Moral Desirability and Rational Decision

Christoph Lumer
(University of Siena)

(Appeared in: Ethical Theory and Moral Practice 13 (2010). Pp. 561-584.)

Abstract: Being a formal and general as well as the most widely accepted approach to practical rationality, rational decision theory should be crucial for justifying rational morals. In particular, acting morally should also (nearly always) be rational in decision theoretic terms. After defending this thesis, in the critical part of the paper two strategies to develop morals following this insight are criticized: game theoretical ethics of cooperation and ethical intuitionism. The central structural objections to ethics of cooperation are that they too directly aim at the rationality of moral action and that they to do not encompass moral values or a moral desirability function. The constructive half of the paper takes up these criticisms by developing a two-part strategy to bring rationality and morals in line. The first part is to define 'moral desirability'. This is done, using multi-attribute utility theory, by equating several adequate components of an individual's comprehensive (rational) utility function with the moral desirability function. The second part is to introduce mechanisms, institutions, in particular socially valid moral norms, that provide further motivation for acting in accordance with morals.

Keywords: justification of morals, ethical internalism, rational decision, multi-attribute utility theory, prudentialistic desirability theory, sympathy, respect, moral desirability, moral obligations, socially valid norms; ethics of cooperation (critique of), contract theory (critique of), intuitionism (critique of).

Most secular ethicists think that ethics and moral action are, or should be, rational. Rational decision theory provides the most fundamental, and widely accepted formal criterion of practical rationality. So, rational decision theory should have something to say about the foundations of ethics: about justifying morals, about the rationality of moral action, and in particular about justifying the kind of moral that makes it rational to act morally. In the present paper this view will be briefly defended; then, in the main part, taking this view as the constructive starting point, a particular conception of morals will be developed and defended.

A straightforward way to turn the insight about the relevance of decision theory for the rationality of ethics into practice is game theoretic ethics of cooperation, as it has been advocated e.g. by Gauthier. However this approach – the above statements notwithstanding – faces several structural and material problems (cf. below, section 1), which have contributed to the rather modest reputation of decision theoretic views of ethics. Because of those problems, a different approach, which has two main parts, is proposed in the following text. The first part is to establish moral desirability functions via multi-attribute utility theory (= MAUT) and strong philosophical theories of prudential desirability. (Multi-attribute utility theory is a branch of rational decision theory that provides tools for assessing the utility of objects in a synthetic way, namely by valuing the several attributes or dimensions of the object separately and then aggregating these values to the object's comprehensive utility.)

When assessing the personal desirabilities of a set of options via multi-attribute utility synthesis, their moral desirability will always make up one relevant attribute – however only one among several. Because the other attributes may and often do favour immoral choices some device is needed to bring rational decisions and moral requirements into line. The second part of the approach deals with this problem of alignment and proposes social norms as such a device.

The paper is structured as follows: (1) First, the claims that rational decision theory is necessary for justifying the rationality of morals and that moral action has to be rational in the decision theoretic sense are defended. (2) Then ethics of cooperation, which try to put this insight directly into practice, are criticized. The most important objection is that they lack a moral desirability function. (3) Rationally motivating moral desirability functions can, however, be determined with the help of multi-attribute utility theory, namely as one dimension of one's comprehensive utility function; the main characteristics of multi-attribute utility theory are explained. (4) Next, intuitionistic ways of finding a holistic moral desirability function are criticized. (5) As a consequence of their failure, an analytic strategy to define 'moral desirability' out of precisely chosen motives is developed and (6) executed. Other motives that do not *define* 'moral desirability' but *support* morality are identified as well. (7) All these motives are still too weak for realizing morals reliably; morally good social norms, which include threats of punishment, have to provide a further and now sufficient piece of motivation. Therefore, finally, some considerations are presented as to how socially valid moral norms can be brought about on the basis of comparatively weak motives only.

Note regarding terminology: In rational decision theory the term "*value*" often refers to quantified representations of preferences over events (or more generally: states of affairs) whose relevant consequences (or non-causal implications) are known. "*Utility*", on the other hand, refers to quantified representations of preferences over events with uncertain prospects, so that utility includes the weighting of the consequences' probabilities. Even though both these usages are somewhat problematic, here I will largely follow them. In addition, I will use "*desirability*" as a generic term, which shall cover values as well as utilities.

1 Some Reasons for a Decision Theoretic Approach to Ethics

Arguably, philosophical justifications of good systems of morals should fulfil, among others, two rather general and formal adequacy conditions, motivational effectiveness and stability with respect to further information.[1] That a justification of morals is "*motivationally effective*" roughly means: if the justification's content is true and epistemically accessible, someone justifiedly believing in the morals' justification and its thesis adopts these morals practically: if he believes that these morals require a certain kind of action he is nearly always motivated to fulfil this

[1] For a more precise version of these adequacy conditions and an extensive justification see: Lumer <2000>/2009, 30-46.

requirement.[2] "*Stability with respect to further information*" shall mean that the motives on which the motivational effectiveness relies are stable and will not or would not vanish if the respective person obtains new (true) information. These adequacy conditions cannot be appropriately defended here (however see: Lumer <2000>/2009, 30-46); but a hint as to the line of argument in their favour can be given. *Motivational effectiveness* of a moral justification is mainly a requirement of practical relevance; that a justification of morals is not motivationally effective implies that it is ignored in practice; so, in the end, the ethicist can also disregard it (cf. Gauthier 1991, sect. III); it may be theoretically impressive and beautiful etc. but it is practically useless. In addition, without the – at least implict – appeal to some motivation and desire required by motivational effectiveness, we usually do not even understand why an argument brought forward for some morals should be a reason or justification of it. Motivation's *stability with respect to further information* on the other hand, firstly, shall guarantee some sort of epistemic rationality and wisdom in the choice of moral principles in the sense that in this choice all relevant information (which could alter this choice) has been taken into account; the choice is not naive and is not due to persuasion. And this implies, secondly, that the founding motives and the resulting principles are intellectually very highly developed. Finally, stability with respect to further information guarantees the durability of the adopted morals.

All current scientific psychology says that in order for someone to act there has to be (or in cases of automatic actions: must have been) some desire or motive or, in decision theoretic terms, preference and a belief that connects desire and action (cf. e.g. Camerer 1995; Koehler & Harvey 2004, part 3). This is the reason why ethicists who adhere to a decision theoretic justification of morals and who accept the two adequacy conditions endorse a *mundane* [3] or *decision theoretic rationalism*, which holds: justification of morals must be based on a mundane, subject-centred form of practical rationality, namely rational decision theory, which in turn takes up the subjects' preferences. More specifically, in order to fulfil the adequacy conditions, (ethically justified) morally required actions (at least nearly always) must also be rational in the decision theoretic sense and for decision theoretic reasons. Apart from being mundane, the type of rationality required here has to be specifically decision theoretic for the following reasons.[4] Decision theoretic rationality takes up and is based on the (moral) agent's preferences; its only aim is to maximize their fulfilment. Therefore, decision theoretic rationality is suitable for fulfilling the motivation requirement under the restrictions of human psychology because it is based on what people already want (or are prone to want). In addition, decision theoretic rationality is a rather minimalistic and

[2] Motivational effectiveness is a special kind of internalism. However, because of the many uses of "internalism" the label "motivational effectiveness" is preferred here.

[3] It is mundane at least compared to Kant's "higher" rationalism of pure reason, which tries to guarantee autonomy by backing decisions on *a priori* reasons.

[4] Rational decision theory has been frequently contested since the 1970s. Most of these criticisms, though, attack its use as a true model of empirical decisions (e.g. Kahneman & Tversky 1979). This kind of critique is irrelevant in the present context, which speaks of *rational* justification.

formal conception, and it is neutral with respect to the contents of people's preferences. Therefore, it should be widely acceptable and it does not risk begging the moral question.

2 Ethics of Cooperation and Their Shortcomings

The purest decision theoretic approach to justifying morals holds that the decision theoretic conception of rationality alone, or more specifically, its game theoretical branch, provides adequate and justified morals (e.g. Gauthier 1986, 2-4).[5] Although a straightforward application of game theory often leads to morally bad selfish decisions, after the introduction of additional premises, which game theoretical ethicists hold to be normally fulfilled in real life, game theory recommends cooperative actions and thus justifies an ethic of cooperation. The paradigm case discussed in such an ethic are prisoner's dilemmata, where a straightforwardly maximizing rationality requires not to cooperate and thereby to end up with a Pareto suboptimal outcome. The initial solutions proposed are: (1) Cooperative game theory: if the possibility of enforced contracts is given both players can covenant that they will cooperate, and in case of defection will be punished by the enforcing institution. This solution, however, turns the situation into one that is only similar to a prisoner's dilemma, it *presupposes* contract enforcing institutions and cannot provide or explain their existence. (2) Another solution presupposes a possible iteration of the prisoner's dilemma. In iterated prisoner's dilemmata, always to defect turns out to be a poor strategy. Tit For Tat, i.e. to begin with cooperation and then always to imitate the other player's last move, is a much better strategy (Axelrod 1984, ch. 2).

