Erschienen in: Game theory : critical concepts in the social sciences ; Vol. 4: Discontents / Varoufakis, Yanis (Hrsg.). - London : Routledge, 2001. - S. 213-241. - ISBN 978-0-415-22244-0

78

HOW TO MAKE SENSE OF GAME THEORY

W. Spohn

Source: W. Stegmüller, W. Balzer and W. Spohn (eds) *Philosophy of Economics*, Berlin: Springer-Verlag, 1982, pp. 239–70

1 A complaint

Game theory and decision theory are congenial, or so at least one would expect from their akin subject matter and their akin basic concepts and methods. And this expectation is justified by first inspection of the standard accounts of these theories: Decision theory investigates rational behaviour of single persons in isolation; game theory is concerned with the rationality of mutually dependent decisions of several persons; thus game theory is the more embracing theory, leaving to decision theory the special case of one-person games or, according to a rather unfortunate phrase, of games against nature.

Upon closer inspection, however, the standard accounts of game theory and its relation to decision theory appear quite unsatisfactory. Of course, decision theory, too, is clouded by problems; but in comparison, I think, game theory is additionally sapped by three connected disconcernments: it is, to put it strongly, confused about the rationality concept appropriate to it, its assumptions about its subjects (the players) are very unclear, and, as a consequence, it is unclear about the decision rules to be applied. Or in other, somewhat paradoxical words: Decision theory may be a specialization of game theory (viewed from game theory), but game theory as presented today is never a generalization of decision theory (viewed from decision theory). Rather, in anticipation, game theory should be viewed as a specialization of decision theory.

This is my complaint. I shall substantiate it in the subsequent sections and explain how I think it should be remedied.

The reader may suspect that the objections are directed to the higher and dimmer regions of game theory such as three-or-more-person games or games in characteristic function form, and then he may perhaps concede them

willingly. But, on the contrary, they address to the seemingly clean and settled base, to two-person zero-sum games. For the sake of perspicuity I shall deal only with games in normal form.[1]

The reader may also suspect a pleading for a Bayesian game theory, and I shall indeed argue from a puristic Bayesian position. However, the label "Bayesian game theory" has become associated most notably with the work of John C. Harsanyi, which seems to me to be still more game theoretic than decision theoretic in spirit and hence criticizable on similar grounds as the standard accounts. Thus, there is a difference here which we have to take up in the last section.

In all that I am not claiming that the position set forth here or any of the arguments for it would be new (though some twists may be). It is only that earlier attacks on game theory guided by the same spirit have apparently been unable to stir up the received theory from its complacency and to set it on a better founded path; and it is this fact which has led me to make another try.

2 How to make sense of game theory

Before substantiating the complaint, it is fair to outline the basic conviction on which it rests. This conviction is an orthodox Bayesian one:

According to it, people have aims and wishes, they like the world to be such and such; they have beliefs, they think the world to be such and such; and, if rational, they act so as to promote their wishes best according to their beliefs. For the sake of definiteness, decision theory formalizes this in quantitative decision models. In such a model formalizing a person's decision situation, this person is assumed to have numeric subjective utilities and probabilities; then rational action is defined as action maximizing expected utility; and as a normative theory decision theory recommends rational action, while as an empirical theory it assumes rational action, well knowing that this is a strong idealization entitled at most to approximative validity. Nevertheless, this is a model which is claimed to be applicable in principle to each and every human action. (This claim is not quite as strong as it may seem, since it is not to be extended to all human behaviour. It must be observed that action is a narrower concept than behaviour, and despite its circular air it is not unreasonable to say that actions just are behaviour to which decision theory is applicable.[2])

It is not really necessary here to go into the details of the decision theoretic formalization. But let us assume, for the sake of precision, that it is done in Savage's well-known way, where probabilities are defined for a set of possible world states and where utilities refer to possible outcomes each of which is uniquely determined by a world state and an action, so that the familiar utility matrix also found in two-person games in normal form ensues. For our discussion this is the most suitable formalization.[3]

By the way, this was not quite the usual story which is more cautious by trying to render the quantitative model as something derivative. It defines rational action as choosing what is most preferred according to rational preferences; preferences are rational, if they satisfy some rather evident conditions such as transitivity etc.; and then, amazingly, it can be proven that rational action is such as if it maximized expected utility. But this "as if" is almost as out of place as saying that bodies move through space, as if they had a mass, as if they were obeying Newton's second law, etc. No, according to Newtonian mechanics bodies move the way they do, *because* they have such and such a mass, because such and such forces are acting upon them, etc., and according to decision theory people act the way they do, because they have so and so strong desires, because they have so and so firm beliefs, etc. Surely, there are a lot of subtleties hidden in this subject, about which philosophers of science are still divided. But there is no doubt that philosophy of science has outmoded operationalism as expressed by the "as if" in physics and anywhere else.[4] Therefore one should treat the quantitative decision model as basic. (This may change the status of all the ingenious metrization theorems backing up the "as if"-story, but does not at all diminish their value.)

Turning now to game-like situations of mutually dependent decisions, is then anything of the above general characterization of decision situations to be revoked? No, nothing at all. Other persons and their behaviour are to us just as much parts of the outer world as anything else, though certainly rather complicated and often very dear ones. Formally, this means that in any player's decision model the possible actions of the other players are but parts of the possible world states. We may further take these possible actions as constituting a small world (in Savage's technical sense; cf. Savage (1954), sect. 5.5) and reduce the model to this small world – in effect, this is the same as reducing a game in extensive form to its normal form. Thus, the reduced model contains the utility matrix of this normal form, and the *right and only* way to complete it is to add the player's subjective probabilities for its possible world states, i.e. for the other players' actions. After all this, the rational thing to do is, as always, to maximize expected utility; and that's it.

