Julian Nida-Rümelin

Entscheidungstheorie und Ethik
Decision Theory and Ethics

2., erweiterte Auflage
2nd, Extended Edition

Herbert Utz Verlag · München

Wir danken Blackwell Publishers, Oxford
und Kluwer Academic Publishers, Dordrecht
für folgende Abdruckrechte:

Bibliografische Information Der Deutschen Bibliothek:
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN 3-8316-0395-2

Printed in Germany

Herbert Utz Verlag GmbH, München
089-277791-00 · www.utzverlag.de

Vorwort

Dieses Buch beschäftigt sich mit dem Verhältnis von Rationalität und Moralität. Der größere, erste Teil enthält meine nur unwesentlich veränderte Dissertation, die im Jahre 1982 fertig gestellt wurde, die also schon geraume Zeit zurückliegt. Der zweite Teil dokumentiert Ergebnisse meiner weiteren Auseinandersetzung mit der gleichen Thematik bis in die jüngste Zeit. Es ist dabei ein mich selbst gelegentlich irritierendes Merkmal meiner philosophischen Arbeiten, dass sich ihre inhaltlichen Grundlinien in rund 25 Jahren nicht verändert haben. Die Darstellungsform hat sich gewandelt. Insbesondere habe ich in späteren Schriften versucht, formale Mittel so sparsam wie möglich einzusetzen, um die philosophische Substanz der Argumentation nicht zu verdecken. In meiner letzten Buchpublikation zu dieser Thematik *Strukturelle Rationalität. Ein philosophischer Essay über praktische Vernunft* (Reclam Universalbibliothek, 2001), habe ich ganz auf formale Darstellungsformen verzichtet. Es muss allerdings hinzugefügt werden, dass dieser Verzicht nicht ohne Kosten ist. Klare theoretische Zusammenhänge, die in der formalen Darstellung, sei es in der Spieltheorie oder in der Theorie kollektiver Entscheidungen (*public choice, collective choice*), elegant in Matrix-Form oder durch formal präzise Kriterien dargestellt und erfasst werden, bedürfen recht umständlicher Umschreibungen, wenn man auf dieses Analyse-Instrumentarium verzichtet. Die normalsprachlichen Umschreibungen erhöhen zudem die Gefahr von Fehlinterpretationen und Missverständnissen.

Der zweite Teil dieses Buches führt vier Texte auf, die das Verhältnis von Entscheidungstheorie und Ethik näher beleuchten.

Der erste dieser Texte, „Rational Choice: Extensions and Revisions“, gibt einen kritischen Überblick über den *rational choice*-Ansatz und geht auf interne Inkohärenzen ebenso ein, wie auf seine systematischen Grenzen.

„Why Consequentialism Fails“ fasst in kompakter Form meine Kritik des Konsequentialismus zusammen, die detaillierter in *Kritik des Konsequentialismus* (München/Wien: Oldenburg 1993, 2. Aufl. 1995) ausgeführt ist.

Der dritte Text, „Rationality: Coherence and Structure“, argumentiert an einem Beispiel, dass die deontologische Dimension ethischen Handelns sich in einen konsequentialistischen Begriffsrahmen von *rational choice* nicht einbetten lässt, sehr wohl aber in einen kohärentistischen.

Der letzte Beitrag, „Why Rational Deontological Action optimizes Subjective Value“, entwickelt für diese These eine präzise Rechtfertigung.

Julian Nida-Rümelin, München 2004

Preface

This book deals with the relation between rationality and morality. The bigger, first part contains my only insignificantly changed dissertation, which was completed a while ago, in 1982. The second part documents the results of some of my studies on the same topic until the recent time. It is a feature of my philosophical work that confuses myself sometimes, that, regarding the content, the basic ideas have not changed over about 25 years. The form of presentation has changed, however. In particular, in my later work I tried to use formal devices as thriftily as possible, in order not to hide the philosophical substance of the argumentation. In my last publication on this topic, *Strukturelle Rationalität. Ein philosophischer Essay über praktische Vernunft* (Reclam Universalbibliothek, Stuttgart 2001), I dispensed even with any formal ways of presentation. It has to be added, however, that this renunciation does not go without costs. Clear theoretical relations which, in a formal presentation, be it in the fields of game theory or collective choice, can be elegantly depicted in the form of a matrix or through formally precise criteria, require quite long-winded descriptions if these instruments of analysis are left out. Ordinary language descriptions also increase the danger of misconceptions.

The second part of this book presents four texts that closely examine the relation between decision theory and ethics.

The first one "Rational Choice: Extensions and Revisions" provides a critical overview of the rational choice approach and discusses its internal incoherences, as well as its systematic limits.

"Why Consequentialism Fails" summarizes my critique of consequentialism, which is set out in more detail in my book *Kritik des Konsequentialismus* (München/Wien: Oldenburg 1993, 2. Aufl. 1995).

The third text, "Rationality Coherence and Structure", argues with an example that the deontological dimension of ethical actions cannot be embedded in a consequentialist frame of *rational choice,* yet in a coherentist one.

The last contribution, "Why Rational Deontological Action optimizes Subjective Value", gives a precise justification of the same thesis.

