

Julian Nida-Rümelin

Why Consequentialism Fails¹

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,

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).

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.

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.

² See Kant (1785), Second part.

³ See Habermas (1981).

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 that an agent is indifferent between two probability distributions over X , if one can be transformed into the other according to the rules of the probability calculus. Secondly, if the rational agent is indifferent between a probability distribution x^* out of X^* and a specific alternative x out of X , then x^* and x can be substituted in any context without changing the preference relation. Thirdly: if one element, let us call it " x_b ", is best and one alternative, call it " x_w ", is worst in X , then for each alternative x out of X there is a probability p for x_b such that the probability distribution $(px_b \& (1-p)x_w)$ and x are regarded as indifferent. The fourth requirement is that if there are two different probability distributions over two alternatives x and y out of X , then the probability distribution, which assigns a higher probability to the better alternative, ranks above the other.

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.

If a preference relation is Ramsey-coherent, then it can be represented by a real-valued function over X^* . This function is unique up to positive linear transformation. Representation here means that whenever an alternative x out of X^* is (weakly) preferred to y out of X^* , then the numerical value of x , which is assigned to x by this real-valued function u , is at least as big as the numerical value of y , which is assigned to y by u .

There has been some discussion whether these coherence-conditions are in fact adequate. For the sake of the argument we will assume that they are. If they are adequate, i. e. are indeed necessary conditions for the rationality of preferences, then the utility-theorem says that the rational action is optimizing an utility-function. And if the rational person is optimizing an utility-function, then it seems that this person is a consequentialist (or teleological) agent.⁵ The possible consequences of the respective action are alternatives x out of X , which have an action-dependent probability greater than 0, these consequences have a specific (subjective) value $u(x)$, and the values of these possible consequences of the respective action are weighed by (subjective) probabilities. Beginning with some minimal requirements of preference-coherence, we get a full-blown consequentialist theory of rational action. If this move from coherence to consequentialism were well-founded, we had a strong argument in favour of consequentialism as a general theory of rational action. Still it is conceivable that after looking at these far reaching implications of the coherence-assumptions we have a closer look and in the end dismiss at least one of these.⁶ I do not exclude this possibility, but my argument is much more radical: This move from coherence-conditions to consequentialism is a *non-sequitur*, in other words: Coherentism about preferences does not justify consequentialism about actions.

Think of some ideal deontologist agent.⁷ Such an agent might accept a variety of *prima facie* duties which in simple cases constrain the optimization of his personal

⁵ See Vallentyne (1987,1988) and Broome (1991).

⁶ For this point cf. McClennen (1990).

⁷ For the following argument, see also Nida-Rümelin (1993), § 51.

interests. In more complex cases, i. e. in cases in which at least two of these *prima facie* duties are in conflict with each other, the agent is forced to apply some weighing procedure or rely on his spontaneous intuitions. It depends from the content of the *prima facie* duties and - in complex cases - from the weighing procedure respectively the moral intuitions, whether this deontologist agent would act consequentialistically rational or not. If we think of David Ross' list of *prima facie* duty, there can be no doubt that a person whose actions are guided by these *prima facie* duties is not acting consequentialistically rational, because it is not the consequences (and their subjective value) alone which are relevant for choosing an action.⁸ For example, if keeping one's promises is one of these *prima facie* duties, the past is intrinsically relevant for deciding. But the past is *per se* irrelevant for consequentialist rationality, since it is only the consequences of the respective actions what counts. If coherentism implied consequentialism, then the deontologist agent's preferences had to violate at least one of the requirements of Ramsey-coherence. But I cannot see any reason why this should be the case. Why should the preferences of a person who is guided by the *prima facie* duty to keep one's promises be incoherent? We can certainly assume that David Ross' ideal moral person would have reflexive, complete and transitive preferences and that - if probabilities came into play - she would fulfill the four additional requirements named above. The same seems to be true for the ideal Kantian agent.