Indeed, very often there is nothing more to game-like situations. In so many of our daily routines we treat other people just as if they were regularly and reliably behaving automata, about which we have rather definite expectations without wasting any further thought; they figure in our decision problems in no other way as do, say, the traffic or weather conditions. (This somewhat heartless talk is but a harmless "déformation professionelle"; fortunately, we do, and are able to, take more interest in some people.)

But this being accepted, what realm is then left as peculiar to game theory? Game theory commences, when we take other people in the outer world seriously as persons, when we give up looking only at their behaviour and start theorizing about them, and in particular, when we discover that decision

theory is approximately the right theory about them, when we try to figure out what their aims and beliefs may be, assuming that they act rationally. Note, however, that on this account game theory does not embrace decision theory, but is rather a specialization of it. Game theory is decision theory about special decision makers, namely about decision makers who theorize decision-theoretically about the other persons figuring in their decision situations.[5]

All this probably sounds very familiar. It is just the orthodox Bayesian stand on game theory and more or less what Harsanyi, for instance, has told us so many times for more than twenty years. But strangely, everyone – standard game theorists anyway, but also Bayesians like Harsanyi (cf. the last section) – seems to have sinned against the pure doctrine, to have shrinked from pushing it to its consequences.

The sinning has its reasons, however. For it seems difficult, if not impossible, to justify within the pure Bayesian doctrine what everyone held justified – that is: to justify equilibrium points as solutions of two-person zero-sum games or generally of non-cooperative games (cf. section 4). Thus we must have a careful look at what can be concretely done with the hitherto sketchy doctrine without betraying it. But let me first inspect the standard game theoretic reasoning for two-person zero-sum games from this Bayesian point of view.

3 How not to make sense of game theory

To this end we should briefly recapitulate the received reasoning. I hope everyone agrees that Luce, Raiffa (1957), ch. 4, and von Neumann, Morgenstern (1944), ch. III, are not only representative for, but still among the most thorough and convincing accounts of this reasoning, so that I can base the recapitulation on them. It consists of four parts.

The standard story

Part 1 (pertinent to all games in normal form): Let a game be given in normal form. The basic problem of game theory then is, very vaguely stated, somehow to find out for each player which choice would be a good one for him. However, this is much too indeterminate a question; it needs specification. So let us first assume that each player is rational either in the loose sense of trying to get out of the play as much as possible (according to his utility function) or in the stricter sense "that, given two alternatives, he will always choose the one he prefers, i.e. the one with the larger utility" (Luce, Raiffa (1957), p. 55). And let us secondly assume that each player has full knowledge of the game in normal form, i.e. that he is aware of every player's possible alternatives (strategies) and that he knows every player's utilities for the outcomes of all possible strategy combinations (which, in general, are already expected utilities with regard to the chance moves of the game).

Without the first assumption game theory could not get off at all; for what general theory could there be about irrational action? And the second assumption is necessary, too; else the problem tackled by the game theorist might be the wrong one, i.e. different from the problem of the players as they subjectively see it. With these assumptions, however, we may hope to have rendered our problem specific enough to be solvable. So let us try to solve it:

Part 2 (pertinent to all non-cooperative games in normal form): A first consideration moves us ahead quite a bit. It says that, if game theory is to be at least potentially public – as it should doubtlessly be –, then it can distinguish only equilibrium strategies as rational choices for the players. (To be sure, so far I am talking only of pure strategies; mixed strategies do not come up until part 4.) Or more fully: Game theory is to find out for each player which choice would be a rational one for him; if it manages to do so, then each player can know as well as the game theorist himself, which choices would be rational for the other players (since, according to the second assumption above, each player sees the game situation in the same way as the game theorist); and since each player is assumed to act rationally, this assumption must not be a reason to any player to deviate from what is rational for him according to the theory; hence only equilibrium points can be rational strategy combinations, and only equilibrium strategies, i.e. strategies leading to some such point can be rational choices.

As is well known, this consideration is of varying force. Some games have no equilibrium point in pure strategies and some have many, in which cases its success is still incomplete. But regarding two-person zero-sum games with an equilibrium point in pure strategies, it is a bull's-eye, since the equilibrium point which such a game has may be proved to be essentially unique (cf. Luce, Raiffa (1957), sect. 4.5). Thus, in this special case we have already solved the basic game theoretic problem.

Part 3 (pertinent only to two-person zero-sum games with an equilibrium point in pure strategies): There is yet another forceful consideration to the same effect in this special case. Call the two players Charlie and Lucy. Charlie might intuitively reason as follows: "Lucy, this rational beast, tries to get out of the play as much as possible. This runs against me. So I better look for how much I minimally get from each of my options and try to make this amount as large as possible, that is, as I have heard someone express it, I better maximize my security level. If this is reasonable, then rational Lucy will do the same, i.e. maximize her security level. Ah, but my security level maximizer is best against her security level maximizer, so I should all the more stick to my choice."

Or in von Neumann's words: Consider Charlie's minorant and majorant game. In the minorant game he has to choose first, and then Lucy may choose, knowing what he has done. In the majorant game it is just the other way around. Obviously, in the minorant game Charlie is at most as well off

as in the actual game, whereas in the majorant game he is at least as well off as in the actual game. And, as is equally obvious, the only rational thing for him to do in the minorant game is to maximize his security level, and the only rational thing to do in the majorant game is to choose what is best against Lucy's security level maximizer (provided she has been so rational to take this choice). But both cases result in the same strategy combination and in the same utility for Charlie. Hence, for the actual game being "between" the minorant and the majorant game exactly this and no other thing is rational.