Julian Nida-Rümelin, München 2004

Inhalt · Contents

[1] in *Ratio (New Series)*, Vol VII, Blackwell Publishers, Oxford 1994, pp. 122–144.

[2] in G. Hilström-Hintikka and R. Tuomela, eds., *Contemporary Action Theory*, Vol II., Kluwer, Dordrecht 1997, pp. 295–308.

[3] in J. Nida-Rümelin and W. Spohn, eds., *Rationality, Rules and Structure*, Kluwer, Dordrecht 2000, pp. 1–16.

Rational Choice: Extensions and Revisions

Abstract

The rational choice paradigm has proved to be a fruitful means of analysis and explanation in various disciplines concerned with questions of practical rationality, but in various attempts to extend its fields of application the deficiencies of the original account of rational choice have become evident. This paper tries to give a systematic overview of ways in which the rational choice paradigm can be extended and modified.

1. The Basic Model

A person acts rationally if her actions appear to make sense with respect to the person's aims. Actions make sense with respect to a person's aims if they seem to be appropriate means for achieving those aims.

These two propositions constitute the core of an influential conception of practical rationality, which has shaped a variety of disciplines in various degrees: theoretical economy, political theory, theoretical historiography, sociology, etc. The application of this conception of rationality in the empirical sciences is based on the assumption, that human behaviour is sufficiently rational for social phenomena to be explicable. The fascination which it has exerted is, however, due to the way it has been made precise through the formalisation of first Ramsey (1931) and later Savage (1954).

The plausible idea that rational agents choose their actions to optimize their aims, from now on called the *rational choice paradigm*, is attractive as an instrument of analysis, and also because of the unsatisfactory explanatory power of macro-sociological theories like historical materialism and functionalism.[1] Although the methodological individualism of the social sciences is logically and systematically independent of the rational choice paradigm as an explanatory model, there is a close historical connection between the two. The political use of the superficial contrast between 'critical' macro-sociological theories of historical-materialist extraction and 'right-wing' or market-oriented rational-choice conceptions has long been obsolete. In particular Jon Elster (1979, 1983) has ensured that ideological controversies no longer mar the

[1] See Wiesenthal in Elster (1987).

discussion. Without a doubt, the rational choice-paradigm is booming.

This paradigm has its roots in theoretical economy and statistics, roots which it has not disowned to this day, and which lead some authors (e.g. G. Becker) to revert to an individualist and egoistic anthropology, which guided economic theory up to the nineteen-thirties. According to this egoistic anthropology, rational agents do not optimize arbitrary personal aims, but their personal well-being or 'utility'. However, from the perspective of the logical empiricism which flourished in this century, a concept like the 'utility' of a person appeared problematic: it did not satisfy the stringent criteria of significance and therefore appeared 'metaphysical'. Thus the elimination of the classical concept of utility and its replacement by modern utility theory has often been interpreted as a mere conceptual clarification. However, on closer inspection it turns out that this perspective is misleading. The transition to modern utility-theory initiated by Ramsey (1931) and improved by Savage (1954) is at the same time a transition from a conception of practical rationality as instrumental and egoist to a coherence-theory of practical rationality. The endeavour to leave micro-economic theory untouched is partly responsible for concealing this fact. The coherence-axioms of modern economic theory not only dispense with *egoistical* assumptions (rational agents optimize their personal interests), but are also independent of an *instrumental* conception of practical rationality.

Let X be a set of alternatives which are ordered hierarchically by the preferences of a rational agent. In that case the coherence theory requires that this sequence forms an *ordering relation* R on X, i.e. the (weak) preferences of the agent are reflexive, complete and transitive. While *reflexivity* is trivial for (weak) preferences – each alternative is at least as desirable as itself – completeness and transitivity often fail to obtain, as empirical studies have shown.[2] *Completeness* requires of a rational agent that he has a preference with respect to any alternatives x, y, i.e. he has a (weak) preference for x over y, or (inclusive) for y over x. *Transitivity* demands that if a rational agent has a weak preference for x over y and for y over z, he also has a weak preference for x over z. While the condition of reflexivity can be seen as a meaning-postulate of the concept 'preference', it seems plausible to view the conditions of completeness and transivity as genuine postulates of rationality. Accordingly,

[2] See Tversky/Kahneman (1984).

it is not conceptually excluded that a person may have intransitive or incomplete preferences.

However, the aforementioned three ordinal conditions do not suffice to pass from the *qualitative* concept of preference to the *quantitative* concept of utility. To do this one must extend the set of alternatives X probabilistically to X*. The preferences of the agent no longer relate exclusively to the alternatives from X, but also to arbitrary probability distributions (lotteries) over X. To begin with, the same conditions as before apply to the set X*: the preferences must be reflexive, complete and transitive. In addition, four axioms of coherence must be postulated, as follows.

Reduction: a rational agent is indifferent as between two probability distributions over X, if one can be transformed into the other according to the probability calculus.
Substitution: if a rational agent is indifferent between a probability distribution x* and a specific alternative x from X, then x* and x can be substituted in any context without altering the preferences.
Continuity: if x_a is the best and x_z the worst alternative in X, then for each alternative x from X there is a probability p for x_a, such that the rational agent is indifferent between the probability distribution [p x_a & (1-p) x_z] and x.
Monotony: of two probability distributions between x and y, the rational agent prefers the one which predicts a higher probability for the better alternative.