There are two groups of problems that have prompted ethicists of cooperation to amend their game theoretic approach. The first group is compliance problems. An institution that will enforce contracts may not exist at all or may not be accessible, or may be too weak in the present situation or may not be willing to enforce every kind of contract (the agreement's contents may be rather unimportant, private or against the institution's own interests etc.); or the institution's functioning presupposes a cooperative spirit of its executives that cannot always be procured by a meta-institution; or it may be impossible to make a contract (because there is no possibility of communicating, no time to make an agreement, etc.); finally, institutions are costly. So the solution by cooperative game theory only helps in a restricted range of cases and needs to be complemented by other solutions, e.g. meta-strategies in iterated cooperation situations. However, not all cooperation situations will be iterated, or iteration will be limited or very unlikely; or it may be unclear if a given situation is an iteration of earlier situations or a new type of situation. Therefore, some ethicists of cooperation add further devices that are intended to lead to compliance, e.g. the reactions of third persons: not only should the cheated cooperator no longer cooperate with the

[5] Many authors have developed their own game theoretic justifications of an ethic of cooperation, e.g.: Binmore 2005; Hoerster 2003; Narveson <1988>/2001; Stemmer 2000. Because the present paper mainly develops a positive alternative, there is no room for doing justice to all these models. Instead Gauthier's theory, which probably is the best known model, will be used here as a representative paradigm case.

defector, other people, too, should take the defection as a sufficient reason not to cooperate with the defector (Gauthier 1986, pp. 15, 167-183). However, this third person may be sure that she herself will not be cheated. In this case the required refusal of cooperation with the defector could be a kind of cooperation with the second, cheated or a fourth person etc. Thus what begins as a clear case of mutually advantageous cooperation gradually changes into a general practice of social sanctioning, which to the greater extent is now sustained by motives quite different from getting more profit by cooperating with others (see e.g. Fehr & Gächter 2001), and into a practice of following social norms, again sustained by motives quite independent from receiving reward from concrete two-party cooperation, namely the motives to avoid sanctions, to generate a positive reputation and to receive social recognition or even moral motives (see below).

The second group of problems that pushed ethicists of cooperation beyond the purely game theoretical approach consists of problems of moral content, i.e. that this approach may lead to recommendations that are too much at odds with intuitive morality. Gauthier, for example, discusses the possibility that slave-holders cooperate with their slaves in giving them some freedom for smoother exploitation. His solution is that the initial bargaining position on the basis of which the terms of cooperation are arranged has to respect the Lockean proviso, i.e. one does not yet take advantage of the presence of others (Gauthier 1986, 15 f., 191-217, 225). Making recourse to the initial bargaining position, of course, reconciles ethics of cooperation with intuitive morality up to a certain point, however (apart from the enormous and completely unresolved problem of determining what the initial bargaining position is) it is a departure from a pure decision theoretically rational conception of morality. Respecting the Lockean proviso may already be a requirement of morals but not of rationality.

Although turning to some initial bargaining position brings ethics of cooperation a bit closer to usual moral standards, this kind of ethics still falls (very) short of fulfilling these standards. As frequently noted, ethics of cooperation lead only to very weak, minimal, business morals (i.e. morals for improving business) (Trapp 1998).[6] 1. The basic idea of these ethics is cooperation: the morally correct action is defined as the cooperative action; and its rationale is *do ut des*, give for receiving: I cooperate with you to the end that you cooperate with me because we both gain. It is part of the very logic of these ethics of cooperation that those who cannot cooperate or make a contract cannot be (direct) beneficiaries of these ethics: small children, the very old, severely handicapped people, future people, animals. Abandoned, severely handicapped people will not only not get food and medical help they are not even protected against assault or murder. 2. In addition, those who can cooperate will receive goods in proportion to their cooperative value for others; those whose cooperation is of little value to other people will receive only small advantages from the system of cooperation ethics; and those whose cooperation is of high value for others will receive great advantages, even if their own contributions cost them little. 3. Furthermore, being based on the rationality of personal advantages, rational cooperation makes sense only if the cooperator gains from it. Only Pareto improvements are possible, i.e. arrangements that imply

[6] For a more detailed critique of cooperation ethics see: Lumer <2000>/2009, 102-116.

improvements for all parties making the contract. This excludes every kind of moral redistribution of desirable goods. 4. Finally, Tit For Tat requires non-cooperation with defectors even if they are in a plight. – All this goes strongly against most people's morals.

This kind of intuitionistic and material critique is well known but argumentatively weak, in the sense of presupposing strong material premises. Ethicists of cooperation can reply that moral intuitions cannot refute rationally justified ethics because intuitions vary interpersonally and because they are too weak in relation to rational justification. However, there is a group of still stronger criticisms of cooperation ethics, namely that this type of ethics lacks the formal, structural requirements of developed ethics.

1. Moral desirability function: Ethics of cooperation do not comprise any kind of *moral desirability* function, i.e. an attribution of supra-personal [7] (rank-ordered or numerical) values to states of affairs – as e.g. in welfare ethics, like utilitarianism, which defines the sum (or the mean) of the individual utilities as the moral desirability, or in Scheler's "Wertethik". According to ethics of cooperation, in a two-person game there are just the two personal desirability functions and the recommendation of decision theoretic rationality to act in a certain way. However, there is no moral desirability that says how good the actions are from an impartial or social or general or universal perspective. According to a fully developed morals, a certain cooperation e.g. can be rational for all parties but nevertheless unjust, or an initial bargaining position may be unfair, or an action could be personally advantageous but morally bad etc.; ethics of cooperation cannot say things like these.

2. Moral emotions: Adopted moral desirability functions are the basis of people's moral value judgements and the moral emotions stemming from them. Indignation for example, begins with perceiving or recognizing or imagining some state of affairs, which then is morally evaluated as being outrageous; this valuation subsequently causes the corporal and feeling parts of the emotion and these, finally, cause motivational tendencies. The moral valuation is a crucial step in this process.[8] (Another way how moral emotions lead to moral motivation is hedonic: subjects may try to cause pleasant moral emotions, like high moral self-esteem, and avoid unpleasant moral emotions, like shame and guilt, by acting in accordance with morals.) Without moral desirability functions these moral value judgements and hence the moral emotions are also lacking. This is so because moral emotions depend on a certain kind of supra-personal authority, which goes beyond personal self-interest. So on the basis of cooperation ethics, someone may feel annoyed about herself because of the personally negative consequences of her own defection and regret it.

[7] "Supra-personal" here roughly means: with an authority above that of a singular person. This may e.g. be the authority of a valuation that reflects everybody's interests or the authority of objective values that are independent of any value subject.

[8] For the general role of valuations and appraisals for emotion see e.g.: Lazarus 1991; Solomon <1976>/1993, 209-212; Lumer <2000>/2009, 456-474; Zeelenberg et al. 2008. For moral emotions being based on moral appraisals see e.g.: Izard 1991, 361. (Many ethicists think moral emotions are based on moral appraisals, e.g. C. D. Broad, Richard M. Hare, John Rawls, Ernst Tugendhat.) Particular moral emotions and their underlying appraisals are analyzed e.g. by the following authors: guilt: Lazarus & Lazarus 1994, 55-59; Montada 1993, 262-266; Solomon <1976>/1993, 258-263; indignation: Montada 1993, 266; Solomon <1976>/1993, 270-272; resentment: Solomon <1976>/1993, 290-295.

However she cannot value herself and her action in terms of moral desirability from a supra-personal perspective and as a consequence feel low (or high) moral self-respect, remorse or guilt – in spite of the personal advantage. The analogue holds for other-regarding moral valuations, so that indignation, resentment, moral anger, moral disgust, abhorrence are excluded. Now emotions also lead to specific motivational tendencies (to punish the bad, to restore justice, to protect victims etc.) and they themselves have hedonic motivational significance (guilt e.g. is an unpleasant emotion, this gives us reason to avoid guilt by behaving correctly). Therefore, not including moral emotions deprives cooperation ethics of an important source of motivation (via moral self-esteem, shame, guilt etc.) to act morally in the first place and (via indignation, moral disgust, moral contempt etc.) to stimulate others to do the same, in the last resort by punishing them. And this source of motivation, of course, is particularly important if doing so goes against one's personal interests.

3. Self-transcendent ego ideal: Apart from its motivational impact, an authoritative supra-personal desirability function adopted by the subject also has an important function for personal identity and self-image. It can lead to going beyond one's limited personal horizon and to adopting a self-transcendent ego ideal, which may have a lot of positive consequences for the subject: providing meaning, participating in great collective human projects, spiritual union with other people, serenity and diminishing personal disquiet (in particular fear of death). All this is not possible in an ethic whose personal aim is to further one's own interests by cooperation. – Summing up points 2 and 3 it may be said: In a certain sense the whole inner, psychic side of a developed moral is missing in ethics of cooperation. And reinforcing this even further, there is rationality and behaviour (often) in accordance with morals, but no morality and moral action in a strong sense, only self-interest.

4. Supra-personal perspective: Morality, at least sometimes, requires acting against one's personal interests. However, if in ethics of cooperation there is no supra-personal moral perspective but only personal interest such a conflict is not possible. Of course, clever cooperation often requires acting against one's short-term interests in favour of one's long-range interests. However, these are still one's personal interests, and there are no moral motives and requirements.

5. Intrinsic value of other beings and their interests: At least more demanding morals require going beyond one's self-interest in another respect as well, namely by demanding care for the needy just because they are in need and independently of whether or not they can repay this care. In ethics of cooperation, however, (active) moral subjects and direct beneficiaries of the respective moral are identical; there are no (direct) beneficiaries beyond the active bearers of this moral.

6. Socially shared, common perspective: A supra-personal moral desirability function can nonetheless be individualistic – e.g. if it has been adopted by only one person and if its realization is independent of other persons sharing it; strong benevolent ethics sometimes are of this kind. However, many people's moral desirability functions are just conceived for being shared by other people. They are supposed to represent a common value base and perspective. As such they can determine a best common social project, guidelines for policies to be realized in cooperation;[9] and

[9] This kind of cooperation differes from the type aimed at in ethics of cooperation. It is a cooperation for realizing a supra-personal project; it is not a cooperation where I buy your cooperation for my personal advantage.

in case of interpersonal conflicts they can determine the just solution, which is sustained by the whole community. Ethics of cooperation cannot serve to this end because, again, the supra-personal perspective is missing.