To summarize: Starting from the assumptions in part 1, we have presented two completely independent reasonings. Each one alone would be telling in the special case considered, and both lead provably to the same result. What better justification could there be?

Part 4 (pertinent to all two-person zero-sum games): Now von Neumann tells us that we can generalize the whole story to all two-person zero-sum games, if we are willing to allow a little trick, i.e. to allow each player to mix his pure strategies. Further arguments were invented to give the last pull to those who felt a bit uneasy about this trick, e.g. the secrecy argument, the consideration of playing a game repeatedly, or the diet argument (cf. Luce, Raiffa (1957), p. 75). But we need not elaborate here on this additional backing, since it would be void without the main reasoning. And this can stand by itself. Indeed, any player is free to choose a mixed strategy; thus, mixed strategies are among the alternatives to be considered, and with respect to them the above reasoning is no less powerful than with respect to pure strategies. So, that is how mixed strategies, maximinimizers and equilibrium points have found one another, and they lived happily ever after.

Sadly, this story is not as sound as it sounds; it is in need of a commentary, critical not of its conclusions, but of the way these are reached.

Commentary

To part 1: One may think that the rationality and knowledge assumptions of part 1 unduly restrict the applicability of game theory. But, in fact, they rather are either not strong or not clear enough. Does it really suffice to assume that the players are rational? It certainly seems that one should also assume that each player believes the other players to be rational. This is particularly clear from part 2 of the story where we have been very sloppy in distinguishing between what the game theorist assumes a player to assume about the other players and what the game theorist himself assumes about the other players. But then, one should presumably also assume that each player believes that the other players think their fellows to be rational. At this point, some may tend to a radical move, i.e. to climb up the whole infinite ladder of iterated mutual rationality assumptions, as some have done in a similar case within the theory of meaning.[6] That is, the game

theorist might assume that the rationality of the players is mutual or common knowledge among the players (in the technical sense of Schiffer (1972), p. 30f., or Lewis (1969), p. 56; cf. also here in sect. 4). Of course, all this applies equally to the second, the knowledge assumption in part 1 of the story. So, what should the game theorist assume? One feels that it does make a difference exactly how much is assumed about the players, but it is hard to see how this is reflected in the received story.

There is another unclarity. What does "rational" exactly mean as used in the rationality assumption? The explanation cited from Luce and Raiffa is of no great help, since preferences and utilities refer only to strategy combinations; nowhere in standard game theory is a preference order or even a utility function established solely for the alternatives of one player. So, one would like to have sharply specified another, more utilizable sense of "rational". Presumably, however, the question was the wrong one. Presumably, standard game theory thinks it preferable or unavoidable to leave "rational" vague in the initial assumptions and explanations, promising to render it precise later on. But for the moment, this is only to say that "rational" is intendedly vague, and this is no improvement.

The crux of the matter is this: Standard game theory does nowhere reason from the initial assumptions in a rigorous way; they are exclusively employed in plausibility arguments. The attitude seems to have been that first the intuitive grounds are to be prepared for the subsequent exact theorizing, and that one need not weigh every word in that preparation. Thus, some nice differences are blurred already at the intuitive level, leaving no chance to the hard theorizing to undo this laxness. From the Bayesian point of view, this is the first decisive slip onto shaky grounds.

To part 2: We have already mentioned that stronger assumptions about the players than those of part 1 are necessary for having the players see the game situation in the same way as the game theorist and thus for part 2 to pass through. But there is another flaw, which is particularly clear in the case of a two-person zero-sum game with exactly one equilibrium point in pure strategies. In this case, part 2 concludes that each player can rationally choose only his equilibrium strategy. But this is premature; what follows is only this: If the game theorist succeeds in distinguishing exactly one choice as rational, then this must be the equilibrium strategy. However, there is no guarantee that the if-sentence is true; perhaps the game theorist's problem is such that he can narrow down the range of rational choice only partially and not to a singleton. More generally: What part 2 shows is that the game theorist cannot establish some choice set as rational to the exclusion of equilibrium strategies, but it has still to be shown on other grounds that a choice only among the equilibrium strategies can be positively established as rational. Part 3 might prepare such grounds; so let us turn to it.

To part 3: There has been a lot written about the decision rule of maximinimizing, and all the essential pros and cons are known. The present

state of discussion is, I feel, a somewhat smoothed one. It seems to be generally accepted that maximinimizing cannot serve as a basic decision rule entitled to general applicability; it leads to absurdities in too many situations. Nevertheless, it is acknowledged as a discussable, respectable, or even convincing decision rule in some types of situations, most notably in two-person zero-sum games, but also for decisions under uncertainty, in statistical decision theory, and more recently in Rawls' original position (cf. Rawls (1971), sect. 26).

From a theoretical point of view, however, this state of affairs is utterly dissatisfying. From this point of view, it simply does not do to find intuitively convincing decision rules for various types of situations, to support the intuitive judgment by some sort of systematic argument, and to leave it at that. No, if different decision rules are really to be accepted for different types of situations, then one would want to know some leading or unifying principles explaining or at least describing exactly under which conditions which decision rules are appropriate in which situations; or, what would be nicer, one would like to have some basic decision rule from which the others may be derived. But in trying to answer this demand with respect to game theory, we obviously run straight into the obscurities found in part 1.

