungskriminalität wächst.[14] In den Phasen höheren Drogenspiegels ist die Person somit zwar disponiert zu den kriminellen Dispositionen, besitzt diese kriminellen Dispositionen jedoch noch nicht (bzw. besitzt sie nur dispositionell). Ähnlich ist es bei Jekyll&Hyde: der kultivierte und moralisch integere Dr. Jekyll ist disponiert, periodisch zum wilden, verbrecherischen Mr. Hyde zu mutieren, ohne darauf Einfluss nehmen zu können, dass und wann diese Transformationen erfolgen. In den Phasen, in denen er Dr. Jekyll ist, ist er nicht zu den für Mr. Hyde spezifischen Handlun-

[14] Gemeint ist nicht, dass das Subjekt bei hohem Spiegel nicht die Disposition hat, kriminelle Taten zu verüben, solange sich keine Entzugserscheinungen bemerkbar machen, sondern dass es die Disposition hat, dies auch dann nicht zu tun, sollte der Spiegel sinken.

gen disponiert, sondern dazu, diese Dispositionen zu erwerben. Moralische Zurechenbarkeit knüpft sich an die situativ existierenden Handlungsdispositionen, denn *diese* konstituieren das Subjekt moralisch, nicht jedoch die dispositionell existierenden Dispositionen. Letztere sind Dispositionen für die *Transformation* der personalen Identität, konstituieren somit nicht die *faktische* situative Identität.

Aber ist die Wandlung, die in einem Süchtigen vor sich geht, im relevanten Sinne analog zu der Wandlung, der Dr. Jekyll unterworfen ist? Dr. Jekyll und Mr. Hyde sind zwei ganz unterschiedliche Personen, der Wandel erfasst sogar ihren Körper, während der Süchtige doch auch in Situationen, in denen er unter dem Einfluss der Drogenwünsche steht, noch so viele Einstellungen mit dem Süchtigen außerhalb dieser Situationen gemein hat, dass man noch von der selben Person sprechen kann.

Hierzu ist zu sagen, dass es nicht wichtig ist, ob der Süchtige über die Zeit hinweg als ein und dieselbe Person angesehen werden kann, sondern ob er die konkreten Eigenschaften, die der Verantwortungszuschreibung zugrunde liegen, noch besitzt. Es ist ja sicherlich nicht sinnvoll, jemanden, der vor 20 Jahren prächtige Haare hatte, heute jedoch kahl ist, mit dem Hinweis darauf für seine Haarpracht zu loben, dass es sich immer noch um dieselbe Person handelt. Entsprechend trifft es zwar zu, dass der Drogensüchtige, wenn er die kriminellen Taten verübt, dieselbe Person ist wie derjenige, der diese Taten hinterher bedauert; aber die bedauernde Person besitzt die für die Taten kausal entscheidenden Dispositionen nicht mehr, und deshalb ist sie nicht in relevanter Weise kausal involviert.

Die Jekyll&Hyde-Analyse ist geeignet, drei Dinge herauszustellen: (1) inwiefern handelt ein Süchtiger irrational; (2) inwiefern ist er für sein Handeln nicht verantwortlich zu machen; (3) inwiefern handelt er unfrei. Irrational handelt er, insofern sein Handeln nicht seinem autonomen Wollen entspricht, wobei das autonome Wollen nicht eines ist, mit dem sich der Handelnde zum Zeitpunkt der Handlung stärker identifiziert als mit dem motivational durchschlagenden, sondern das Wollen, mit dem er sich zu einem *späteren* Zeitpunkt identifiziert. Während der Handlung mag er sich durchaus mit dem Suchtwollen identifizieren.

Zu (2): *Verantwortlich* zu machen ist der Akteur nicht, insofern sein okkurenter Wille und seine gegenwärtigen volitiven Dispositionen nicht in die relevanten, vergangenen Handlungen involviert sind. Inwiefern fehlt es ihm (3) an *Freiheit*, darauf Einfluss zu nehmen, ob die Suchtwünsche handlungskausal wirksam werden oder nicht? Ich schlage vor, die relevante Fähigkeit konditional zu analysieren, bezogen auf die kausale Abhängigkeit

der Ereignisse vom Willen.[15] Der Prozess, der die volitive Identität oszillieren lässt, ist nicht willentlich steuerbar. Angenommen, der Süchtige weiß, dass sein volitives System die Disposition besitzt, zu fluktuieren, indem es nämlich periodisch intrinsische Drogenwünsche hervorbringt. Was immer er auch wollen mag, nichts wird es verhindern, dass die Fluktuation geschieht, wie auch Dr. Jekyll nicht fähig ist, die Verwandlung in Mr. Hyde aufzuhalten. Die diachrone Willenskontrolle, die hier fehlt, lässt sich wie folgt konditional analysieren:

> J&H-Kontrolle
> S kann der Transformation T seiner volitiven Dispositionen genau dann widerstehen, wenn gilt: wenn S T widerstehen wollen würde, so würde S T widerstehen.

Nun muss es einem Süchtigen nicht gänzlich an J&H-Kontrolle mangeln. Es könnte sein, dass es irgendetwas gibt, für das gilt: wenn er dies wollen würde, so würde er die Fluktuation unterbinden. Wenn er um dieses Kontrollmittel wüsste, wäre er nicht völlig hilflos. Unter Umständen können die geeigneten Maßnahmen rein mentaler Art sein. Der Süchtige müsste sich einfach zusammenreißen, sich konzentrieren, bestimmte Gedanken vermeiden. Möglicherweise sind jedoch Körperbewegungen nötig, etwa muss ein Medikament eingenommen werden.

3. Willensschwäche und Zwang

Wann nun ist ein Süchtiger, der irrational handelt, *willensschwach*, und wann unterliegt er *psychischem Zwang*? Eine Möglichkeit, die Trennungslinie zu ziehen, rekurriert auf das ungraduierte Kriterium der J&H-Kontrolle: ist sie gewährleistet, so handelt der Süchtige willensschwach – er «konnte ja anders» –, wenn nicht, so unterliegt er motivationalem Zwang. Alternativ könnte man, inspiriert von Gary Watson, *Grade* der J&H-Kontrolle unterscheiden und motivationalen Zwang auch dann zuschreiben, wenn J&H-Kontrolle zwar existiert, es jedoch *nicht gefordert* wird, dass das Subjekt sie ausübt, etwa weil dies mit hohen Kosten psychischer oder sonstiger Art einherginge. Man würde hier mit normativen Maßstäben operieren, die

[15] Zu den Gründen, die für eine Konditionalanalyse von «Können» sprechen, vgl. Schälike, op. cit. (Fn. 2).

festlegen, welches Maß an Selbstkontrolle zu fordern ist.[16] Wer irrational handelt, weil er J&H-Kontrolle in einer Situation nicht ausübt, in der eine «normale» Person dies getan hätte, gilt als willensschwach; wem diese Kontrolle fehlt oder wer sie zwar ausüben kann, es aber ebenso unterlässt, wie es jeder «normale» Mensch getan hätte, der unterliegt motivationalem Zwang. Dies bringt einen relativistischen Zug in die Analyse hinein. Wie Watson konstatiert, könnte jemand, dessen Verhalten *wir* als zwanghaft bezeichnen würden, in einer Gesellschaft von Yogis als frei und somit willensschwach gelten.[17]

Wie verhält sich diese Analyse zur Frage der Verantwortung? Ich bin der Auffassung, für die ich hier allerdings aus Platzgründen nicht argumentieren kann, dass moralische Verantwortung Freiheit voraussetzt und dass Freiheit konditional analysierbar ist.[18] Wenn J&H-Kontrolle gewährleistet ist, so ist das Subjekt moralisch verantwortlich dafür, wenn seine praktische Identität fluktuiert. Es hätte dies verhindern können. Es ist auch für die Taten verantwortlich, von denen es voraussieht, dass es sie vollziehen wird, nachdem die Fluktuation einsetzt. Angenommen, wir kritisieren diese *Taten*. Sind wir berechtigt, auch das *Subjekt* dafür zu tadeln, dass es keine J&H-Kontrolle ausgeübt hat? Nicht unbedingt. Dies hängt davon ab, ob die Kontrollausübung zumutbar gewesen ist. Wenn es etwa ein Medikament gäbe, das das Aufkommen von Suchtwünschen verhindern würde, dieses Medikament aber furchtbare Nebenwirkungen hätte, so würden wir womöglich nicht verlangen, dass jemand diese Kosten der Selbstkontrolle trägt. Verlangen kann man jedoch, dass andere Vorsichtsmaßnahmen getroffen werden. Im Extremfall kann man etwa erwarten, dass jemand, der sehr gefährliche Fluktuationen voraussieht, sich in die Obhut anderer begibt, die ihn an den gefährlichen Taten hindern.

Ob ein Süchtiger verantwortlich für die Taten ist, die er irrationalerweise begeht, hängt davon ab, ob er J&H-Kontrolle in dem *ungraduierten* Sinne besitzt. Verantwortung für eine kritikwürdige Tat impliziert jedoch nicht Tadelnswürdigkeit des Akteurs. Letztere hängt davon ab, ob Kontrolle im *graduierten*, normativ aufgeladenen Sinne vorliegt.

16 Vgl. Gary Watson: *Scepticism about Weakness of Will* (1977), in *Agency and Answerability. Selected Essays* (Oxford: Oxford University Press, 2004) S. 33-58.
17 Ibid. S. 51.
18 Vgl. Schälike, op. cit. (Fn. 2).

Einschränkungen von J&H-Kontrolle scheinen mir einen Teil dessen darzustellen, worunter Süchtige oftmals tatsächlich leiden. Sie sind jedoch sicher nicht das einzige Problem, mit dem sie zu kämpfen haben. Hinzu kommt ein Problem, das man als *Intentionsschwäche* bezeichnen kann. Zur Analyse dieses Phänomens haben Richard Holton[19] und Neil Roughley[20] bereits hilfreiche Schritte unternommen, die Grundidee geht auf Aristoteles zurück. Auch bei Intentionsschwäche geht es nicht darum, dass das, was das Subjekt situativ als seinen eigentlichen, autonomen Willen betrachtet, von *in derselben Situation* als heteronom erachteten motivationalen Kräften überwältigt wird. Die Analyse ist daher nicht mit den m. E. unüberwindlichen Schwierigkeiten befrachtet, die sich der Annahme solcher Phänomene stellen. Vielmehr geht es darum, dass das Subjekt unter dem Einfluss einer Versuchung seine vorherigen, wohlerwogenen Intentionen entweder ändert oder vergisst. Sein Wille ist nicht stark genug.

Das Phänomen der Willensstärke und sein negatives Pendant Willensschwäche sind in der philosophischen Tradition seit Sokrates intensiv diskutiert worden.[21] Der Fokus lag dabei allerdings auf einem Teilaspekt dieser Phänomene, den ich hier ausklammere. Das Hauptinteresse galt nämlich der Frage nach dem Zusammenhang zwischen evaluativem Urteil und Motivation. Ist es möglich, dass die Handlung, die situativ am höchsten bewertet wird, eine andere als die ist, zu der man am stärksten motiviert ist? Wenn hier eine Diskrepanz auftritt, spricht man von Akrasie. Ob Akrasie möglich ist, ist kontrovers; an anderer Stelle habe ich es bestritten.[22] Akrasie ist von dem Phänomen der *Intentionsschwäche* unterschieden. Beide Phänomene lassen sich unter dem Begriff «Willensschwäche» fassen. Wie Richard Holton feststellt, charakterisieren Nicht-Philosophen – anders als Philosophen – Willensschwäche typischerweise nicht als Handeln wider das situativ bessere Urteil, sondern bezeichnen denjenigen als willensschwach, der einmal ge-

19 Richard Holton: *Intention and Weakness of Will*, in *The Journal of Philosophy* 96 (1999) S. 241-262.

20 Neil Roughley: *Willensschwäche und Personsein*, in *Personalität. Leipziger Schriften zur Philosophie* 18, hg. von Frank Kannetzky, Henning Tegtmeyer (Universitätsverlag Leipzig, 2007) S. 143-161.

21 Vgl. Julius Schälike: *Willensschwäche. Ein Forschungsbericht*, in *Information Philosophie* 5 (2006) S. 18-29.

22 Vgl. Julius Schälike: *Willensschwäche und Selbsttäuschung. Über die Rationalität des Irrationalen und das Verhältnis von Evaluation und Motivation*, in *Deutsche Zeitschrift für Philosophie* 52 (2004) S. 361-379; Schälike, op. cit. (Fn. 2).

fasste Vorsätze zu leichtfertig revidiert, dessen Wille also nicht hinreichend stabil ist.[23] Willensschwäche gilt im Alltag somit weniger als Akrasie, denn als Intentionsschwäche.

4. Intentionsschwäche

Vorsätze (Intentionen, Absichten) haben eine nützliche Eigenschaft: Sie «speichern» gewissermaßen die Rationalität, zu der ein Subjekt in einer bestimmten Situation fähig ist, und transferieren sie in Situationen, die weniger rationalitätsförderlich sind. Beispielsweise kann es sinnvoll sein, sich schon jetzt Gedanken darüber zu machen, was man morgen tun sollte, weil man jetzt mehr Zeit, Ruhe und besseren Zugang zu Informationen hat. Wenn ich dann schon heute entscheide, was ich morgen tun werde, werde ich morgen rationaler handeln, als ich es könnte, wenn ich morgen entschiede. Dies setzt allerdings voraus, dass die im Voraus getroffene Entscheidung zum relevanten Zeitpunkt auch noch motivational wirksam ist. Das ist nicht ohne weiteres gewährleistet, denn in der Zwischenzeit können Ereignisse auftreten, die zum Verlust des Vorsatzes führen, etwa wenn das Subjekt den Vorsatz einfach vergisst oder ihn revidiert. Letzteres muss nicht unbedingt irrational sein: Oftmals gelangt das Subjekt an neue Informationen, die die Situation in einem anderen Licht erscheinen lassen. Manchmal jedoch führen Motive, die längst berücksichtigt und als irrelevant verworfen wurden, dazu, dass das Subjekt die gefasste Intention erneut in Frage stellt. Selbst wenn eine solche Revision tatsächlich einmal *ex post* vorteilhaft erscheint, zeigt sich in ihr eine *irrationale Tendenz*: Wer einmal gefasste Vorsätze ohne guten Grund fallen lässt, setzt seine Zeit und Kraft nicht vernünftig ein. Wenn wir etwa einmal entschieden haben, in welchem Restaurant wir speisen wollen, so ist es vernünftig, davon abzusehen, erneut in die Abwägung der Vor- und Nachteile der Alternativen einzutreten, auch wenn dies im Einzelfall zur Folge hat, dass wir nicht die beste Wahl treffen. Als endliche Wesen müssen wir mit unseren Ressourcen haushalten, und der Intentionsschwache wird – ohne es zu bemerken, also nicht absichtlich – den diesbezüglichen Normen nicht gerecht. Holton nennt – ohne Anspruch auf Vollständigkeit – fünf vage «Faustregeln», die bestimmen, unter welchen Bedingungen die Revision einer Intention irrational ist.[24] Neil Roughley schlägt vor, diese Bedingungen durch die Unterscheidung von zwei Parametern zu strukturieren:

23 Holton, op. cit. (Fn. 19) S. 241.
24 Ibid.

Die Standards der ersten Sorte betreffen die *relative Überlegungsförderlichkeit* der Bedingungen, unter denen die Absicht gebildet wurde, und der Bedingungen, unter denen ihre Revision zur Debatte steht. Sie lassen sich wie folgt formulieren:

S1
Es ist *pro tanto* unvernünftig, eine Absicht unter Bedingungen aufzugeben, die im Vergleich mit den Bedingungen, unter denen die Absicht ursprünglich gefasst wurde, größere Einschränkungen relevanter Informationen oder der eigenen Fähigkeit, klar zu denken, mit sich bringen.