The structural deficits just outlined may be more or less important, and it may turn out that at least some of them cannot be fulfilled in rationally justified ethics. However, one should at least try to fulfil them in a rationally justified ethic. All the criticisms advanced so far against ethics of cooperation do *not* imply that these ethics are a bad thing. On the contrary, these ethics lead to some very important results. But in any case they are very weak and their morals probably only the *beginning* of morality – perhaps even historically.

3 Moral Desirability Functions and Multi-attribute Value Theory

The straightforward solution to most of these problems is to introduce a moral desirability function that ascribes supra-personal values to all social states of affairs (cf. problem 1), which then can be adopted by moral subjects and, as a consequence, make up one – more or less important – part of their comprehensive desirability function besides the self-interested components of comprehensive desirability. The moral desirability function then could be the basis of moral emotions (cf. problem 2) and of a self-transcendent ego ideal (cf. problem 3). It could be rational to follow moral or altruistic requirements against one's self-interest because the moral part of one's desirability function could make the difference (cf. problem 4). Of course, such moral desirability functions could take up the interests e.g. of all sentient beings, in particular of those who are not able to cooperate (cf. problem 5). Finally, one and the same moral desirability function could be interpersonally shared and thus also be the basis for deciding between common social projects (cf. problem 6). In fact, most people have such moral components in their total desirability functions, only psychopaths do not.

Advocates of cooperation ethics might reply to this suggestion: rational people always decide according to their total (or comprehensive) desirability function; a moral component does not have any privilege to be decisive.[10] Although this is true it does not refute a decision theoretic approach that bases morals on moral desirability functions. At least one could examine what possibilities the separation of the moral component of our desirability function from the other components offers. The hope would be that the formal characteristics of a supra-personal and perhaps interpersonal moral desirability cause a dynamic that confers it a particular if not decisive power. And what has been said in the last section about the crucial role of moral desirability functions for several psychic (moral emotion and motivation, self-transcendent ego ideal) and social mechanisms (common value assumptions as basis of social projects) nourishes this hope. Of course, this has to be elaborated further. In any case a rational morality that makes use of motivationally relevant moral desirability functions probably will lead to morally stronger results

[10] David Gauthier, personal communication (12 December 1994).

than deliberately (Gauthier 1986, 7 f., 11, 327 f.) doing without – no matter how weak the moral component may be. Cooperation ethicists may now reply: But the strength of cooperation ethic is that it relies on the rules of practical rationality alone (which may even be taken as *a priori*), it does not rely on contingent empirical foundations like our sympathetic inclinations (Gauthier 1986, 103, 327-329; Narveson 2010). However, this reply is weak. It is true that, being a purely rational approach, this kind of cooperation ethic equates morality with long-term rationality and thus is based on the rules of rationality alone. However, whether following this ethic leads to something at least similar to moral behaviour as it is intuitively conceived depends again completely on contingent empirical facts. If others cannot cooperate at all or only a little, or if the rational subject is not interested in cooperation (she lives or prefers to live in isolation), or if the rational decider as compared to other beings is extremely strong, the recommendations of (long-term) rationality are anything but moral (in an intuitive sense). Rationality by itself does not lead to morality; the rationality of morality always depends on favourable empirical conditions: our and other beings' motives, capacities, needs, resource situation etc. A purely formally rational approach to morals fails, morals have particular moral contents. In this respect cooperation ethics is based on empirical premises just like the moral desirability centred approach.

The straightforward way in rational decision theory to technically capture a separate moral desirability component in one's comprehensive desirability function is the use of multi-attribute utility theory, or more precisely: multi-attribute value theory.[11] Because the moral dealing with risk is not the topic of this paper and because measuring values is much easier – it does not require quizzing about complicated preferences between lotteries over morally relevant outcomes – we can restrict the present considerations to certain consequences (or non-causal implications) and thus to value theory. Instead of determining values or utilities on the basis of the decider's holistic preferences over entire options,[12] multi-attribute desirability theory first tries to measure value functions of various aspects or attributes or dimensions (here we will not distinguish between these things) of a certain kind of options and then to synthesize the value of these options according to their empirical qualities and their respective desirabilities. This is done for example in consumer counselling; the options may be different models of cars; their attributes may be exterior dimension, interior dimension, baggage space, power, fuel consumption, safety, etc.; the empirical qualities in

[11] Important foundational contributions to multi-attribute utility theory are: Krantz et al. 1971; Keeney & Raiffa <1976>/1993. Comprehensive overviews are provided in: Clemen & Reilly 2001; French 1986; Raiffa 1968; Watson & Buede 1987. A simple method is presented by: Edwards 1977. For recent developments see: Figueira et al. 2005.

[12] Determination of desirability functions on the basis of (really) holistic preferences over options with complex consequences is questionable anyway. If such complex holistic preferences have to be taken as the uncriticized basis of the desirability determination why can't we immediately take preferences over the present options as part of the preference basis on which the desirability is calculated? But if we already dispose of the preferences over the options in question (and have no way of criticizing or altering them) then the whole undertaking of determining the options' desirabilities is superfluous. (Lumer 1998, pp. 33-37) Therefore, only a synthetic determination of the available options' values makes sense. And the straightforward way to do so is the use of multi-attribute utility theory.

all these dimensions are measured; an (axiological) value is attributed to the single qualities; finally the overall value of the respective option (e.g. Volkswagen type xy) is determined by adding all of its dimensional values. The aim of this exercise is to assess the options' values much more precisely than is done on the basis of holistic preferences: people usually aim at a precise assessment even in their holistic value judgements, they consider various aspects of their options, however they usually do not bring this analysis to a systematic end but interrupt it and proceed to the holistic assessment. So multi-attribute value analysis merely completes, in a systematic and elaborated way, what we begin intuitively. But the deeper rationale – which is already the basis of the intuitive procedure – is this. Actions usually do not have any intrinsic value, they are executed for bringing about intrinsically good consequences. However, actions normally have many different kinds of intrinsically relevant – good and bad – consequences. These are aggregated, namely their intrinsic values are added up, to the actions' overall or total or comprehensive value. So when we choose the action with the highest overall value, by our action we seek to attain the world with the highest intrinsic net value.

Now the various attributes of an option usually correspond to the various ways by which this option produces an intrinsically relevant result. The attribute may stand for exactly one intrinsically relevant consequence of a certain type, e.g. the taste experience of eating a crumb of bread, or, much more frequently, a series of or many of similar ways of producing a certain type of result. The safety aspect of a car, for example, means that there are more or fewer ways to cause intrinsically relevant harms to the passengers via accidents, from loss of life to having fewer possibilities of consumption as a consequence of having to pay damages.

In multi-attribute value theory many methods of establishing value functions for the single types of attributes – apart from asking for the decider's preferences over lotteries – have been proposed (see e.g. Fishburn 1967). One method is direct *ranking*: x_i^* may be the highest and best – that is, best among the objects to be compared – quality or quantity in dimension X_i (x_i^* may be an objectively measured quantity like maximum speed), to which the (axiological) value 100 is attributed; x_i^0 may be the lowest and worst quality in dimension X_i, to which the value 0 is attributed; now the decider has to say which value – between 0 and 100 – she wants to attribute to a certain value x_{ij} between x_i^0 and x_i^*. (Watson & Buede 1987, 198 f.) Another method, *bisection*, relies on comparing preference strengths: the decider is asked for that point x_i^j that for her is as much better than x_i^0 as x_i^* is better than x_i^j; in the next step the value means between x_i^0 and x_i^j as well as between x_i^j and x_i^* are elicited; this may already be sufficient for interpolating a continuous value function $V(x_i)$ from x_i^0 to x_i^*. (Watson & Buede 1987, 195 f.) The easiest way to determine the value function $V(x_i)$ is to assume *linearity*, i.e. that: $V(x_i) = a \cdot x_i + b$ – with a and b being constant (Edwards 1977, 328 f.). This, of course, will often not be very exact but in many circumstances sufficiently exact for not inverting the final preference order (ibid. 329). If these rankings and bisections and even more the linearity assumption serve to measure things such as the value of maximum speed, fuel consumption etc. they are still rather holistic because in the end they are only estimates of how these objective qualities affect really intrinsically relevant consequences;

and there are very many ways to do this. A philosophically much more ambitious valuing methodology tries to find out what the really intrinsically relevant consequences of the respective aspects are; then a value function for this kind of consequences has to be established – using methods as those just described –; and finally it is necessary to assess how these intrinsically relevant states of affairs are affected by some quality x_i^j. The last step often will only be a rough estimate. However, this methodology has the philosophically important advantage of stimulating the decider to deep reflection about what is really at stake with a given attribute in terms of *intrinsic* relevance. And this is probably what we do in ethics when we determine moral desirability functions.

The values in the various dimensions can be calibrated in such a way that they are directly comparable, for example, a 10 point improvement in dimension 1 is as good as a 10 point improvement in dimension 2. However, in multi-attribute value theory the values within the various dimensions mostly are all normalized within the 0 to 100 interval, with 0 being the value of the worst result in this dimension in the given set of options, and 100 being the value of the best result. With this kind of normalization, in addition to measuring the value within the dimension, the dimensions themselves have to be weighted. These weights w_i can be attributed in two steps. First, a preference order or order of importance of the dimensions is established by asking: 'Which improvement from value 0 to value 100 is more important, the one in dimension *i* or the one in dimension *j*?'. Supposing that the answer was *i* one can determine the relative weight of dimension *j* with respect to that of dimension *i* by asking: 'For which quantity x_i^a in dimension X_i are you indifferent about the combinations $\langle x_i^a; x_j^0\rangle$ and $\langle x_i^0; x_j^*\rangle$?'; i.e. the equivalent in dimension X_i to the optimum in dimension X_j is sought. The value of x_i^a (i.e. $V(x_i^a)$), established before, then is identical to the relative weight of dimension X_j. (Watson & Buede 1987, 200-202; other method: ibid. 202 f.) The resulting relative weights, finally, will be normalized in such a way that the sum of all the dimension weights is identical to 1. The total value of an option *i* with the qualities x_j^i then is:

$V(i) := \sum_j w_j \cdot V_j(x_j^i)$.