To be sure, all I am doing here is to appeal to theoretical awareness. But I would like to make this appeal more pressing by the following argument.

It has to do with Savage's small words – a subject whose theoretical importance, I think, has only insufficiently been recognized, and which is concerned with the fact that the description of one and the same decision situation may be based on different worlds. Here, a world is – loosely speaking, we need not really go into technical details – the collection of all the items which are *explicitly* to be considered in the description as relevant to the decision situation. Savage's observation was now that there seems to be no good way of telling which is the right world on which to base the description of a given decision situation. Prima facie, it may seem plausible to put into a world each item which is in fact relevant, but in general this would yield unmanageably large worlds. So, instead of looking for the right world we should rather try to find out when two descriptions based on different worlds may be said to be equivalent. To this end Savage developed a method of reducing a description based on a large world to a description based on a small world which may be warrantedly said to be equivalent to the first one. The essential feature of the reduction method is how it ascribes utilities to the possible consequences included in the small-world description, and Savage does this in the following way: Viewed from the large-world description, there are certain probabilities p_i with which a less detailed small-world consequence, say c, shapes to various, more detailed large-world consequences c_i having certain utilities u_i; then the utility of c in the small-world description is to be the expectation value $\sum p_i u_i$.[7]

One might perhaps envisage other reduction methods (though one need not, I think); but what is important for us now is that, whatever reduction method is chosen, it must be such that the decision rule adopted is compatible with it. This means that, when the decision rule is applied to the large-world description, the same decision (in fact, the same preference order among the alternatives) must result as when the decision rule is applied to the reduced small-world description.[8] Actually, it is somewhat misleading to say only that reduction method and decision rule must be compatible. Rather, the reduction method is the *basic* thing to be chosen, and then the decision rule ensues as a mere special case; for the decision rule effects nothing but a maximal reduction to the minimal description which explicitly considers only the alternatives of the decision maker and nothing else.

The next point to observe is that the reduction method which is the natural generalization of the decision rule of maximinimizing is a wild one, indeed. According to it, the utility of a small-world consequence would be the minimum of the utilities of the large-world consequences to which it might shape up; and it need not be demonstrated that this leads to all sorts of absurd and intuitively unacceptable results. In fact, nobody, even no adherent of maximinimizing, has ever seriously considered this reduction method. That is, maximinimizing was held to be reasonably applicable only to small-world descriptions of a decision situation, which are already obtained by Savage's reduction method of forming expected utilities. Or more briefly, what is maximinimized are always expected utilities (with respect to some large-world description). This is particularly clear in game theory where the utilities contained in the normal form usually are expected utilities derived from the extensive form.

Thus, the theoretical muddle turning up with the decision rule of maximinimizing is profounder than it seemed. First, the muddle was that various decision rules seemed to be appropriate to various decision situations without there being any unifying principles. But now, when decision rules are seen to be special cases of reduction methods, we have the muddle within single decision situations, since to maximinimize expected utilities is in effect to apply two different reduction methods within one decision situation. There are urgent questions then. Which reduction method is appropriate exactly to which items of the decision situation? And why? Why first reduce by taking expected utilities and then reduce by considering minimum utilities? Why not the other way around? (This makes a difference; the two methods are not commutative.) And so on. All this is very awkward, and we should try everything to avoid this muddle.

A final word: Von Neumann's version of 3, the "betweenity"-argument, has more the air of being rigorous than Charlie's intuitive reasoning. But it is not. In the minorant game Charlie knows that Lucy will know what he will do, and in the majorant game he will know what Lucy will do and he

also knows that Lucy knows this, etc. In the real game situation he has no such knowledge, i.e. he is epistemically worse off than in both the minorant and the majorant game. (This also means, however, that in terms of expected utility he may be better off than in the other two games.) In this respect the real game is not "between" the minorant and the majorant game, and there seems to be little chance to render the "betweenity"-argument correct (as is also argued by McClennen (1976)).

To part 4: This part of the received story is still the clearest symptom to me that something must have gone wrong with it; somehow all the little slips seem to have led us completely astray. For a mixed strategy simply cannot be *the* rational or optimal choice. This need not be argued anew, I think; the ineffectiveness of such compelling reasoning as that of Chernoff (1954) can only be explained by the fact that (the other parts of) the standard story had too strong a hold on people. Let me just repeat a brief version of such reasoning: It starts by assuming that the players have some sort of preference ordering among their alternatives. Though game theory does not establish such an ordering, as already mentioned, to deny its being possible or making sense just for game situations would be a strange claim indeed. Now a mixture of two comparable alternatives obviously cannot be better than both of them; and if the ordering should not be complete or connected, if there should be two incomparable alternatives, then a mixture of these would in turn be incomparable to both of them. Hence, in no case can a mixture be better than what it is mixed of, and there is no need to consider mixed strategies as options of the players.

In fact, it is not clear whether anyone has really claimed a mixed equilibrium strategy to be *the* rational choice, since there is the following inherent counter-argument which is well known. If a player is firmly convinced that his opponent plays his mixed equilibrium strategy, then all the pure strategies mixed in his own equilibrium strategy (and all other mixtures of them) have the same maximal expected utility. That is, if either of the players is faithful to game theory, the other need not be and is justified to neglect mixed strategies, and if either of them is not faithful to game theory, then game theory is suspended anyway for the present. This instability of equilibrium points in mixed strategies (which indicates that part 2 of the story, even if unobjectionable, cannot be smoothly carried over to part 4) has also worried Harsanyi in his (1973) article to which we shall return.