Most consequentialists might argue that moral deontology is incompatible with consequentialist rationality and therefore ill-founded. But this argument is convincing only if we assume more than modern utility theory does. If we already presuppose that the rational agent chooses his decisions exclusively on the basis of his subjective valuation of consequences (which are states of affairs causally - or probabilistically - connected with the respective action) weighing the consequences by the measure of his subjective probabilities, then the incompatibility with the deontologist agent is obvious. Therefore, if consequentialist rationality were implied by Ramsey-coherence, then we had to assume

⁸ Cf. Ross (1930).

that the preferences of the ideal deontologist agent would violate at least one of the coherence-conditions. But since consequentialist rationality is not implied by Ramsey-coherence, the incompatibility of consequentialist rationality and deontology does not imply that the deontologist agent is Ramsey-incoherent.

The incompatibility of consequentialist rationality and acting on deontological reasons can be made more explicit if we take a closer look at the notion of a 'consequence'. Loosely speaking, every property of an action can be interpreted as a possible consequence. If an action fulfills a promise, say, this fact can be understood as one of its consequences. Such a wide notion of 'consequence' would make the idea of consequentialist rationality trivial or empty. The intuition on which the theory of consequentialist rationality is based is that we look at the causal effects which an action has on the history of the world and that we value an action as right or wrong exclusively with regard to these causal effects. This intuition can only be saved if the notion of consequence applied in this context implies that all relevant aspects of consequences can be described using unhistoric predicates. Classical utilitarian consequentialism is a good example. The evaluation of the consequences are exclusively determined by the sum of individual well-being, whereas well-being is to be interpreted in such a way that it characterizes a subjective state of mind. Historical information might be necessary to form rational expectations over consequences, but consequences themselves can be described without using historic predicates. History has a causal influence on consequences, but the value of a given consequence is independent from history. This explains in more abstract terms why deontological agency in many cases cannot be consequentialist. To help a miserable person to feel better is a good reason for action, and there might exist a duty to help the miserable. If we have only this duty in mind, the only relevant information to judge the adequacy of an action would concern consequences, more precisely the consequences regarding the subjective state of the miserable person. But we have good reasons to follow other duties such as being truthful, keeping promises etc., which require more informations than those concerning the consequences of the respective action only.

III. Consequentialism and the Moral Point of View

In the last section I have shown that we have to discriminate sharply between a coherentist theory of rational preference and a consequentialist theory of rational action. For the latter, there is a primacy of the (subjective) valuation of consequences which motivates the (consequentialist) person to choose an action out of her options if and only if this action optimizes the consequential value regarding her subjective probabilities. If moral agency is to be rational, morality can only be constituted by a specification of how consequences are to be evaluated. It is a constitutive trait of consequentialism that if the subjective valuation of consequences is given, then the rational action is well-determined except for cases of indifference. Aspects of universalizability e. g. are irrelevant for consequentialist rationality.

The Kantian theory of practical reason says that whatever your interests are and however you value consequences on the basis of these interests, you are acting irrational if following these interests is not universalizable. The Kantian theory of practical reason includes two steps: first you form the subjective rules (maxims) on which you want to act, second you check the interpersonal compatibility, i. e. universalizability of these maxims. The second step imposes constraints on following your personal interests. Consequentialism on the other hand is a one-step theory of practical rationality, and this is one of the main strengths of this account. You have to know how the person values consequences, and then you can tell the person what is rational for her to do (dependent on her subjective probabilities). Therefore, within a consequentialist conceptual framework the whole burden of moral evaluation rests on the individual value-function over consequences.