None of these axioms refers to specific substantive motives of the agent or makes assumptions about what is of subjective value to the person concerned. The plausibility of these axioms is independent not only of all substantive aims (e.g. of an economic or egoistic kind), but also of whether the rational actions optimize the subjective aims of the agent. There seems to be no logical connection between the instrumentalist egoist theory from which we started and the coherence theory of practical rationality.

It can now be demonstrated that a preference-relation R which satisfies the above axioms of coherence can be represented by a real valued function. There is such a function u over X* which is uniquely determined (i.e. can only vary within the range of linear transformation) and for which the following holds: $\forall x,y \varepsilon X^*$: $u(x) \geqslant u(y) \longleftrightarrow xRy$.[3] This so-called *utility-theorem* is primarily nothing but a harmless *representation theorem* which transforms the

[3] See Luce/Raiffa (1957), ch. 2.

qualitative notion of preference into a quantitative notion, a real-valued function u. The possibility of representing preferences through such a function is a necessary and sufficient condition for a given preference relation R satisfying the axioms of coherence. Initially there is no reason to suspect that this function u is connected with the personal interests of the agent, as traditional economic theory assumed, or, more generally, with his subjective aims (consequentialist interpretation).

We have left completely open what the preferences of the agent are or how they can be determined empirically. The so-called *revealed preference*-model of rational choice theory identifies actions with the manifest preference for specific elements of X* in view of the given options. If X is interpreted as a set of states of the world or possible worlds, it is plausible to interpret u over X as representing the subjective evaluation of world-states and u over X*|X as the expected value, for it can be shown that u is linear as concerns the probability distributions, i.e. the following holds: $U(x^*) = p\ u(x) + (1\text{-}p)\ u(x')$, if $x^* = [p\ x\ \&\ (1\text{-}p)\ x']$.

Since the representation theorem can be logically deduced from the axioms which have been stated informally above, anybody who considers the axioms of coherence as an adequate condition for the preferences of a rational agent must, on this interpretation, also recognize that the decisions of a rational agent maximize the utility function u over X (or X*).

The consequentialist interpretation of the utility-theorem states that a rational agent evaluates states (this evaluation need not be explicit, it can simply be expressed by the decisions of the agent), and chooses its actions such that the expected value of the (chosen) action which is constituted by the probability distribution of the evaluation of states of the world is maximized. If one makes the additional assumption, that this expected value-maximization is the *motive* of a rational agent, the transition from a coherence theory of practical rationality to a consequentialist theory of optimization is complete. The original theory, according to which a person is rational if his preferences are coherent, has been transformed into a theory according to which an agent is rational, if he maximizes the expected value of subjective evaluations of states of the world. That this transition is not logically compelling can be shown by the following consideration.

Imagine a Kantian agent whose preferences (at least insofar as they concern moral issues) are guided by the moral law (the Categorical Imperative). Imagine further that you are dealing with a 'rational' Kantian who chooses an action which violates the Categorical Imperative in cases in which abiding by the Imperative would have catastrophic consequences. For our argument it suffices that the Categorical Imperative occasionally requires actions which do not maximize the expected value with respect to the agent's subjective evaluation. The Kantian agent will occasionally opt for actions the consequences of which are not optimal (even considering moral expansions), or in the probabilistic case do not maximize the expected value. Consider for example a situation in which a single murder could prevent two murders. A world in which only a single murder takes place is – *ceteris paribus* – better than one in which two murders take place. A Kantian will nevertheless refrain from committing the single murder, and thereby opt for an action the consequences of which are not optimal. Could the preferences of a Kantian agent nevertheless fulfil the conditions of the coherence theory of practical rationality?

If X is the set of possible worlds and the preferences of actions are identified with preferences of the probability distributions assigned to these actions then the decisions of the Kantian agent do not conform to the conditions of coherence. This can be illustrated as follows: Two actions h and h′ can have the same probability distribution of states of the world, and yet h may be permitted and h′ prohibited by the Categorical Imperative. For the place of an action within the preference relation of a Kantian agent is precisely not determined exclusively by the probability distributions of consequences or states of the world induced by the action, but by another criterion, the Categorical Imperative.

This incompatibility of coherence-theory and Kantian rationality is removed, however, if one reinterprets the set of alternatives. Since for the coherence theory of practical rationality the coherence of preferences is fundamental (which allows simultaneous assignments of degrees of belief and desirability), it is plausible to test in the first instance the coherence of the preferences of the Kantian agent. So the preferences to be ordered are indeed options for action. And the preferences of the Kantian agent over such options still ideally satisfy the coherence conditions of practical rationality. A first step is to identify X with a set of options and X* with a set of lotteries over these options: depending on the circumstances, by doing h I might perform either an action h′ or h″, etc. The preferences of a Kantian agent over a set of alternatives thus interpreted are coherent in terms of consequentialist amelioration, even without reverting to step functions or transfinite values, namely through a simultaneous attribution of the necessary belief and desirability-functions.[4] On the other hand, in this case the transition from coherence to optimization-model is precluded. The two decisive steps of that transition cannot be realized: firstly the actions cannot be correlated in the standard fashion with the probability distributions over possible worlds (or possible states of the world); secondly u, the quantitative representation of the preferences cannot be interpreted as representing the motives for action. The attribution of quantitative values to the elements of X* does not represent the subjective desirability of possible worlds in the ordinary sense.