The arguments usually added are of no help here. The secrecy argument that randomizing is good for hedging against clever opponents[9] is a nonstarter, since, as (normal form) game situations are usually described, the players simply cannot know or find out before their choice what the other players will do, unless they have telepathic or similarly exotic capacities. They may have more or less well evidenced beliefs about the others, but again, according to the usual description, the unobserved process of choosing in the situation at hand cannot be part of that evidence. To put it

somewhat polemically: the intriguing point in game theory is not the fear of the advent of knowledge, but rather the certainty of the absence of knowledge.

Another line often found in textbooks, whether for illustratory or for justificatory reasons, is to imagine a game being played very or infinitely many times. If this is taken, however, as one playing of the supergame constructed from the original game, this line is no advance, simply because all the theoretical trouble we had with the original game turns up anew in the supergame. But even if we suppose that a statistically unexploitable random sequence of pure strategies of the original game (showing up in the appropriate proportions) is a reasonable choice in the supergame (which it is, of course) and that there would be a theoretically unobjectionable justification of this, we do not get ahead. There is no strict inference from that to what is rational when the original game is played only once.[10]

The secrecy argument makes more sense in this context of repeated playing, because randomizing in earlier plays may be used for becoming incalculable in later plays. But all this misses the point. The plausibility and the practical value of such considerations is uncontested. The point, however, is that as such they do not contribute to foundation-oriented theorizing. And there mixed strategies taken as possible choices of the players can be safely neglected for the reasons mentioned.

4 How to make sense of game theory (continued)

We could have evaded all this trouble by strictly sticking to the decision theoretic position. Then we would have to spell out full decision models for the players which force us to explicitly state all our assumptions, in particular the epistemic ones, about the players and to rigorously deduce the rational choices from them by the rule of maximizing expected utility instead of reasoning by plausibility. Thus, part 1 of the story would be as precise as desired. Part 2 would be still in force, though in its weakened form stated in the commentary to it. The muddle of part 3 would be cleared up at once. And we would never have the idea of resorting to mixed strategies.

Very well then. But what does the positive Bayesian theory look like? And does it not run into new trouble? Let us see. We should first introduce some terminology. In this section, being *rational* is precisely to mean maximizing expected utility and nothing else; this is important. That a person *firmly believes* that p, is to mean that its subjective probability for p is 1. With respect to two persons 1 and 2 we define recursively: Person i ($i = 1, 2$) has a *belief of first order* in p if it firmly believes that p; person i has a *belief of* $n+1$*-st order* in p if it firmly believes that person j ($j \neq i$) has a belief of n-th order in p; and p is *mutual knowledge of the n-th order* among the two persons iff p is true and both have beliefs of all orders up to n in p (though, strictly speaking, it need not be knowledge what they have, but rather only true beliefs).

Let us turn now to the simplest case, to two-person zero-sum games in normal form with exactly one equilibrium point in pure strategies, where Charlie (the row-chooser) and Lucy (the column-chooser) are our two opponents. Part 1 of the received story and the commentary to it suggest to start the analysis by assuming that the rationality of Charlie and Lucy and their *u*tilities as given by the game matrix are *m*utual knowledge of some order still to be specified among them. If this order is n, let us call this assumption RUM_n. Does some RUM already solve these games? Unfortunately no. What the RUMs do is to eliminate alternatives which are strictly dominated from the beginning or after some alternatives have been eliminated in this way. For instance, the following game is solved by RUM_5 (which, of course, implies $\mathrm{RUM}_4, \ldots, \mathrm{RUM}_1$):

	b_1	b_2	b_3	b_4
a_1	4	5	6	7
a_2	3	2	7	5
a_3	2	4	1	8
a_4	1	3	4	0

Because of RUM_1, Lucy firmly believes that Charlie will never do a_4; because of RUM_2, Charlie firmly believes that Lucy firmly believes this and will hence never do b_4; and in the same way, a_3 is eliminated by RUM_3, b_3 by RUM_4 (that solves the problem for Charlie), and finally a_2 by RUM_5 (that solves the problem for Lucy, too).

To generalize: If the RUMs effect to eliminate all but one alternative of a player, then the alternative left can only be his equilibrium strategy. Unfortunately, the games in which RUM is so effective are of rather special character. For example, all RUM together is powerless in the following typical game:

	b_1	b_2	b_3
a_1	2	3	3
a_2	1	0	4
a_3	1	4	0

Here, RUM_1 does not eliminate anything, and then no RUM can.

There is the snag of the Bayesian position. According to the standard story, the somewhat vague assumptions of part 1 seem to justify equilibrium or maximin strategies for all two-person zero-sum games in quite convincing a way. Now, under a decision theoretic exactification these assumptions condense to RUMs; but the RUMs are weak and do not knock down but

the most special cases. For non-Bayesians this settles the matter, and even Bayesians start to roll at this point. But in my view, any departure from the decision theoretic path is theoretically disastrous for the reasons mentioned. Thus, as equilibrium strategies seem and are widely held to be reasonable, the task can only be to strengthen RUM by some plausible assumption from which the equilibrium strategies may be proved to be rational. The assumption I am going to state is, I think, the one which is closest to the spirit of standard game theory; in fact, it will be so trivial an adaption that you will be disappointed:

The trouble with our second example was that, according to RUM, Charlie's and Lucy's epistemic states concerning the other's actions were not restricted at all, and that each of his or her alternatives was optimal with respect to some epistemic state. Thus, we should introduce some restriction concerning these epistemic states. One way to do this is to strengthen RUM_n to RUE_n, i.e. the assumption that not only the rationality of Charlie and Lucy and their *u*tilities, but also their *e*pistemic states concerning the other's actions are mutual knowledge of some order n among them.

