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.