Multi-attribute value theory provides a completely formal, neutral way of valuing, which, of course, can also be applied to moral decisions of rational persons. In this case at least one dimension will be dedicated to the action's moral value, as it is weighted by the respective agent, and a second dimension can comprise the action's non-moral value for the agent.[13] In a more complex model the moral part will be split into several dimensions like "objective" moral value, social retribution, intrapersonal sanctions. In the following, first, the simple approach will be presented and this more complex model will be dealt with subsequently.[14]

[13] Margolis (1982) distinguishes the selfish and the altruistic component of one's comprehensive option utilities. However, he does so in a quite different framework, specifically for determining the game theoretically adequate weights of the two components from an evolutionary perspective.

[14] How does this approach relate to other approaches? Some other uses of a decision theoretic framework in ethics can be understood as an application of multi-attribute utility theory as well, though as a quite different application. Every additive social welfare function of the form $U_{mor}(p) := \Sigma_i\, w_i{\cdot}U_{morg}(U_i(p))$ – where w_i is the

4 Two Kinds of Intuitionistic Moral Desirability Functions – Some Problems

The simplest way to construct moral desirability functions with the help of multi-attribute value theory may be called "*motivational intuitionism*": the moral component of the actual total desirability function is determined and taken as the moral desirability function. This method is *intuitionistic* in the fairly general sense of justifying moral judgements, criteria, principles etc. – in our case: moral valuations – on the (ultimate) basis of certain moral attitudes or behaviour, which, as far as the theory goes, are taken to be primitive (and not further justified or analysed).[15] *Motivational* intuitionism uses motivational preferences as its basis. The method of motivational intuitionism, more precisely, is this: The decider first has to declare what, according to him, the moral features of the single actions are: is an action morally forbidden, permitted or obligatory, has it a certain value, would it be virtuous, neutral, vicious to realize the action etc.? The spectrum of these features m^1, m^2, ..., m^n makes up the moral dimension M; all the other dimensions may be summarized as the complimentary set C. Next, the subjective (motivational) value of these moral qualities has to be determined by asking – following the methods described above – for preferences or preference strengths between the realization of actions of the types $\langle m^i; c^j\rangle$ or $\langle m^k; c^l\rangle$. The result is the moral value function $V(m)$. It reflects the moral part of our motivational and decisional tendencies.

Motivational intuitionism works, but it is rather crude. Although the moral component of our motivation evolves, among other things, with the help of cognitive operations, motivational

moral weight attributed to person i's welfare stemming from p and U_{morg} is a dimensional moral value function of a person's utility of p – can be interpreted as a multi-attribute utility function with the various individuals' utilities being the attributes of the object p to be valued (this is an extension of: Keeney & Raiffa <1976>/1993, 515-547). The simplest welfare function of this kind is utilitarianism, which assumes $w_i=w_j$ for all i and j and $U_{morg}(x)=x$; in prioritarianism e.g. U_{morg} is concave. Another example of the ethical use of multi-attribute utility theory is extended preferences (cf. Arrow (<1951>/1963), Harsanyi (<1955>/1976; 1977) and Sen (<1970>/1984, ch. 9).

Although multi-attribute utility theory is used (or could be used) in these theoretical approaches the way it is used is quite different from its application in the present approach. First, the just mentioned approaches try to formalise only soical welfare functions, i.e. a special type of *moral* desirability functions; they do not try to capture the *comprehensive* (moral plus amoral) desirability of options from the point of view of a given subject – as is done in the present approach. So they consider (at best) the *sub*-dimensions of one dimension (or even sub-sub-dimensions of a sub-dimension) of the option, namely of its moral dimension, but they do not consider and compare the other main dimensions of that option. Second, the theoretical framework of these approaches is judgemental intuitionism: they try to provide a specification of intuitive moral judgements (or to derive moral maxims from intuitive moral axioms); Harsany, for example, calls the extended preferences "ethical preferences" (Harsanyi <1955>/1976, 13 f.). Judgemental intuitionism does not make sure that these preferences have any motivational force at all, and even less does it use multi-attribute utility theory for resolving the problem of rational motivation to act morally (by exploiting the dynamics between moral and amoral dimensions of our comprehensive desirability) – as the present approach tries to do.

[15] In the literature many more specific notions of 'ethical intuitionism' are used (e.g. by Audi, Dancy, Huemer, McCann, Stratton-Lake). The general notion of 'ethical intuitionism' just defined includes nearly all of them. Here I do not use one of these more specific notions for not unnecessarily restricting the range of my argument.

intuitionism does not reconstruct how the (motivational) moral value is synthesized; the moral features m^i are already complex constructs subject to ethical theorizing. In addition, different kinds of moral distinctions, namely deontic, axiological and aretic qualities, are conflated into a single quantitative dimension. This is unsatisfactory from a theoretical viewpoint which wants to preserve the structural features of common morality.

Another approach may be called "*judgemental intuitionism*" – where “intuitionism” is meant in the general sense explained above, and “judgemental” means that certain moral judgements or doxastic intuitions are taken as the justifying basis. Most of ethical theorizing and moral arguing deals with moral judgements. A theoretically satisfying approach to justifying morals should give moral judgements a prominent role. However, for these judgements – in order to satisfy the adequacy conditions of motivational effectiveness and stability with respect to further information (cf. above, sect. 1) – to be motivationally and decision theoretically relevant they have to be "translated" in some way into motivational moral values $V(m^i)$, where "(correct) translation" means that: moral value judgements lead to proportional (motivational) moral values; deontic judgements about moral forbiddenness lead to a low, about moral obligatoriness to a high moral value. Motives that start with moral judgments, which then are "translated" into a proportionate motivation to act, here will be called "*moral motives*" (in the strict sense) because it is the judgement that determines the motivational force in the end. Ethics is mostly silent about this extremely important question of how the "translation" works, i.e. what moral motives consist in. Moral psychologists say there is only a weak positive correlation between moral judgement and moral action (Nunner-Winkler <1993>/1999). However this may be due to the non-moral dimensions C of our decisions, and there may be some "translation" mechanism nonetheless. Actually, there is a primary, proactive mechanism, which converts moral evaluations of our possible actions into a corresponding motivation, and there are secondary, reactive mechanisms, which convert moral evaluations of other persons’ actions or of our own past actions into a corresponding motivation to prevent or punish or to promote or reward the action. Here we can deal only with the primary, proactive mechanism. The best hypothesis about this mechanism is that it works mainly via moral self-esteem (cf. Taylor <1985>/1995; Lumer 2002, 181; criticism of competing hypotheses: Lumer 2002/2003, sect. 2-7). Moral judgements about our moral obligations or the (judgemental) moral value of our actions are the basis of our moral self-appraisals, i.e. judgements about our personal moral performance. Low self-appraisals lead to low self-esteem and self-reproaches, which are very unpleasant feelings (cf. Rawls <1971>/1999, 388-391). High self-appraisals lead to high self-esteem and satisfaction with oneself, which are very pleasant feelings. These feelings can be anticipated, and this gives rise to the respective hedonic valuations of the actions. This rather complicated mechanism usually functions nearly automatically – we form the moral judgement about the action, which almost instantaneously leads to the corresponding attribution of the respective (motivational) value. However, we can recognize its functioning in considerations of the following type: "I didn't think I could live with that knowing that I could have done something" or: "I could not stand that thought. I never would have forgiven myself" (Oliner &

Oliner 1988, 168).[16] We may summarize this by saying that moral self-esteem is the most important proactive moral motive in the strict sense.

So in the end there is a proactive mechanism that "translates" moral judgements into (motivational) moral valuing, namely moral self-esteem. However, where do the moral judgements come from, how could they be justified? Here again the mainstream of present ethics is silent in that at present the dominant methodology in normative ethics is intuitionistic; and this means taking moral judgements as they come or trying to make them coherent in a kind of reflective equilibrium but not questioning their origins (cf. e.g. Rawls <1971>/1999, §§ 4 and 9).[17] However the moral principles, in the sense of "(general) normative premises", in these mainstream methods are always taken from moral intuition and are never justified.

This judgemental intuitionism is problematic in several respects.[18] 1. Above we identified moral self-esteem as a mechanism that "translates" moral judgements into motivation. However, so far there is no guarantee that this mechanism works for every kind of moral judgement. And because judgemental intuitionism does not care about motivation at all it may be, and empirically is often, the case (cf. Nunner-Winkler <1993>/1999) that the intuitive moral judgements do not motivate – thus violating our initial adequacy condition of motivational effectiveness (cf. above, sect. 1). 2. Most moral intuitions are not stable with respect to further information, and those that are stable cannot be identified on an intuitionistic basis. This violates the second adequacy condition for a justification of morals. We acquire our (ontogenetically) first moral convictions in a completely heteronomous way by adopting them from the socialization agents. New peers, acquaintance with new principles, awakening of critical reflection, search for coherence, experiences with impressive cases, and non-moral motives – such as sympathy or respect – that nonetheless are close to moral motives lead to changes in and the development of our moral convictions. For some people these changes come to an early end (sometimes only because of intellectual indolence), for other people never. Judgemental intuitionism has no means to establish that a certain system of moral convictions will be stable e.g. because it satisfies certain rationality or motivational conditions etc. 3. Foundationalism in normative epistemology holds that justified beliefs must have been acquired by valid and sound inferences or as the result of the working of a reliable mechanism. The most interesting moral intuitions are, of course, non-derived moral premises or principles. Judgemental intuitionism, *qua intutionism*, views these principles as

[16] The quotations are explanations given by recognized rescuers of Jews during the Third Reich when they were later asked, why they had taken the risk to help the Jews.