I have been talking about 'possible worlds' in the 'ordinary sense', since it is of course possible to attribute probability distributions over possible worlds to the actions in a 'formal way', insofar as the descriptions of these states contain features like 'the agent's action satisfying the Categorical Imperative'. However, with this modification the coherence model can no longer be transformed into the consequentialist optimization model, for in that case it is not states which are optimized; rather, certain features of the action themselves acquire an intrinsic value.

Expanding the application of the rational choice model from the original egoistic paradigm of subjectivist consequentialism to the coherence theory involves a fundamental revision of the conception of practical rationality. The consequentialist and coherentist conceptions are logically independent, even if the writings of economists erroneously cite the utility-theorem as proof of their equivalence.

2. *Interactions*

Rational choice theory not only deals with situations of risk and uncertainty, but also with situations of interaction. Since this branch of rational choice theory emerged historically from the mathematical analysis of games, it has up to the present day been labelled, misleadingly, as 'game theory'.[5] Game theory and decision theory in the narrow sense investigate individual rational decisions. From this narrow perspective the rational choice is determined relative to a given value function of consequences and the subjective probability function of the circumstances. Of what

[4] This is discussed in greater detail in Nida-Rümelin (1993), §51.

[5] An authoritative introduction is Luce/Raiffa (1957); see also Holler/Illing (1991).

III. Why Consequentialism Fails[1]

I. Introduction

The paradigm of consequentialism is an ethical theory: utilitarianism. The traditional critique of (act-)utilitarianism confronts some practical implications of this theory with our moral intuitions. I think that this traditional critique is indeed quite successful. It is successful because normative theories cannot be justified *more geometrico*: A good normative theory can be developed out of one principle alone, but the justification of this principle is the successful role of this principle to make our normative judgements coherent. If some principle is unable to do this, then it fails, even if it may have appealing qualities in terms of explicitness, simplicity and universality. Even if the traditional critique of utilitarianism and – more generally – ethical consequentialism is successful, it has one major weakness: it overlooks the fact that ethical consequentialism can be understood as a specification of consequentialism (with regard to actions) in general. To put it in more explicit terms: If there were good reasons to adopt a consequentialist theory of rational action, ethical consequentialism – even if not necessarily in the form of classical utilitarianism – is a natural consequence. The specific strengths of ethical consequentialism are based on the forcefulness of a consequentialist theory of action.

In the critical literature on contemporary utilitarianism this close linkage between the theory of rational action and utilitarianism has mostly been neglected. The reason is partly that most theorists criticizing utilitarianism are either deontologists, who quite seldom take rational-choice theories to be adequate, or (anti-normative) contextualists, rejecting the idea of moral action as rational action. My personal point of view can instead be characterized by the following assumptions:

1) Moral agency constitutes one kind of rational agency.
2) An action which has best reasons in its favour, cannot be irrational.
3) There are moral actions which are not ‘rational’ in the sense of the consequentialist theory of practical rationality.
4) An adequate theory of rational action is not consequentialist.

[1] The critique of consequentialism presented in this paper is developed in much greater detail in Nida-Rümelin (1993). The formal parts are presented in Kern/Nida-Rümelin (1994).

This point of view differs from the deontologist's critique insofar as it does not accept the idea that moral action is based on a specific form of rationality which has nothing to do with ordinary (instrumental or pragmatic) rationality. This deontologist separation of instrumental and moral rationality has its roots in Kant's distinction between pragmatic and moral imperatives.[2] Jürgen Habermas' communicative action vs. strategic rationality is a contemporary version of this Kantian dichotomy.[3] I am convinced that contrary to this dichotomy there is a unity of practical reason embracing moral and extra-moral reasons from a complex variety of different types of reasons partly based on principles, partly on obligations, partly on duties, partly on self-interest, partly on institutions etc., respectively, but I will not have the space here to delineate the constitutive types of good reasons for rational agency in general. Nevertheless, I hope that the argument is sufficient to show that consequentialism as a general theory of rational action fails and that this does not force us into deontologist dichotomies. The unity of practical reason can be saved without consequentialism.

II. Decision Theory and Consequentialism

Let us begin with the most forceful argument in favour of consequentialism. This argument says that if one accepts some quite minimal requirements of coherence constraining preferences, one is forced by deductive logical means to accept consequentialism. These requirements are well-known as the conditions of the utility-theorem. Let us call a preference-relation which meets these requirements 'Ramsey-coherent'[4]. Ramsey-coherent preferences are reflexive (as weak preferences: 'at least as good as'), they are complete, i. e. there is a (weak) preference between any pair of alternatives, and they are transitive, i. e. if x is preferred to y and y is preferred to z, then x is preferred to z. These three requirements are not sufficient to pass from coherence to consequentialism. The next step is to extend the assumed set of alternatives (or outcomes) X by including all probability distributions (or lotteries) over X. Let us call this extended set X^*. Ramsey-coherence now requires additionally to the three just mentioned conditions that the preferences are reflexive, complete and transitive not only on X, but also on X^*, and that four additional requirements are met. The first is

[2] See Kant (1785), Second part.

