*First publ. in: Croatian Journal of Philosophy 2 (2002), 6, pp. 249-264*

# The Many Facets of the Theory of Rationality

*Wolfgang Spohn*

*Department of Philosophy*
*University of Konstanz*
*78457 Konstanz*
*Germany*

## 1. Introduction

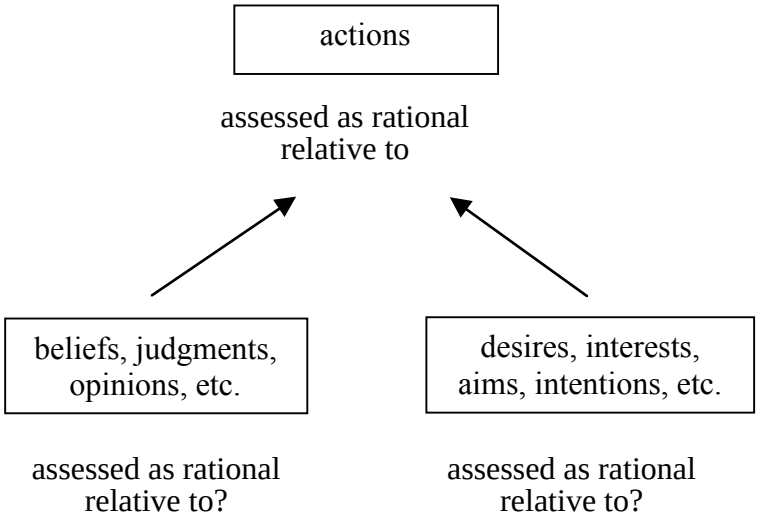
Modern theory of rationality has become large and rich. The search for the most general principles is driven forward as much as the countless specializations in countless branches. Often, the questions lie far apart. The methods to answer them are often disparate and none of the questions is exhausted. The theory of rationality has truly grown into a science of its own. Many details have become so special that the philosophical relevance is lost. It is, however, evident that the general topic is genuinely philosophical.

This essay is concerned with giving a brief overview over the theory of rationality. Section 2 explains the common and an improved version of the fundamental scheme of all rationality assessments. With its help, we can give a schematic order of the main questions concerning the theory of rationality, and we shall find that some questions are addressed by a wealth of literature, whereas other questions find astonishingly little response. Section 3 discusses the fundamental issue that the theory of rationality seems to be both a normative and an empirical theory. Section 4, finally, shows how the unity of the theory of rationality can nevertheless be maintained. The purpose of this essay is to serve as a kind of guide for the reader.[1]

## 2. The Fundamental Scheme of Rationality Assessments

Any map of the theory of rationality must start with a *fundamental scheme* showing what we assess as rational with respect to what. Usually, the scheme looks like this:

[1] In Spohn (1993), I have given a similar overview with similar core assertions. It is about twice as long. Therefore much is presented in more detail there and hence better comprehensible. Some things, however, are more muddled and others are not up to date.




This scheme expresses that it is foremost a subject's actions which are assessed as rational, relative to her empirical beliefs or judgments about the happenings in the world, and relative to her interests, desires, values etc., and thus according to her subjective standards. Arguably, these beliefs and desires can themselves be assessed as rational. The relevant subjects are usually human beings or perhaps also higher developed animals – this should not be dogmatically excluded – and perhaps even groups, organisations or institutions, insofar as we can assign opinions and aims to them as wholes.[2]

This form of rationality figures under many more or less suitable terms: it is called 'instrumental rationality' – which is perhaps the most appropriate term term – or ends-means-rationality – which I avoid since the scope and complexity of this form of rationality assessment is severely underestimated by this term – or local rationality – a somewhat derogatory term – or strategic rationality – which should better be reserved for a more restricted technical meaning. The list is certainly not complete. For instance, what is nowadays called belief-desire psychology, which is the conventional and usually quite rationalistically designed philosophical psychology, is covered by these terms as well.

Modern decision theory is clearly *the* fundamental (although certainly not sacrosanct) theory working out this scheme. In the past century it has, particularly in the hands of economists and statisticians, developed into a most ramified and

[2] How it is possible to aggregate the wishes or interests of groups from the wishes and interests of their members is object of the theory of social choice or theory of group decision; cf. e.g. Kern, Nida-Rümelin (1994).

elaborate theoretical edifice.[3] There were a few early philosophers who observed and advanced these developments. But only with Jeffrey (1965) decision became again a large topic in philosophy, which has since contributed an enormous amount to decision theory.