Bedingungen der zweiten Sorte betreffen den *Inhalt der Absicht* und können in folgender Form wiedergegeben werden:

S2
Es ist *pro tanto* unvernünftig, eine Absicht aufzugeben, wenn seit der Bildung der Absicht die Kosten ihrer Realisierung nicht wesentlich gestiegen sind oder die Wahrscheinlichkeit, dass sie sich überhaupt realisieren lässt, nicht wesentlich gesunken ist.[25]

Holton und Roughley übernehmen eine Idee von Michael Bratman, der zufolge es nur dann rational ist, eine Intention zu überdenken, wenn in diesem Akt eine rationale Tendenz zum Ausdruck kommt.[26] Rational ist die Tendenz, einmal gefasste Intentionen nur unter Bedingungen, wie sie in S1 und S2 *ex negativo* zum Ausdruck kommen, zu revidieren. Wer die entsprechenden Dispositionen hat, besitzt Willensstärke: die Fähigkeit, Versuchungen zu widerstehen.[27] Diese Fähigkeit ist einem Muskel vergleichbar: sie auszuüben erfordert Anstrengung, führt kurzfristig zu Ermüdung, bei Wiederholung jedoch zur Stärkung.[28]

25 Roughley, op. cit. (Fn. 20) S. 13-14.

26 Michael Bratman: *Intention, Plans, and Practical Reason* (Cambridge: Cambridge University Press, 1987) S. 68.

27 Mangelnde Stabilität von Intentionen ist eine, jedoch nicht die einzige irrationale Disposition, die als Form von «Willensschwäche» bezeichnet werden kann. Roughley hat eine weitere herausgearbeitet: die Tendenz, miteinander inkompatible Vorsätze auszubilden, sowie es zu versäumen, untergeordnete, instrumentelle Intentionen auszubilden, die erforderlich sind, um dem übergeordneten Vorsatz zur Erfüllung zu verhelfen. Auch hier wird der Willensschwache unabsichtlich einer Rationalitätsnorm nicht gerecht (vgl. Neil Roughley: *Three Ways of Willing Weakly*, MS 2005; Roughley, op. cit. [Fn. 20]).

28 Vgl. Richard Holton: *How is Strength of Will Possible?*, in *Weakness of Will and Practical Irrationality*, hg. von Sarah Stroud, Christine Tappolet (Oxford: Clarendon Press, 2003) S. 39-67.

Diese Überlegungen sind unmittelbar für die Frage nach diachroner Willensfreiheit relevant. Die Instanz, um deren Freiheit von Hinderung es in diesem Kontext geht, ist das relativ aufgeklärte Wollen Wa1 zum Zeitpunkt t1. Verfügt das Subjekt über die Dispositionen, einmal gefasste Absichten über die Zeit hin so lange stabil zu halten, wie dies gemäß S1 und S2 rational ist, so ist es besser gegen Hindernisse gefeit, die sich ihm in Gestalt von Versuchungen entgegenstellen. Noch ungehinderter allerdings wäre Wa1, wenn das Subjekt die Disposition hätte, einmal gefasste Entschlüsse niemals aufzugeben. Dies würde man jedoch kaum als Ausweis von Willensfreiheit auffassen, sondern als Zeichen von Starrsinn. Starrsinn jedoch ist durchaus freiheitseinschränkend. Dies liegt daran, dass die Instanz, welche den Bezugspunkt der Rede von Hinderung und Freiheit darstellt, nicht der mehr oder weniger aufgeklärte Wille zu einem bestimmten Zeitpunkt ist, sondern der jeweils aufgeklärteste Wille des Subjekts. Wenn somit auf Wa1 zu einem späteren Zeitpunkt t2 der Wille Wa2 folgt, welcher insofern in höherem Maße aufgeklärt ist als Wa1, als er im Lichte von umfangreicheren bzw. relevanteren Informationen gebildet wurde, so übernimmt Wa2 die Funktion, diejenige Strebensinstanz zu konstituieren, welche für das Streben *des Subjekts* steht. In diesem Streben schlägt sich seine optativische Stellungnahme stärker nieder als in allen anderen, weniger informierten Willenshaltungen, die es zu anderen Zeitpunkten bildet. Den Übergang von einem Wollen zu einem weniger defizienten Wollen erfährt das Subjekt nicht als freiheitseinschränkend, sondern -steigernd, während die Erwartung oder die retrospektive Diagnose einer Einbuße an Aufklärung als Hindernis erlebt wird. Erwartung und retrospektive Diagnose erfolgen vom Standpunkt eines Strebens aus, das durch die Einbuße an Aufklärung in seinen Realisationsaussichten bedroht wird.

Einem Subjekt kann ein Mangel an diachroner Willensfreiheit bezüglich einer Versuchung V attestiert werden, wenn es nicht in der Lage ist, V zu widerstehen. Die Widerstandskraft lässt sich nun wie folgt analysieren:

> WK
> S kann bezüglich der Absicht, A zu tun, der Versuchung, V zu tun, genau dann widerstehen, wenn gilt: A und V sind inkompatibel; S zieht V nur unter Bedingungen eingeschränkter Information volitiv vor;[29] wenn S zu t1 die Absicht bilden würde, A zu tun, und zu t2 mit V konfrontiert würde, täte S A.

29 Diese beiden Punkte machen V zu einer Versuchung.

Für diachrone Widerstandsunfähigkeit ergibt sich:

WU
S kann bezüglich der Absicht, A zu tun, der Versuchung, V zu tun, genau dann nicht widerstehen, wenn gilt: A und V sind inkompatibel; S zieht V nur unter Bedingungen eingeschränkter Information volitiv vor; wenn S zu t1 die Absicht bilden würde, A zu tun, und zu t2 mit V konfrontiert würde, täte S ¬A.

Meine These nun ist, dass Phänomene der Sucht, der Phobie und der Manie nicht allein, aber auch darum die Freiheit der betroffenen Subjekte bedrohen, weil ihnen dieselbe Struktur zugrunde liegt wie der Versuchung, gegen den aufgeklärten Willen ein Eis zu essen, wenngleich die Versuchung in jenen Phänomenen sehr viel stärker ist. Zunächst werde ich das Phänomen der Sucht genauer betrachten.

Der Zwang, den ein Suchtwunsch ausübt, wirkt nach dem Mechanismus der Versuchung: Das Subjekt wird, wie Watson feststellt, nicht überwältigt, sondern verführt:

> Recalcitrant cravings for nicotine and heroin are not like internal tensions, sometimes mounting to a breaking point. The circumstances of the seriously unwilling addict seem rather more like those of the exhausted climber. The discomfort both inclines one to give up the project and leads one not (in the end) to resist the desire to do so. Unlike external obstacles (or internal pressure), motivational obstacles work in part not by defeating one's best efforts but by diverting one from effective resistance. [… Addiction] enslaves by appeal, rather than brute force.[30]

Jemand, der «widerwillig» heroinsüchtig ist, erliegt ebenso einer Versuchung wie jemand, der seinem Vorsatz, abzunehmen, untreu wird, wenn sein Blick auf das Sortiment eines Eiscafés fällt, wenngleich die Versuchung, der letzterer ausgesetzt ist, sehr viel schwächer ist. Eine Diskrepanz zwischen dem, was das Subjekt situativ für sein autonomes Wollen hält, und seinem faktischen Wollen gibt es hier nicht. Vielmehr führen bestimmte Umstände dazu, dass ein Wunsch vehement ins Bewusstsein tritt, die Aufmerksamkeit in eine bestimmte Richtung lenkt und dadurch abzieht von den Gründen, die gegen die Erfüllung dieses Wunsches sprechen.[31]

Diese Analyse wirft folgende Frage auf: nicht jeder, der im Sinne von WU unfähig ist, einer Versuchung zu widerstehen, wird im Alltag als jemand angesehen, der willentlichem Zwang unterliegt. Vielmehr unterscheidet man

30 Watson, op. cit. (Fn. 3) S. 65-66.

31 Suchtwünsche sind «sources of a good deal of ‹noise› – like a party next door» (ibid. S. 72).

hier zwischen Willensschwäche und willentlichem Zwang. Zwanghaft nennen wir das Handeln etwa eines Kleptomanen oder eines Klaustrophobikers: Diese, so sagt man, können nicht anders, als zu stehlen bzw. enge Räume zu meiden. Hingegen sagen wir von einem Willensschwachen, der seine Absicht, kein Eis zu essen, irrationalerweise revidiert, *nicht*, er hätte nicht widerstehen können. Im Sinne von WU ist er jedoch ebenso unfähig, der Versuchung zu widerstehen, wie der Kleptomane. Auch hier lässt sich an Gary Watson anknüpfen. Man kann den Unterschied, der im Alltag gemacht wird, so verstehen, dass wir die Fähigkeit der Akteure zur Selbstkontrolle an unseren eigenen normativen Maßstäben messen: der Willensschwache verliert die Kontrolle bereits angesichts von Versuchungen, denen ein «normaler» Akteur widerstehen könnte,[32] während der Zwanghafte von Motiven überwältigt wird, denen ein typischer Erwachsener unserer Gesellschaft nicht standhalten könnte.[33] Die Unterscheidung zwischen Zwang und Schwäche ist dann relativ zu den Standards, die in der Gesellschaft gelten. Relativität ist eine Eigenschaft, die den Begriff der Schwäche generell kennzeichnet, etwa den Begriff der physischen Schwäche. Willenskraft liegt in graduierter Form vor, einer Versuchung, der der eine erliegt, hätte ein anderer widerstanden. Der Willensschwache unterscheidet sich nur graduell vom Zwanghaften. Obwohl beide im Sinne von WS unfähig sind, zu widerstehen, ist ersterer insofern in der Lage, zu widerstehen, als er widerstehen würde, wenn er das Normalmaß an Willenskraft besäße.

Dieser Vorschlag passt auch gut zu den reaktiven Haltungen bezüglich Intentionsschwäche und Zwang. In beiden Fällen kritisieren wir die Subjekte nicht moralisch, da der Informationsverlust dazu führt, dass die Übel nicht intentional angerichtet werden: An ihrem Wertesystem bzw. den moralischen Prinzipien, die sie leiten, ist nichts falsch, sie scheitern lediglich situativ daran, zu erkennen, welches Verhalten richtig ist. Die angemessenen Reaktionen sind Scham und Mitleid, nicht jedoch Schuldgefühl, Groll und Tadel.[34] Allenfalls könnte man die Subjekte moralisch wegen Fahrlässigkeit kritisieren, falls sie den Kontrollverlust vorausgesehen haben oder voraussehen konnten, es jedoch unterlassen haben, ihm vorzubeugen.

Wenn die Person in der Änderung ihrer Absicht Intentionsschwäche zeigt, so erfolgt die Änderung – anders als bei Jeckyll&Hyde-Phänomenen – nicht, weil sich ihre praktische Identität geändert hätte. Vielmehr hat sie immer

32 «Könnte» – im Sinne von WK.
33 Watson, op. cit. (Fn. 2) S. 48ff.; op. cit. (Fn. 3) S. 72-73.
34 Watson, op. cit. (Fn. 2) S. 51.

noch dieselben intrinsischen Wünsche, in deren Licht die Intentionen gebildet wurden. Jedoch gelingt es ihr nicht, angemessen zu erkennen, was sie tun muss, um diese Intentionen zu realisieren, bzw. zu erkennen, dass die gegenwärtige Handlungssituation Aspekte hat, die in Bezug zu einigen ihrer Intentionen steht. Vielleicht vergisst sie auch die Intention selbst. Ein Rationalitätsproblem stellt dies deshalb dar, weil es rationaler wäre, die Intentionen zu realisieren. Dieser Rationalitätsverlust fällt dem Subjekt hinterher schmerzlich auf. Es erlebt sich als daran gehindert, sein Leben rational zu führen. Intentionsschwäche wird deshalb als Freiheitsdefizit erfahren.

5. Schluss

Freiheits- und Rationalitätsdefizite, die traditionell nach dem synchronen Modell interpretiert werden, lassen sich, wie sich gezeigt hat, plausibel diachron analysieren. Die synchrone Analyse stellt die richtige Diagnose, dass solche Fälle von Rationalitäts- und Freiheitsdefiziten die Struktur haben, dass ein eigentliches, autonomes, rationales Streben von motivationalen Faktoren an seiner Entfaltung gehindert wird. Der Fehler liegt darin, die Interaktion der beiden Faktoren, des autonomen und des heteronomen Wollens, so zu fassen, dass beide gleichzeitig okkurent bestehen und gegeneinander kämpfen, wobei derjenige gewinnt, den das Subjekt *in der Situation* als den Falschen betrachtet. Tatsächlich wandelt sich der praktische, vom Subjekt als autonom eingeschätzte Standpunkt jedoch über die Zeit hinweg, zunächst ist er vom autonomen, dann vom heteronomen, und schließlich wieder vom autonomen Wollen eingenommen, wobei das Subjekt in der Situation jeweils meint, autonom zu handeln. *Beim Handeln* erfährt der Süchtige sich hierbei nicht als überwältigt und fremdbestimmt, wenngleich er natürlich ambivalent sein kann. Erst hinterher, wenn sich das Wollen wieder verändert hat oder das unter dem Einfluss der Wünsche temporär entglittene Wissen wieder restauriert wurde, erscheint das Wollen kritikwürdig, fremd, irrational.

Studia philosophica 68/2009

Würdigung / Mémoire

Hans Saner

Die Transzendenz als Opferlamm des virtuellen Todes. Zur Auseinandersetzung von Hans Kunz mit der Transzendenz bei Jaspers*

In a 1934 review of Jaspers's Philosophie*, Hans Kunz says that there is one «point» in which he must refuse to follow Jaspers. It is no longer possible for him to absolutise transcendence, «Because for us the hidden God is dead too.» It appears to him, anyway, that what Jaspers calls «transcendence» can to a large part be interpreted immanently, namely in view of the unique position «that death has within humans». Only in the nineteen fifties did Kunz attempt to interpret transcendence as «immanent constituent of being human». He explicated death in two ways, as «real» death and as «possible» or «virtual» death. Real death, he interpreted as the singular factual end of being that excludes any repetition; and virtual death as the a priori in the origin of thought in which death repeatedly announces itself anew as possible and eventually real. From both manners of death he then infers immanently, as workings of death, all functions and achievements that emanate from purportedly absolute transcendence and of which we become conscious in Existenzerhellung. He sacrifices transcendence to virtual death.*
Jaspers reacted indignantly. He considered the attempt methodically lacking, scientifically fruitless and philosophically untenable – in brief: as «an example of the voidness of this regularly appearing type of psychological hypothesising». In the «reinterpretation of transcendence» he saw merely a hostile attitude towards his own philosophising, which he traced back to Kunz's denial of freedom. The radicality of the defensive reaction itself becomes a question.

Im Sommersemester 1927 besuchte Hans Kunz – er war 23-jährig – eine Vorlesung von Jaspers in Heidelberg, die unter dem Titel *Grundriss der philosophischen Weltanschauung* angekündigt war und inhaltlich das We-

* Vortrag, gehalten am 14. Oktober 2008, anlässlich der Ausstellung «Hans Kunz. Philosoph und Naturwissenschaftler» an der Universität Basel.

sentliche dessen enthielt, was Jaspers später in seiner *Weltorientierung* und *Existenzerhellung* breiter entfaltete. Kunz kannte bereits Jaspers' Schriften zur Psychopathologie und auch die *Psychologie der Weltanschauungen*. Jaspers war ihm also ein bekannter Autor, aber eine unbekannte Gestalt, die er nun gleichsam im Mitphilosophieren kennenlernte.