[17] Though Rawls does not call his method “intuitionistic”, reflective equilibrium is intuitionistic in the general sense explained above: even if this method does not simply accept initial moral judgements but requires to make them coherent by a critical reflection, which among others has to consider the relevant facts, the resulting considered judgements are not supposed to rely on a justification or any other controllable and reliable process. Rather they are taken as the theory’s basis iff the subject, after the critical reflection, accepts them, period. Thus the considered judgements are taken as primitive and not as something that ethics can or must further analyse, explain or justify.

[18] Some of the following criticisms have been advanced by Mackie (1977, sects. 5.1 and 5.5). A more extensive critique is developed in: Lumer <2000>/2009, 77-89.

intuitions and does not analyze their origins. Therefore, it cannot – and does not try to – show that these intuitions stem from a reliable mechanism. Consequently, we should treat these intuitions as well as all inferentially acquired beliefs relying on them as unjustified. Of course, being unjustified runs counter to any attempt to justify morals. As a consequence, moral intuitions have no authority with respect to divergent morals or for determining one's future morals. 4. Moral intuitions interpersonally diverge to a considerable degree. This makes them unsuitable as a basis for shared social morals (cf. above, end of section 2).

5 A Multipart Strategy to Determine Motivationally Strong and Rational Moral Value Functions

The first two, i.e. intuitionistic, attempts to go beyond a pure rationalistic ethic of cooperation and to include a moral dimension within the total desirability function have turned out to be problematic. One general source of the problems is that the moral dimension considered so far is holistic and undifferentiated. This leads to instability and motivational weakness because the internal dynamic of the moral area, which by a rational ethic may be redirected to reliable processes and amplifying effects, is not captured. Therefore, now several subdimensions of moral motivation will be distinguished and their interactions analyzed.

A detailed analysis of the various motives for acting morally or in agreement with morals (Lumer 2002) [19] suggests splitting the moral dimension of actions into four main groups of subdimensions.

SR: motives close to morals: 1. 'moral desirability' defining motives: A first group of dimensions, which, as we shall soon see, immediately has to be split into two groups, may be called "motives close to morals" (or more exactly: this group of dimensions consists of the attributes that are the targets of motives close to morals). Such motives are sympathy, respect, love, friendship and creative expansion. These motives are *close* to moral motives because their aims largely coincide with those of morals; they all are e.g. altruistic in a broad sense. However, they are not *moral motives* in the strict sense because they are not based on *moral* judgements but on seemingly inborn and thus stable value criteria (which however have to be unfolded during ontogenesis and are only unfolded if attention is devoted to their objects, e.g. the well-being of other beings). To the contrary, an idea to be developed below is that motives close to morals can and shall be the basis of our judgemental moral valuations and of the moral desirability searched for. Precisely because they are *not* based on moral judgements they are suited as a basis for defining 'moral desirability'. However, for reasons of universality – which will be expounded later – not all motives close to morals are formally suitable for defining 'moral desirability'. So the group of motives close to morals has to be split into (i) 'moral desirability' defining motives and (ii) (mere) friendly motives.

[19] An earlier study of mine (Lumer 2002) provides the psychological material for the present and the following section as well as a rough analysis of which motive may have which function for morals and its justification. The ethical utilisation of this material for the multi-attribute utility approach is new.

'Moral desirability' defining motives are above all sympathy and respect for other beings and valuable objects (which leads to the abbreviation "SR"). Their targets are the well-being of sentient beings (for sympathy) and the preservation as well as – where this is possible – autonomous development of evolutionarily highly developed, complex, precarious and fragile entities (for respect).

F: motives close to morals: 2. mere friendly motives: The other half of motives close to morals because of their formal properties, in particular their idiosyncratic nature, is not suitable for defining 'moral desirability'; they are mere friendly motives that are altruistic in the broad sense and sustain morals without defining it. Love, friendship, creative expansion are such merely friendly motives.

I: internal sanctions: Internal sanctions, positive as well as negative ones, are feelings like positive and negative moral self-esteem, moral pride, shame or pangs of conscience. They are based on *moral* assessments of one's actions or of the whole proper person. They "translate" moral judgements into motivation, namely the desire to receive positive "sanctions" and to avoid negative sanctions (cf. above, sect. 4). Because they translate moral judgements about one's own behaviour into moral motivation they are moral motives in the strict sense.

E: external sanctions: External moral sanctions, again positive as well as negative ones, are rewards or punishments given to or inflicted on the subject by other people for (adequately) fulfilling or not fulfilling the moral standards of these people. Rewards include material goods, symbolic goods (e.g. medals) as well as acknowledgements. Moral sanctions may be informal, i.e. administered by anybody who wants to, or formal, i.e. administered by particularly authorized persons. External sanctions are important means for "translating" a social moral into motivation. Law and its sanctionative possibilities are usually rather strong and effective means for realizing social morals.

Because the valuative criteria underlying motives close to morals are independent of moral convictions, whereas internal and external sanctions are based on an (individually and socially, respectively) adopted morality these subdimensions of moral motivation (this time, of course, in the broad sense) may have completely different functions – at least this is the strategy developed in the following. Because motives close to morals lead to valuations that intuitively can be considered as moral even though they rely on stable natural criteria and not on moral convictions, the decisional value function springing from them could define the sought-after moral desirability function. However if we enter into further details (cf. below), for more formal reasons only a part of the motives close to morals is suitable as a basis for defining 'moral desirability' (dimension *SR*). This means the decisional value of the qualities (sr^i) of an object i in this dimension (*SR*) would (at least roughly) define the moral desirabilities of these qualities and of the object i itself: $D_{mor}(i) = D_{mor}(m^i) \approx V(m^i) := V(sr^i)$. Because of the stability of the respective value criteria this move could lead to a stable moral desirability function, thus fulfilling the adequacy condition 'stability with respect to further information' and providing the reliability needed for a justification. Mere friendly motives often will support 'moral desirability' defining motives. However they will not do so in all cases; in addition, they have their own enduring amoral aims. Internal and external sanctions, on the other hand, which "translate" the judgements stemming from an adopted moral into motivation could exactly take up the moral desirability function thus defined and lend it more motivational force. So, according to this strategy, the 'moral desirability' defining motives would provide the moral signal so to speak, and the sanctions would function as motivational amplifiers (whereas mere friendly motives only frequently support the moral decision motivationally). Compared to the ‘moral desirability’ defining motives, both kinds of sanctions are rather strong motives [20] – even in ethics of cooperation external sanctions are the means to resolve prisoner's dilemmata. Therefore this part of the strategy, i.e. adding sanctions to the 'moral desirability' defining motives, could resolve the problem (raised e.g. by Gauthier, cf. above, sect. 3) of the weakness of moral motivation: the value of the four moral dimensions (*SR*, *F, I*, *E*) together could nearly always outweigh the perhaps opposite preference in the complementary and in part selfish dimensions (*C*). In such a case, acting according to morals would be unselfish and nonetheless rational.

Sanction values are only very roughly amplifiers of values deriving from 'moral desirability' defining motives. An important tendency towards proportionality notwithstanding, there are several discrepancies. 1. The domain of the sanction values is only a subset of the domain of the value function on the basis of motives close to morals. Sanctions are agency-centred; sanctions react to our actions, and the anticipation of such reactions constitutes motives for our actions. Therefore, the domain of sanction values is our *own actions*. Motives close to morals and in particular sympathy and respect, on the other hand, are beneficiary-centred; we react (emotionally and) motivationally to other beings' fates. (Cf. Margolis 1982, 21 f.) This beneficiary-orientation is the central prerequisite for being the basis of defining 'moral desirability'. Beneficiary-orientation implies that value on the basis of motives close to morals can be attributed to all states of affairs that are identical with or contain or can influence the other beings' fates – beyond our own actions, e.g. the fate of individual beings, the whole world, other persons' actions. For example, we may feel sympathy for distant or dead people or people well cared for by other persons or people within reach, though we do not think of our intervention. In these cases we can attribute a sympathy value to these states of affairs but, because no action of ours is involved, there is no sanction value. 2. Value arising from sympathy and respect, according to the strategy adopted here, shall correspond to moral *desirability*. Sanction values, on the other hand, derive from anticipated sanctions, which

[20] J. C. Flügel (1925) conducted a psychological study, in which the subjects (N=9) kept an emotion diary for a period of 30 days, inserting type, intensity (rating from -3 to +3) and duration (in minutes) of their emotions. This led to nearly 10,000 entries altogether (mean: 33 entries per person per day). From these data the extents (absolute intensity integrals, i.e. absolute value of the intensity multiplied by the respective duration) of the various kinds of emotions can be calculated, representing something like the actual importance of these types of emotions or, in the terminology of multi-attribute value theory, the relative weight of the emotional dimension. The mean (mean of the subjects) extent of positive self-feelings made up 3.14% of the extent of the positive emotions or 2.11% of the extent of all emotions; the mean extent of negative self-feelings made up 4.05% of the negative and 1.33% of the extent of all emotions; finally, pity made up 0.04% of the negative and 0.014% of the extent of all emotions; positive sympathy and respect were not mentioned. (Flügel 1925, 345 f.) Though moral self-esteem makes up only a part of "self-feelings" this part probably is still much more important than sympathy.

in turn mostly will be reactions to *deontic* assessments of our actions. However, although moral obligations are instruments to realize moral improvements, – at least according to most people's morals – they are not simply identical to a duty to maximize moral desirability. Apart from moral desirability, moral obligations also reflect reasonableness for the agent, rights of other people, personal relations etc.[21] The relation between moral desirability and moral obligation is complicated after all. 3. External moral sanctions are based on socially adopted morals. These are not necessarily identical to the agent's moral. There will always be some concordance, but also nearly always some discordance.