According to the above scheme, it is not only the subject's actions, but also her beliefs and desires which can be assessed as rational. I have not yet said anything about the latter – because the scheme needs modification. In my opinion, actions do not belong to the primary objects of rationality assessments. It is rather decisions or intentions etc. (which need not be understood as conscious mental acts – this would clearly overpopulate our mental life – but may well be implicit or dispositional). The reason is that there is a causal path or mechanism – which may or may not work – leading from the intention to the action which cannot by itself be assessed as rational. This point is obfuscated by the fact that by calling a piece of behavior an action we already imply that this behavior is appropriately caused by an appropriate intention – which conceals, but does not eliminate the difference between actions and intentions.[4]

If this point is recognized, only two *primary* kinds of objects of rationality assessment remain. But apart from them, there are many, as I call them, *secondary* kinds of objects to be assessed as rational. These are all the objects which lie within the realm of influence of the primary objects. Amongst them are quite directly the subjects' actions, but also their outcomes. A knife can be unreasonable insofar as it is cumbersome or dangerous to handle and hence does not accord with our interests. The placement of the goods in a supermarket can, from the manager's point of view, be rational insofar as it stimulates the consumer and hence enhances the gross product. A law can be rational insofar as it furthers its aims. And so on. According to this view, even emotions are secondary objects of rationality assessments. For instance, irrational views concerning the way of preserving relationships might govern a certain specimen of exaggerated jealousy; in this respect the jealousy is irrational, too. But this is a philosophically problematic point, on which I do not wish to dwell.[5]

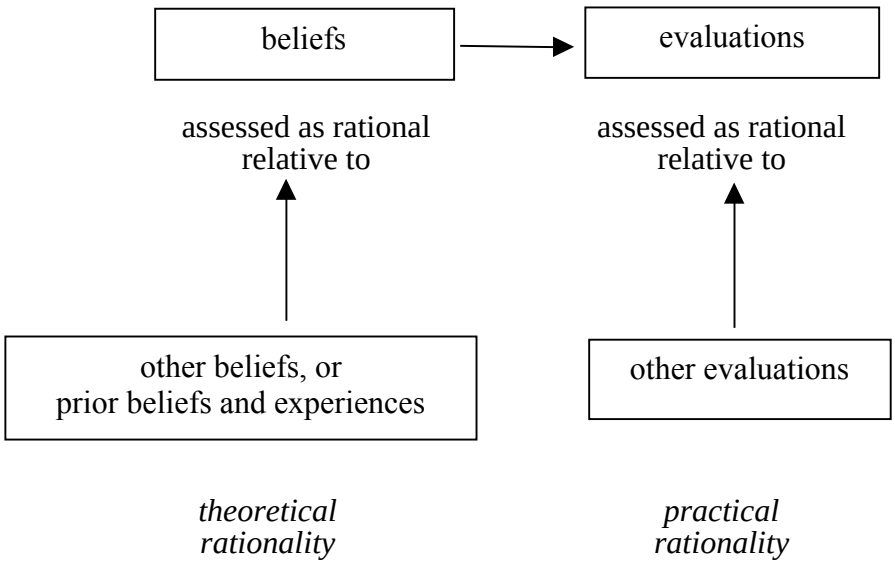
---

[3] The beginnings of decision theory reach back to the 17th century. For the contemporary discussion, Savage (1954) is doubtless the most influential work. Comprehensive and actual information can be found in Camerer (1995). Beautiful text books are Raiffa (1973) and Resnik (1987).

[4] Of course, these assertions already assume a certain position concerning the foundations of the philosophy of action, namely the causalist position which seems dominant in the philosophy of action since Hempel (1961/62) and the essays of Davidson (1980, 1-5).

[5] See, e.g., DeSousa (1987) or Nussbaum (2001).

Let's rather turn to the primary objects; they are our primary concern. On the one hand, there are the opinions, judgments or beliefs – which I always understand as empirical beliefs about how the actual world is; the word "belief" is certainly broader in ordinary language. Note that beliefs usually have degrees of strength which can be modelled, e.g., as subjective probabilities.[6] On the other hand, we find a mixed bag of wishes, aims, ends, interests, norms, utilities or utility functions, intentions, and similar things more. Here, I will uniformly speak of (the subject's) evaluations – which is neutral, indicates that our evaluations may have many sources, not only individual self-interest, and makes clear that evaluations are weighed and come in degrees. In particular, action intentions are of this sort; the intention to perform a certain action is the conclusive over-all evaluation of this action as better than or at least as good as the possible alternatives. The uniform term "evaluation" gives voice to the hypothesis that there is also a uniform theory for our evaluations. This hypothesis can, of course, be criticized, but then it is the task of the critic to present a theory in which different roles are assigned to the different sorts of evaluations. No such theory is known to me; hence I shall, for the time being, stick to this hypothesis. Thereby we obtain an improved scheme of primary rationality assessments:




[6] However, in the meantime there exists an astonishing number of alternatives to the probabilistic paradigm (see, e.g., Gabbay et.al. 1994).