In einer Rezension der *Philosophie* von 1934 beschrieb Kunz Jahre später die Wirkung, die damals von Jaspers und seinem Kolleg auf ihn ausgegangen war. Die Vorlesung war «die für mich entscheidende und philosophisch einzigartige Begegnung, von der ich glaube, dass ich ihr im Hinblick auf die hohen Möglichkeiten des Menschseins das Tiefste verdanke».[1] Das damals Gehörte habe «auch heute bei immer wiederholtem Lesen nichts von seiner ergreifenden und erweckenden, zu sich selbst bringenden Kraft verloren».[2] Die Bewunderung galt vorab der Gestalt, der «konkreten Existenz»,[3] die ein Philosophieren in statu nascendi nicht bloß verbalisierte, sondern vorlebte und darin eine «einzigartige Adäquatheit von Person und Werk»[4] entstehen ließ, so dass es dem jungen Studenten schien, er begegne gleichsam «*dem* philosophierenden Menschen».[5]

Das für Kunz «in der philosophischen Welt alles überragende Ereignis des Jahres»[6] 1927 war aber das Erscheinen von *Sein und Zeit*. Er las das Werk erst im Winter, aber dann in wenigen Tagen. In seiner *Selbstdarstellung* von 1972 verglich er diese erste Lektüre mit «einem Sturm»,[7] der über ihn hereingebrochen sei. Als er von Prinzhorn hörte, dass Scheler von dem Buch «erschüttert» worden sei und ihm davon ganz erregt, «wie von einer Tarantel gestochen»[8] erzählt habe, konnte Kunz dies nachfühlen. Er besuchte aus anderem Anlass mit Prinzhorn Heidegger in seinem Freiburger Seminar. «Die Begegnung hinterließ in mir ein merkwürdiges, jedoch eindeutiges Misstrauen, das, von Heideggers Haltung her gesehen, ganz unberechtigt

1 Hans Kunz: Rezension von Karl Jaspers: *Philosophie*, in *Zentralblatt für die gesamte Neurologie und Psychiatrie* (= ZNP) 72 (1934) (Berlin: Julius Springer) S. 450-460, hier S. 459.
2 Ibid.
3 Hans Kunz: *Selbstdarstellung* (= SD), in *Psychologie in Selbstdarstellungen*, hg. von Ludwig J. Pongratz, Werner Traxel und Ernst G. Wehner (Bern, Stuttgart, Wien: Hans Huber, 1972) S. 126ff., hier S. 135.
4 Ibid.
5 Ibid.
6 Ibid.
7 Ibid. S. 136.
8 Ibid.

war.»[9] Kunz meinte eine 'Diskrepanz' zwischen «dem großen Werk eines Dichters und seiner menschlichen Substanz»[10] zu verspüren. Deshalb erstaunte ihn «Heideggers spätere politische Entgleisung»[11] kaum. «Indessen wurde meine Überzeugung nie auch nur vom leisesten Zweifel berührt, dass *Sein und Zeit* die philosophische Leistung schlechthin unseres Jahrhunderts darstellt.»[12]

Vielleicht darf man sagen, dass Kunz in der Begegnung mit Jaspers als Gestalt einen Weg zum *Philosophieren* als möglicher Verwandlung menschlichen Daseins in Existenz im Rahmen der Identität von Person und Werk gefunden habe, seinen Weg zur *Philosophie* aber in der Begegnung mit Heideggers Daseinsanalytik als systematischer Lehre, die unabhängig ist von der Person. Beide Wege haben sich Kunz nahezu zur gleichen Zeit eröffnet. Aber man kann auf die Länge nicht beide zugleich gehen. Der eine transzendiert so exzessiv, dass Dasein und Existenz vom Ursprung her zwei unvereinbare Bereiche sind. Der andere bindet sich so eng an das Seiendsein des Seienden, dass er nur noch die *Frage* nach der Transzendenz zulassen kann, aber nicht die *Behauptung* ihres Seins und Eigenseins. Also ist die künftige Auseinandersetzung absehbar: Der Eine wird dem Anderen die Transzendenz ausreden wollen, indem er ihm etwas anempfiehlt, das alle Funktionen der Transzendenz für die Existenz und die Existenzerhellung übernehmen kann, aber ein unbezweifelbar Wirkliches ist. Und der Andere wird sich über die Nichtigkeit ebendieser Philosophie beklagen.

Schon in der Rezension der *Philosophie* von 1934 sagt Kunz, es sei Jaspers zwar gelungen, das «Suchen und Fragen der großen Denker der Vergangenheit nach der metaphysischen Transzendenz»[13] in einer Sprache zu erneuern, die unserer geschichtlichen Situation angemessen sei. Aber damit werde ein Punkt erreicht, an dem er Jaspers die Nachfolge versagen müsse. «Denn für uns ist auch der verborgene Gott tot; […]».[14] Zwar möge ihm als Idee oder als Bild des Geborgenseins noch eine gewisse «Macht über die Seele»[15] zukommen. Aber diese Macht «zur Transzendenz zu verabsolutieren, ist uns nicht mehr möglich».[16] Und dann folgt der programma-

9 Ibid.
10 Ibid.
11 Ibid.
12 Ibid.
13 ZNP (Fn. 1) S. 459.
14 Ibid.
15 Ibid.
16 Ibid.

tische, über viele Jahre vorausgreifende Satz: «Uns scheint das, was Jaspers Transzendenz nennt, um ein gutes Stück immanent interpretierbar zu sein, nämlich im Hinblick auf die einzigartige Stellung, die der Tod im Menschen innehat, der nicht nur das Ende seines Daseins ist, obzwar er dies im absoluten Sinne auch ist […]».[17] 'Programmatisch' nenne ich diesen Satz, weil Hans Kunz, soweit ich sehe, sich eben diese Entbindung von einer verabsolutierten Transzendenz und ihre Ersetzung durch das gewisseste, basalste und weitestreichende ubiquitäre Ereignis zur künftigen Aufgabe gemacht hat; und 'vorausgreifend' nenne ich ihn, weil dieser Arbeitsprozess erst 20 Jahre danach an ein Ende gekommen ist, als Kunz seinen *Versuch einer Auseinandersetzung mit der Transzendenz bei Karl Jaspers* als Beitrag für den Schilpp-Band[18] abgeschlossen hatte. Die Hauptarbeit bestand nicht in der Abnabelung; diese war offenbar 1934 schon vollzogen. Sie bestand vielmehr im Überzeugend-machen-Können, dass die Funktionen und Leistungen, die von Jaspers der Transzendenz zugedacht worden sind, durch das Wissen um die je eigene Sterblichkeit und ihre Auslegung ohne große Verluste übernommen werden können.

Wir wollen in die Auseinandersetzung zwischen Kunz und Jaspers über die beiderseitige Analyse des Erkennens einsteigen, weil sie ein klares Bild von den unterschiedlichen Positionen gibt.

Für Kunz geschieht alles Erkennen in Situationen, die den reinen Vollzug des Erkenntnisaktes übergreifen, an dem immer auch Antriebe nichtkognitiver Art beteiligt sind. Selbst wenn diese am Anfang des Erkennens stehen sollten, sind sie doch nicht der Ursprung desselben, der das Wesen des Erkennens konstituiert. Der Erkenntnisakt gleicht vielmehr dem Verhalten des sehenden Auges, das gewaltlos, distanziert, staunend das Seiende sein lässt, was es ist. Der primäre Sinn des Erkennens liegt für Kunz «in der offenen Hinnahme des Begegnenden, so wie es sich von sich her zeigt».[19] Nichts von vornherein ausklammern und nichts heranbringen, für Ergänzungen und Korrekturen aber stets offen sein: das ist seine phänomenologische Haltung zu allem, «was aus der Welt her und in der eigenen Innerlichkeit begegnet».[20] Der «reinen Wissensintention»[21] ist alles Einwirken – zumal das destruk-

17 Ibid.
18 *Karl Jaspers*, hg. von Paul Arthur Schilpp (= SCH) (Stuttgart: Kohlhammer, 1957) S. 493-514.
19 Ibid. S. 496.
20 SD (Fn. 3) S. 139.
21 SCH (Fn. 18) S. 496.

tive – zuwider. Sie geht «auf die Bewahrung des zu erkennenden Seienden und Geschehens».[22] Von sich selbst sagt Kunz, dass er sich schon frühzeitig «den Grenzen des Wissens gebeugt und den sie überschreitenden illusionären Auskünften misstraut»[23] habe.

Jaspers dagegen hat ein Leben lang auf die *Unterschiede* zwischen der wissenschaftlichen und der philosophischen Erkenntnisweise aufmerksam gemacht, die, seiner Meinung nach, heute klarer gesehen werden können, als jemals zuvor:

Wissenschaftliches Erkennen ist immer immanentes und partikuläres Erkennen. Es hat einen zwingenden Charakter und ist methodisch bewusst erschlossen. Es ist unabschließbar und damit für alle Korrekturen und Ergänzungen offen, sofern auch sie einen wissenschaftlichen Charakter haben.

Philosophisches Denken dagegen transzendiert. Es übergeht das wissenschaftliche Wissen nicht, sondern nimmt von ihm Kenntnis, aber übersteigt es. Wo es an die Grenzen des Wissbaren stößt, bricht es seine Bemühungen nicht ab, sondern denkt im Wissen des Nichtwissens weiter. Dieses Denken ist für Jaspers ein Erhellen, z.B. der Existenz, der Transzendenz, des Ursprungs, des Ziels, des Wesens, des Glaubens – kurz: aller Kategorien, die selber zum Transzendieren hin offen sind.

Negativ bedeutet das: Es gibt kein wissenschaftliches Wissen vom Ganzen des Menschseins, der Welt, des Geschichtsverlaufs und kein positives Wissen über die Transzendenz und das Transzendente.

Positiv bedeutet es: Die Grenze der Immanenz ist nicht die Grenze des *Denkens*, sondern nur des *Erkennens*. Das erhellende Denken und mit ihm die philosophische Kommunikation ist auch über diese Grenze hinaus möglich. Das philosophische Denken ist insofern unbegrenzt offen, das wissenschaftliche Erkennen aber ist endlich. Die Trennung der Philosophie von der Wissenschaft hat dem philosophischen Denken seine ursprüngliche Offenheit und Weite zurückgegeben.

Hätte man Kunz gefragt, was in seinem Erkenntnis-Konzept das philosophische Erkennen noch vom wissenschaftlichen unterscheide, hätte er vermutlich geantwortet: «je länger je weniger». Denn er meinte zu sehen, dass das Erkennen aus einem eigenen philosophischen Ursprung kontinuierlich unter dem Druck der positiven Wissenschaften schwindet, die gleichsam die Gebiete annektieren, die zuvor der Philosophie vorbehalten waren. In diesem epistemischen Prozess nahm für ihn Jaspers eine zunehmend

22 Ibid.
23 SD (Fn. 3) S. 134.

exzentrische Position ein: «Gelegentlich kann man sich des Eindrucks nur schwer erwehren» – so schreibt er –, «dass Jaspers gleichsam auf dem letzten – vielleicht verlorenen – Posten eines aus den eigenen Ursprüngen sich nährenden Philosophierens steht und ihn gegen die andrängenden Ansprüche der positiven Wissenschaften zu halten versucht […]»[24] Das Motiv des Unzeitgemäßen klingt an und wird mehrfach erhärtet: Jaspers konserviere einen Wissenschaftsbegriff, angesichts dessen es fraglich sei, ob er noch allen Bemühungen gerecht werde, die sich heute als «wissenschaftliche» verstehen. Die enge Fassung des Begriffs ermögliche es ihm aber, andere Konzepte als 'unwissenschaftlich', 'wissenschaftszerstörend' und 'wissenschaftsabergläubisch' zu verwerfen, wie vor allem die Psychoanalyse. Es frage sich überdem, ob Jaspers nicht an einer «Idee der Wissenschaftlichkeit» festhalte, die sich allzu sehr am naturwissenschaftlichen Erkennen orientiere. Den Prinzipien dieser Idee könnten aber nicht einmal die beschreibenden Naturwissenschaften genügen, und für das Erfassen des geschichtlich-gesellschaftlichen Geschehens sowie des menschlichen Erlebens und Verhaltens seien sie «grundsätzlich unangemessen»,[25] weshalb sie für die Forschung «faktisch überholt» und nicht mehr «maßgebend»[26] seien. – Ein anderer Weg, «dem Philosophieren einen legitimen Raum neben der gegenständlichen Forschung zu sichern und zu retten», sei die «radikale Verwerfung schon der Möglichkeit eines zwingend begründbaren ontologischen Wissens um das Sein […] und insbesondere um den Seinscharakter des Menschen».[27] Mit dieser Verwerfung habe er nicht nur den größeren Teil der abendländischen philosophischen Tradition gegen sich, sondern auch Heideggers Entwurf einer Fundamentalontologie, der sich schwerlich kurzweg als «prinzipieller philosophischer ‹Irrweg›»[28] abtun lasse. Und schließlich könnte man fragen, ob seiner Existenzerhellung nicht eine verborgene, «unausdrücklich gebliebene ‹Ontologie› des Menschseins zugrunde»[29] liege.

Darf man aus dieser präliminierenden Kritik an Jaspers indirekt schließen, was Kunz selber gesucht hat: ein zwingend begründbares ontologisches Wissen, insbesondere um den Seinscharakter des Menschen, der ausdrücklich «festgelegt», «unveränderlich» und damit «bestimmbar»[30] ist, und ein

24 SCH (Fn. 18) S. 494.
25 Ibid.
26 Ibid.
27 Ibid.
28 Ibid. S. 495.
29 Ibid.
30 Ibid.

entsprechendes Wissen um das Sein, bzw. die Transzendenz, die eine völlig andere Auslegung als bei Jaspers bekommt und mit dieser nur noch den Namen gemein hat? Man kann sich schwerlich ein philosophisches Bemühen denken, das demjenigen von Jaspers radikaler entgegengesetzt ist. Und doch muss Kunz die Hoffnung gehabt haben, wenn auch nicht Jaspers zu überzeugen, so doch ihn für seine Fragen und für sein Projekt zu interessieren. Jaspers aber tat sich schwer damit und verwarf es schließlich mit ebensolcher Radikalität.

Zu Beginn seiner Umdeutung der Transzendenz erinnert Kunz an ihre funktionale Bedeutung bei Jaspers. Die Verwandlung des bloßen Daseins in das eigentliche Existieren ist ohne Gründung des Menschen in der Transzendenz nach Jaspers nicht möglich. Die Kurzformel dafür heißt: «Existenz ist nicht ohne Transzendenz.»[31] Dieser Satz ist vielleicht das einzige Dogma in Jaspers' Philosophieren. Aber niemand weiß exakt, was er bedeutet. Denn die Transzendenz hat viele Namen. Sie heißt auch «das namenlose Eine» oder «das Sein» oder «der Seinsgrund» oder «das Umgreifende» oder «das Umgreifende der Umgreifenden» oder «die eine Wahrheit» oder «die Gottheit» oder «Freiheit». Kunz wählt die Übersetzung, die ihn gleichsam am meisten ärgert: «Gottheit». Was immer über sie gesagt werden mag: Es ist für Jaspers bloß Chiffre – Symbolik, die auf Transzendenz hinweist, aber sie nicht vergegenständlicht oder gar mit ihr zusammenfällt. Alles kann auf Transzendenz hinweisen – keineswegs allein das Denken oder die Wortsprache, sondern auch gegenständlich Seiendes: eine Blume, ein Kunstwerk, das Meer, ein Gesicht.

Kunz will nun zeigen, «dass und warum sowohl eine Aneignung der Jaspers'schen Existenzerhellung möglich wie die Verwirklichung des eigentlichen Existierens vollziehbar wird ohne Bezug auf die Transzendenz, so wie Jaspers sie interpretiert».[32] Das bedeute, dass er eine andere «Interpretation» für das gebe, was Jaspers «Transzendenz» nenne, und zwar «eine solche, die den von ihm [Jaspers] gemeinten Sinn wenigstens partiell in das genaue Gegenteil verkehrt: die Transzendenz soll als ein immanentes Konstituens des Menschseins ‹aufgewiesen› werden».[33] Er werde versuchen, über eine

31 Karl Jaspers: *Der philosophische Glaube angesichts der christlichen Offenbarung*, in *Philosophie und christliche Existenz. Festschrift für Heinrich Barth zum 70. Geburtstag am 3. Februar 1960*, hg. von Gerhard Huber (Basel, Stuttgart: Helbing & Lichtenhahn, 1960) S. 30.