Most practical philosophers agree that the moral point of view requires some form of impartiality, but there are many different opinions about the content and the range of morally grounded impartiality. Deontologists accept the variety of different subjective

points of views if these are action-guiding only within some impartial constraints. 'Impartiality by universally acceptable constraints' might be taken as the basic formula of the deontologist theory of practical reason. For the consequentialist the impartiality of the moral point of view must become part of the value-function over consequences which motivates the rational person's actions. The most radical consequentialist inclusion of the moral point of view requires each rational and moral person to adopt the same subjective value-function over consequences. The right action is then defined as optimizing that interpersonal invariant value-function over consequences. The question of how this interpersonally invariant value-function is to be constituted has different answers, ranging from the classical utilitarian theory, which identifies this value-function with the total sum of happiness, to modern theories, identifying this value-function with an aggregate of individual 'personal' preferences, giving each individual preference-relation equal weight.⁹

Let us call a theory of this kind requiring an interpersonally invariant value-function as the only adequate expression of the moral point of view 'strict ethical consequentialism'. The theoretical design of strict ethical consequentialism is simple and elegant: it adheres to consequentialism as the general theory of rational action on the one hand and incorporates the moral point of view by means of withdrawing interpersonal differences regarding the action-guiding evaluation of consequences. The problem with strict ethical consequentialism is that it obviously requires too much. Not only that it is supererogatory because nobody can be expected to choose his actions such that everybody's interests in the world are equally fostered, but - more fundamentally - the idea of a society in which every individual maximizes the same value-function disregards the necessary and valuable complexity based on individual differences. A good society is constituted by individuals with different projects and bindings, different life-plans and virtues, different values, hopes and fears.

⁹ See e.g. Harsanyi (1955).

If strict ethical consequentialism requires too much and the ordinary *homo oeconomicus* model of optimizing personal interests is incompatible with the moral point of view, then it seems reasonable to choose a 'middle road'. The middle road adheres to the consequentialist theory of practical rationality and to the idea that moral agency is rational agency, but it dismisses the interpersonally equilibrizing tendency of strict ethical consequentialism. Samuel Scheffler's agent-centered prerogative is an example. It is still a consequentialist theory insofar as an action is right if and only if it optimizes the value of (expected) consequences, but it allows the agent to give unequal weight to the interests of different persons. It is allowed (not required) that the agent gives more weight to his personal interests than to the interests of others. The moral point of view has adopted a more humane stance. We can think of a continuum beginning with the *homo-oeconomicus* model of self-orientated maximization and ending with strict ethical consequentialism. Far from being a 'rejection of consequentialism', this proposal - even if it might imply many problems in detail - seems to save consequentialism as a general theory of rational action because it incorporates the moral point of view without destroying individual differences.

To summarize: consequentialism as a general theory of rational action is incomplete if it is not able to account for good moral reasons for action. One possible account is the one of strict ethical consequentialism. Since strict ethical consequentialism is incompatible with interpersonal differences, it is inadequate for anthropological reasons. But there seems to be a vast space between the self-oriented *homo oeconomicus* on the one hand and the strict ethical consequentialist on the other to find an adequate reconciliation of consequentialism and the moral point of view.

IV. Consequentialism and Differences

The consequentialist *homo oeconomicus* accepts every kind of interpersonal differences. Within this paradigm these differences are differences of interests, however, they might be laden by moral values. On the theoretical level this account excludes moral

reasons for action. Strict ethical consequentialism on the one hand includes good moral reasons for action in the most demanding form of prescribing one interpersonally invariant value-function (over consequences) for each individual. Weaker forms of (ethical) consequentialism try to include good moral reasons for action in a weaker form, such as allowing every person to give special weight to her own interests, but limiting this 'agent-centered prerogative' in some way or another. Think for example of a weighing-factor which allows to aggregate personal interests only within a range in which no-one's interest counts more than five times as much as the interest of any other person.¹⁰ Proposals of this kind result in different personal value-functions over consequences. They allow for different personal projects and plans of life in general. But when consequentialism tries to include good moral reasons in a general account of rational action without going all the road to strict ethical consequentialism, the failure becomes apparent. If a normative theory allows for interpersonally different valuations of consequences, it has to say how to solve problems of individual rights, collective decisions and cooperation. None of these three problems can be adequately answered within the consequentialist framework. Some well-known decision theoretic results can help to make this argument transparent and short.