Let us consider first the left side which represents theoretical or epistemic rationality. It contains two forms of rationality assessments of beliefs and degrees of beliefs, a static and a dynamic one. The static assessment consists, e.g., in assessing one of my present beliefs in the light of my other present beliefs. Here, my beliefs are examined for consistency and coherence. Consistency is a clear but very weak criterion[7], whereas coherence is a strong but very unclear criterion, which can be explicated in very different and for different reasons problematic ways.[8]

The dynamic assessment consists of assessing one of my present beliefs in the light of past beliefs and experiences which I had in the meantime. Here, it is examined whether I arrived at my present belief in a rational manner. There is a rich and highly interesting probabilistic theory elaborating the dynamic issue, in which various principles like conditionalization rules, the principle of minimizing relative entropy or van Fraassen's reflexion principle play key roles.[9] The non-probabilistic theorizing about the dynamic issue which has vigorously developed in the last 20 years is perhaps even more interesting.[10] And if in this scheme beliefs and experiences are replaced by theories and data, then it is obvious how deeply one can digress at this point into the philosophy and history of science and into questions of the dynamics and the confirmation of theories.[11]

Let's turn now to the right side which represents practical rationality. We were dealing with it already when talking about the assessment of the conclusive overall evaluation of actions which derives from our evaluations of things beyond actions and our beliefs how likely we are to realize our values by the actions. But this form of an instrumental rationality assessment holds for evaluations in general. In planning a holiday trip, to use a trivial example, we evaluate the alternative destinations with respect to their probable satisfaction of the basic values for holidays, and this evaluation of destinations enters itself into the evaluation of specific alternative holiday plans. In this respect, my improved scheme embodies not only conventional action rationality, but value rationality in general.

However, this kind of value rationality is still a relative rationality and hence not what is commonly understood under value rationality; I shall return to this point. Presently it is important to see that usually both, beliefs and further evalua-

---

[7] It is as much or, rather, as little disputed as deductive logic.

[8] See, e.g., Bartelborth (1996, in particular ch. IV, V and IX) and also Spohn (1999b).

[9] See, e.g., Earman (1992), Hunter (1991), Hild (1998, 2002), Maher (1993), Spohn (1999a), and van Fraassen (1984, 1995). The reflexion principle may already be found in Spohn (1978, p.162).

[10] See, e.g., Gärdenfors (1988), Spohn (1988), and Gabbay et al. (1994).

[11] See Stegmüller (1973a, b).

tions, have to be consulted for the assessment of the rationality of a given evaluation. In this respect, it is appropriate to say – with Rott (2001, ch.6) and many others – that the theory of practical rationality embraces the theory of theoretical rationality; the former involves both, evaluations and beliefs. But I prefer it to say that the theory of practical rationality presupposes the theory of theoretical rationality. There is something like epistemic decision theory.[12] But these attempts appear to me unsuccessful; it should be conceded that the theory formation in this realm of theoretical rationality is of a different kind than in the realm of practical rationality – which is obscured by speaking of the latter as embracing the former.

So far, the assessment of the rationality of evaluations contained in my scheme is of a static nature. Is there, in analogy to the epistemic side, also a dynamic assessment? This is a most interesting question. Of course, changes of our evaluations are entailed by changes of our beliefs; this is trivial. But the question is whether a genuine change of our desires or evaluations can be assessed as rational – which would allow rationality assessments of our desires with respect to their genesis. As far as I can see, the literature has great difficulties with this question.[13]

Above, I have emphasized the relativities of rationality assessments. This suggests the question whether and how one might obtain a more absolute point of view. This question opens up many highly interesting perspectives. It is obvious that we call a thing (belief, evaluation) rational in the absolute sense, if it is rational relative to other things which are rational in turn. However, this presupposes an absolute concept of rationality; the absolute rationality assessment is thereby only shifted and not established. Can we do more than shift the question?

We can. First, consider the theoretical side. In a first step, the shift refers us first to experiences and prior beliefs. That question of the rationality status of experiences leads us further to the fundamental epistemological debate between foundationalism and coherentism: foundationalists look upon experiences as givens from which the justifications of other beliefs take their departure. The things given, however, are neither justifiable nor in need of justification. Coherentists, on the other hand, deny experiences this special status.[14] The question of the ration-

---

[12] This starts with Levi (1967) and is still pursued, e.g., by Levi (1991) himself and by Maher (1993).

[13] See, e.g., the introduction of Millgram (2001) and within this volume in particular the contributions by Schmidtz, Kolnay, and Wiggins but also Fehige (2000, in particular ch.5 and 6) and Kusser, Spohn (1992), Fehige (2001). The question how to behave rationally under the expectation that one's utilities or evaluations will change is treated in dynamic decision theory (cf. McClennen 1990) and in reflexive decision theory (cf. Spohn 2002b).

[14] See, e.g., Bartelborth (1996) and Spohn (1997/8).

ality status of prior beliefs, by contrast, refers us ever further back to the question whether there are beliefs or judgments a priori from which all belief formation departs. If this question is answered in the positive there remains the question which beliefs and judgments have this special status. This question, too, opens up a wide field of discussion, which is, alas, in a quite unsatisfactory state.[15]

How does the practical side fare in this respect? Here, the question of rationality is first deferred from evaluation to evaluation and finally arrives at intrinsic evaluations, ultimate aims or values and ends in themselves. Do they exist? Yes, of course; everybody must have them – at least according to her own standards; not everything can be valuable only with respect to other things. These values in themselves should not, of course, by mixed up with absolute values in the sense that they cannot be weighed against and balanced out with other values; it is doubtful whether there are any such values.[16]