6 The 'Moral Desirability' Defining Component of the Rational Desirability Function

So far the concept "moral desirability' defining motive' here has been used without much explanation. This concept will now be specified in such a way as to make its extension fit the strategy just explained, i.e. the motivational values stemming from these motives should be suitable for defining 'moral desirability'. In addition, the distinction between 'moral desirability' defining motives and mere friendly motives has to be explained and defined. This clarification implies a specification of the dimension *SR* of the rational value function, too.

The strategy followed in this section is this. Instead of formally defining ‘‘moral desirability’ defining motive’, first, several adequacy conditions that such motives should fulfil are set out and briefly justified. In a second step, these adequacy conditions are used to filter out the ‘moral desirability’ defining motives; in addition, the roles of the motives that did not pass the test are specified.

The adequacy conditions for ‘moral desirability’ defining motives are the following:

AQ1. Stable motivation: Motives on whose basis 'moral desirability' shall be defined should fulfil the initial adequacy conditions for a rational justification. This means they should lead to a motivation that is stable with respect to further information and reliable in general. – As a consequence, because of the subjects' cognitive errors, values attributed on the basis of these motives, cannot simply be obtained by asking the subjects for their preferences – if these are not most basic, really intrinsic preferences. If they are not intrinsic, the values have to be *rationally synthesized* from intrinsic preferences and true empirical information about the value objects leading to the intrinsically preferred states of affairs. If the subject e.g. intrinsically prefers another person's well-being *A* to her own well-being *B* to degree 1; and if an action a_1 objectively leads to the change from *B* to *A* then a_1 in this dimension has the value 1 (or a positive-linear transformation of 1) – independently of what the subject thinks about the consequences of a_1.

AQ2. Independence of moral convictions: If the motives searched for are intended to define 'moral desirability' and to rationally justify this part of morals the respective motives must not be moral

[21] These and many other considerations have been discussed in the debate about the limits of morals (cf. e.g. Scheffler <1982>/1994; Williams 1973).

motives in the strict sense, i.e. motives that are based on moral convictions. (Because, as we have seen above, intuitive moral convictions are not stable, independence of moral convictions is also required for stability.)

AQ3. Altruism: The value functions implied by the 'moral desirability' defining motives should give positive value to other beings' flourishing and negative value to their harm; in this sense the value functions should be altruistic. – Altruism in this sense does not exclude the *agent's* flourishing from being positively valued as well; however the agent should be considered in the value function only as one being on a par with others. Altruism provides the supra-personal quality, self-transcendence (cf. above, sect. 2) and other-centeredness of moral desirability. In addition, altruism guarantees some correspondence with intuitive morals.

AQ4. Moral adoptability: The first three adequacy conditions guarantee that the resulting value function defines a rationally justified and motivational concept of 'moral desirability'; however they do not yet guarantee that this kind of moral desirability leads to inner sanctions. In order to guarantee this, the 'moral desirability' defining value function has to be morally adoptable: it has to be suited to be adopted as the basis of our moral self-assessments that lead to moral self-esteem.

AQ5. Subject universality: The adequacy conditions introduced so far may already be sufficient for an individual moral, which is adopted by one subject only or which need not be shared by other people. But if the moral desirability function shall also be amplified by social sanctions it has to be shared by other moral subjects who administer sanctions based upon it. This means, that in order to lead to this kind of motivational support the 'moral desirability' defining motives must be the same for the subjects of this moral and imply interpersonally (roughly) identical value functions and in this sense must be *subject universal*. – 'Identical' here means that the functions give (roughly) the same value to the same things – and not only to analogous things. If I rate my pleasure positively and you yours, this is not identity of valuings; identity requires that we *both* rate your (or my) pleasure positively. The domain of subjects may be more or less extended; a small domain could be one's tribe, the most ambitious domain would be all mankind. Subject universality means that this value function is part of all the *subjects*' multi-attribute value functions. Subject universality does not analytically imply *beneficiary* universality, i.e. that all beneficiaries (of a certain domain) are considered equally. However, beneficiary universality will quite naturally be an empirical consequence of subject universality.

Mere friendly motives are always altruistic at least in a broad sense (AQ3) and independent of moral convictions (AQ2). However, they leave at least one of the other three adequacy conditions unfulfilled. Below we will consider some of them that violate stability (AQ1) or subject universality (AQ5).

The adequacy conditions for 'moral desirability' defining motives just developed can now be used to filter out the adequate motives. I have done so in a study in which all important motives for acting in correspondence with moral requirements have been scrutinized (Lumer 2002). Here only the most significant results of that study can be summarized.

The two most important motives that passed the adequacy test, i.e. that are appropriate for defining 'moral desirability', are sympathy and respect, or more precisely: sympathy optimization and respect optimization. Let us consider sympathy first, which is by far the more important of the two. Sympathy, precisely spoken, is an emotion, not a motive: Someone feels good because she believes that somebody else is well, and she feels bad because she believes the other to be in a bad way. (The well-being attributed to the other being can but need not consist in his present emotion, feeling or mood; it can consist for example also in his long-term prospects.) There are two ways how sympathy can lead to motivation. *1. Sympathy optimization:* Pity, after all, is an unpleasant emotion, and positive sympathy, i.e. to be pleased about another person's positive well-being, a pleasant emotion. So one can develop hedonic desires to optimize one's sympathy by improving the fate of other people. *2. Acting from sympathy:* Like other strong emotions, sympathy can induce intrinsic desires, in this case an intrinsic desire to change the situation that led to sympathy. Pity induces the intrinsic desire to improve the other's situation; positive sympathy induces the intrinsic desire to maintain, protect and possibly to increase the other's well-being.[22] – Acting from sympathy is also altruistic in a strong sense because improving the others' situation is its intrinsic aim. However, acting from sympathy depends on the current presence of sympathetic emotions, which first must ignite us to arouse the intrinsic helping desire, which again vanishes when the emotion is over. Therefore, acting from sympathy does not fulfil the stability condition (AQ1) and cannot be a basis for defining 'moral desirability'; it is a mere friendly motive. Sympathy optimizing, on the other hand, – if it goes via improving the situation of other beings and does not include avoiding pity by selective perception – is altruistic only in a weak sense, but nonetheless is sufficient for fulfilling the respective adequacy condition (AQ3). In addition, sympathy optimizing is independent of moral convictions (AQ2) and stable (AQ1) because stability is exactly one big advantage of hedonic motivation, which makes it suitable as a basis for rational long-term planning. Because we timelessly desire to avoid unpleasant feelings and to obtain pleasant feelings we can act today on the basis of this desire with the aim of improving the feelings we will experience tomorrow, in a week or in thirty years. Most socially adopted morals have some component based on sympathy, and there are many well-known cases where people deplore or even feel guilty about not having acted according to the demands of sympathy. This means that a moral desirability function defined according to sympathy optimization is morally adoptable (AQ4): it is apt to be adopted in such a way as to be one criterial basis of our moral self-assessments that leads to moral self-esteem. However, only a part of sympathy optimizing fulfils the condition of subject universality (AQ5), namely that part that has as its object the fate of beings more or less personally unknown to us and whom we can meet – or not meet – with approximately equal likelihood**.** The amount of sympathy with beings personally well-known to the subject, instead, will differ for

[22] Psychological hedonists have contested that improving another person's well-being can be the content of an *intrinsic* desire. However psychologists have shown that subjects experiencing sympathy help the commiserated person even when the subjects believe they will not receive any hedonic gain from their helping (Batson et al. 1983). – For a general model of hedonic as well as non-hedonic emotion-induced intrinsic desires see: Lumer 1997; <2000>/2009, 477-493.

different subjects because the amount of sympathy increases the more often we are confronted with the other beings' fate. So only what may be called "universal sympathy optimizing" is a 'moral desirability' defining motive, whereas particular sympathy optimizing is a mere friendly motive.

There may be doubts as to whether (universal) sympathy optimizing rather than acting from sympathy is the right motive for establishing the rational moral desirability function. First, however, although the instability of acting from sympathy makes it really unsuited for *defining* 'moral desirability' this does not prevent it from being one of the important friendly motives for acting morally; it simply does not define 'moral desirability'. Second, once having found out that there is this way to hedonic improvement via sympathy optimizing this kind of improvement is rationally desirable – independently of whether people recognize this desirability or not. In addition, because of its independence from present sympathy emotions and thus also from acting from sympathy, this motive can influence *every* decision relevant for other beings; and thus we find it actually present when people deliberate, for example, about public policy in this way: 'We do not want to have people begging / starving, let's give them a positive outlook for the future.'

Respect as it is understood here is also an emotion: On the basis of a subject's judgement that an object she is confronted with is evolutionarily highly developed, complex, precarious and fragile, perhaps also autonomous and caring about himself, which for the subject implies a positive intrinsic valuation of this object, the subject experiences an emotion from the group of respect emotions: respect, reverence, veneration, admiration, fascination, (admiring) astonishment, amazement, marvelling, (admiring) recognition. If simultaneously with the respect experience the subject is aware that the object of her respect (probably) will be impaired, damaged, violated, destroyed or killed secondary respect emotions come up: sorrow, grief, anger, rage, indignation, outrage, fear. Primary and secondary respect emotions lead to motivation in two ways – like sympathy. First, primary respect emotions induce an intrinsic desire that the respected object shall persist, live, act or function unimpaired or undisturbed; this desire can motivate actions to protect the object, to abstain from damaging it and to treat it carefully. Secondary respect emotions induce an intrinsic desire to protect the respected object or, depending on the situation, to punish the person who has damaged or destroyed it. This kind of motivation and action may be called "*acting from respect*". Second, primary respect emotions are pleasant, secondary respect emotions are unpleasant; therefore, we can have hedonic desires to optimize our respect emotions by protecting and conserving the objects of emotion and by threatening possible destroyers with punishment and if necessary by inflicting the punishment. This kind of motivation and action here is called "*respect optimization*". Acting from respect, like acting from sympathy, is bound to the presence of the respect emotions; if these emotions fade the intrinsic desire to protect the respected object vanishes too. So acting from respect does not fulfil the stability requirement (AQ1), whereas respect optimization does. Furthermore, respect optimization is independent of moral convictions (AQ2), it is altruistic in a broad sense (AQ3), and many of the respected objects seem to be cherished by all sensible people so that a sufficiently big set of respected objects could exist for which subject universality is fulfilled (AQ5). There are documented cases of moral conversion, where people, impressed by deep emotions of respect, radically change their moral convictions and subsequently act, even with strong engagement, on the basis of a moral that includes strong respect components (Weinreich-Haste 1986, 396). It is quite plausible that the new strong moral engagement is, among other factors, also sustained by a moral self-assessment that is based on the new moral convictions. This would show that the moral adoptability condition (AQ4) is also fulfilled for respect optimization.