1. Individual rights. Consequentialism is bound to collective choices which are pareto-inclusive. Some social state is better than another if at least one person is better off in the first and no person is worse off in the latter. Even the traditional *homo-oeconomicus* version of consequentialism is closely linked to the theory of the ideal market in which individual optimizers realize a social state which is pareto-efficient.¹¹ Versions of consequentialism which are not bound to one single type of good reasons for action which are based on self-interest, i. e. versions of consequentialism which try to include genuine moral reasons for individual action, *a fortiori* cannot accept pareto-inefficient social states as the result of ideal rational behaviour. Whatever hybrid versions of consequentialism offer as a compromise between the personal point of view and impartiality, the resulting

¹⁰ Cf. e.g. Scheffler (1982), the contributions in Scheffler (1988), and Margolis (1982).

¹¹ See the idea of the market as a morally free zone in Gauthier (1986), ch. IV.

normative collective choice rule has to be pareto-inclusive. But whenever individual differences are allowed for, it seems reasonable to attribute individual rights, too. Individual rights protect the personal sphere, they are part of those conditions which are needed to live a personal life. But as Amartya Sen has proven in 1970, it is not possible to have both conditions fulfilled by a reasonable collective choice rule. The deontologists' insistence on constraints for optimization seems to make sense again. It is not possible to combine consequentialism, interpersonal differences, individual rights and efficiency.¹²

2. *Collective Action.* Collective action requires some commonly accepted procedure of aggregation. Ideal rational consequentialists optimize expected value of consequences in every single decision. To participate in a collective action and the form how to participate are decisions like any other. A well-known theorem of collective choice, independently proven by Allan Gibbard and Mark Satterthwaite¹³, has far-reaching implications for hybrid versions of consequentialism, since we cannot assume that some kind of simple interdependence of preferences, as it is proposed by hybrid consequentialism, would suffice to make a reasonable collective choice rule strategy-proof. But if reasonable collective choice rules are not strategy proof in a collective of hybrid consequentialists, then we face a basic instability in a society of ideal moral and rational consequentialists. Collective action requires agents who under certain circumstances refrain from optimization and stick to the rules of the game.

3. *Cooperation.* Genuine cooperation is a type of action for which the following holds: If each of the (two or more) participants optimized their respective value-function over consequences, each participant would be worse off than if all participants would

¹² For a formal statement of Sen's theorem (commonly known as the 'Liberal Paradox' - the original proof was published as Sen (1970)) see the appendix to this paper. Additional material is contained in Kern/Nida-Rümelin (1994), Ch. 11.

¹³ See Gibbard (1973) and Satterthwaite (1975).

choose the cooperative action.¹⁴ If we characterize cooperation in this way, genuine cooperative action is bound to situations of the prisoners' dilemma type. But if we use the usual characterization of the prisoners' dilemma, cooperative action is rendered irrational, i. e. Ramsey-incoherent. If a two-person prisoners' dilemma is defined by the following preferences: A (row-chooser): $DC > CC > DD > CD$ and B (column-chooser): $CD > CC > DD > DC$, then to decide for action C is obviously incompatible with these preference-relations for A and B respectively if their decisions are independent from each other. Strict ethical consequentialism excludes such combinations of preference relations. It excludes the possibility of a prisoners' dilemma-situation because the action-guiding preferences of ideal strict ethical consequentialists are interpersonally invariant. But the whole rest of the consequentialist spectrum allows for interpersonal differences and cannot exclude PD-situations. If the consequentialist theory of rational action includes moral reasons - other than the standard *homo oeconomicus* model -, it cannot at the same time exclude one of the most central moral reasons, and this is cooperation.

How is it possible to cooperate without having incoherent preferences? The answer is that we have to enrich the description of the situation in which cooperative action is possible. Let us even grant for a moment that the action itself has no intrinsic value whatsoever and that the only relevant properties of the feasible actions concern the consequences of these actions. The only additional assumption which we need for our argument is that the preference structure of the situation is part of the information on which the individual reasons for action rest. Within the orthodox consequentialist framework we would be forced to describe an agent willing to act cooperatively under the (necessary) condition that she can expect the other agent to be willing to cooperate conditionally, too, by the following preference structure: $CC > DC > DD > CD$, which constitutes the assurance game in case the agent B has analogous preferences. These

¹⁴ This is not - as it might seem - a circular definition, because the predicate 'cooperative' in the last sentence has no other function than to identify the type of action. One could substitute 'cooperative' by 'respective'.