32 SCH (Fn. 18) S. 498.

33 Ibid.

«angemessenere Auslegung» eines «Sachverhalts», der für die Existenzerhellung wesentlich sei – nämlich des Todes – «eine sich im Raum der Existenzerhellung bewegende Auseinandersetzung durchzuführen».[34]

Die erste Auseinandersetzung betrifft folgerichtig den Tod. Jaspers bedenke den Tod als «Grenzsituation». In ihr bekomme der Tod zwar existentielle Relevanz dadurch, dass ich mich zu ihm verhalte. Aber dieses Verhalten verändere sich mit der Dynamik meiner Existenz. Meine Geschichtlichkeit fließe so in den Tod ein. Er ist «nicht endgültig, was er ist».[35] Aber gerade darauf käme es an. Nur als wirklicher Tod: als meine künftige finale Faktizität, ist er ganz das, was er ist, und das, von dem ich mit einer «unbedingten singulären Gewissheit weiß»,[36] dass er einmal kommen wird. Mit der Gewissheit des Dass des Todes, der nur die «Daseinsgewissheit» die Waage halten kann, sei aber die «Unfassbarkeit» des Was des Todes verbunden: eine inhaltliche Leere und Bodenlosigkeit, in der uns das Wesen des Todes entgleite. Dieser Zwiespalt von Gewissheit und Unfassbarkeit hat Kunz zur «Vermutung» veranlasst, «im Ursprung des Wissensaktes selbst könnte sich der Tod […] bekunden».[37] Das wäre aber nicht mehr der wirkliche Tod, sondern der mögliche oder auch «inständige» oder «virtuelle» Tod. Kunz hat diesen Tod in die «Herkunftsdimension» des Wissens, Denkens und Erkennens vorverlegt. Dort ist er die Bedingung der Möglichkeit, dass wir überhaupt vom Tod sprechen und an ihn denken können, oder auch die apriorische Bekundung des Seinsverlustes in allen Denkakten.

Den beiden Toden entspricht ein zwiefaches Enden des Menschen: «einmal» im «faktischen Enden des zeitlichen Daseins», also im «zu sterbenden Tod», dann «jederzeit» als mögliches Enden «im inständigen Tod als Ursprung des Denkens».[38] In diesem jederzeitigen Enden öffnen sich «die Weisen der inständigen Zeitlosigkeit»,[39] also das, was wir auch «immanente Ewigkeiten» nennen könnten: die «inständige Zeitlosigkeit», die «Ewigkeit des Augenblicks», die «Überzeitlichkeit der Denkgehalte», die «punktuelle Zeitlosigkeit der Denkakte»[40] u.a.m.

34 Ibid.
35 Ibid.
36 Ibid. S. 499.
37 Ibid.
38 Ibid. S. 500.
39 Ibid.
40 Ibid.

Sofern es aber im Menschen so etwas wie ein «geschehendes Hereinstehen» des virtuellen Todes gibt, könnte man darin den ständigen Einbruch des «Jenseits» sehen, im wirklichen Tod aber die «Faktizität des ganz Anderen» oder «das Jenseits des Menschen und des Weltseins».[41] In diesen Entfremdungsweisen sieht Kunz «vom Menschen her» die einzigen Antriebe zum *Fragen* und *Suchen* nach der Transzendenz in der doppelten Bedeutung als «Ursprung» seines Woher und als «jenseitiges Ziel» seines Wohin.[42] Es ist nicht mehr ein Fragen aus der subjektiven Willkür des philosophierenden und glaubenden Menschen, sondern ein Fragen, das zur Natur des Menschen gehört – oder, wie Kunz abermals sagt: zu seinem «unveränderlichen Seinscharakter».[43] Schließlich sei der inständige Tod der Grund, wenn auch nicht der einzige, für den spezifischen Möglichkeitscharakter des Menschseins, dank dem der einzelne Mensch sein bloßes Dasein in Existenz zu verwandeln vermag.

Jede solche Verwandlung hatte für Jaspers einen Doppelcharakter. Sie war als Bemühen Freiheit und als Gelingen Geschenk. «Als *Existenz* bin ich, indem ich mich durch Transzendenz mir geschenkt weiß.»[44] Das «Durch-mich-Sein ist mir ein in meiner Freiheit Geschenktsein.»[45] Beide Hinweise gingen Kunz zu weit: der Hinweis auf die Freiheit, weil er zu voluntaristisch war, und der Hinweis auf die Hilfe durch die Transzendenz, weil er in ihm einen «Ausläufer der selbstentfremdeten dämonologischen Auslegung des Menschseins»[46] sah. Er selber interpretierte, dass der Einzelne seine eigentliche Existenz «gleichsam abstoßend und aufschwingend»[47] angesichts des inständigen Todes gewinnt, der sich als «möglicher Seinsverlust» offenbart.

Zur Frage wird nun, inwiefern sich die existentiellen Bezüge zur Transzendenz immanent interpretieren lassen. «Immanent» meint: «auf das Menschsein beschränkt unter Einbezug seiner Grenze im faktischen Tod».[48] Es versteht sich von selbst, dass diese immanente Auslegung der Transzendenz nicht Jaspers' Gedanken verdeutlicht, sondern sie unter einen anderen

41 Ibid.
42 Ibid.
43 Ibid.
44 Karl Jaspers: *Der philosophische Glaube. Gastvorlesungen, gehalten auf Einladung der freien akademischen Stiftung und der philosophischen Fakultät der Universität Basel im Juli 1947* (Zürich: Artemis, 1948) S. 20.
45 Ibid.
46 SCH (Fn. 18) S. 501.
47 Ibid.
48 Ibid. S. 503.

Aspekt rückt, wie Kunz sagt, in der Absicht, «damit etwas an den Tatbeständen als solchen [und nicht nur in Worten über sie] aufzudecken».[49]

Bei Jaspers gibt es im späteren Philosophieren gelegentlich den Satz: «Dass Gott ist, ist genug.»[50] Er hat eine Reihe von Wandlungen durchgemacht. In der *Philosophie* hieß es noch: «Es ist genug, dass Sein ist.»[51] Dann im Vortrag *Vom europäischen Geist:* «Es ist genug, dass Transzendenz ist.»[52] Und später eben: «Dass Gott ist, ist genug.»[53] Kunz fragt nun, ob innerhalb des Menschseins ein Phänomen nachweisbar sei, das an die Stelle der Transzendenz zu treten vermöchte, so dass man sagen könnte: dass x ist, ist genug. «In der Tat» – so Kunz – «scheint uns das auf den bevorstehenden Tod zuzutreffen.»[54] «Wenn alles versinkt» – Kunz persifliert damit die Schlusspassage des Vortrags *Vom europäischen Geist*[55] – «Wenn alles versinkt, wenn die Tradition, aus der wir sind, radikal ausgerottet und das Lebendige um uns von Grund aus zerstört würde, dann bliebe allein noch der je eigene Tod.»[56] Dieser sei aber offenbar an eine Voraussetzung gebunden, der dieselbe Gewissheit zukomme: «das gegenwärtige Da-Sein und die Seinsgewissheit des daseienden Menschen».[57] – Da dieser aber um das eigene Verschwinden wisse, könne das Bleiben der Transzendenz nicht im endlichen Menschsein sein Vorbild haben. Damit stellt sich Kunz die äußerste Frage schlechthin: «Woraus entspringt denn das ‹ewige Sein› der Gottheit?»[58]

Sie werden es schon erraten haben: «Hier bietet sich nun der Rückgriff auf den möglichen Tod an.»[59] Als solcher bekundet er sich nicht jenseits des Menschen, sondern stets «nur inmitten des existenten Menschen».[60] Er ist insofern auf den existenten, d.h. da-seienden Menschen «angewiesen», auf den er zugleich «verweist».[61] Wenn nun dieses Hinweisen aber «aus der

49 Ibid.
50 Karl Jaspers: *Einführung in die Philosophie*, (Zürich: Artemis, 1950) S. 38.
51 Karl Jaspers: *Philosophie*, Bd. III (= Ph III) (Berlin: Julius Springer, 1932) S. 236.
52 Karl Jaspers: *Vom europäischen Geist. Vortrag, gehalten bei den Rencontres Internationales de Genève, September 1946* (München: Piper 1947) S. 31.
53 Jaspers, op. cit. (Fn. 50).
54 SCH (Fn. 18) S. 504.
55 Jaspers, op. cit. (Fn. 52).
56 Ibid. S. 54.
57 SCH (Fn. 18) S. 504.
58 Ibid.
59 Ibid.
60 Ibid.
61 Ibid.

Negativität des möglichen Seinsverlustes» erfolgt, wird es zum «radikal Anderen seiner selbst». Als radikal Anderes seiner selbst umfasst es als selber inhaltlich Leeres virtuell alles Seiende. Es ist das «umfassende Sein als Gegenwurf des Nichts».[62] Da «überdies das Menschsein als geschichtliches zeitlich-endlich begrenzt ist und der faktische Tod dessen Ende bedeutet»,[63] lässt sich nach Kunz nun verstehen, dass und warum dem Sein «Ewigkeit» eignet. Seine Ewigkeit ist die Zeit-losigkeit im Enden der gelebten Zeitlichkeit. Im Tode ragt gewissermaßen die «Ewigkeit» als Zeitlosigkeit in das Dasein – und dieser zugleich virtuelle wie faktische Sachverhalt kann auf verschiedene Weise ausgelegt werden, u.a. auch so, dass dem Sein im Enden seiner Zeitlichkeit 'Ewigkeit' zukommt.

In der Frage nach der Transzendenz als Ursprung scheint es Kunz, dass sich Jaspers (wie so oft) dem Erkennen «entzieht».[64] Aber wie kann er sich dem Erkennen «entziehen» oder auch «stellen», wo es nichts zu erkennen gibt? Er bewegt sich ja nicht, wie Kunz, in den Gefilden einer immanenten Transzendenz, sondern der ganz und gar transzendenten, die Kunz wohl eher für ein Gebiet der imaginierten «Ausgeburten»[65] hält.

Gemeinsam ist beiden nur, dass sie die Frage nach der Transzendenz «als zum Menschsein wesensnotwendig gehörendes Konstituens»[66] bewahren möchten – Kunz betont: «aber nur als solches».[67] Jaspers möchte mehr: die Vergewisserung, dass Transzendenz ist, die kein Wissen werden kann. Denn sie muss «ohne Bestimmung»[68] sein und bleiben: Die Transzendenz ist, was sie ist. Er darf sich also dem Nichtwissen nicht entziehen, das mit dem hier einzig möglichen Wissen verbunden ist: mit dem Wissen des Nichtwissens. Für Kunz ist das ein Weg des Denkens, der für die «anthropologische Wahnhaftigkeit» offenbleibt. Für Jaspers ist dieses Urteil aber eine Blindheit für das, was Transzendenz als Ursprung sein könnte.

Drei Formen der absoluten Transzendenz sind nach Jaspers historisch in einer unvergleichlichen Machtfülle wirksam geworden: das namenlose Eine, die eine Wahrheit und der eine Gott.[69] Für Kunz wird zur Frage, woraus ihre faszinierende Macht zu erklären sei. Die Antwort ist schnell zur Hand:

62 Ibid.
63 Ibid. S. 505.
64 Ibid. S. 506.
65 Ibid.
66 Ibid. S. 507.
67 Ibid.
68 Ph III (Fn. 51) S. 67.
69 Ibid. S. 116ff.

«dass hier der Tod als das Paradigma der Einmaligkeit eine untergründige Rolle spielt».[70] Zwar ist alles Seiende und alles Geschehen im strengen Sinne singulär. Aber unser Erkenntnisinteresse nimmt es nicht als solches wahr. Es orientiert sich nicht am Seinscharakter des Seienden, sondern an seiner pragmatischen Relevanz, der wir «Einmaligkeit» zuerkennen, wenn sie in besonderer Weise über das Durchschnittliche herausragt. Lebensgeschichtlich im strengen Sinne einmalig sind das Geborenwerden und der Tod, wobei die Singularität des Todes insofern die radikalere ist, als sie alle Formen der Wiederholung ausschließt. Diese «eigentlichste» Singularität des faktischen Todes muss sich auch im virtuellen Tod bekunden: «Sie scheint im sich aufschwingenden Denken des Einen, im Suchen der einen Wahrheit und im Glauben an den einen Gott manifest zu werden.»[71] Da sich aber in der innerweltlichen Erfahrung die Vielheit des Seienden, der Wahrheiten und der Götter streiten, stellt sich im Hinblick auf die Gestalt des Einen die Frage nach deren Ursprung im Menschsein. «Hier eben bietet sich der Tod an, der das, was er ist, nicht nur als einmaliges Ereignis, sondern als der je eine des je individuierten Lebens ist.»[72] Den drei Gestalten des Einen wachse ihre «unbedingte berückende Macht» aus der «verborgenen Gewalt des inständigen Todes»[73] zu. Zwar sei der Tod an die Bedingungen des Daseins geknüpft, d.h., man muss vorerst leben, um sterben zu können –, aber durch den Eintritt hebe der Tod die Bedingung seiner Selbst auf. Dieser Verlust des Bedingtseins räume als Möglichkeit dem Menschen «die Fähigkeit des unbedingten Sichentscheidens ein».[74] Von daher wäre zu zeigen, dass «die menschliche Freiheit […] im inständigen Tode wurzelt»[75] und dass auch der Mensch «zum Träger des Einen werden» kann, «das ich bedingungslos bejahe»[76] – was natürlich bedeutet, dass es des Gottes nicht bedarf, um das Eine zu verehren.

Das wohl Seltsamste an der menschlichen Existenz war für Jaspers das «Grundfaktum», dass der Mensch sich tief zu ängstigen vermag und dazu, wenn er die Wirklichkeit sieht und bedenkt, auch allen Grund hat – und dennoch fähig ist, den Sprung aus der Angst in die Ruhe und in die Gebor-

70 SCH (Fn. 18) S. 507.
71 Ibid. S. 508.
72 Ibid.
73 Ibid.
74 Ibid.
75 Ibid. S. 508-509.
76 Ibid. S. 508.

genheit, nicht aus Blindheit, sondern angesichts der Wirklichkeit, zu tun. Dass ihm dieser Sprung gelingt, so dachte sich Jaspers, muss seinen Grund über die Existenz des Selbstseins hinaus in der Transzendenz haben. «Die Gottheit ist Ursprung und Ziel, sie ist die Ruhe. Dort ist Geborgenheit. Es ist unmöglich, dass dem Menschen die Transzendenz verloren geht, ohne dass er aufhört Mensch zu sein.»[77]

Kunz hat diesen Passus als letzten gewählt, um ihn zurechtzurücken, und er beginnt damit in folgender Weise: «Wir lassen hier die Angst aus dem Spiel, nicht weil sie für die Erfahrung des Nichts und des Seins bedeutungslos wäre […], sondern [weil] wir glauben, dass das Sein sowohl wie das Nichts zunächst und zumeist im Denken vernommen wird und dass ihre stimmungshafte Bekundung in der ursprünglichen Angst nur gelegentlich durchbricht – als solche dann allerdings ihren wesentlichen Bezug zum möglichen Tode bezeugt.»[78]

Da es Jaspers um den Sprung aus der Angst in die Ruhe geht, Kunz aber die Angst «aus dem Spiel lässt», beschreibt er weitgehend anderes, als Jaspers angedacht hat: nämlich Formen der Ruhe, der Bewegungslosigkeit, des Ausfalls auch der inneren Bewegtheit: also des Fühlens, Wollens und Denkens; Stadien der Nachdenklichkeit, Formen der Stummheit – und beachtet sie als vorgestellte Todeszustände, deren reale Ereignisse, die Weisen des Endens, er als «momentweise» Ausbreitung des inständigen Todes deutet, der gleichsam aus seiner Verborgenheit im Ursprung des Denkens heraustritt. Ob nicht darin, so fragt Kunz, der Grund des Sprungs in die Ruhe liege, den Jaspers «in das Sein der Transzendenz»[79] verlege? Er bestätigt sein Verstehen: «So gibt die Ruhe dem denkenden Menschen einen Halt, der an der Gewissheit des künftigen Todes partizipiert.»[80]