Are there general propositions about intrinsic evaluations? Yes, without doubt. There is at least the hedonistic insight, that my sensations, feelings, moods, etc. are ipso facto evaluated by me; the traditional pleasure/pain scale was supposed to represent this. However, the stronger hedonistic claim that my intrinsic evaluations are thereby exhausted should be rejected – especially since it is obscure how further evaluations are obtained from the evaluation of pleasure and pain.[17] There is, though, the interesting idea that other persons' feelings are also intrinsically evaluated by me in a concordant way.[18] It is clear that we are again moving on most difficult ground, all the more as the question is not only which intrinsic evaluations we have, but also in how far they can be characterized as rational; this is the problem of value rationality as it is traditionally understood. The problem is accompanied by the additional difficulty that it easily changes into the question of objective values, which are usually understood as objective moral standards. At this point we quickly slide into the problem of justifying morality. If morality could be reduced to rationality, objective values could be shown to this extent to be rational; if not, the problem of value rationality remains open.

---

[15] How much the concept of the a priori is under dispute – without any result – can, e.g., be seen in the collection of Casullo (1999).

[16] Human dignity or human life are often presented as absolute values in this sense. But there are, of course, conflicts in which life stands against life, and then a comparison *must* be made. Moreover, in traffic policy, e.g., life is permanently weighed against money with public consent.

[17] See again Kusser, Spohn (1992) and Fehige (2001).

[18] See Fehige (2002), who defends the thesis that we a priori have such evaluations, i.e., a priori empathy.

Certainly, I won't solve the problem here. But I hope to have carried out my cartographer's task so far in an illuminating way. Have I even completed it? Two kinds of doubt are well entertainable.

The first doubt is a rather internal one. So far I have just explained (with a metatheoretical intention) which kinds of rationality assessment there are. But the theory of rationality gets its content only from making specific claims about each kind of assessment. If the theory is built up in this way, then various conflicts might result: the theory might mark many things as rational which are intuitively considered to be irrational, and vice versa. One of the most infamous examples is certainly that cooperation in the single-shot and in the finitely iterated prisoners' dilemma is irrational according to the standard theory, but not according to our intuitions. Similar doubts prompt, for instance, Nida-Rümelin (2001) to oppose local rationality – as he calls what he considers to be the standard form of instrumental rationality – and to propose structural rationality as an alternative. This prompts two questions: first, what precisely the standard theory is and whether its resources are indeed exhausted, and second, what the alternative theory of structural rationality really looks like. In both respects, structural rationality stands on shaky grounds.[19] The dissent is, however, rather of an internal kind, insofar as the categorical frame of our scheme above is not disputed in principle.

The external sort of doubt disputes this frame. The main example I am thinking of here is Habermas' communicative rationality which he wants to add to the epistemic and the instrumental or strategic rationality as a third pillar[20] and which seems to explode my scheme above. Here we meet a very complex dialectic situation. Certainly it is not obvious from the outset that we can subsume the rationality of a speaker's linguistic utterances under general action rationality; likewise, it is not obvious whether the role of the audience can be understood within the model of epistemic rationality. On the other hand, so many illuminating and promising things have been said by the standard theory on the rationality of communication[21] that the case for a fundamental supplementation remains weak. In particular, the notion of evaluation as explained above is wide enough to include

---

[19] Cf., e.g., Fehige (2000, ch. 5 and 6) to the issue how a subject's punctual standpoint can incorporate his (alleged) past and future desires and hence how it can cast off its punctuality. Concerning the prisoners' dilemma, in particular, there are suggestions from Kreps et al. (1982) and Spohn (2002b) for the rationalization of cooperation, which are well within the standard theory. On the other hand, the structural rationality is in a quite programmatic condition, still.

[20] See Habermas (1982) and most recently (1999, ch. 2).

[21] See the extensive literature initiated by Grice (1957) and Lewis (1969) and discussed by Habermas – though, of course, critically in part.

both, the use of other people to egoistic ends and the respect for other persons as values in themselves, which is so important to Habermas.[22] Finally, we have the problem that the standard theory and the alternative theory are in such an incomparable state of elaboration that already from this side the dialectic situation is hard to judge.

I do not want to built up dogmas. But I very much want to recommend to immanently test the power of the above scheme and the theorizing which stands behind it. Before this is done, claims of incompleteness cannot be properly demonstrated.

## 3. The Theory of Rationality as a Normative and as an Empirical Theory

After this short survey of the landscape of the theory of rationality we must ask, still with metatheoretical intentions, what sort of theory the theory of rationality is. Of which kind are its claims? The answer is not clear. But the first and best answer is that they are normative claims within a *normative theory*. I am criticized for my past irrationalities. When I permanently speak and act irrationally, people stop listening and taking me serious. And the question "what shall I do now?" is – if it is not a moral question – tantamount to "what is best to do?" or "what shall I rationally do?". The theory of rationality is created in normative discourse.