(assurance game-) preferences are compatible with the assumption that the individual preferences over consequences have not changed at all.¹⁵ In other words: The PD-structure can survive if we look at the participants' preferences over consequences whereas at the same time they reveal a cooperative attitude by assurance-game-preferences regarding the possible strategy combinations. This cooperative attitude might be motivated exclusively by two bits of information: the own and the other person's evaluation of consequences (which constitute a PD) and the knowledge of the others' conditioned cooperative attitude. If we go one step further and assume an agent who cooperates unconditionally (other-regarding preferences), we can even drop the second bit of information. Obviously we have to drive a wedge between choice and preference in order to reconstruct cooperative action. The cooperative agent does not optimize (individually) her evaluation of consequences.

V. Weighing Reasons

In many everyday situations we have good reasons to respect individual rights, stick to some accepted rules of collective action or to cooperate. In these cases we refrain from optimizing consequences. But still we try to be intra- and interpersonally coherent. Ramsey-coherence is a minimal requirement of being coherent, i. e. living a reasonable life. Living a reasonable life requires weighing reasons. Different types of reasons refer to different properties of actions and their contexts. In some rare cases reasons exclusively refer to consequences. In other cases they refer to obligations which have arisen due to past actions, duties which are bound to social roles, individual rights which protect autonomy and universal principles as constraints. To include these reasons adequately, it is necessary to enrich the conceptual frame of the theory of rational choice beyond

¹⁵ The idea of analyzing metapreference structures in the context of the prisoners' dilemma was introduced by Sen (1977), see also Nida-Rümelin (1991). For a formal definition of the prisoners' dilemma see also the appendix to this paper.

erschienen in: G. Holmström-Hintikka/ R. Tuomela (Hrsg.). Contemporary Action Theory. Vol. II. Kluwer: Dordrecht 1997, S.295-308.

consequentialism. Ramsey-coherence is the result of weighing reasons, not weighing goods. Ramsey-coherence is a core element of the unity of reason.

Appendix

The purpose of this appendix is to state more formally some of the technical concepts and results referred to and used in the arguments of the main body of the paper.

1. Ramsey-coherence

Let $X=\{x_1,\dots,x_n\}$ be a set of outcomes and let $R \subset X \times X$ be a preference relation of some individual over X (where " $(x,y) \in R$ " is to be read as "the individual weakly prefers x over y "). R is called an *ordering* iff the following three conditions are satisfied:

Reflexivity. $\forall x \in X : (x,x) \in R$. - Every outcome is at least as good as itself.

Completeness. $\forall x,y \in X : (x,y) \in R \vee (y,x) \in R$. - Every pair of outcomes can be ordered according to the weak preference relation.

Transitivity. $\forall x,y,z \in X : ((x,y) \in R \wedge (y,z) \in R \rightarrow (x,z) \in R)$. - If an outcome x is weakly preferred over y and y is weakly preferred over z , then x is weakly preferred over z .

Let $p_1, \dots, p_n \in [0,1]$ be nonnegative real numbers smaller or equal than 1 with $p_1 + \dots + p_n = 1$. Then the n -tupel $(p_1/x_1 \ \& \ \dots \ \& \ p_n/x_n)$ is called a *prospect* (hence, a prospect is a probability distribution over outcomes). X^* denotes the set of all prospects and we assume the preference relation R to be extended over X^* . R is called *Ramsey-coherent* iff it is an ordering and, additionally, the following four conditions are met:

Continuity. If for the individual the best element from X is x_b and the worst element is x_w (and the individual strongly prefers x_b over x_w - i.e. is not indifferent between all elements from X), then for each x in X there is a probability $p \in [0,1]$ such that the individual is indifferent between x and the prospect $(p/x_b \ \& \ (1-p)/x_w)$.