[3] See Habermas (1981).

[4] For a formal treatment of this notion see the appendix to this paper. Cf. also Nida-Rümelin (1993), § 8, and Kern/Nida-Rümelin (1994), Ch. 2.

Rationality: Coherence and Structure[*]

ABSTRACT

In this paper, it is argued that the standard decision theoretic axioms provide core principles action-guiding preferences of rational agents should be required to satisfy. This, however, does not involve a commitment to consequentialism.

It can be rational to be polite. To behave politely requires to follow certain rules. A. Sen has shown that following these rules is incompatible with acting as a consequentialist. Sen proposes, therefore, to give up some principles of decision theory. I argue instead that a 'comprehensive' description allows for both: being polite *and* acting in accordance with the principles of decision theory. This is possible because these principles should be interpreted as an expression of coherentism, not of consequentialism.

1. INTRODUCTION

The ideal rational person has coherent preferences, i.e., preferences satisfying the decision-theoretic axioms of rational choice. A person with, say, intransitive preferences cannot be perfectly rational. A similar statement seems to be true for every single axiom out of the set of requirements constituting the theory of expected utility maximization. This theory is a weak theory of rationality. It is unable to discriminate between good and bad reasons for acting. It only qualifies revealed preferences as coherent or incoherent.

Revealed preferences are the result of practical reasoning. In the last resort, rationality is constituted by actions guided by good reasons. Therefore, we do not know whether a person is rational if we know that the person maximizes expected utility. If a person does not maximize expected utility we know that she has incoherent preferences and is irrational.

[*]I am grateful for having had the opportunity to discuss my arguments with John Broome, Ned McClennen, Michael Slote, Wolfgang Spohn and others during the conference. I thank Martin Rechenauer and Thomas Schmidt for reading this paper and proposing valuable amendments.

Let us call this view the *compatibility assumption*: Whatever are the reasons which guide a person's actions, action-guiding rational preferences fulfill the standard axioms of rational choice if the person is to be considered rational. Rational agency must be compatible with a description under which the person maximizes expected utility.

The intuitive underpinning of the compatibility assumption is not consequentialist but strictly coherentist. It is not based on the idea that rational action means to optimize consequences, because there are many types of good reasons for action which cannot be reconstructed within a consequentialist theory of rationality.[1] Certainly, this is only a prima facie argument against consequentialism. It does not prove that consequentialism is wrong. Since there are good prima facie reasons against consequentialism, consequentialism could be taken to be true only as a result of some convincing theory. In this paper, I do not exclude the possibility of such a convincing theory. What I take to be granted is that there are prima facie arguments against consequentialism such that consequentialism cannot be the starting point of the argument.

Two examples might suffice for rendering the assumption plausible that many of our good reasons for action – moral and non-moral ones[2] – are non-consequential.

First Example. The fact that Tom has promised (at time *t*) to come can be a good reason for Tom to come at some later time *t'*. It is not necessary to spell out the conditions under which a given promise in fact *is* a good reason to keep it in order to see that these conditions cannot be confined to considerations regarding outcomes alone. If Tom keeps his promises only in case he expects the outcome to be optimal, he has not really understood what it means to give (and to keep) a promise. If Suzan knows that Tom keeps his promises only in case he expects the outcomes to be optimal she will not trust him. In this case, promise-giving would lose its power to coordinate intentions and actions. If we assume that forming the intention to keep the promise is part of the act of (genuine) promise-giving – as John Austin assumed in his speech-act analysis[3] – then a consequentialist intention would even be *incompatible* with giving a promise. This does not exclude that certain or probable consequences of keeping the promise vs. not keeping it are relevant for determining whether Tom has a good reason to come. It does exclude, however, that the outcomes of keeping the promise vs. not keeping it are the only aspect relevant for determining whether Tom has a good reason to come.[4]

Second Example. Tom might have the intention to cooperate in a specific situation with Suzie, expecting that Suzie has the same intention.[5] Let us assume that cooperation can be defined with regard to a matrix of outcomes which shows the structure of a Prisoner's Dilemma. Tom might intend to cooperate although he is aware of the outcome structure of the game he plays. It is not necessary that Tom attributes some kind of intrinsic value to the act of cooperation itself. He might deliberately choose a dominated strategy in cooperating. In this case Tom would not maximize the expected value of outcomes. Nevertheless, Tom could have coherent preferences. The intention to cooperate should be compatible with having coherent preferences. According to the definition above a coherent preference relation satisfies the decision-theoretic axioms. It should, therefore, be conceptually possible that Tom (in cooperating) maximizes expected utility, although he chooses (in cooperating) the dominated action regarding outcomes. This might seem strange at first sight, but it leads to the core of an adequate analysis of rational action which drives a wedge between consequentialism and coherentism.

The observation that there are good reasons for action which are not consequentialist, i.e., which recommend actions which do not maximize the expected values of their outcomes, can lead to quite different reactions:

(1) One might assume that those prima facie reasons which cannot be integrated are no genuine good reasons (this account could be called 'consequentialist'). The problem with this account is that some of our most central types of reasons would have to be excommunicated.