But how do we arrive at judgments about what is rational? It is obviously problematic to say that we strive for truth in this connection. But we accept certain normative propositions and reject others; and this acceptance and rejection is not different in kind from the case of empirical judgments. In the latter we collect data and build theories which permit us to ask questions about further relevant data. We modify the theories accordingly until theories and data fit together – though it is not always the theories which are adjusted to the data; it can also be the other way round. Moreover, there are not only the level of data and the level of theory; there can be more levels with different degrees of generality. How judgment formation and justification of theories work in detail is a very complicated matter about which philosophy of science and epistemology have a lot to say. In all its vagueness, the vivid metaphor of the reflective equilibrium is appro-

[22] Indeed, often a much too narrow concept of evaluation is read into to the standard theory. This is an important, but certainly not the only, reason why it is constantly underestimated.

priate here; the totality of opinions is subjected to changes until finally the internal tensions are minimized and the coherence is maximized.

The metaphor of the reflective equilibrium is no less apt for normative theorizing. There are, as it were, data, too, elementary data which concern individual cases, e.g. that it is – under normal conditions – not reasonable to start shouting very loudly and that something must therefore be wrong with a theory if it marks such shouting as rational. A more interesting claim, which all of us endorse, is that it would be unreasonable – as long as we care about money – to accept bets in which we lose in any case whatsoever. The famous so-called Dutch Book argument attempts to derive from this and some seemingly innocent assumptions that your degrees of belief must rationally obey the laws of mathematical probability.[23] This result can be accepted, or the additional premises can be doubted, or one might seek to undermine the normative premise, etc. To give another example: games in the sense of game theory can be represented in so-called extensive form and normal form and many are inclined to postulate that these are equivalent representations. This is already a rationality postulate. But there are counterexamples which seem to show a relevant difference. What now? Some attempt to dissolve the power of these counterexamples, others claim that we can stick to the equivalence of extensive and normal form at least in the context of decision theory, etc.[24] And so the normative discourse develops, ramifies, and becomes more and more sophisticated.

The drive for systematization plays an all-important role in this process. Without the utmost attempt to systematize all the many rationality claims neither coherence nor reflective equilibrium would ensue – only a mess. The successful attempts at systematization are so far most noteworthy but in no way sufficient. Quite a lot remains outside the scope of the two main theories of practical rationality, decision theory and game theory – not necessarily so, but in any case in their present states. And even the relationship between decision theory and game theory is not sufficiently clear. The agenda is long, the field of activity rich and interesting and the search for a normative reflective equilibrium is always in a transitory state; it can never be taken to be completed.

However, the situation becomes even more complicated through the fact that the theory of rationality is not only a normative theory. We permanently use it as

---

[23] Cf. Stegmüller (1973b, p.436ff.) for his standard version, Skyrms (1990, ch.5) for his dynamic version, and Hild (1998) for an interesting restriction.

[24] See, e.g., the discussion in Myerson (1991, ch.2, 3, 4) and McClennen (1990, ch.7, 11).

an *empirical theory* for the prediction and explanation of human behavior.[25] Why did the student leave my class just now? Because she wanted to call somebody at 11.30 and rather wanted to miss part of the lecture than to disturb it. And then certainly a longer story could be told about the reasons for her wishing to call somebody at exactly 11.30. In this way, I give a rational explanation of her behavior – an explanation in which the premise that the actions of the student are rationally guided by her preferences and beliefs is an essential part without which the explanation would have to remain incomplete. For the micro-economical reduction of macro-economics the assumption that the economic subjects are rational is central.[26] Virtually in all cases in which sociology or economics seek an individualistic approach, the individual theory of rationality is fundamental.[27]

Under this perspective the above scheme of rationality assessments and the relevant theories deal with the fundamental laws, ways of functioning or mechanisms of those propositional attitudes for which I have chosen the title "beliefs and evaluations". As the physicists explain how a TV works, what happens between aerial plug and screen so that such a beautiful picture results, the rationality theorist tells us how the human mind works and why this and that behavioral output of a person occurs after a given perceptual input. Of course, the rationality theorist does not have a complete theory for this. He cannot say anything about why the mind needs rest and sleep. Also memory and the details of the perceptual process are not his subject. And so on. Yet his domain is the central module of propositional attitudes.

Doubtlessly I present matters as much better than they actually are. The brain has not at all been investigated by the rationality theorist. Speaking of the module of propositional attitudes may be quite unwarranted. The theory of rationality as an empirical theory is indeed refuted in many ways. Psychologists and economists have constructed many situations and experiments in which human beings drastically diverge from the claims of decision theory and game theory. The negative results are astonishingly abundant here.[28] How, then, can we wish to stick to the theory of rationality as an empirical theory?

---

[25] This has been paradigmatically discussed already in Hempel (1961/62).

[26] This aim at reduction is, of course, the decisive motivation behind the very intensive and far-reaching attempt during the last decades (manifested, e.g., in Friedman 1991 and Ordeshook 1986) to reconstruct economics and related disciplines strictly on the basis of decision theory and game theory.

[27] See Coleman (1990).

[28] Cf., e.g., Kahnemann et al. (1982) and Rapoport (1989).