Monotonicity. Let $A = (p/x \ \& \ (1-p)/y)$ and $A' = (p'/x \ \& \ (1-p')/y)$ be two prospects. If the individual strongly prefers x over y , then $(A,A') \in R \leftrightarrow p \geq p'$.

Reduction. Prospects can be manipulated according to the probability calculus without changing the individual's preferences between them.

Substitution. If the individual is indifferent between an outcome x and a prospect A , then x is substitutable by A in every prospect A' in which x occurs without changing the individual's preferences.

It can be shown that if the individual preference relation over X^* is an ordering and additionally meets these four constraints, then there is a linear function u (which is - modulo positive linear transformation - uniquely determined) from the set of prospects into the real numbers which cardinally represents the individual's preferences:

$$\forall x,y \in X^*: u(x) \geq u(y) \leftrightarrow x \geq y$$

2. The Liberal Paradox

Let again $X = \{x_1, \dots, x_n\}$ be a set of alternatives and let a set $K = \{1, 2, \dots\}$ be a group of individuals (the "collective"). A *preference structure* is a function g mapping individuals into preference relations: $g: K \rightarrow \text{Pot}(X \times X)$ (where $\text{Pot}(X \times X)$ denotes the power set of X). For each individual $i \in K$, $g(i)$ is the weak preference relation of i over X . Let the set of all preference structures be denoted by G . The strong individual preference relation for individual i is denoted by $\dot{g}(i)$.

A choice function f mapping preference structures (in which all the individual preference relations are orderings) into collective orderings is called a *social decision function*. We define two properties a social decision function may or may not have.

Strong Pareto Principle. $\forall g \in G \forall x, y \in X: [\forall i \in K: (x, y) \in g(i) \wedge \exists j \in K: (x, y) \in \dot{g}(j) \rightarrow (x, y) \in \dot{f}(g)]$. - If every individual regards an alternative x to be at least as good as an alternative y and at least one individual strictly prefers x over y , then x should be collectively strictly preferred over y .

Liberalism. $\forall g \in G \forall i \in K \exists x, y \in X (x \neq y): [(x, y) \in \dot{g}(i) \rightarrow (x, y) \in \dot{f}(g) \wedge ((y, x) \in \dot{g}(i) \rightarrow (y, x) \in \dot{f}(g))]$. - For every individual there are at least two alternatives over which the individual is decisive (when "decisive" means that the preferences of the respective individual are to become the social preferences).

Sen's theorem states that there is no social decision function simultaneously satisfying the Liberalism condition and the Strong Pareto Principle.

3. The Theorem of Gibbard and Satterthwaite

Let X be a set of alternatives, K the collective, G the set of all possible preference structures, G^S the set of preference structures consisting only of strict individual preferences, and g a preference structure.

If $g = \langle g(1), \dots, g(n) \rangle$ and $g' = \langle g'(1), \dots, g'(n) \rangle$, then we define $g|g'(i) := \langle g(1), \dots, g(i-1), g'(i), g(i+1), \dots, g(n) \rangle$. For any individual i let $g^S(i)$ be the set of possible (strict) individual preference relations for i . A function $f: G \rightarrow Pot(X \times X), g \mapsto R = f(g)$ mapping preference structures into preference relations over X is called a *rule of aggregation*. A *choice function* is a function $a_{f(g)}$ mapping subsets of X into choice sets according to the collective preference ordering given by an aggregation rule f applied to a preference structure g . A choice function is said to fulfill the condition of *strategyproofness* iff

$$\neg \exists g \in G^S : \exists i \in K : g^S(i) \in g^S(i) \wedge \exists g' \in G^S : g' \neq g \wedge a_{f(g)}(X) \neq a_{f(g')} \wedge g^S(i) \neq g^S(i)$$

If a choice function is strategyproof, then it is not rational for an individual to feed other preferences into the collective aggregation process than his real ones.