Gegen Ende seiner Abhandlung gibt Kunz zu bedenken: «Wenn die Transzendenz als Ursprung und Ziel des Menschen aus dem möglichen Tode ‹hergeleitet› wird, so liegt die Frage nahe, ob daraus nicht als unvermeidliche Konsequenz ein ‹Nihilismus› resultiere.»[81] Die Frage mag naheliegen, aber nicht unbedingt die Konsequenz. Ich wüsste jedenfalls nicht, wo man im Schaffen und in der Person von Hans Kunz auch nur eine Spur von Nihi-

77 Karl Jaspers: *Vom Ursprung und Ziel der Geschichte* (= UZG) (Zürich: Artemis, 1940) S. 280.
78 SCH (Fn. 18) S. 509.
79 Ibid. S. 511.
80 Ibid.
81 Ibid. S. 512.

lismus ausfindig machen könnte. Im Gegenteil: Er ist der selbst auferlegten Verpflichtung nachgekommen: «Eine Weile Bewahrer des entgleitenden Seienden und damit der Menschlichkeit zu sein».[82]

Zum Schluss kommt Kunz auf den heikelsten Punkt seiner Auseinandersetzung mit Jaspers in beeindruckender Redlichkeit zu sprechen: Sein Versuch, die Transzendenz zu opfern, beruhe auf der Annahme, dass in der «seinsmäßigen Konstitution des Menschen […] der virtuelle Tod als Ursprung des Denkens dem faktischen Tod ‹vorausgehe›».[83] Kunz «unterstellt» – wie er ausdrücklich sagt –, dass diese Voraussetzung zutrifft, «ohne sie ‹beweisen› oder auch nur als in sich einsichtige dartun zu können».[84] Ja, er gesteht, dass er für die Annahme «keinen unbezweifelbaren Sachverhalt aufweisen kann»[85] – und dass deshalb seine Interpretation als Möglichkeit nur wahrscheinlich sei. Obwohl er wisse, dass «die These vom inständigen Tode dem Charakter des Jaspers'schen Denkens von Grund aus widerstreitet, danke sie ihm doch ihre Entstehung».[86] Deshalb habe er den Versuch gewagt, ihre «vielleicht erhellende Kraft in der Konfrontation mit einigen Zügen der Transzendenz zu prüfen – diese zwar, aber nicht die hohen Möglichkeiten des Menschseins opfernd».[87]

Kunz sieht also seine Lage ganz genau: Er spricht zwar gelegentlich von «Sachverhalten» – aber er hat nur Hypothesen. Von der nahezu die ganze Auseinandersetzung tragenden Hypothese weiß er, dass sie unbeweisbar ist, ja, ohne Evidenz. Er weiß, dass er Jaspers' Philosophie Gewalt antut, weil seine Hypothese ihr von Grund auf widerspricht – kurz: Er müsste eigentlich sehen, dass er in diesem Kampf von vornherein auf verlorenem Posten steht. Und dennoch geht er die Auseinandersetzung ein.

Etwas muss ihn dazu gedrängt haben. Es ist vermutlich die über viele Jahre dauernde meist stumme Auseinandersetzung mit Jaspers, über den für diesen so zentralen und für Kunz so marginal gewordenen Begriff einer reinen Transzendenz, den Kunz ohne große Bedenken opfert, weil er auf die «erhellende Kraft» einer These hofft, die man auch für abstrus halten könnte. Jaspers ist mit ihr erbarmungslos umgegangen, was andererseits vielleicht zeigt, dass Kunz eine empfindliche Stelle in seinem Philosophieren getroffen hat.

82 Ibid. S. 514.
83 Ibid.
84 Ibid.
85 Ibid.
86 Ibid.
87 Ibid.

Wenn man Jaspers' Antwort auf die Umdeutungen der Transzendenz von Kunz auf einen Satz bringen wollte, könnte er in Stichworten etwa lauten: methodisch mangelhaft, wissenschaftlich unergiebig, philosophisch nicht haltbar, in der Gesinnung feindlich und im Gehalt nichtig.

Die methodische Mangelhaftigkeit zeigt sich für ihn darin, dass bei Kunz schon im Ansatz «Tatbestand und Hypothese nicht scharf getrennt»[88] werden. Sein Verfahren, so kritisiert Jaspers, sei nicht Beobachtung, sondern «Sinnverstehen», das er als «Tatbestand» und «Sachverhalt» ausgebe. Er unterscheide überdies nicht die beiden Wege des Sinnverstehens: den Weg, der zur psychologischen Beobachtung und zur empirischen Erkenntnis führe, und den Weg des appellativen Denkens, das sich an Freiheit wende. «Kunz leugnet diese Einsicht.»[89] Nur deshalb vermute er, dass sie beide von den gleichen Tatbeständen reden und sie nur anders interpretieren. Bei ihm, Jaspers, gehe es aber überhaupt nicht um Tatbestände und Sachverhalte, sondern um Freiheit und ihre möglichen Gehalte. Dass dagegen Kunz die philosophische Seite sinnverstehender Gedankenbewegungen außer acht lasse, zeige sich in seinen Aussagen immer wieder.

Eng mit dem Methodologischen hängt die Theorienbildung zusammen. Kunz liefere «ein neues Beispiel» einer psychologischen Theorie über das Grundgeschehen der Seele. Diese «Theorie» sei «ein Entwerfen ohne echte Verifizierbarkeit». Es handle sich bloß um «mehr oder weniger große Plausibilität eines Soseinkönnens im an sich unzugänglichen Zugrundeliegenden»,[90] womit Jaspers vermutlich die Auslegung sowohl des faktischen wie des möglichen Todes meint. Hier werde «nichts Tatsächliches gewonnen»[91] und die Forschung der Psychologie nicht gefördert, weil kein Angriffspunkt für empirische Untersuchungen gegeben werde. Der Versuch sei eine gedanklich schwer zu vollziehende, unanschauliche und in der Konstruktion quälende Theorie, die wissenschaftlich unergiebig sei.

Zugleich ist für Jaspers Kunzens Versuch «philosophisch nicht haltbar». Denn: «Kunz leugnet in der Tat die Freiheit oder versteht unter dem Wort etwas anderes […]».[92] Diese Leugnung spreche er in der These aus: «Die Freiheit der möglichen Existenz und der sich verhaltende und erlebende

88 Ibid. S. 817.
89 Ibid. S. 818.
90 Ibid. S. 820.
91 Ibid.
92 Ibid. S. 818.

Mensch»[93] seien seinsmäßig nicht zwei getrennte Bereiche, sondern nur eine einzige Wirklichkeit, nämlich die des in bestimmten Situationen lebenden konkreten Menschen. Diese Einheit als Faktizität aber sei für Kunz «das Seinmüssen dieses Ganzen»,[94] das auch das «Freisein» umfasse. Die Einheit des Menschseins sei ihm «Gegenstand der Psychologie»[95] und nicht Idee. Dies aber sei mit der Möglichkeit der Freiheit nicht vereinbar.

Schließlich meint Jaspers, in den Umdeutungen der Transzendenz durch Kunz herauszuhören, dass sie sich grundsätzlich gegen sein Philosophieren wenden: «Sie deuten hypothetisch aus einem Grundgeschehen, wo ich appellierend an Möglichkeiten denke. Sie bringen eine Lehre in Thesen, wo ich durch Gedankenbewegungen etwas erzeugen möchte im Denkenden. Sie unterlassen und verwerfen das, was mir der Sinn des Philosophierens ist.»[96] Die Umdeutungen der Transzendenz scheinen ihm «wie Mehltau auf den Gehalt meiner existenzerhellenden Versuche zu fallen».[97] Die eigenen Bemühungen verschwinden unter den Umdeutungen durch Kunz. So oft er aber diese bedenkt, scheinen sie ihm «wie eine Chiffer des Nichts»[98] zu sein, «fast wie ein Beispiel der Nichtigkeit dieser oft auftretenden Art von psychologischer Hypothetik».[99]

Ein freundliches Wort findet Jaspers allein für den philosophierenden Menschen Hans Kunz, der «die unersetzliche Kostbarkeit jedes Tuns und jedes Seienden» erfahren möchte und darin die Forderung erkenne und anerkenne, «eine Weile Bewahrer des entgleitenden Seienden und damit der Menschlichkeit zu sein».[100] Von ihm sagt er: «Darin liegt die philosophische Energie dieses bedächtigen Gelehrten und verborgenen Philosophen, gegen die ich keinen Widerspruch erhebe.»[101]

93 Ibid.
94 Ibid.
95 Ibid. S. 819.
96 Ibid. S. 817.
97 Ibid. S. 820.
98 Ibid.
99 Ibid.
100 Ibid.
101 Ibid.

Buchbesprechungen / Comptes rendus

Laurent Cesalli : **Le réalisme propositionnel. Sémantique et ontologie des propositions chez Jean Duns Scot, Gauthier Burley, Richard Brinkley et Jean Wyclif** (Paris : Vrin, « Sic et Non », 2007) 496 pages.

Pour le nombre de philosophes de la fin du moyen-âge, les propositions constituent un élément essentiel de notre rapport au monde, de notre capacité à formuler son organisation, voire à la penser : dans la perspective médiévale, qui place la relation entre mots, concepts et choses au cœur du langage, le niveau mental joue un rôle-clé. Le livre de Laurent Cesalli nous plonge dans une famille de théories sur la sémantique des propositions, nées à Oxford au XIV^e^ siècle, qui partagent une conception remarquable de la frontière entre le mental et le réel, une frontière où vivent d'étranges entités hybrides, suffisamment mentales pour avoir leur être dans l'esprit, suffisamment réelles pour que Gauthier Burley puisse faire voler un oiseau entre le sujet et le prédicat (voir pp. 196-201). Il faudra quatre cents pages et tout le talent de Laurent Cesalli pour nous donner à saisir la portée philosophique de telles entités, plus faciles à trahir ou moquer qu'à restituer et comprendre. L'auteur y parvient en nous fournissant, systématiquement et progressivement, les contextes et les problématiques qui leur donnent un sens, les motivent et les encadrent. Ce faisant, il montre que le Réalisme propositionnel constitue une catégorie à la fois historiographique et philosophique à part entière, et ouvre un champ inédit d'interrogations sur la conception qu'avaient les logiciens médiévaux de ce que nous appelons « sémantique » et sur la façon de la restituer.

La méthode choisie par Laurent Cesalli, qu'il présente au début de l'ouvrage, consiste en une «reconstruction analytique contextuellement éclairée». Il s'agit pour l'auteur de ne pas se laisser enfermer dans le dilemme que l'on présente parfois à celui qui veut étudier les philosophes du passé : les saisir hors de tout contexte, comme s'ils étaient nos confrères, ou à l'inverse les aborder comme on aborde les protagonistes d'une culture différente, incommensurable, que nous devons reporter le plus littéralement possible. Un problème philosophique de la sorte de ceux dont traite le Réalisme propositionnel est un ensemble articulé d'enjeux et de concepts dont la cohérence interne peut être saisie hors des contextes particuliers de ses instanciations historiques. Le point est important, dans la mesure où il permet de se donner une certaine liberté dans la restitution les théories médiévales qui instancient le problème du réalisme propositionnel ; de proposer des grilles de lecture, de reconstruire leur cohérence et de reconstituer les parties inexprimées mais logiquement impliquées. Dans le même mouvement, cette restitution produit les contextualisations historico-philosophiques du problème indispensables à la mesure de sa portée philosophique ; elles fournissent les articulations théoriques complexes qui en expriment les ques-

tions et justifient les solutions. Les contextes éclairent l'analyse, l'analyse oriente l'éclairage. L'ouvrage de Laurent Cesalli constitue une illustration en profondeur de cette approche. Plutôt que de produire une reconstruction abstraite du problème du réalisme propositionnel tel qu'il serait s'il était formulé par un philosophe de notre époque, et plutôt que de se contenter de décrire par le menu des situations historiques, l'auteur va s'attacher à reconstruire ces situations de sorte de faire apparaître les motivations philosophiques du réalisme propositionnel.

Le premier mouvement consiste à présenter l'histoire des trois notions qui, traditionnellement, couvrent les trois aspects de la sémantique des propositions : *propositio, res* et *dictum*. L'étape suivante fournit un aperçu des modèles théoriques au sein desquels ces notions ont été impliquées, tels que ceux construits autour de la problématique de la création comme acte de langage, des objets de la connaissance divine et de l'isomorphie entre le langage et le réel. D'Anselme à Simon de Faversham, en passant par Guillaume d'Auvergne, Thomas d'Aquin, Raoul le Breton, Martin et Boèce de Dacie, des problématiques sémantiques des grammairiens modistes aux interrogations sur les objets de la foi, Laurent Cesalli éclaire les raisons linguistiques, philosophiques et théologiques qui peuvent conduire à faire de morceaux de complexité réelle des protagonistes (ultimes) de la chaîne sémantique.

La reconstruction orientée que fait Laurent Cesalli de ces théories a ceci remarquable qu'elle rend manifeste le rapport étroit, largement causal, qu'il peut y avoir entre certains concepts et certains enjeux, considérés selon une certaine perspective, et le développement de certains objets théoriques. Cela va de pair avec une conception axiale des problématiques philosophiques, lesquelles possèdent un point de départ, des étapes et une destination théoriques – en l'occurrence, les propositions réelles. Les solutions proposées par les auteurs médiévaux peuvent ainsi être évaluées en fonction de leur position sur l'axe ; dans un moment d'enthousiasme téléologique, Laurent Cesalli estime que des auteurs tels que Raoul le Breton, Simon de Faversham et Boèce de Dacie s'en allaient dans la direction des « propositions réelles » et que, s'ils ne s'étaient pas arrêtés en chemin, ils auraient poursuivi jusque là où arrivera Gauthier Burley un peu plus tard (pp. 64-65). Si l'on peut ne pas être entièrement d'accord avec le principe d'une telle fatalité théorique (qui affecte aussi Scot, voir p. 166), l'on doit reconnaître que l'identification de motivations structurelles constitue l'un des moyens les plus efficaces de donner à comprendre le statut et la portée d'un problème philosophique qui ne nous est pas familier. Cela permet par ailleurs de positionner en chemin des objets théoriques exotiques tels que les *intentiones secundae in concreto* ou *in praedicamento*, les *enuntiabilia* (ou du moins certains d'entre eux), la c*onstructio realis* ou la *locutio rerum* – ce qui représente un gain collatéral non négligeable. Enfin, le dynamisme inhérent à une telle perspective axiale confère au livre de Laurent Cesalli un côté haletant rarement rencontré à ce niveau de technicité: nous sommes plongés dans une traque, celle de la proposition réelle, que l'on croit toujours être sur le point d'atteindre et qui toujours nous échappe, au dernier moment, d'un cheveu théorique. Jusqu'au coup de théâtre final, diligemment fourni par Wyclif et son monde de propositions.

Le panorama problématique une fois dépeint et les conclusions tirées, l'auteur focalise son propos d'un cran, en sélectionnant une question, celle du signifié pro-

positionnel, qu'il pose aux principaux auteurs de la période qui l'intéresse, le quatorzième siècle. Un rapide survol de 23 philosophes représentant toutes les couleurs du spectre nominaliste-réaliste permet de tracer une carte qui servira à mieux positionner les notions et arguments collectés dans la seconde partie du livre, où la traque des propositions réelles se concentre sur les auteurs qui sont les plus susceptibles d'avoir été jusqu'au terme du processus : Jean Duns Scot, Gautier Burley, Richard Brinkley, Jean Wyclif. Une battue systématique s'organise alors, auteur par auteur. Elle passe tout d'abord par une mise en cohérence des éléments pertinents de la philosophie de l'auteur examiné, puis, à partir de tous les indices théoriques disponibles, puis par la détermination de la position de l'auteur quant au signifié et à la « véri-faction » des propositions. Ils ne l'expriment pas toujours de manière explicite : la reconstitution systématique permet d'avancer des hypothèses et de combler les vides – c'est le point fort de l'approche analytique éclairée par le contexte prônée par Laurent Cesalli, qui permet de faire ressortir un objet théorique tel que le réalisme propositionnel en reconstituant la structure philosophique qui le présuppose.