All these situations are, of course, in need of interpretation. Perhaps some things really demand a change in our theory of rationality, e.g. Allais' paradox[29], the prisoners' dilemma tournaments of Axelrod (1984) or the empirical investigation of the ultimatum game.[30] But there might also be possibilities of evading these problems. The standard manœuvre is to conceive of the theory of rationality as an idealization which holds only approximately. This is not a mere immunization manœuvre if we can specify the idealizations and can at least indicate pertinent correction theories with which reality is better modelled.

There are many attempts in this direction, in particular the rich literature which goes under the label 'bounded rationality'.[31] There the point is that the execution of our rational capacities makes use of our resources, and our temporal, computational, and motivational resources, and whatever else we need for deciding, are always limited. The theory of bounded rationality has much to say about our ways of dealing with these limitations.

The real reason why we respond to the many difficulties in the empirical use of the theory of rationality in this way is quite obvious. In a rough form, this theory is so pervasively at the bottom of our everyday practice that we would be totally at a loss if we had to dispense with it entirely. In particular, we want to understand ourselves as at least approximately rational. It simply cannot be that we are mostly irrational or that rationality is a category which can scarcely be applied to us.

Again, many kinds of arguments are at work here and again we can generally apply the metaphor of the reflective equilibrium to the result of the attempts to establish an empirical theory of rationality.

## 4. The Unity of the Theory of Rationality

This description seems to me fairly appropriate despite its overview character. Obviously I have described two very different theories, even two very different forms of theory. Why are both called 'theory of rationality'? How do both fit together?

At first sight, they do not fit together at all. On the one hand, it seems impossible to embody the normative theory within the empirical theory; the latter would then loose its empirical character. On the other hand, the empirical theory seems

---

[29] See Allais, Hagen (1979).

[30] See, e.g., Roth (1995) and Güth (2001).

[31] See, e.g., Rubinstein (1998).

to rob the normative theory of its point: if the laws of the movement of the mind and of reason are empirically determined and established by the empirical theory of rationality, which sense is there in normative discourse? It seems as useless as telling the stars what they should do. Both qualms are substantial; nevertheless they do not obstruct the unity of the theory of rationality.

First: If we demand rationality from our fellow men and discuss with them what is rational to do and to believe, we do not argue against natural laws. On the contrary, this is useful in two respects.

On the one hand, it can be useful to enforce the rational powers against various non-rational factors which may gain room. I have already said that the empirical theory of rationality can endure only as an idealization which is to be supplemented by correcting theories for various disturbing factors. Amongst them are: simple flaws in reasoning or computing, fatigue, habitualization, psychological problems, drugs, etc. The normative discourse, which may even reduce to shouting “pay attention!”, can have exactly the function of pushing back the non-rational factors.

The other purpose is at least as important. In order to understand it, we must first understand that there is indeed an unbridgeable gap between the normative and the empirical theory of rationality. This gap consists of the normative theory striving towards objectivity while the empirical theory cannot be adequate without the utmost relativization to a subject. What I have in mind here has indeed several dimensions:

If somebody rebukes my actions and beliefs as irrational, then he usually does not want to doubt their rationality on my subjective basis. He wants, perhaps, only to point out to me that it rests on a false belief and that I cannot stick to it as soon as I have recognized my error. Here the external rationality assessment works relative to the objective truths and not relative to my subjective beliefs. In so far as objective values can be characterized as rational, his reproach could also refer to my orientation along my unreasonable subjective values and not along the objective ones.

But even if my subjective standards are left untouched, it will often not be the case that I take into account everything that is relevant – even if relevance is taken according to my subjective standards. What is relevant even according to my subjective standards is potentially infinite. Usually only a small part of this enters my picture, and the rest is indeed often negligible, but sometimes it is not. My circumspection, my thinking-of-everything is simply limited – often at the cost of

my rationality. And the admonitions of other persons can be helpful to lessen my limitations.

Finally, if my beliefs and desires are evaluated as rational even with respect to my limited powers of circumspection, then this judgment still refers to the content of my beliefs and desires. Since, however, these contents are accessible to me not directly, but only via certain representations (often of a linguistic and often of a non-linguistic nature), the judgment has an objective rational component of which I need not have subjective insight. I mean here something very simple for which the problem of consistency is paradigmatic: Surprisingly you claim that I have inconsistent opinions and hence that I am irrational. How could that happen to me? If certain of my opinions are conjoined, I naturally hold this conjunction to be true as well. As it happens, the content of this conjunction is a logical falsehood, and this is what you criticize. The contradiction was, however, very well hidden, and hence I did not notice it (and perhaps I am not to blame for failing to notice it). The reason is simply that the contradiction is accessible to me in some representations and not in others.[32]

The general moral is as follows: the normative discourse always aims at such an objectivization; exactly because of this it is useful. And the normative theory always presupposes at least the last two objectivizations. Neither the power of circumspection nor the accessibility of the contents via representations is really a matter of rationality. In this respect, the normative theory cannot account for these subjective matters. An empirical theory about our thinking and behavior must, however, account for them. Exactly this is the unbridgeable gap: the normative theory must propose a normative ideal, which can only be stated to be insufficiently attained by the empirical theory.[33]

This resolves one of our two qualms and explains why the normative theory is not made obsolete by a fine-tuned empirical theory. But thereby we have not yet come closer to the unity of normative and empirical rationality. However, this seems to me to be not so obscure. We deal with quite a complex two-fold reflective equilibrium of theory formation. On the one hand, we form – as described above – our normative theory of rationality as best as we can. Then we feed this into the empirical theory as an idealized point of reference, which is to be sup-

---

[32] The literature on this problem of logical omniscience is endless; cf., e.g., Stalnaker (1999, ch.13, 14).