Respect is mostly prohibitive and conserving. It protects things that already exist, it does not aim at creating things that should be respected in the future. The personal value function stemming from respect and, consequently, the moral value function defined via the respect optimizing dimension gives a high positive value to the persistence, to the proper-functioning and to autonomous actions of the respected objects. According to sympathy optimizing, the positive well-being of sentient beings has a positive value, whereas negative well-being has negative value, and both values increase monotonously as well-being increases. However pity is a stronger emotion than positive sympathy. Therefore, well-being and sympathy are not linearly correlated. Improving the situation of someone who is badly off contributes more to sympathy improvement than an equally great improvement of the situation of someone who is well off. Therefore, the value function in the sympathy optimizing dimension and, accordingly, the moral desirability defined via this dimension are prioritarian: they give more value to improvements for beings who are worse off.[23]

Some motives that are very important motives for acting in accordance with morals but, according to the strategy developed in the foregoing section, are not suitable as the basis for defining 'moral desirability' are the following. The objects of *love* and *friendship* are too specific and thus do not fulfil the condition of subject universality (AQ5). The same holds for *creative expansion*, i.e. the desire to produce works, in the broadest sense, with which one wants to enrich and shape the world beyond one's own existence; although creative expansion is often very altruistic it is bound to a certain project chosen by the subject. *Identification with* and pride in being part of *a collective* one identifies with provides the desired personal satisfaction only if the collective is not all-embracing; such identification needs the contrast with other collectives. So this motive again does not fulfil the condition of subject universality (AQ5) – if the "universe" is ambitiously chosen.

So we have identified sympathy optimizing and respect optimizing as the 'moral desirability' defining dimensions of the individual comprehensive desirability function. We have identified internal and external sanctions as moral amplifiers. And we have identified a set of strong and altruistic but particularistic motives that sustain moral action.

[23] The moral desirability function based on sympathy optimizing has been determined much more exactly and quantified: Lumer <2000>/2009, sects. 7.2-7.3 (= pp. 589-632).

7 Putting Moral Norms into Effect

There is still one, big open question with regard to the strategy adopted here for reconciling rationality and morality. Part of the strategy was to use external sanctions as an amplifier of moral requirements; e.g. the adequacy condition subject universality (AQ5) for defining 'moral desirability' has been chosen among others for this purpose. This however requires that the definition of 'moral desirability' filtered out by the adequacy conditions, in particular by subject universality, with a certain probability and at least in the long run be used as a basis for installing new social norms, which then oblige the subjects to contribute to at least some improvements in terms of the just defined 'moral desirability'. The implementation of social norms implies that the norm is propagated and that sanctions are announced and executed; someone has to do that. Let us call such actions "*norm sustaining actions*". In addition to norm sustaining actions, *norm political effort* is necessary to put new – which? – norms into effect. How can morally good social norms, first, be put into effect and, second, sustained? And will the actions necessary for this aim be rational? Let us call these two problems the "*norm realization problem*".

The norm realization problem really seems to be a *problem* because of at least two difficulties. 1. If there is already a motivation problem in the first place in following moral requirements that diverge from self-interest and if this problem shall be resolved, among others, by the help of external sanctions deriving from social norms, how can it be rational to support politically and later to sustain moral – and not self-interested – norms? 2. In economics of policy there is a familiar problem that ordinary people's own contribution to putting new policies into force, to installing a new government by voting etc. is marginal; this holds in particular for general elections where one's own vote, except for extremely rare cases where there is just a one vote majority, will not change the result. How can it ever be rational to sustain any policy at all, and even more so, a morally good policy?

I cannot offer a theory to resolve these two problems, but only some observations. In any case however, it is reassuring that implementing moral norms seems to work in some way, even if very slowly and with relapses. What we do not know in detail is only whether this kind of moral policy is rational and how we could make it more effective and efficient. In addition, the norm realization problem may seem larger than it is.

1. The efforts required by morals are not heroic – no continuous moral optimizing is required – but are always a compromise between moral optimizing and pure self-interested actions.
2. A certain level of moral norms has already been realized; now, only further moral improvements are at stake.
3. The moral way can be an easy solution in cases of pure coordination games and constant-sum games. In such situations, from a pure self-interested standpoint there are many weak equilibria; adding moral considerations may just make the difference in favour of one strong and moral equilibrium.

4. Although committing oneself particularly to moral politics seems to add yet another issue to the well-known problem of marginal influence, it is actually the approach to a solution. If one wants to implement something socially, one has a chance of success only if these interests (be they egoistic or already moral) at least are couched in general terms and thus already close to moral terms. This necessity then develops its own dynamics in that what was only self-interest in disguise will probably, and rather soon, be unmasked as such with the consequence of eliminating any chance of realization. So new proposals tend to be realizable only if they are really close to some moral solution and do not only appear to be so. In addition, all this is done by public discussion, which, of course, leads to coordination and to improved individual reasons.

5. Actions required by moral obligations at least in part are quite different from norm political efforts and from sustaining norms. So the motivation required for the latter may be quite different too. First, moral obligations demand compliance in every situation and moment. For making norm political and sustaining action sufficiently effective, sporadic efforts, preferably when they are "cheap", may be sufficient. Second, the nature of the actions required is quite different. Speaking and voting in favour of new norms or sanctioning in favour of old ones is often less costly than following these norms. Third, new motives come in, in particular motives of politicians. Politicians often – depending on the political system – may have more advantages from successfully putting through some policies near to morals (advantages in terms of acknowledgement, of being re-elected etc.) than from trying to put through immediate self-interests.

6. Strong informal norms, which rely on sanctions administered by anybody who is interested in doing so, are more and more difficult to be implemented and sustained as the collective increases in size, and as more exchange of social positions takes place within it. This has led to the disappearance of many informal norms. On the other hand, at the same time strong executives have developed, and this makes it relatively easy to introduce norms in one go.

These features have led to a process of moral norm implementation with the following characteristics. The fundamentally good news is that moral norms are and have been put through socially, thus providing subjects with strong additional motivation to act morally and, consequently, morally improving social actions. However, the resources to politically put through new moral norms are rather limited. A relatively strong force in their favour can be the social subjects' self-interests if they are coordinated for cooperation and lead to Pareto improvements. But the criticisms brought forward in section 2 show that self-interested cooperation often falls short of being morally good. The only reliable resource for putting through precisely moral norms is the group of moral motives itself (SR, I), which in favourable moments get support from other motives, socially harmonize and are coordinated in campaigns to establish a new social norm with moral content. Once such a norm is effective its maintenance requires much less energy so that a new moral norm can be put through on the basis of the freshly available excess capacities of moral dedication. This leads to a process of morally improving and strengthening the stock of moral norms, in historical dimensions however and not without relapses. This progress notwithstanding, the resulting complete set of moral norms at any time is far from requiring the subjects to morally optimize (as do many forms of utilitarianism, though verbally only). This is because, on the individual level, the new socially valid norms are always a compromise of moral dedication and self-interest and, socially, they arise from a power play between morally progressive and other forces. The blend of individual motives and social forces out of which the single norms grow also leads to many socially valid norms not being moral at all; socially valid norms that are moral differ from amoral or immoral norms only with respect to the fact that the former, according to the moral desirability function, are morally good and better than the latter. Again, the decisive force for targeted moral progress, which also includes repeal of immoral or morally bad norms, is the group of moral motives.

The six features listed, along with the mechanisms mentioned in the just told intuitive story make it rather probable that there is also a game theoretic solution to the norm realization problem, which explains this or a similar story. However, such a precise game theoretic reconstruction is still missing.[24] It would be the copestone for the strategy sketched here to bring moral and rational action into line by determining a rational and motivationally effective moral desirability function as well as by enforcing actions on its basis by social norms. However, we have made progress by developing a strategy to resolve the rationality and motivation problem of moral acting, in determining the moral desirability function with the help of multi-attribute utility theory, and by gathering some features and sketching an outline of the still missing game theoretical solution to the realization problem.[25]

References

Arrow, Kenneth J. (<1951>/1963): Social Choice and Individual Values. New York: Wiley 1951. 2nd edn.: New Haven, CT: Yale U.P. ^{2}1963.

Axelrod, Robert (1984): The Evolution of Cooperation. New York: Basic Books.

Baliga, S.; T. Sjöström (2008): Mechanism Design. Recent Developments. In: Lawrence E. Blume; Steven N. Durlauf (eds.): The New Palgrave Dictionary of Economics. Second Edition. Basingstoke, Hampshire; New York: Palgrave Macmillan. Vol. 5.