Jean Duns Scot, premier dans l'ordre chronologique, ouvre le bal. L'analyse permet d'établir que le signifié des propositions est une *compositio rerum*, une entité qui a un *esse obiectivum* dans l'intellect – un type d'être distinct de l'*esse subiectivum* que possèdent aussi bien les choses extra-mentales que les qualités mentales, et distinct des *ficta*. L'*esse obiectivum* est un *esse intentionale*, c'est l'être des choses «en tant qu'elles sont intelligées», ce qui fait que le signifié d'une proposition devrait correspondre à quelque chose comme *compositio rerum* (*ut intelliguntur*) *ut intelligitur* (p. 137). En dépit du « ut intelligitur », dont on pourrait croire qu'il présuppose qu'il y a une composition à intelliger du côté des choses, Duns Scot se refuse à s'engager du côté ontologique de la frontière. Du moins pour ce qui est des signifiés des propositions ; car en localisant ces derniers dans l'esprit, il implique qu'ils soient distingués de la cause de la vérité – et de la fausseté – des propositions qui, aussi longtemps que ces dernières portent sur un état contingent du monde et non sur une relation nécessaire, peut difficilement ne pas être dans le réel extra-mental (on peut d'ailleurs hésiter à accepter la proposition de Laurent Cesalli de substituer au véri-facteur réel, quand il fait défaut, par exemple pour les propositions au passé et au futur, un véri-facteur mental (pp. 143sqq.) : la proposition « Socrate courait » ne saurait être vérifiée par un concept que dans la mesure ou un concept est susceptible de courir). Le véri-facteur réel est une chose ou un complexe de choses réelles à quoi la proposition doit être « conforme » pour être vraie – à quoi les propositions fausses ne sont pas conformes. C'est dans la relation étroite entre cette complexité réelle et la *compositio rerum* mentale que réside la particularité sémantique de toutes les approches relevant du réalisme propositionnel. De fait, c'est elle qui pousse Gauthier Burley à proposer l'une des meilleures illustrations de ce que peuvent être ces entités-frontière, *via* la notion de *propositio in re*. Les *res* qui composent ce genre de propositions sont, comme chez Duns Scot, des choses « en tant qu'intelligées » ; des objets mentaux, mais qui constituent l'intellection d'objets extra-mentaux avec la disposition réelle en vertu de laquelle le prédicat leur est attribué. Ainsi, c'est en un sens bel et bien la réalité que l'on compose, en accord avec sa propositionnalité – que l'on reflète ou que l'on explicite mentalement. Une composition mentale fondée dans

le réel. Richard Brinkley, pressé lui aussi de fournir ses signifiés et ses véri-facteurs propositionnels, semble aboutir à une conclusion similaire. Ainsi non seulement l'ordre des parties du signifié d'une proposition est pertinent (pp. 286-288), mais la véri-faction d'une proposition se fait sur la base de sa signification du monde *sicut est*, alors que sa falsi-faction intervient si elle signifie *sicut non est*. Autrement dit, elles signifient la réalité telle qu'elle n'est pas. Dans le même ordre d'idées, les propositions négatives vraies ne signifient pas la réalité *sicut est*, puisqu'elles affirment la non-existence réelle d'une certaine composition. La proposition « nullus homo est asinus » ne signifie pas *sicut est*, puisqu'il n'y a justement rien qui soit homme-âne. D'où l'idée que ce que signifient les propositions affirmatives (vraies), n'est pas réductible à la liste des signifiés de leur termes, que cela doit être cette liste « plus quelque chose » (p. 295). Le raisonnement proposé par Laurent Cesalli est brillant ; il n'est cependant pas exclu que l'idée brinkleyienne soit plus triviale. Le signifié premier des propositions est un objet mental : Lorsque nous formulons une proposition, nous composons mentalement quelque chose, dont nous affirmons ou nions l'existence. Selon que les choses réelles signifiées par les termes de la proposition sont ou non dans la relation qui correspond à ce que l'on a composé mentalement, la proposition sera ou non *sicut est*. Selon que l'on affirme ou nie le composé, elle sera vraie ou fausse. Nul besoin d'introduire dans l'ontologie une entité qui soit plus que ce que signifie le sujet de la proposition. S'il en ressort l'image un peu frustrante d'un Brinkley moins tenté qu'on le voudrait de franchir la frontière du mental pour aller jalonner le réel d'unités émergentes, c'est largement compensé par celle d'un Wyclif qui la franchit au pas de gymnastique, avec armes et bagages. Pour le Doctor Evangelicus, plus question de minauder : les propositions sont non seulement dans le réel, mais le réel tout entier n'est que propositions. Toute chose créée est composée, après tout, et toute composition est le résultat d'une prédication. Cette ontologisation radicale n'empêche pas Wyclif de déployer une entité intermédiaire, *partim in anima, partim extra animam* : l'*ens logicum*, qui, à l'instar de ses confrères en réalisme propositionnel, est une entité mentale fondée dans le réel – l'équivalent wycliffien de la proposition réelle de Burley (voir p. 389).

Ainsi, nous étions partis à la recherche de complexes bêtement réels, et nous avons trouvé beaucoup mieux : les objets théoriques fascinants que Laurent Cesalli appelle entités-α et qui constituent la marque de fabrique du réalisme propositionnel. Le tableau de chasse présenté vers la fin de l'ouvrage (p. 395) rend claire leur situation particulière d'entités mentales complexes fondées dans le réel d'une façon plus étroite que ce qu'offre une simple correspondance. La direction habituelle du lien sémantique qui unit le langage au monde se voit compléter d'un mouvement inverse, qui va du métaphysique vers la complexité propositionnelle et fait de notre esprit le lieu de rencontre d'où émerge l'unité. L'ouvrage de Laurent Cesalli apporte ainsi au triangle langage-esprit-réel un éclairage intensément original, et montre que l'étude de l'histoire de la philosophie, quand elle est à ce point maîtrisée, peut être l'occasion d'aventures théoriques qui sont autant d'ouvertures conceptuelles.

Frédéric Goubier (Freiburg im Breisgau)

Thomas Raeber: ***Ja und Aber.*** **An Grenzen der Wahrheit. Tagebücher 1992-2007, mit einer kompakten Darstellung der beim Schreiben entstandenen Philosophie und einem Beitrag von Martin Götz** (Bern: Stämpfli, 2007) 535 Seiten.

Thomas Raeber ist ein Außenseiter der helvetischen Philosophie, ohne dies explizit sein zu wollen. Das Fach hat er in Basel, Zürich und Freiburg i.Üe. gründlich studiert und sein Studium mit einer soliden Dissertation über die mehrfache Bedeutung des Daseins-Begriffs in Jaspers' *Philosophie* abgeschlossen.[1] Er wurde danach Assistent bei Bochenski für die Redaktion der *Formalen Logik*[2] und Stipendiat des Nationalfonds. Alles sah danach aus, als würde er den üblichen Karriereweg eines Professors der Philosophie gehen. – Aber dann wurde er plötzlich Buchhändler, danach Zentralsekretär der Europa-Union, Schweiz, dann Mitglied der Direktion für «Entwicklungszusammenarbeit und humanitäre Hilfe» im Departement des Äußern, danach Botschafter in Tansania, Sambia, Botswana, Madagaskar und Mauritius und schließlich ständiger Vertreter der Schweiz beim Europarat. Und nun, im Alter von 85 Jahren, legt er seine «Tagebücher» vor (in Wahrheit einen Teil der Tagebücher), die in mehrfacher Hinsicht bemerkenswert sind: sprachlich durch ihre Präzision; geistig durch ihre Weltoffenheit; ästhetisch durch die Subtilität der Wahrnehmungen eines guten Auges und philosophisch durch die ökologische und ontozentrische Wende seines Denkens.

Raeber hat seine Notate nicht systematisch, sondern chronologisch geordnet. Er hat aber aus den etwa 90 000 (!) Texten eine Auswahl getroffen, die sich an die ursprüngliche Chronologie hält. So folgt z.B. auf den Text 77 177 der Text 78 003 und auf diesen der Text 78 026. Der Sinn dieses Vorgehens liegt nicht nur darin, die Auslassungen kenntlich zu machen, sondern vor allem darin, das Subjekt des Denkenden zurückzunehmen. Alle Philosophie, die das Ich als bestimmtes Subjekt zum Ursprung des Denkens macht, ist ihm verdächtig, und zwar aus der ganz anderen eigenen Denkerfahrung. Aus ihr setzt er auf das unbestimmte Subjekt «Es»: «Es denkt in mir.» Oder vielleicht auch: «Es denkt aus mir.» Schon Leibniz und Lichtenberg haben auf diese Denkerfahrung aufmerksam gemacht. Das philosophische Denken folgt nicht dem Projekte-Voluntarismus des Herstellens. Es ist vielmehr ein schweifendes Geschehen-Lassen von geistigen Vollzügen, die nicht unbedingt die Meinung oder Überzeugung des Denkenden zum Ausdruck bringen. – Wer so denkt und auch danach handelt, muss sich allerdings die Frage gefallen lassen, wer denn für das Denken des Es die Verantwortung trage.

Es ist nicht möglich, der Fülle der Tagebücher in einer Besprechung gerecht zu werden. In den konkreten Passagen bestehen sie aus Beobachtungen des Wirklichen und Skizzen des Möglichen sowie aus Reflexionen über Ereignisse und Erlebnisse. In ihnen spiegelt sich die breite Weltkenntnis des Autors und ein oft etwas narzistischer Hang zur Selbstbetrachtung. In einer anderen Tiefendimension fragt er zugleich immer wieder nach den Fundamenten und Grenzen der Wahrnehmung und des philosophischen Denkens sowie nach dem Aufbau der Natur und der Bedeutung des Seins. Weltverständnis, Selbstbetrachtung und Seinsvergewisserung sind die drei

1 Thomas Raeber: *Das Dasein in der «Philosophie» von Karl Jaspers* (Bern: Francke, 1955) 204 S.

2 Josef M. Bochenski: *Formale Logik* (Freiburg, München: Alber, 1956).

Ebenen seines Denkens, wobei ihm die dritte, seine eigene Philosophie, zunehmend zur wichtigsten geworden ist.

Diese eigene Philosophie ist den Tagebüchern nur in Bruchstücken vorausgegangen. Die Konturen eines offenen und gleichsam «fließenden» Ganzen sind ihr erst beim Schreiben während zweier Jahrzehnte zugewachsen und Raeber allmählich auch bewusst geworden. Er hat sie in einer «kompakten Darstellung» den Tagebüchern als Nachwort beigegeben. Sie ist einerseits eine wichtige authentische Verständnisquelle, die zeigt, was er mit seiner Philosophie bewirken möchte, und andrerseits ein Quell der wachsenden Zweifel am ganzen Projekt, weil sowohl seine Wünschbarkeit zum Teil fraglich ist und sein metaphysisches Fundament einem Zurück vor Kant gleichkommt.

Das Fazit seiner Philosophie, so Raeber am Schluss seiner Kompakt-Fassung, sei:

«Diese Philosophie sucht einen Weg von der aufklärerischen, anthropozentrischen liberalen und humanistischen, sozialen Gesellschaftsauffassung» zu einem ontozentrischen, naturozentrischen und ökozentrischen Weltverständnis, in dem der Mensch zwar immer noch in der unaufhebbaren perspektivischen Täuschung bleibt, im Mittelpunkt der Welt zu sein, aber nun weiß, dass er nicht der Mittelpunkt ist. Im Hinblick auf Raeber hat das zur Folge, dass er das «in den Vordergrund stellen» möchte, was «unabhängig vom Menschen existiert». Ihm möchte er alles unterstellen, was es ohne den Menschen nicht gäbe, «was durch ihn und für ihn entstanden ist», und insbesondere alles, was er «planmäßig» hervorgebracht und hergestellt hat.

Das Bild, das dem Passus zugrunde liegt, ist das der zwei Wege, die sich trennen. Man muss den einen verlassen, um den anderen gehen zu können. Sie führen nicht nebenher an naheliegende Orte, sondern in verschiedenen Richtungen an weit entfernte Ziele. Und diese entfernten Ziele heißen: *Entweder* eine anthropozentrische, aufklärerische und humanistische *oder* eine ontozentrische, ökozentrische, mundozentrische Gesellschaftsauffassung. Beides zugleich ist nicht zu haben, weder vom Ziel her noch als Weg. Das ist eine typisch religiöse Propaganda-Strategie, die das Problem, vor dem wir stehen, aus den Augen verliert. Die Frage ist nicht, ob wir *entweder* eine aufklärerische, humanistische, anthropozentrische *oder* eine ontozentrische, ökozentrische und mundozentrische Welt wollen, sondern wie wir *zugleich* eine humanistische *und* ökologische, eine aufklärerische *und* naturschonende Welt bauen können. Wenn man, wie Raeber, nach einer Welt der «Erträglichkeit zwischen Menschen und allem Lebendigen und der Natur» sucht, darf man ihre Differenzen nicht zu unverträglichen Alternativen stilisieren. Im Übrigen: Wer möchte um der Natur oder um des Seins willen eine inhumane, unliberale und asoziale Gesellschaftsauffassung anpreisen?

Aber eigentlich meint Raeber gar nicht, dass letztlich der Mensch diese neue Welt baut. Die Veränderung der Welt und mit ihr des Menschen hat vielmehr einen Geschehens-Charakter. Zum Geschehen gehört «das absichtliche und unabsichtliche, ‹gute› und ‹böse›, moralisch konforme und abweichende Tun und Lassen des Menschen». Die «fortschreitende geschichtliche Entwicklung» selber «bringt die Werte und moralischen Grundsätze hervor. Sie sind nicht Dogmen, sondern Ereignisse der Natur». Man könnte diese Überzeugung für einen naturalistischen Fehlschluss halten.

Aber wenn das «Es» die Normen «denkt», sie «weiß», sie «sagt», sie «mitteilt», und der Mensch auch noch gleich «im Es drin» ist, hat es die Kritik schwer.

Eine ontozentrische Philosophie muss vielleicht vom Sein ausgehen. Für Raeber jedenfalls ist das Sein der zentrale Gedanke seiner Philosophie. Dieses Sein ist weder der Ursprung noch eine besondere Qualität des Seienden, sondern sowohl der Raum, in welchem alles Seiende ist, als auch das Insgesamt der Seienden in diesem Raum. Der Raum ist prallvoll, und außerhalb des Raumes ist nichts.

Das Sein kann man zwar nicht definieren, weil außer ihm nichts anderes ist. Aber man kann Seinserfahrungen machen. Es sind die Erfahrungen der einen und einzigen Wirklichkeit, der das eine und einzige Sein entspricht. Es gibt unabsehbar viele Arten des Wirklichen wie des Seienden, aber nur *eine* Wirklichkeit und *ein* Sein. Wirklich sind für diesen Sprachgebrauch auch alles Virtuelle, unsere Träume, Utopien, Gedanken. Raeber hat deshalb auch keine Probleme mit der Frage nach der Wirklichkeit oder der Existenz Gottes. Dass Gott ist, sagt er ohne Bedenken und präzisiert: Gott ist das Sein als «Gestalt». Ob dieser Gott nun «bloß» ein Gedanke sei oder eine Erscheinung, gilt ihm gleich viel. Was soll denn an einer Gestalt oder Erscheinung wirklicher sein als an einem Gedanken?

Nur in einem Punkt muss er sich widersprechen. Er kann oder will dem Nichts, das ja auch ein Gedanke ist und vielleicht nur ein Gedanke (und darin dem Gott zum Verwechseln ähnlich ist), kein Wirklich-Sein zusprechen. Es gibt kein absolutes Nicht-Sein und kein wirkliches Nichts. Und deshalb ist das Sein alles – alles in steter Bewegung und Wandlung, weshalb ihm auch nur ein «fließendes» Denken gerecht werden kann.

So wie es für Raeber die Erfahrbarkeit der einen Wirklichkeit und des einen Seins gibt, gibt es für ihn auch eine Gotteserfahrung. Er nennt sie die «Unmittelbarkeit zu Gott». Er weiß, dass diese Formulierung eigentlich der Philosophie davonläuft und verwandelt sie deshalb zuweilen in die bloße Formel «U.z.G.» oder ersetzt sie, der Philosophie näher, mit der Metapher «Berührung des Grundes».