(2) One might confine the range of application of rational choice and exclude moral and other types of good reasons for action (this could be called the 'narrow account of rationality'). Rational choice theory would then not be an all-embracing theory of practical rationality any more.

(3) One might give up some of the axioms constitutive for standard rational choice theory as it is done e.g., by McClennen (1990) in his theory of resolute choice (this account could be called 'revisionist').

(4) One might redesign the conceptual framework, i.e., reinterpret the basic concepts of rational choice such that non-consequentialist reasons for action can be integrated (I call this the account of 'structural rationality' for reasons which I hope to become clearer later on).

The aim of this paper is to provide some arguments lending support to the fourth option.

2. WEAK COHERENTISM

I shall call a preference relation 'weakly coherent' or 'Ramsey-coherent' if it satisfies the standard decision-theoretic axioms. This is a 'weak' requirement since although the rationality of an action depends on a complex framework of conative and epistemic attitudes, this framework is not to be considered when testing whether a preference relation satisfies the decision-theoretic axioms. An action is said to be rational *in the full sense* if, additionally, the action and the framework are coherent.

Preferences which do not satisfy the standard decision-theoretic axioms are not coherent. However, preferences which do satisfy the standard decision-theoretic axioms can still be incoherent in relation to the broader frame of conative and epistemic attitudes. Preferences which satisfy the standard decision-theoretic axioms are weakly coherent. Taken together, the standard decision-theoretic axioms are necessary and sufficient criteria for weak coherence, but they are only necessary criteria for coherence in the broader sense.

We can think of many different types of conative attitudes (e.g., desires, wishes, hopes, intentions, preferences) and of many different theoretical possibilities for systematizing this plurality. It seems, however, that, independently from the type of the respective conative attitudes and their systematization, there is an initial plausibility of (weak) coherentism or Ramsey-coherentism, because (and insofar as) conative attitudes lead to singular actions. There might occur irresolvable conflicts between different desires, hopes and wishes, there might even be genuine moral dilemmas, but in the end we have to integrate our conative attitudes such that we are able to act. If, for example, our action-guiding (overall) preferences were incomplete, then in some situations we could not act, or, to put this in a different way: our actions would lose their basis in our conative attitudes. The violation of each singular axiom would have the same devastating result for the agency of individuals trying to decide rationally. Preferences – however they may be constituted – have to be Ramsey-coherent if they can be thought to guide rational action.

Irrationality of action-guiding preferences has two main sources, depending on

(1) *how* they are based on conative (and epistemic) attitudes and

(2) on *which* conative (and epistemic) attitudes they are based.

Even if the action-guiding preferences cannot be criticized for being inadequately based on conative attitudes or for being determined by inadequate conative attitudes (because the determining conative attitudes are inadequate), they can still be criticized if they are (weakly) incoherent.[6]

Since Ramsey-coherent preferences can be represented by a real-valued utility function (determined up to positive linear transformation), quantitative representability (in this sense) is a minimal requirement for ideal rationality. This explains the use of the term 'compatibility assumption' which I have introduced above: the plurality of our reasons for actions has to be made coherent such that the action-guiding overall preferences conform with the axioms of rational choice. If they conform we do not know whether these preferences are rational, but if they do not conform we know that at least one of these preferences is irrational. Compatibility is a relation holding (or not holding) between reasons for actions on the one hand and coherence of action-guiding preferences on the other. If our reasons for actions are such that they are incompatible with Ramsey-coherent action-guiding preferences, they have to be modified, whatever their content may be.[7] It cannot, e.g., be rational to have action-guiding (overall) preferences which are not transitive. No type of reason for action can be held responsible for violating reflexivity, completeness, transitivity, continuity, reduction, substitution or monotonicity.[8]

Some familiar arguments in favour of the rational choice axioms like the money pump argument or the dutch book argument might inhibit a proper understanding of my argument. I postulate an initial plausibility of the axioms of rational choice as minimal coherence requirements of action-guiding preferences, whereas the traditional defenders of the axioms of rational choice use consequentialist arguments. However, since many of our good reasons for actions are not consequentialist we should not rely on consequentialist, but on coherentist intuitions. Our coherentist intuitions do favour the compatibility assumption. There is a prima facie justification of the rational choice axioms which does not depend on consequentialism.

Although standard arguments in favour of the axioms of rational choice are based on consequentialist intuitions, there is a kind of critique of the axioms of rational choice which also is consequentialist in spirit. If an agent, who maximizes expected utility, is in the end worse off than an agent who does not maximize, conseqentialism seems to require to disregard the axioms of rational choice. To avoid confusions, it is helpful to distinguish terminologically between *general consequentialism* as the view that consequences and their value determine exclusively what is rational and the more specific view that an individual action is rational if and only if its consequences are optimal – let us call this latter view *agency consequentialism.*

At first sight, it seems that this distinction does not make much sense. Agency consequentialism obviously is a specialization of general consequentialism. It seems that a general consequentialist must be agency consequentialist, too. The reversal, though, does not hold. If one adopts the plausible stance that only concrete actions of natural individuals can bear the predicate 'rational', the consequentialist account had to be confined to agency consequentialism anyway. Every different usage would on this account be at its best acceptable as merely metaphorical. However, even if there are other entities to which the predicate 'rational' can be applied (e.g., plans of life), the rationality of these entities is not implied by the rationality of singular actions.[9]

The distinction between agency consequentialism and general consequentialism is useful to understand self-defeating arguments against general consequentialism. Tom, who judges the course of his life up to now using the time integral of some value function which – say – attributes to his mental state at every point of time a real number representing its subjective quality, realizes that he would have done much better if he had refrained from choosing every singular action such that it optimized its causal (including probabilistic) consequences. If he is right – and I suppose he is –, consequentialism applied to whole courses of life would be incompatible with agency consequentialism: general consequentialism would be self-defeating.