A bit more formally, this amounts to the following theorems which in fact apply to all two-person games in normal form. Denote the set of Charlie's alternatives by A_1 and that of Lucy's by A_2, and let us consider the following propositions:

(1) Charlie is rational,
(1′) Lucy is rational,
(2) his utility function for $A_1 \times A_2$ is U_1,
(2′) her utility function for $A_1 \times A_2$ is U_2 (not necessarily $-U_1$),
(3) his subjective probability function for A_2 is P_1,
(3′) her subjective probability function for A_1 is P_2,

(4), (5), (6) } he firmly believes that { (1′), (2′), (3′),

(7), (8), (9) } he firmly believes that she firmly believes that { (1), (2), (3)

(4′)–(9′) as (4)–(9) with the roles of Charlie and Lucy interchanged,
(10) Charlie's mixed strategy $s_1 = P_2$ and Lucy's mixed strategy $s_2 = P_1$ are in equilibrium,
(11) he chooses a pure strategy which is best against $s_2 = P_1$,
(11′) she chooses a pure strategy which is best against $s_1 = P_2$.

Then we have the following "intrapersonal" theorem (in the sense that it speaks only about one person), that (1)–(9) imply (10) and (11), and the "interpersonal" theorem, that (1)–(6) and (1′)–(6′) imply (10), (11), and (11′).

The proof hardly deserves stating: Let $B_i \subseteq A_i$ $(i = 1, 2)$ be the set of all pure strategies which are best against s_j $(j \neq i)$ according to U_i. Denote by $M(B_i)$ the set of all mixtures of strategies in B_i. Then, of course, each mixed strategy in $M(B_i)$ is best against s_j. Now (3)–(6) imply that $s_2 \in M(B_2)$. Similarly, it follows from (6)–(9), or from (3′)–(6′), that $s_1 \in M(B_1)$. Hence, s_1 and s_2 are in equilibrium, and finally, (1) and (1′), respectively, entail (11) and (11′).

I hasten to add that we have just used mixed strategies only as a formal device (as which they are still useful, of course); P_1 is here considered only as something that Charlie has and s_2 not as something that Lucy may do, though they may be formally equated. Let me also add that these "theorems" may be quite trivially generalized to all n-person games in normal form.[11]

My reason for stating the "theorems" was that I think their form to be paradigmatic of game theoretic theorems. They characterize a player by a full decision model in which both his desires and his beliefs are described as detailed as is needed; and they uncompromisingly take maximizing expected utility as the only decision rule. Thus, they are strictly Bayesian, and as such they conform to all demands arisen from the criticism in the preceding section.

And they should not be blamed for their conclusions (11) and (11′), I think, though this conclusion is not completely determinate for games without an equilibrium point in pure strategies. Standard game theory is equally unspecific with respect to pure strategies, and it was already clear that within a Bayesian account we cannot achieve more specific results by allowing mixed strategies. Thus, this much indeterminateness is unavoidable, and there is no ground for disappointment in this respect.

But, presumably, you will blame them for their premises, though you will certainly grant that the premises accord to the spirit of standard game theory. Referring to the "intrapersonal" theorem, the premises (1), (2), (4), (5), (7), and (8) are part of RUM_2, which is accepted in game theory anyway; (3), (6), and (9) also conform to the general tendency to assume publicness of its assumptions, and, in particular, they account for the fear of being transparent to the opponent, which game theory imputes to the players.

However, one will retort, it is not at all in the spirit of, but rather a caricature of game theory to take (3), (6), and (9), though true of rational players, as premises, because thereby it is outright presupposed what game theory does, or strives to, establish by showing that s_1 and s_2, respectively, are the rational things to do for Charlie and for Lucy (which entails (3), (6), and (9) because of the assumed mutual knowledge of rationality). I could now repeat arguing that something like (3), (6), and (9) is not at all rigorously established in standard game theory, and here we are again. Where is the rub here? I think, even if one grants me what I have said so far, there remains the definite feeling that I have not done full justice to standard

game theory. The fact that the Bayesian renarration produces such a triviality when taking the apparent aim of the standard story, i.e. that of establishing rational action, at face value clearly indicates that the standard story intends something more that we have not yet grasped. But let us defer this crucial point for the moment; we shall get clearer about it when we approach it from an abstracter level later on.

Another blame might be that (3), (6), and (9) are much more implausible assumptions than the others (though this is rather the opposite of the preceding blame that (3), (6), and (9) were presupposed instead of proved). Three remarks are pertinent here:

First, all of (1)–(9) are idealizations, of course. But there is no reason at all, why (3), (6), and (9) should be graver idealizations than the other assumptions. Thus, this cannot be the point this blame is directed to. (And the problematic nature of idealizations in general is not a subject we need to engage with.)

Secondly, it is hard to say generally, whether (5) or (6), or whether (8) or (9), is the more critical assumption of our theorems, since it seems to be impossible to make any general, substantial assertion as to whether beliefs or desires of other persons are more easily knowable; and this need not be argued, I think.

And a third thing to note is that it would not be quite correct to say that the surplus of RUE as against RUM consists in the mutual knowledge of the players' epistemic states, since usually *some* such thing is already contained in RUM. This is, if a game has chance moves, then the players' epistemic states concerning these chance moves are assumed by RUM to be mutually known, because RUM then requires expected utilities to be mutually known.