Der Gott der U.z.G. meint «die lebendige Gegenwart des Ganzen, des einen Seins», aber nicht als abstrakten Begriff oder vage Empfindung, sondern als überzeugende Einsicht, als Zustimmung zum Lauf der Natur, zum Freisein, das ich erfahre, und zur Wahrheit, in der ich lebe. Die Empfindung dieser Harmonie, die die Welt, den Gott als Gestalt des Seins und das Ich in Wahrheit und Freiheit ungehindert vereint, könnte man «Glück» nennen, das, wie ein Funkenlicht, in seiner Unbeständigkeit aufleuchtet und wieder erlischt, zum Zeichen dafür, dass man das Glück nicht machen und erzwingen kann, sondern dass auch es einen Geschehens-Charakter hat.

Es wäre noch über viele Texte aus den Tagebüchern zu reden, die nicht von der Grundlegung einer eigenen Philosophie handeln, sondern von alltäglichen politischen, ästhetischen, soziologischen Fragen. Sie sind insgesamt in der Überzahl. Die Prägnanz, die Differenzierungsfähigkeit, das sprachliche Niveau müssten hervorgehoben werden, und das gute Auge müsste beim Betrachten der Fotographien, die Raeber selber geschossen hat, noch einmal gelobt werden. – Die Leistung als Ganze ist eindrücklich, und sie verdient es, gewürdigt zu werden, auch wenn man ihre metaphysisch-religiösen Höhlengänge nicht immer mitmachen kann.

Hans Saner (Basel)

Christine Clavien, Catherine El-Bez (éds): **Morale et évolution biologique. Entre déterminisme et libre arbitre** (Lausanne: Presses polytechniques et universitaires romandes, 2007) 354 pages.

L'ouvrage *Morale et évolution biologique*, issu du groupe de recherche «Déterminisme et libertés» de l'Université de Lausanne et de son cycle de conférences «L'éthique, l'inné et l'acquis» (2005-2006), se propose, par le biais d'un engagement et d'un dialogue interdisciplinaires, de traiter des questions liées à l'interprétation naturaliste, plus précisément évolutionniste, des comportements moraux et des normes morales. Ce livre couvre un grand nombre de problèmes liés à la morale et à l'évolution biologique, à travers une répartition thématique sous la forme de cinq chapitres traitant, respectivement, de l'émergence de la vie morale, du lien entre évolution et émotions morales, du rôle du langage dans la constitution et dans l'effectuation de l'action morale, de la confrontation entre science évolutionniste et philosophie morale, ainsi que, finalement, des rapports entre déterminisme et responsabilité morale.

Les travaux formant cet ouvrage couvrent un vaste champ scientifique, en ceci qu'ils mettent à contribution les investigations actuelles en psychologie cognitive et en biologie, ne cessant de dialoguer avec les résultats issus de la recherche scientifique contemporaine. Interrogeant tant l'histoire de l'espèce humaine comprise comme entité biologique évoluant au gré de ses stratégies adaptatives (Ronald de Sousa), que l'action morale particulière et les rapports qu'entretiennent dans ce cadre les mécanismes de contrôle (Luc Faucher) ou les émotions de honte et de culpabilité (Julien A. Deonna) avec la norme éthique, cet ouvrage circonscrit le champ général des questions que la théorie de l'évolution pose à l'émergence et à la pratique du comportement moral, tout en critiquant divers aspects de l'évolutionnisme éthique. En effet, d'une part, certaines contributions défendent les thèses de l'évolutionnisme ou du naturalisme moraux, se fondant notamment sur l'analyse empirique des comportements humains (Christine Clavien) et sur le rejet du subjectivisme ou de l'anthropologie linguistique (Fabrice Clément). D'autre part toutefois, et contrairement à ce que laisse croire le titre, l'approche évolutionniste pure n'est pas unilatéralement défendue dans cet ouvrage, les auteurs réunis dans le cadre de cette recherche mettant ladite approche à l'épreuve de concepts issus de domaines philosophiques aussi variés que le contractualisme (Nicolas Baumard), l'anthropologie d'inspiration wittgensteinienne (Yves Erard), la philosophie du sens commun (Christine Clavien) ou encore le stoïcisme (Catherine Dekeuwer). Il découle de cette pluralité de points de vue un anti-dogmatisme manifeste, qui se révèle dans des contributions opposées à l'éthique évolutionniste, soit qu'elles accusent l'éthique évolutionniste d'être incapable de rendre compte de la particularité de l'action morale, celle-ci étant éminemment fondée sur l'organisation personnelle de la vie de l'agent, et non sur un ensemble de phénomènes compris quantitativement (Julien A. Deonna, Philippe Huneman), soit qu'elles en décèlent les apories d'un point de vue logique, identifiant la circularité du discours évolutionniste, qui se doit de présupposer l'origine des jugements moraux dans l'analyse génétique avant d'en constater l'émergence (Philippe Huneman). Cette multiplicité d'approches marque à nouveau un souci d'ouverture et de dialogue présent tout au long de ce livre, souci dont l'interdisciplinarité se fait le témoin dans les contributions finales portant sur les

liens entre déterminisme biologique et responsabilité morale. En effet, la sociologie montre comment la culture populaire, encline à admettre en principe un déterminisme génétique moniste, adopte un dualisme explicatif lorsqu'il s'agit de considérer la responsabilité pénale des agents sous l'angle du libre arbitre (Catherine Dekeuwer), tandis que le droit pénal permet de repenser les relations logiques qu'entretiennent entre elles les sphères du descriptif et du normatif, mettant en garde contre le causalisme naïf (Alain Papaux).

Finalement, chacun des cinq chapitres thématiques de cet ouvrage se conclut par un texte des modérateurs de séance du cycle de conférences (Nicolas Perrin, Jacques Dubochet, Anne-Claude Berthoud, Lazare Benaroyo et Alain Kaufmann). Ces textes ne se contentent pas de résumer les contributions des auteurs de l'ouvrage, mais mettent celles-ci en résonance avec des thèmes permettant l'élargissement des interrogations liées à l'évolution biologique, notamment du point de vue politique du biopouvoir ainsi que des rapports entre société civile et recherche scientifique (Alain Kaufmann), dimensions essentielles à la délimitation des enjeux de l'évolution biologique dans son lien à la morale, mais qui, malheureusement, ne restent qu'à l'état de suggestion dans le cadre du présent ouvrage et ne connaissent pas de développement tangible.

Il est dès lors à regretter, malgré la diversité des recherches présentes dans *Morale et évolution biologique*, l'absence de réel questionnement quant à l'émergence historico-sociale des comportements et des normes éthiques, ainsi que la mise au ban du caractère herméneutique de ceux-ci, autrement dit de leur inscription dans l'expérience vécue. Cette absence et cette mise au ban tracent la bordure d'un ouvrage qui, malgré sa volonté d'ouverture, omet de considérer la densité propre à l'étude des phénomènes moraux du point de vue de la constitution de soi du sujet éthique enchevêtré dans un cadre historique donné, se maintenant par conséquent fermement ancré dans la philosophie analytique, soit dans un mode de questionnement réducteur en dernière instance.

Hamid Taieb (Genève)

Joachim Fischer: **Philosophische Anthropologie. Eine Denkrichtung des 20. Jahrhunderts** (Freiburg/München: Alber, 2008) 684 Seiten.

Ein Jubiläum hat stets rituelle Züge der vorbehaltlosen Bestätigung. Hier aber geht es darum, etwas allererst *sichtbar* zu machen, und darum, es zu revitalisieren: die «Philosophische Anthropologie». Achtzig Jahre nach der Veröffentlichung der beiden Werke, welche die Geburt der Philosophischen Anthropologie in der deutschen Philosophie bedeuten (*Die Stellung des Menschen im Kosmos* von Max Scheler und *Die Stufen des Organischen und der Mensch* von Helmuth Plessner), erscheint nun – als Jubiläumsarbeit gewissermaßen, aber auf längerfristige, von Jubiläen unabhängige philosophische Aktualität angelegt – die umfangreiche Arbeit von Joachim Fischer: *Philosophische Anthropologie. Eine Denkrichtung des 20. Jahrhunderts*. Beachtliche Studien zur Einordnung dieses Denkens und zur Vertiefung in dieses Denken gibt es sicherlich bereits. Es fehlte aber das 'Manifest' (wie ich es bezeichnen möchte), das nun vorliegt: die mit Verve vorgetragene Idee einer tiefgreifenden Identität im Denken zwischen ansonsten zuweilen erbitterten Rivalen, die Überzeugung vom besonderen Potential dieser spezifischen Philosophischen Anthropologie gerade im 'biologischen Zeitalter'. Ein 'Manifest', das keiner, der sich für das Thema Mensch interessiert, wird ignorieren können, ein 'Manifest', so reich an Informationen wie eine (mehrfache) intellektuelle Biographie und von so weit reichenden Reflexionsanreizen wie nur ein philosophisches Werk.

Der derartig mit Information und Gedankenschärfe angereicherte Manifestcharakter des Werkes wird evident, wenn man die Zielsetzung zur Kenntnis nimmt. Das Vorhaben ist, zu zeigen, dass es neben den bisher als philosophische Paradigmen verstandenen Denkweisen im 20. Jahrhundert – der Phänomenologie, der Existenzphilosophie, dem Positivismus, dem Neokantianismus und anderen – noch eine präzise und eigenständige philosophische Richtung gibt, die «Philosophische Anthropologie» heißt und die der philosophischen Frage nach dem Menschen eine spezifische Strategie vorschlägt. Die Identität der «Philosophischen Anthropologie» als einer philosophischen Denkweise wird also von den anderen Konzeptionen des Menschen (die jede philosophische Theorie mehr oder minder explizit enthält) unterschieden, aber auch unterschieden von allen Hauptströmungen des zeitgenössischen Denkens. Beides sind Schritte derselben Operation: Philosophische Anthropologie ist eine Konzeption des Menschen, aber eine besondere Vorgehensweise in dieser Konzeption, die eine ganze neue (nichtcartesianische) Philosophie entfaltet – auch eine Philosophie der Religion, der Moral, des Rechts, der Politik usw.

Joachim Fischer nimmt den fundamentalen philosophischen Anspruch, den Max Scheler und Helmuth Plessner selbst für die Philosophische Anthropologie erhoben, damit erneut auf und macht ihn zugleich erst sichtbar. Denn wegen gegenseitiger Plagiatsvorwürfe zwischen Scheler, Plessner und Gehlen, des plötzlichen Tods von Scheler und der erzwungenen Emigration Plessners war die Sichtbarkeit des gemeinsamen Ansatzes erheblich erschwert. Eine Schulbildung konnte so explizit nicht stattfinden. Während Scheler und Plessner je für sich das Erstverwertungsrecht dieses philosophischen Projekts reklamierten (und Gehlen sich als der Nachfolger vor allem

von Herder darstellte), zeigt Fischer in stupender Detailarbeit, dass und inwiefern alle diese Denker und weitere Autoren – etwa Adolf Portmann – einer einheitlichen Denktechnik folgen. *Ex negativo* zeige zudem der Kampf um die Exklusivität des je eigenen Werkes gerade die Gemeinsamkeit der Philosophischen Anthropologie, die nicht zuletzt darin liegt, die Philosophische Anthropologie in Auseinandersetzung mit Darwin zu konzipieren.

Der erste, umfangreichere Teil des Bandes wertet eine akribische Quellenrecherche aus, um die wechselhafte, zuweilen tragisch anmutende «Realgeschichte» der Philosophischen Anthropologie zu rekonstruieren. Daraus ergibt sich ein Bild eines bisher so nicht sichtbar gewordenen Denk- und Forschungsvorhabens und darüber hinaus ein neues Szenarium der deutschen Philosophie des 20. Jahrhunderts, das dokumentiert, wie sich die Philosophische Anthropologie trotz akademischer Streitigkeiten und individueller Schicksale festigen und schließlich durch hochrangige 'Schüler' (wie Hans Blumenberg, Hans Jonas und selbst Niklas Luhmann) eine einflussreiche philosophische Strömung im deutschen Denken der Nachkriegszeit werden konnte: einflussreicher, als es vielen scheinen mag.

Die ersten Phasen der «Realgeschichte des Denkansatzes» reichen mit den parallel entstandenen Werken Schelers und Plessners bis zum ersten großen Werk Arnold Gehlens: *Der Mensch. Seine Natur und seine Stellung in der Welt* (1940). In der Rekonstruktion der Werkgenesen wird deutlich, welchen großen Anteil die Biologen und Verhaltensforscher Friedrich J. J. Buytendijk, Hans Driesch, Johann Jakob von Uexküll und Adolf Portmann an der Entfaltung und Fortentwicklung des Denkansatzes hatten und welchen geringen dagegen die Philosophen Ernst Cassirer, Karl Löwith, Günther Anders. Nach 1945 ist eine «Konsolidierungsphase» zu beobachten, später, in den 1960ern, das Wechselspiel von zunächst vielfältigen Anschlüssen in Soziologie, Psychologie, Pädagogik, Medizin und dann das Vergessen des Denkansatzes in der deutschen intellektuellen Welt bis 1975.

Philosophiehistorisch sind alle Phasen, nicht zuletzt sicher die bisher weniger beachteten Jahrzehnte von 1945 bis 1965 äußerst interessant. Folgt man der von Fischer rekonstruierten Rezeption der Philosophischen Anthropologie, so zeigt sich, dass fast alle wichtigen deutschen Autoren sich an den Themen gemessen haben, die die Philosophische Anthropologie auf den Tisch legte: Sei es, um sie weiter voranzutreiben (wie Blumenberg, Anders, Löwith, Marquard, Straus oder Binswanger), sie neu auszuarbeiten (wie Apel, Schmitz oder Jonas) oder um sie trotz eng verwandter Interessen scharf zu kritisieren (wie Habermas und die Frankfurter Schule). Nicht zuletzt sind zwei hochkarätige Debatten um die Philosophische Anthropologie zur Kenntnis zu nehmen: von Dahrendorf angestoßen, über den Begriff der Rolle, und von Lorenz angestoßen, über den Begriff des Instinktes. Dies korrigiert die nicht nur im Ausland gängige Auffassung, nach der die dominanten philosophischen Paradigmen im Nachkriegsdeutschland einerseits das an Heidegger anschließende hermeneutisch-phänomenologische Paradigma und andererseits die an Wittgenstein und den *linguistic turn* anschließende analytische Denkweise seien.

Eine ebenso aufregende Klärung leistet der zweite Teil («Zur Philosophiegeschichte des Ansatzes»), der die Theoriestruktur der Philosophischen Anthropologie ausbuchstabiert: ihre «philosophiegeschichtliche Lage», ihre «Denkungsart», ihren «Denkort». Dieser Aufweis der spezifischen Denkweise ist für das Gelingen des Gesamtvorhabens zentral: nicht nur wegen der faktischen Rivalität, also der tragischen «Realgeschichte» des Ansatzes, sondern eben auch, weil sich die philosophische Anthropologie mit Kant als eine allgemeine – vieles umfassende – philosophische Disziplin etabliert hat und weil sich in der zeitgenössischen Tendenz eines durchdringend 'nachmetaphysischen' Denkens und infolge der radikalen Historisierung der menschlichen Phänomene nahezu alles 'Anthropologie' nennt.