All forms of consequentialism assume that there is no intrinsic value of rules or types of actions and that in the last resort it is utility or preference fulfillment which renders an action rational or not. In this weaker sense one can be consequentialist without being in the stricter sense consequentialist, i.e., agency-consequentialist. A person, acting rationally in the sense of agency-consequentialism, can be overall irrational regarding overall individual preference fulfillment. This difference is used, e.g., by Edward McClennen in his theory of resolute choice which turns general consequentialist arguments against the specific form of agency-consequentialism. I agree with McClennen in maintaining that practical rationality is not the result of pointwise ('myopic') maximization: agency consequentialism is no convincing general theory of practical rationality. But (as opposed to McClennen) I am convinced that this does not force us to give up the theoretical core of decision theory, if only we rely on a strictly coherentist interpretation (or foundation) of its axioms.

The observation that there are rational actions which do not maximize the expected value of their outcomes leads McClennen to give up axioms constitutive for standard rational choice theory. My proposal, against this move of McClennen, is to redesign the conceptual framework, i.e., to reinterpret the basic concepts of rational choice such that non-consequentialist reasons for action can be integrated (this is the point of the *compatibility assumption*). The theory of resolute choice favours the third reaction of the different options mentioned above, whereas I think we should opt for the fourth alternative.

The justification of the compatibility assumption consists of two elements. The first is to stress its initial plausibility and the second is to show that arguments against it are not convincing. In a recent illuminating article Sen[10] has developed arguments which, if they were stringent, would undermine the initial plausibility of the compatibility assumption. In order to refute them as arguments against the compatibility assumption I delineate an alternative interpretation of the rational choice framework. I hope that in doing so it will become transparent how arguments of this type in general are to be answered.[11]

3. COMPREHENSIVE DESCRIPTION

Let us sketch the idea of a 'comprehensive' description of decisions using one of Sen's examples: you arrive at a garden party and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you to that chair. However, if the matter is left to your own choice, you may refuse to rush to the chair. You select a 'less prefered' one. The reason for this behaviour is that you want to be polite and that you take it as a politeness-constituting-rule not to take the most comfortable chair (if there is only one available) in such situations. It is not the case that you merely want to give the

Why Rational Deontological Action Optimizes Subjective Value

I. Introduction

Does rational deontological action optimize subjective value? Can one be a deontologist and at the same time adhere to decision theory as an all-embracing theory of practical rationality? I think the answer to both of these questions is *yes*. The reasons for it are given in this article.

There are two basic intuitions that frame the bigger part of practical philosophy and which seem to be incompatible. One intuition is *teleological* or, more specifically, *consequentialist* according to which rational action optimizes its consequences. The other intuition is *deontological* or *rule-oriented* according to which rational action is guided by certain rules. I am a deontologist, I think that consequentialism is an inadequate theory of ethics and rationality alike, but at the same time I am convinced that rational action maximizes subjective value.

The reader will probably think that the following two assumptions cannot be true simultaneously

(1) Consequentialism as a theory of *rationality* is false
(2) Rational action maximizes subjective value

Respectively

(1') Consequentialism as a theory of *morality* is false
(2') Moral action maximizes subjective value.

In this article, I try to show that this alleged incompatibility does not exist, and that the teleological intuition can be upheld not in the original *consequentialist*, but in a weaker *coherentist* form.

If one takes moral actions to be rational actions, there is a close link between ethics and the theory of practical rationality. The best developed theory of rationality, however, is decision theory (including game theory), and decision theory is a consequentialist theory of rationality, or so it seems. Most ethical deontologists share the common philosophical belief that decision theory is about prudence (or strategic action), whereas morality establishes constraints on prudential optimizing. Deontologists, therefore, tend to reject the idea of decision theory being the core of a general theory of practical rationality and the attempt to integrate moral actions into its conceptual framework.

If the decision theoretic framework cannot be dismissed in order to define rational action, it seems that ethical consequentialism has strong arguments in its favour. Ethical axiology would be about moral values and ethical rationality about how to optimize moral values. Utilitarianism and decision theory which are both interlinked in the history of ideas would then still be dependent on each other from a systematic perspective, too. I will instead argue that rational deontological action is compatible with standard decision theoretic axioms. *Rational deontological action optimizes subjective value.*

II. The Wedge between Choice and Preference

Decision theory works with the assumption that the rational person reveals two basic propositional attitudes in acting: *subjective preferences* and *subjective probabilities*. It attributes two real-valued functions to the rational actor: the function of *subjective probability* and the function of *subjective value*. This attribution is based on the comparative concepts *preferring* and *expecting* – person *i prefers p to q* and *i takes p to be more probable than q*. It is not possible to attribute one of these two functions independently from the other, and insofar the attribution of the two functions is interlinked.