The theorem proven by Gibbard and Satterthwaite states that if a choice function is strategyproof, then there exists an individual the preferences of which are decisive (i.e. a dictator). In other words: there exists no strategyproof choice function for which there is no dictator.

4. The Prisoners' Dilemma

A prisoners' dilemma (of the standard type) is an interaction situation involving two agents (labelled A and B) each facing two possible alternatives, C and D ("C" for "cooperation" and "D" for "defection"):

		B	
		C	D
A	C	CC	CD
	D	DC	DD

The *prisoners' dilemma* (PD) is defined by the preference structure

A: $DC > CC > DD > CD$

B: $CD > CC > DD > DC$.

The *assurance game* (AG) is defined by the preference structure

A: $CC > DC > DD > CD$

B: $CC > CD > DD > DC$.

The *other-regarding game* (OR) is defined by the preference structure

A: $CC > DC > CD > DD$

B: $CC > CD > DC > DD$.

erschienen in: G. Holmström-Hintikka/ R. Tuomela (Hrsg.). Contemporary Action Theory. Vol. II. Kluwer: Dordrecht 1997, S.295-308.

Bibliography

Broome, J.: 1991, *Weighing Goods*, Basil Blackwell, Oxford.

Gauthier, D.: 1986, *Morals by Agreement*, Oxford University Press, Oxford.

Gibbard, A.: 1973, 'Manipulating Voting Schemes: A General Result', *Econometrica* **41**, 587-601.

Habermas, J.: 1981, *Theorie des kommunikativen Handelns*, Suhrkamp, Frankfurt a.M.

Harsanyi, J.C.: 1955, 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economy* **63**, 309-321.

Kant, I.: 1785, *Grundlagen zur Metaphysik der Sitten*, transl. by H.G. Paton: *Groundwork of the Metaphysics of Morals*, New York 1956.

Kern, L./Nida-Rümelin, J.: 1994, *Logik kollektiver Entscheidungen*, Oldenbourg, München/Wien.

Margolis, H.: 1982, *Selfishness, Altruism, and Rationality*, Cambridge University Press, Cambridge 1982.

McClennen, E.F.: 1990, *Rationality and Dynamic Choice*, Cambridge University Press, Cambridge.

Nida-Rümelin, J.: 1991, 'Practical Reason or Metapreferences? An Undogmatic Defense of Kantian Morality', *Theory and Decision* **30**, 133-162.

Nida-Rümelin, J.: 1993, *Kritik des Konsequentialismus*, (paperback ed. 1995), Oldenbourg, München.

Ross, D.: 1930, *The Right and the Good*, Oxford University Press, Oxford.

erschienen in: G. Holmström-Hintikka/ R. Tuomela (Hrsg.). *Contemporary Action Theory*. Vol. II. Kluwer: Dordrecht 1997, S.295-308.

Satterthwaite, M.A.: 1975, 'Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions', *Journal of Economic Theory* **10**, 187-217.

Scheffler, S.: 1982, *The Rejection of Consequentialism*, Clarendon Press, Oxford.

Scheffler, S., ed.: 1988, *Consequentialism and its Critics*, Oxford University Press, Oxford.

Sen, A.K.: 1970, 'The Impossibility of a Paretian Liberal', *Journal of Political Economy* **78**, 152-157.

Sen, A.K.: 1977, 'Choice, Orderings, and Morality', in: S. Körner, ed., *Practical Reason*, Oxford University Press, Oxford, 54-67.

Vallentyne, P.: 1987, 'The Teleological/Deontological Distinction', *The Journal of Value Inquiry* **21**, 21-32.

Vallentyne, P.: 1988, 'Teleology, Consequentialism, and the Past', *The Journal of Value Inquiry* **22**, 89-101.

Prof. Dr. Julian Nida-Rümelin
Georg-August-Universität Göttingen
Philosophisches Seminar
Humboldtallee 19
D-37073 Göttingen
Tel. (0551)39-4722/4721
Fax (0551)39-9607