Was kennzeichnet also «Philosophische Anthropologie» mit großem P? Zunächst, dass sie sich keiner Denktradition verbunden fühlt: der Verzicht auf die großen *récits* der Philosophie und das Bedürfnis nach dem Austausch mit den empirischen Wissenschaften. Sehr viel genauer aber kennzeichnet den Denkansatz dann der Versuch, den Begriff des 'Geistes' neu zu verhandeln: ihn der Alternative von Dualismus oder Monismus zu entziehen. Aus diesem Ansatzpunkt erklärt sich die Zentralität der Biologie für die Philosophischen Anthropologen: Sie ist ihre Leitwissenschaft, mit der sich die dualistische Trennung des Menschen in Geist und Körper verhindern lässt. Es geht um eine Untersuchung, welche die menschlichen Monopole – die Kultur – im Ausgang von der menschlichen *Natur* als lebendiger Organismus in den Blick nimmt. Dieses Prinzip eines inneren Zusammenhangs von Natur und Kultur des Menschen (als die *Grammatik* des menschlichen Lebens) wird bei allen Autoren des Denkansatzes nachgewiesen. Und nicht nur dies, vielmehr ist bei allen Autoren dieses Denkansatzes das Bemühen zentral, den Menschen weder monistisch noch reduktionistisch zu denken. Wohl *drückt sich* seine Artikulation, also die *Syntaxis* des Lebens dualistisch *aus*. Aber diese Dualität ist kein Faktum, sie ist eher eine Gegenüberstellung zweier zutiefst mit einander verschränkter Sphären als die Spaltung, die Descartes zuerst eingeführt hat und die nach wie vor eine unvordenkliche Rolle spielt. Plessner nennt diese 'Gegenüberstellung' die «exzentrische Positionalität» des menschlichen Lebens.

Ein zentrales Ausgangsmoment des philosophisch-anthropologischen Denkens wird von dem Biologen Jakob von Uexküll übernommen: Jeder lebendige Körper ist nur in Korrelation mit seiner «Umwelt» denkbar. Diese Korrelation verdankt sich einer ersten Differenzierung: einer kontrastierenden Distanz zum Außen. 'Lebendig' sind diejenigen Körper, die eine Grenzlinie zwischen Innen und Außen «besitzen» und diese gegen ihre Umwelt behaupten. Lebendige Körper sind, in Plessners Worten, «grenzrealisierende Dinge». Fällt nun diese Grenze nicht mehr mit dem Körper zusammen, sondern wird sie über Kleidung, Schmuck, Behausung, Denkstile und Lebensformen 'hinausgeschoben', dann wird sie zur Möglichkeit, die stets neu zu definieren ist. Dann eröffnet sich jener virtuelle Raum, jenes «Nichts», jener Blick von Nirgendwo, den wir Selbstbewusstsein, Reflexivität oder Freiheit nennen. Diese hören nicht auf, Bewusstsein von etwas, das uns *affiziert,* zu sein: Reflexion über Gegebenes, Freiheit gegenüber etwas und jemandem. Deshalb besteht methodisch für die Philosophische Anthropologie ein Primat des Objekts

vor dem Subjekt, der *physis* vor der *noesis*, ein Primat, das die Stabilisierung und Kompensation der «Krise» der menschlichen Natürlichkeit ist, die durch die instinktuale Imperfektion des 'Mängelwesens' Mensch hervorgebracht ist. Mit einer glücklichen Formulierung spricht Fischer statt vom Primat des Objekts auch vom «flankierenden Blick» der Philosophischen Anthropologie (S. 522): jenem indirekten, dezentrierten Blick, jenem «Standpunkt eines Dritten», den wir imaginieren und der unsere Art, Erfahrungen zu sammeln, begleitet: immer prekär auf der Grenze positioniert, im Bereich der Umkehrung zwischen Geist und Körper, Sinnen und Sinn, Schließung der Umwelt und kosmischer «Öffnung». Die Begründer der Philosophischen Anthropologie (neben Scheler, Plessner, Gehlen werden dazu auch der Mediziner Paul Alsberg, der Biologe Adolf Portmann und der Philosoph Erich Rothacker gezählt) haben auf je eigene Weise diese Grenzposition dekliniert. Und sie haben dabei vor allem je andere Aspekte der menschlichen Situation in den Vordergrund gerückt: die Sprache, die Expressivität, die Institutionen, die Konstruktion von Mikro-Welten, die Transzendenz und die Religion. Aber das Vorgehen ist dasselbe: die Welt des Geistes in der Grammatik des Lebens zu verankern. Und gleich ist auch die Methode: der flankierende Blick, der die Dopplungen reflektiert, den Zwang zur Objektivation, zur Distanz, die notwendig sind für das menschliche Leben – der Mensch lebt nicht einfach, sondern «führt» sein Leben, steht vor ihm als einer zu realisierenden Aufgabe: die «exzentrische Positionalität» ist von Natur aus künstlich.

Zur Bestätigung der These – des identischen Denkansatzes, der sich als prägnant und bedeutsam gegenüber den bereits bekannten Denkansätzen im 20. Jahrhundert erweist – schließt der Band mit einer knappen Differenzierung dieser Denkströmungen des 20. Jahrhunderts, deren Berührungspunkte und Differenzen in Relation zur Philosophischen Anthropologie markiert werden.

Obwohl folgerichtig für das Vorhaben, weckt eine solche Akribie – im Unterschied zur notwendigen Akribie in der Rekonstruktion der «Realgeschichte» im ersten Teil – doch Verwunderung. Fischers 'Manifest' riskiert am Schluss, die Philosophische Anthropologie zu isolieren, sie übermäßig einzuengen. Um der Wiedererkennbarkeit des Denkansatzes willen droht am Ende eine vielleicht zu starke Eingrenzung und infolge dessen womöglich eine Beeinträchtigung der Effizienz dieser Denkart: vor allem wenn zu deren Charakteristika – so möchte ich es eher verstehen – die methodologische und ebenso die interdisziplinäre Offenheit gehört. In dieser Offenheit sieht Fischer offenbar eine Gefahr. Man kann in ihr aber auch einen Vorteil, ein Potenzial sehen, das es erst noch auszuschöpfen gälte. Die Auswahl der «offiziellen» Repräsentanten der Philosophischen Anthropologie ist hierfür ein Beispiel: Autoren der Größenordung von Cassirer, Löwith und Blumenberg werden als Vertreter des Ansatzes ausgeschlossen (nicht aber als von ihm Beeinflusste). Natürlich hat Fischers These ihre eigene Kohärenz, auch philologisch. Aber es ließe sich die ebenso starke These Odo Marquards geltend machen, nach der die Philosophische Anthropologie in der frühen Moderne als Analyse der menschlichen «Lebenswelt» aus naturalistischen Voraussetzungen entspringt: einer Welt, die weder mit den Mitteln der positiven Wissenschaften noch mit denen der traditionellen Metaphysik erfassbar war. Ich persönlich bevorzuge weiterhin diese Lesart, zumal

sie dem *modernen* philosophischen Thema des Menschlichen mehr historische Tiefe und größere begriffliche Nuancen verleiht. Dies verspielt nicht die Originalität der Philosophischen Anthropologie des 20. Jahrhunderts. Vielmehr beweist es die Neuartigkeit dieses Denken, das bleibende Fragen und Antworten zu formulieren vermochte (selbst das 'moderne' Wort «Anthropologie» ist ein Zeichen dafür). Originalität und Autonomie einer Theorie verlieren sich nicht, wenn sie ohne festen Wohnsitz bleibt. Im Gegenteil …

Marco Russo (Salerno)

Adressen der Autoren / Adresses des auteurs

Urs Allenspach, ETH Zürich, Departement Umweltwissenschaften, CHN H 73.2, Universitätsstraße 16, CH-8092 Zürich

Johannes B. Balle, Dr. phil., Universität Köln, Philosophisches Seminar, Albertus-Magnus-Platz, D-50923 Köln

Yves Bossart, M. A., Paulstraße 21, D-50676 Köln

Georg Brun, Dr. phil., ETH Zürich, Institute for Environmental Decisions, CHN H 73.1, Universitätsstraße 16, CH-8092 Zürich

Christoph Calame, 53, avenue de Rumine, CH-1005 Lausanne

Christine Clavien, Dr. phil., Université de Lausanne, Département d'écologie et évolution, UNIL – Sorge, Biophore 3112, CH-1015 Lausanne

Marcello Ostinelli, Dr. phil., Dipartimento della Formazione e dell'Apprendimento, Scuola Universitaria Professionale della Svizzera Italiana, Piazza San Francesco 19, CH-6600 Locarno

Hans Rott, Prof. Dr. phil., Universität Regensburg, Institut für Philosophie, D-93040 Regensburg

Hans Saner, Dr. phil., Wanderstraße 10, CH-4054 Basel

Julius Schälike, PD Dr. phil., Universität Konstanz, Fachbereich Philosophie, Postfach 5560 D 16, D-78457 Konstanz

Peter Schulte, Dr. phil., Universität Bielefeld, Abteilung Philosophie, Postfach 100 131, D-33501 Bielefeld

Daniel Schulthess, Prof. Dr. phil., Université de Neuchâtel, Institut de philosophie, Espace Louis-Agassiz 1, CH-2000 Neuchâtel

Thomas Sturm, Dr. phil., Departament de Filosofia, Universitat Autònoma de Barcelona, Edifici B, E-08193 Bellaterra (Cerdanyola del Vallès)

Hartmut Westermann, Dr. phil., RWTH Aachen, Philosophisches Institut, Eilfschornsteinstraße 16, D-52062 Aachen

Redaktion / Rédaction

Grundriss der Geschichte der Philosophie. Begründet von Friedrich Ueberweg
Völlig neu bearbeitete Ausgabe. Herausgegeben von Helmut Holzhey

Die Philosophie des 18. Jahrhunderts

Band 2: **Frankreich**
Herausgegeben von Johannes Rohbeck und Helmut Holzhey
2008. XXXVIII, 1044 Seiten. 2 Halbbände. Gebunden. Leinen. Schutzumschlag. Mit CD-ROM
sFr. 245.– / € 170.–
ISBN 978-3-7965-2445-5

Mit diesen zwei Halbbänden liegt eine umfassende Darstellung der Philosophie des 18. Jahrhunderts in Frankreich und in der französischen Schweiz vor. Die behandelte Periode reicht von der Krise des absolutistischen Regimes Ludwigs XIV. über die Französische Revolution bis zur napoleonischen Ära. Der Band beginnt mit einer Darstellung der Institutionen der philosophischen Bildung. Erste Schwerpunkte bilden die frühe Aufklärung, insbesondere die Clandestina, sowie die vergleichende Kulturgeschichte und die politische Philosophie mit ihrem bedeutendsten Vertreter Montesquieu. Es folgen Kapitel über Voltaires Kultur- und Religionskritik und das kollektive Unternehmen der «Encyclopédie». Berücksichtigung finden ferner die Naturwissenschaften wie auch die aufkommenden Gesellschaftswissenschaften und die politische Ökonomie. Weitere Schwerpunkte sind Erkenntnistheorie und Sprachphilosophie (Condillac), der Materialismus (La Mettrie, Diderot, Helvétius, d'Holbach) und seine Kritik durch die katholischen Apologeten. Ausführlich behandelt wird sodann Rousseaus Kritik an Kultur und Gesellschaft. Nach einem Kapitel zur Ästhetik folgen die Darstellungen der sich neu formierenden Geschichtsphilosophie (Turgot, Condorcet), der kontroversen Debatten um die Revolution und schließlich der «Idéologues» (Destutt de Tracy, Cabanis, Maine de Biran, Degérando).

Die Herausgeber

Johannes Rohbeck, geb. 1947, studierte Philosophie, Germanistik, Politologie und Soziologie in Bonn und an der Freien Universität Berlin. Seit 1993 ist er Professor für Praktische Philosophie an der Technischen Universität Dresden.

Helmut Holzhey, geb. 1937, studierte Evangelische Theologie, anschließend Philosophie und Soziologie. Von 1978 bis 2004 war er Professor für Philosophie, besonders für Geschichte der Philosophie, an der Universität Zürich.

Grundriss der Geschichte der Philosophie. Begründet von Friedrich Ueberweg
Völlig neu bearbeitete Ausgabe. Herausgegeben von Helmut Holzhey

Die Philosophie des 18. Jahrhunderts

Band 3/1: **Italien**
Herausgegeben von Johannes Rohbeck und Wolfgang Rother
2010. Ca. 450 Seiten. Gebunden. Leinen. Schutzumschlag. Mit CD-ROM
Ca. sFr. 140.– / € 98.50
ISBN 978-3-7965-2599-5

Der Band bietet die erste umfassende Darstellung der italienischen Philosophie des 18. Jahrhunderts in deutscher Sprache. Zu Anfang werden die institutionellen Bedingungen der Philosophie behandelt: Gelehrtenzeitschriften, Bildungswesen, Akademien, Zensur, Buchmarkt und Bibliotheken. Einen ersten thematischen Schwerpunkt bildet die durch Geschichtsdenken geschärfte philosophische Reflexion über Staat und bürgerliche Gesellschaft in der ersten Jahrhunderthälfte. Ein eigenes Kapitel ist Giambattista Vico gewidmet, der mit seiner neuen Methode zu den Begründern des Historismus und der modernen Geistes- und Kulturwissenschaften gehört. Es folgen Darstellungen der philosophischen Implikationen und methodischen Fragestellungen der mathematisch-naturwissenschaftlichen Disziplinen und der vielfältigen philosophischen Argumentationsstrategien der katholischen Apologetik. Ausführlich wird schließlich die Philosophie der Aufklärung in der zweiten Jahrhunderthälfte behandelt. Im Zentrum stehen die Metropolen in der Lombardei und im Süden des Landes: Mailand mit Pietro Verri und Cesare Beccaria sowie Neapel mit Antonio Genovesi und Gaetano Filangieri; behandelt werden aber ebenfalls die toskanischen und piemontesischen Denker sowie die vor allem in Venedig geführten Debatten über das Verhältnis von Staat und Kirche.

Die Herausgeber

Johannes Rohbeck, geb. 1947, studierte Philosophie, Germanistik, Politologie und Soziologie in Bonn und an der Freien Universität Berlin. Seit 1993 ist er Professor für Praktische Philosophie an der Technischen Universität Dresden.

Wolfgang Rother, geb. 1955, studierte Philosophie, Theologie und Germanistik in Marburg, Tübingen und Zürich. Er ist Lektor im Verlag Schwabe und Privatdozent für Philosophie unter besonderer Berücksichtigung der Geschichte der Philosophie an der Universität Zürich.

Das Signet des 1488 gegründeten Druck- und Verlagshauses Schwabe reicht zurück in die Anfänge der Buchdruckerkunst und stammt aus dem Umkreis von Hans Holbein. Es ist die Druckermarke der Petri; sie illustriert die Bibelstelle Jeremia 23,29: «Ist nicht mein Wort wie Feuer, spricht der Herr, und wie ein Hammer, der Felsen zerschmettert?»

Eine der großen Herausforderungen der Philosophie als Sachwalterin der Vernunft ist die Tatsache, dass Menschen sich nicht immer rational oder sich gar irrational zu verhalten pflegen. Zu dem von Menschen hervorgebrachten Irrationalen gehören Phänomene wie Irrtum und das Versäumnis der Irrtumsvermeidung, Selbsttäuschung, Wunschdenken und Willensschwäche. Diese Phänomene sind von unterschiedlicher Natur. Während man von Irrtum und Täuschung sagen kann, dass sie uns unterlaufen oder wir ihnen erliegen und dass sie sich auflösen, sobald wir sie erkennen, scheinen Phänomene wie Willensschwäche und Selbsttäuschung nicht einmal auf konsistente Weise beschreibbar, geschweige denn erklärbar zu sein. Denn was sind das für mentale Zustände, in denen jemand etwas glaubt, was er nicht glaubt, oder genau das tut, was er für falsch hält? Dieses Problem verfolgt die Philosophie seit ihren Anfängen, und es ist in jüngerer Zeit – nicht zuletzt wohl im Zuge der erhöhten Rationalitätsansprüche der analytischen Philosophie – erneut virulent geworden.

Anton Hügli, geb. 1939, studierte Philosophie, Psychologie, Germanistik/Nordistik und Mathematik in Basel und Kopenhagen. Er war von 1981 bis 2001 Direktor des Pädagogischen Instituts Basel-Stadt und ab 1981 Privatdozent, dann außerordentlicher Professor und von 2001 bis 2005 vollamtlicher Professor für Philosophie und Pädagogik an der Universität Basel.

Curzio Chiesa, geb. 1953, studierte Philosophie in Genf, Paris und Cambridge. Er ist seit 1978 Maître d'enseignement et de recherche für antike und mittelalterliche Philosophie an der Universität Genf.

Schwabe Verlag Basel
www.schwabe.ch

ISBN 978-3-7965-2596-4