Yet despite these defensive remarks, (6) and (9) still seem to be more problematic than (5) and (8) – at least in the usual examples for two-person games (and this cannot be dismissed by saying that the examples would be biased). This is supported by the following considerations:

Firstly, the assumption that the players' epistemic states concerning chance moves are mutually known seems to be innocuous in many (though not in all) situations – e.g. for chance moves like throwing of dice, but also when the subjective probabilities concerning a chance move cannot be so easily taken as reflecting the knowledge of the objective probabilities for that chance move, and even when there are no objective probabilities for the chance move in question. For instance, a chance move might be whether Snoopy is just searching for the Red Baron, and then we might imagine Charlie to reason as follows: "Snoopy has started searching yesterday, and usually it takes him days. So, very probably, say to a degree of .9, he is still on search. Now, since Lucy and I together observed him mounting his Sopwith Camel yesterday, I know her, and she knows me, to know that Snoopy has

started yesterday. Moreover, she knows him almost as well as I do, and she knows how well I know him; thus she will guess my probability about Snoopy correctly, and she herself will have about the same probability." Whenever such considerations are appropriate, at least second order mutual knowledge of the players' beliefs about a chance move may be plausibly assumed.[12]

Similarly, mutual knowledge of utilities often seems unproblematic. Thus, imagine Charlie and Lucy playing matching pennies; here is another easy reasoning of Charlie establishing (2), (5), and (8) for this play: "I give no quarter, I want to win. So, my utilities stand firm. Now, Lucy knows human nature quite well, and mine in particular. Men are after money, and I am not so different, after all. Thus, she will know my preferences. But she is not so different, too, she has proved it often enough. So, her utilities should be contrary to mine."

In contrast to these two reasonings, let us see whether there is a similar reasoning for (3), (6), and (9). This is how Charlie might deliberate: "How probable are the various alternatives of Lucy? In order to find this out, I should examine my evidence about her." – Pause. – "Well, whatever my evidence is, I have gathered it with her knowledge; there is nothing peculiar or mysterious about it. Thus, (a) she will approximately know what evidence I have about her. But then (b) she will also correctly guess my probabilities; after all, our ways of thinking are not so different. In the same way, she will probably think that I correctly guess her probabilities about me." – Pause. – "Look, now it follows with RUM (RUM_3, to be precise) that (c) my probabilities must be P_1 and hers P_2 [provided this is the only equilibrium point]. And hence, (d) she also thinks me to have P_1. Wasn't that smart?"

No, it was a bit fishy, compared with the first two reasonings. In contrast to the Snoopy case, the evidence about Lucy remained in the dark. The really bad thing, however, is that the reasoning to (d) was sort of self-defeating. For, (a) was the ground for (b), but (b) led to (c) and then to (d) without reference to any evidence; thus, (a) did not become operative at all, and this deprives (b) of its grounds.

The obvious way of rendering Charlie's third reasoning sound seems to be to explicitly state some evidence which Charlie may plausibly have and which directly induces him to have the desired P_1; his inference to (d) then passes through. (Note that Charlie has then P_1 *not* because P_1 is the only probability function compatible with Lucy knowing which probabilities he has, as was suggested by his reasoning; rather, he has P_1 because of the evidence he has, and then P_1 additionally, though not accidentally, proves to be so compatible.)

However, as the discussion in section 6 drives us exactly to the same point, I shall take it up in more detail later on. Thus, for the moment we have to admit that we are still lacking grounds for (3), (6), and (9) which are

as natural as those for (2), (5), and (8), and hence, that both blames against allowing (3), (6), and (9) as additional premises are not yet fully answered.

5 The real issues: action rationality and epistemic rationality

So far, we have presented and compared the standard and the decision theoretic story, and I hope I have made clear where the definite merits of the Bayesian story lie to my view and why they lie where they lie. But we just found also some loose ends of the Bayesian story, and it may seem as if, in order to tie them up, we might be forced to fall back on the received story. So let me belabour the whole subject once more at a somewhat deeper level, i.e. by considering the conceptions of rationality on which the different positions are based, and let me take up the standard story first:

In fact, there does not seem to be a very definite conception of rationality lying behind standard game theory. Another way of developing a concept of rationality was much preferred in decision and game theory and related fields at least during the fifties and sixties. The first rule, born from a sensible suspicion of any grand picture, was not to prejudge the subject by any comprehensive, but rash conception of rationality. Rather, a cautious step-by-step reasoning should lead to a reflective equilibrium, as Rawls (1971), pp. 48ff., termed it, of intuitive and systematic arguments. Thus, one started with some intuitively very compelling assumptions, displayed their deductive consequences, scrutinized whether any of these consequences were intuitively unreasonable, eventually dropped the weakest assumptions, tried to add new assumptions, checked them in the same way, distinguished basic and derived assumptions, and so on. In this way, a stock of basic principles such as the transitivity of preferences and the sure thing principle (and of less basic principles like those of the maximin variety which were tailored to more special situations) emerged which could then very confidently be claimed to characterize rationality; and though these principles were never supposed to exhaust the concept of rationality, they proved to be quite powerful. Indeed, for decisions under certainty and under risk this method has yielded complete success; for decisions under uncertainty the results were illuminating, though not unanimously agreed upon; and at least the simpler game situations were satisfactorily dealt with.