[33] Pollock and Cruz (1999, ch.4) attempt to bridge the gap by reducing the normative demands to what is subjectively satisfiable. This seems to me wrong as I have elaborated in Spohn (2002a).

plemented by correcting theories of many kinds, until a reflective equilibrium is found there, too.

The influence may certainly be mutual. Perhaps we do not find a good empirical reflective equilibrium without changing the normative reflective equilibrium, and then we accept the normative change rather than a bad empirical theory. To give an example: the empirical observation that in the case of the Allais' paradox many people digress from Savage's (1954) normative theory (and are not worried at all after getting explained their digression) could shake our normative trust in this theory. Nevertheless, it is usually the normative theory that wears the pants. For, without the normative discussion we would not have any ideal serving as the point of reference for the empirical theory.

In this way, normative and empirical theory are bound up to a unity. But this dramatizes the first qualm. How can a normative theory be the guide for an empirical theory? Well, if I have described matters correctly, the normative *is* guiding in this way. This is not a riddle. It is rather an important insight. As long as the concept of rationality is essential for psychology, psychology is, at its roots, normatively constituted and cannot be a purely empirical discipline. And as long this is so, there can be no reduction of psychology to physiology, neurology or biochemistry. The only alternative would be so-called eliminative materialism, according to which the science of man will in the (more or less distant) future no longer speak about beliefs, desires or their underlying rationality and hence does no longer have the form of a bundle of correction theories conjoined to a basic theory of rationality.

I have no principled argument against eliminative materialism. But obviously, it is not realized at present. And as a prediction it is most implausible. We will continue to work on the theory of rationality. We will continue to discuss what we rationally should believe and desire. And this will continue to be our point of reference to psychology and for our dealings with other people.

## References

Allais, M., O. Hagen (eds.) (1979), *Expected Utility Hypotheses and the Allais Paradox*, Dordrecht: Kluwer.

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

Bartelborth, T. (1996), *Begründungsstrategien. Ein Weg durch die analytische Erkenntnistheorie*, Berlin: Akademie-Verlag.

Camerer, C. (1995), „Individual Decision Making", in Kagel, Roth (1995), pp. 587-703.

Casullo, A. (1999), A priori Knowledge, Aldershot: Ashgate.

Coleman, J.S. (1990), *Foundations of Social Theory*, Cambridge, Mass.: Harvard University Press.

Davidson, D. (1980), *Essays on Actions and Events*, Oxford: Oxford University Press.

DeSousa, R. (1987), *The Rationality of Emotions*, Cambridge, Mass.: MIT Press.

Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, Cambridge, Mass.: MIT Press.

Fehige, C. (2000), *Ends and Means*, unpublished Habilitationsschrift, University of Leipzig.

Fehige, C. (2001), "Instrumentalism", in Millgram (2001), pp. 49-76.

Fehige, C. (2002), *Soll ich?*, Stuttgart: Reclam, to appear.

Friedman, J.W. (1991), *Game Theory with Applications to Economics*, Oxford: Oxford University Press.

Gabbay, D.M., C.J. Hogger, J.A. Robinson (eds.) (1994), *Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 3, Nonmonotonic Reasoning and Uncertainty Reasoning*, Oxford: Clarendon Press.

Gärdenfors, P. (1988), *Knowledge in Flux. Modeling the Dynamics of Epistemic States*, Cambridge, Mass.: MIT Press.

Grice, H.P. (1957), "Meaning", *Philosophical Review* 66, 377-388.

Güth, W. (2001), „How Ultimatum Offers Emerge – A Study in Bounded Rationality", *Homo Oeconomicus* 18, 91-110.

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

Habermas, J. (1999), *Wahrheit und Rechtfertigung. Philosophische Aufsätze*, Frankfurt a.M.: Suhrkamp.

Hempel, C.G. (1961/62), „Rational Action", *Proceedings and Addresses of the APA* 35 (1961/2) 5-23.

Hild, M. (1998), "The Coherence Argument Against Conditionalization", *Synthese* 115, 229-258.

Hild, M. (2002), "Auto-Epistemology and Updating", to appear in *Philosophical Studies.*

Hunter, D. (1991), "Maximum Entropy Updating and Conditionalization", in: W. Spohn, B.C. van Fraassen, B. Skyrms (eds.), *Existence and Explanation*, Dordrecht: Kluwer, pp.45-57.

Jeffrey, R.C. (1965), *The Logic of Decision*, Chicago: University Press, $^{2}$1983.