Is decision theoretic rationality consequentialist? The standard applications of decision theory are indeed consequentialist insofar as *subjective value* is defined as *subjective value of consequences*, and insofar *rationality* is defined as *maximizing expected subjective value of consequences,* or in short as: *optimizing consequences.* Conceptually, this consequentialism is made explicit in Savage's model, for example, not however in Jeffrey's holistic account[16]. If we think that moral actions should be rational actions, morality should be defined as maximizing *expected subjective value of a specific kind.* It would be the task of *ethical theory* to develop criteria under which a subjective value function seems morally acceptable.

[16] cf. Leonard J. Savage, *The Foundations of Statistics*, New York 1954 and Richard C. Jeffrey, *The Logic of Decision*, New York 1965.

What then, if you are a deontologist; that is, if you think that deontological criteria cannot be dismissed in ethical theory? Under the assumptions that the decision theoretic framework can not be dismissed either and that decision theory commits one to a consequentialist view on rational agency, the application of deontological criteria seems to suggest the following options: either the idea that moral action is rational has to be abandoned, or two kinds of rationality have to be adopted: one for moral and one for extra-moral action – in the latter case, decision theory would have to be taken as dealing with extra-moral actions. The first option appears unattractive: “You should do x, but it is irrational to do x!” is not a convincing moral imperative. The second option would imply that decision theory is not a general theory of rational action. In this article, I shall not be concerned with the first option but focus on the second one: Are we to accept that there are two quite different kinds of rationality? Is it true that deontological action does not fit into the decision theoretic frame?

Decision theory does not begin with postulating probability and value functions. Rather decision theory begins with describing *basic properties* of preference relations. Some of these properties can be interpreted as analytically tied to the notion of preference, others are synthetic (and normative insofar as they, taken together, constitute rationality or are at least necessary conditions of rationality.). There is no clear-cut borderline. On the one hand, it seems that *reflexivity* of preferences is an analytic property, whereas *transitivity* of preferences rather appears to be a synthetic (and normative) requirement. On the other hand: if a person has preferences regarding three well-determined, well-described and well-known alternatives, which are intransitive, one may well ask the person, whether she *really* has those preferences or not. If we think of preferences as being revealed in action and in verbal expression, violations of transitivity or other properties of rational preference relations may result in uncertainty concerning the reliability of the attribution of these preferences in the first place. This phenomenon indicates that there is no precise demarcation line between analytic and synthetic properties of rational preference relations.

Is there any reason to assume that deontological action violates one of these properties of rational preferences (whatever their status may be), e.g. transitivity? This formulation is an elliptic version of: is there any reason to assume that actions which are performed for deontological reasons reveal preferences that violate the transitivity condition? It makes things easier if we circumvent the notion of *action* by asking: is there any reason to assume that *preferences* that are based on deontological rules violate the transitivity condition? Let us illustrate the point by referring to the deontological rule *THOU SHALT NOT STEAL*. The conflict between deontology and consequentialism is apparent: Even if stealing in some cases might have overall good consequences, you are not allowed to steal. This remains true also if we concede that in cases when not stealing leads to extremely bad consequences, e.g. starving, stealing might be justified.

Now think of the following case. There are three possible states of the world *x,y,z*. We assume that these three states of the world are certain[17] consequences of three alternative feasible actions *u,v,w* respectively. The person in question prefers *x* to *y* and *y* to *z*. The subjective value function represents these preferences adequately. Since probability is not involved, maximization of the expected subjective value would result in choosing action *u*. But let us assume that *u* is an act of stealing, whereas *v* and *w* are no acts of stealing (and are not forbidden by other deontological rules). In this case it seems that *v* should be the preferred action. The situation is such that it is not possible to optimize consequences and at the same time conform with the deontological rule in question.

The person has *transitive* preferences regarding states of the world as well as regarding actions. *Reflexivity* of weak preferences is given trivially (in this case there are no weak preferences involved) and *Completeness* is a reasonable postulate for a deontologist. Since the rational deontologist (like the consequentialist) cannot refuse to act, she reveals her preferences regarding any feasible alternative. There is no reason to assume that a rational person with deontological motivations sometimes gets caught in holes of inactivity. The *Completeness Axiom* should be extended to preferences motivated by deontological reasons.

Standard decision theory identifies actions with prospects, i.e. probability distributions over the sets of their consequences. Therefore it identifies *u* with *x*, *v* with *y* and *w* with *z*. Within this conceptual framework there is no logical space for deontological rationality. It is conceptually excluded that a rational person has preferences as we assumed above. It seems to me that *this conceptual exclusion of deontological rationality is unacceptable*. It makes perfect sense that a person, when asked which state of the world she prefers, answers that she prefers *u* to *v* and *v* to *w* while she respects the Seventh Commandment and therefore prefers action *y* to action *x*. There is no reason to think that everybody who respects this commandment is irrational. A rational person can prefer the state of the world *x* to *y* and *y* to *z,* and at the same time prefer the action *v* to the action *u*, because *u* would violate the Seventh Commandment and *v* wouldn't.

[17] "Certain" in the sense that the subjective probability is 1.