[24] At a first glance, implementation theory (e.g. Baliga & Sjöström 2008; Corchón 2007) may seem to provide the desired solution because its aim is to design mechanisms that implement a social choice rule (which result shall be realised from a social point of view in which situation?) in such a way that even the non-cooperative strategies of the players via carefully designed institutions (incentives, taxes, warranty for contracts etc.) and game theoretic equilibria lead to the corresponding social state designated by the social choice rule. More carefully considered, however, implementation theory does not provide the desired game-theoretical reconstruction of establishing morally good social norms: Implementation theory is a device for designing social institutions from the standpoint of a social planner; as such it presupposes the social choice rule and the mighty social planner and, thereby, the main parts of the solution to the norm installation problem, namely the problem to get a sufficient number of people engaged in putting through the same morally ambitious norm in a coordinated manner.

[25] I would like to thank three anonymous referees for their valuable comments.

Batson, Daniel C.; Karen O'Quinn; Jim Fultz; Mary Vanderplas; Alice M. Isen (1983): Influence of Self-Reported Distress and Empathy on Egoistic Versus Altruistic Motivation to Help. In: Journal of Personality and Social Psychology 45. Pp. 706-718.

Binmore, Ken (2005): Natural Justice. Oxford [etc.]: Oxford U.P.

Camerer, Colin [F.]: Individual Decision Making. In: John H. Kagel; Alvin E. Roth (eds.): The Handbook of Experimental Economics. Princeton, NJ: Princeton University Press 1995. Pp. 587-703.

Clemen, Robert T.; Terence Reilly (2001): Making hard decisions with Decision Tools. Duxbury: Pacific Grove.

Corchón, Luis C. (2007): The Theory of Implementation. What did we learn? Working Paper 08-12, Economic Series (07), December 2007. Madrid: Universidad Carlos III de Madrid, Departamento de Economía 2007. 40 pp. Web: <http://e-archivo.uc3m.es:8080/bitstream/10016/2172/1/we081207.pdf> (23.10.09.)

Edwards, Ward (1977): How to Use Multiattribute Utility Measurement for Social Decision Making. In: IEEE Transactions on Systems, Man, and Cybernetics, SMC-7. Pp. 326-340.

Fehr, Ernst; Simon Gächter (2001): Fairness and Retaliation. In: L. Gerard-Varet; Serge-Christophe Kolm; Jean Mercier Ythier (eds.): The Economics of Reciprocity, Giving and Altruism. (International Economic Association (= Iea) Conference Volumes, Vol 130.) Basingstoke: Macmillan. Pp. 153-173.

Figueira, José; Salvatore Greco; Matthias Ehrgott (eds.) (2005): Multiple criteria decision analysis. State of the art surveys. New York: Springer.

Fishburn, Peter C. (1967): Methods of estimating additive utilities. In: Management Science 13. Pp. 435-453.

Flügel, J. C. (1925): A Quantitative Study of Feeling and Emotion in Everyday Life. British Journal of Psychology 15. Pp. 318-355.

French, Simon (1986): Decision Theory. An introduction to the Mathematics of Rationality. Chichester, West Sussex, England; New York: Ellis Harwood, Ltd.; Halsted Press.

Gauthier, David (1986): Morals by Agreement. Oxford: Clarendon.

Gauthier, David (1991): Why Contractarianism? In: Peter Vallentyne (ed.): Contractarianism and Rational Choice. Essays on David Gauthier's Morals by Agreement. New York: Cambridge U.P. Pp. 15-30.

Harsanyi, John C. (<1955>/1976): Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. In: Journal of Political Economy 63 (1955). Pp. 309-321. – Reprinted in: Idem: Essays on Ethics, Social Behavior, and Scientific Explanation. Dordrecht; Boston: Reidel 1976. Pp. 6-23.

Harsanyi, John C. (1977): Rational Behavior and Bargaining Equilibrium in Games and Social Situations. Cambridge: Cambridge U.P.

Hoerster, Norbert (2003): Ethik und Interesse. Stuttgart: Reclam.

Izard, Carrol E. (1991): The Psychology of Emotions. New York; London: Plenum Press.

Kahneman, Daniel; Amos Tversky (1979): Prospect theory. An analysis of decision under risk. In: Econometrica 47. Pp. 263-291.

Keeney, Ralph L.; Howard Raiffa (<1976>/1993): Decisions with Multiple Objectives. Preferences and Value Tradeoffs. New York [etc.]: Wiley 1976. Reprint: Cambridge: Cambridge U.P. 1993.

Koehler, Derek J.; Nigel Harvey (eds.) (2004): Blackwell Handbook of Judgment and Decision Making. Oxford: Blackwell.

Krantz, David H.; R. Duncan Luce; Patrick Suppes; Amos Tversky (1971): Foundations of Measurement. Vol. I: Additive and Polynomial Representations. New York; London: Academic Pr.

Lazarus, Richard S. (1991): Cognition and Motivation in Emotion. In: American Psychologist 46. Pp. 352-367.

Lazarus, Richard S.; Bernice N. Lazarus (1994): Passion and Reason. Making Sense of Our Emotions. Oxford: Oxford U.P.

Lumer, Christoph (1997): The Content of Originally Intrinsic Desires and of Intrinsic Motivation. In: Acta analytica. 18. Pp. 107-121.

Lumer, Christoph (1998): Which Preferences Shall Be the Basis of Rational Decision? In: Christoph Fehige; Ulla Wessels (eds.): Preferences. Berlin; New York: de Gruyter. Pp. 33-56.

Lumer, Christoph (<2000>/2009): Rationaler Altruismus. Eine prudentielle Theorie der Rationalität und des Altruismus. Osnabrück: Universitätsverlag Rasch 2000. 2nd, complemented edition: Paderborn: mentis 2009.

Lumer, Christoph (2002): Motive zu moralischem Handeln. In: Analyse & Kritik 24. Pp. 163-188.

Lumer, Christoph (2002/03): Kantischer Externalismus und Motive zu moralischem Handeln. In: Conceptus 35. Pp. 263-286.

Mackie, John Leslie (1977): Ethics. Inventing Right and Wrong. Harmondsworth: Penguin.

Margolis, Howard (1982): Selfishness, altruism, and rationality. A theory of social choice. Cambridge: Cambridge U.P.

Montada, Leo (1993): Moralische Gefühle. In: Wolfgang Edelstein; Gertrud Nunner-Winkler; Gil Noam (eds.): Moral und Person. Frankfurt, Main: Suhrkamp. Pp. 259-277.

Narveson, Jan (<1988>/2001): The Libertarian Idea. Philadelphia: Temple U.P. 1988. – Reprint: Peterborough, Ont.: Broadview 2001.

Narveson, Jan (2010): The Relevance of Rational Decision Theory for Ethics. In: Ethical Theory and Moral Practice. This issue.

Nunner-Winkler, Gertrud (<1993>/1999): Moralische Motivation und moralische Identität. Zur Kluft zwischen Urteil und Handeln. (1993.) In: Detlef Garz; Fritz Oser; Wolfgang Althof (eds.): Moralisches Urteil und Handeln. Frankfurt, Main: Suhrkamp 1999. Pp. 314-339.

Oliner, Samuel P.; Pearl M. Oliner (1988): The Altruistic Personality. Rescuers of Jews in Nazi Europe. New York: Free Press; London: Collier Macmillan.

Raiffa, Howard (1968): Decision Analysis. Introductory Lectures on Choices Under Uncertainty. Reading, MA: Addision-Wesley.

Rawls, John [B.] (<1971>/1999): A Theory of Justice. Cambridge, Mass.: The Belknap Press of Harvard U.P. 1971. Revised ed.: Oxford [etc.]: Oxford U.P. 1999.

Scheffler, Samuel (<1982>/1994): The Rejection of Consequentialism. A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions. (1982.) Revised edition. Oxford: Clarendon ^{2}1994.

Sen, Amartya K[umar] (<1970>/1984): Collective Choice and Social Welfare. San Francisco: Holden-Day 1970. Reprint: Amsterdam; New York; Oxford: North-Holland 1984.

Solomon, Robert C. (<1976>/1993): The Passions. Emotions and the Meaning of Life. (1976.) Revised edition. Indianapolis; Cambridge: Hackett 1993.

Stemmer, Peter (2000): Handeln zugunsten anderer. Eine moralphilosophische Untersuchung. Berlin; New York: de Gruyter.

Taylor, Gabriele (<1985>/1995): Shame, Integrity, and Self-Respect. (1985.) In: Robin S. Dillon (ed.): Dignity, Character, and Self-Respect. New York; London: Routledge 1995. Pp. 157-178

Trapp, Rainer Werner (1998): The Potentialities and Limits of a Rational Justification of Ethical Norms. Or: "What Precisely is Minimal Morals?". In: Christoph Fehige; Ulla Wessels (eds.): Preferences. Berlin; New York: de Gruyter. Pp. 327-360.

Watson, Stephen R.; Dennis M. Buede (1987): Decision Synthesis. The principles and practice of decision analysis. Cambridge [etc.]: Cambridge U.P.

Weinreich-Haste, Helen (1986): Moralisches Engagement. Die Funktion der Gefühle im Urteilen und Handeln. In: Wolfgang Edelstein; Gertrud Nunner-Winkler (eds.): Zur Bestimmung der Moral. Frankfurt, Main: Suhrkamp. Pp. 377-406.

Williams, Bernard (1973): A critique of utilitarianism. In: J[ohn] J[amieson] C[arswell] Smart; Bernard Williams: Utilitarianism for and against. Cambridge: Cambridge U.P. Pp. 75-150.

Zeelenberg, Marcel; Rob Nelissen; Rik Pieters (2008): Emotion, Motivation, and Decision Making. A Feeling-Is-for-Doing Approach. In: Plessner, Henning; Cornelia Betsch; Tilmann Betsch (eds.): Intuition in Judgment and Decision Making. New York: Lawrence Erlbaum 2008. Pp. 173-189.