Kagel, J.H., A.E. Roth (eds.) (1995), *The Handbook of Experimental Economics*, Princeton: Princeton University Press.

Kahneman, D., P. Slovic, A. Tversky (eds.) (1982), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

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

Kreps, D.M., P. Milgrom, J. Roberts, R. Wilson (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma", *Journal of Economic Theory* 27, 245-272.

Kusser, A., W. Spohn (1992), „The Utility of Pleasure is a Pain for Decision Theory", *Journal of Philosophy* 89 (1992) 10-29.

Levi, I. (1967), *Gambling With Truth. An Essay on Induction and the Aims of Science*, New York: Knopf.

Levi, I. (1991), *The Fixation of Belief and Its Undoing*, Cambridge: Cambridge University Press.

Lewis, D. (1969), *Convention. A Philosophical Study*, Cambridge, Mass.: MIT Press.

Maher, P. (1993), *Betting on Theories*, Cambridge: University Press.

McClennen (1990), *Rationality and Dynamic Choice*, Cambridge: University Press.

Millgram, E. (Hg.) (2001), *Varieties of Practical Reasoning*, Cambridge, Mass.: MIT Press.

Myerson, R.B. (1991), *Game Theory*, Cambridge, Mass.: Harvard University Press.

Nida-Rümelin, J. (2001), *Strukturelle Rationalität. Ein philosophischer Essay über die praktische Vernunft*, Stuttgart: Reclam.

Nussbaum, M. (2001), *Upheavals of Thought: The Intelligence of Emotions*, Cambridge: Cambridge University Press.

Ordeshook, P.C. (1986), *Game Theory and Political Theory*, Cambridge: Cambridge University Press.

Pollock, J.L., J. Cruz (1999), *Contemporary Theories of Knowledge*, Lanham: Rowman & Littlefield.

Raiffa, H. (1973), *Einführung in die Entscheidungstheorie*, München: Oldenbourg.

Rapoport, A. (1989), *Decision Theory and Decision Behavior. Normative and Descriptive Approaches*, Dordrecht: Kluwer.

Resnik, M.D. (1987), *Choices. An Introduction to Decision Theory*, Minneapolis: University of Minnesota Press.

Roth, A.E. (1995), "Bargaining Experiments", in Kagel, Roth (1995), pp. 253-348.

Rott, H. (2001), *Change, Choice, and Inference*, Oxford: University Press.

Rubinstein, A. (1998), *Modeling Bounded Rationality*, Cambridge, Mass.: MIT Press.

Savage, L.J. (1954), *The Foundations of Statistics*, New York: Wiley, $^{2}$1972.

Skyrms, B. (1990), *The Dynamics of Rational Deliberation*, Cambridge, Mass.: MIT Press.

Spohn, W. (1978), *Grundlagen der Entscheidungstheorie*, Kronberg/Ts.: Scriptor.

Spohn, W. (1988), "Ordinal Conditional Functions. A Dynamic Theory of Epistemic States", in: W.L. Harper, B. Skyrms (Hg.), *Causation in Decision, Belief Change, and Statistics, vol. II*, Dordrecht: Kluwer, pp. 105-134.

Spohn, W. (1993), „Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein?", in: L. Eckensberger, U. Gähde (eds.), *Ethik und Empirie. Zum Zusammenspiel von begrifflicher Analyse und erfahrungswissenschaftlicher Forschung in der Ethik*, Frankfurt a.M.: Suhrkamp, pp. 151-196.

Spohn, W. (1997/8), „How to Understand the Foundations of Empirical Belief in a Coherentist Way", *Proceedings of the Aristotelian Society, New Series* 98 (1997/98) 23-40.

Spohn, W. (1999a), „Lewis' Principal Principle ist ein Spezialfall von van Fraassens Reflexion Principle", in: J. Nida-Rümelin (ed.), *Rationalität, Realismus, Revision. Vorträge des 3. internationalen Kongresses der Gesellschaft für Analytische Philosophie 1997 in München*, Berlin: de Gruyter, pp. 164-173.

Spohn, W. (1999b), „Two Coherence Principles", *Erkenntnis* 50, 155-175.

Spohn, W. (2002a), „A Brief Comparison of Pollock's Defeasible Reasoning and Ranking Functions", *Synthese* 131, 39-56.

Spohn, W. (2002b), „Dependency Equilibria and the Causal Structure of Decision and Game Situations", to appear in *Homo Oeconomicus*.

Stalnaker, R.C. (1999), *Context and Content*, Oxford: University Press.

Stegmüller, W. (1973a), *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band II, Theorie und Erfahrung, 2. Halbband, Theorienstrukturen und Theoriendynamik*, Berlin: Springer.

Stegmüller, W. (1973b), *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band IV, Personelle und Statistische Wahrscheinlichkeit*, Berlin: Springer.

van Fraassen, B.C. (1984), Belief and the Will, *Journal of Philosophy* 81, 235-256.

van Fraassen, B.C. (1995), Belief and the Problem of Ulysses and the Sirens, *Philosophical Studies* 77, 7-37.