I hope this was not too distorted a description of the actual procedure, whose only weakness is, I think, that it seems to be lacking a bit of conceptual clarity; it is not fully transparent exactly what is there driven to a reflective equilibrium. This has come to bear particularly on game theory, or so at least I try to argue in the sequel.

In order to get a bit clearer, we have, I think, to observe two or three rather obvious facts about rationality. First of all, we must strictly distinguish between the rationality of actions, the rationality of beliefs, and perhaps the rationality of desires and separately discuss them.

Let us consider action rationality first, which is the declared subject of decision and game theory. The important thing here is that whether an action of a person is rational or not can only be determined relative to the subjective desires *and* beliefs of that person. This is clear from everyday experience; whenever we happen upon an action which plainly seems to be irrational, we might have to withdraw our judgment, when the actor, or somebody familiar to him, explains us his reasons for this action. And it is clear from philosophical literature which has repeatedly pointed out this fact.[13] Now one may call an action rational, only if it is rationally linked with beliefs and desires which themselves are rational. But this is merely a terminological question. There is a certain relation between actions on the one hand and beliefs and desires, whatever they may be, on the other hand; and it seems preferable, and I shall do so here, to call an action rational, whenever it bears this relation to the given beliefs and desires. Which action exactly is rational in this sense, usually is the result of a big weighing in which each of the given beliefs and desires may in principle become relevant. This is vague, of course; but it is a well-defined task to clear this up, and it is quite a different task (which is not yet our topic) to investigate the rationality of beliefs and desires.

This one observation has two consequences for us. One is that, when dealing with action rationality, we really should entertain a subjectivistic interpretation of probability. For there will not be very much what can be said about action rationality independently of a person's subjective beliefs; decisions under uncertainty as well as game situations as characterized in the standard story simply seem to be *under-determined* problems from this point of view. But if a person's beliefs are to be explicitly regarded, then we have somehow to conceptualize these beliefs; and probability measures are a good way, to put it weakly, of such conceptualization. This goes without saying in philosophy, I think, but, strangely, it still seems to need some stressing among game theorists and economists.

In fact, the aversion to subjective probabilities is present in all of standard game theory. It is apparent in the conception and handling of chance moves, it shows up in the fact that the actions of others are not considered as subject of a player's probabilities, and it finds general expression in the stepchild-like treatment of the whole epistemic make-up of the players. There is no doubt that standard game theory has tided over this lack of the unloved subjective probabilities by brilliant substitutes, but it is equally clear, I think, that this aversion is the main cause for the incoherencies present in the standard story. And it has obscured the "reflective equilibrium"-approach to rationality sketched above.

The second consequence is that, if we are keen on capturing action rationality in a mathematical model, we are almost automatically led to decision theory. For the most natural way to mirror that big weighing of subjective beliefs and desires is to conceptualize them in some quantitative way; the

practically unrivaled candidates for such a quantitative conceptualization are, of course, probability measures and utility functions; and then the Bayesian rule of maximizing expected utility is the most plausible and mathematically simplest model of that weighing process and its outcome. Of course, this consideration alone cannot establish decision theory; but since the solid "reflective equilibrium"-groundwork has already done all to support this mathematical model, it may be put that simply.

What is important now is that this model gives us a *complete* account of an action being rational relative to given beliefs and desires. That is, any other account working within a comparable conceptualization is either entailed by or contradicts the decision theoretic account. (Strictly speaking, this is not quite true; there may be ties according to decision theory; and in these cases, but only in these cases, another account may be compatible with decision theory without being entailed by it.)

All that comes to this: We might perhaps quarrel with the received conceptualization of subjective beliefs and desires. But if we do not, then we cannot do full justice to action rationality when working with less than full decision models, and we have all we need for a complete characterization of action rationality when working with full decision models. Hence, from this general perspective too, we have no good choice but to keep a strict decision theoretic course in game theory as everywhere else where rational action is at issue.

Now it is at last time to submit the conjecture that game theory was interested not so much in action rationality in the weak sense discussed just now, but in the stronger sense of being also based on rational beliefs and perhaps on rational desires. The rationality of desires, however, is a very dim subject. There exists a not totally unclear notion of a desire being rational relative to other given desires, according to whether the first might be inferred from the latter by rational beliefs. But whether there is also some way of judging the rationality of desires absolutely – this is an open question reminiscent of the grave ethical problem whether there are such things as objective values. In this situation it is wise not to presuppose absolutely rational desires, and this is, of course, what every decision or game theorist has done by taking preferences and utility functions as subjectively given. Thus, we have only to discuss epistemic rationality to which we finally turn.

6 The real issues (continued)

First, I should briefly mention a familiar point (in order to forget about it subsequently), namely that the decision theoretic account of action rationality already assumes a formal minimum of epistemic rationality, i.e. that subjective probabilities behave like mathematical probabilities. But this was taken for granted all the time; of course, we now have in mind a material property which goes beyond this.

Actually, it is not so clear that standard game theory really is concerned with epistemic rationality and not only with action rationality. At least, I could not find good evidence for this in the standard references (like von Neumann, Morgenstern (1944) or Luce, Raiffa (1957)); this may also have to do with the somewhat undifferentiated "reflective equilibrium"-approach to explicating rationality. But the impression from the end of section 4 that our Bayesian story somehow did not do full justice to the standard story points to this concern. The issue becomes much clearer, when we look at what Harsanyi has written from his kind of Bayesian approach to game theory. For instance, in (1965), p. 450, he says:

"The basic difficulty in defining rational behavior in game situations is the fact that in general each player's strategy will depend on his expectations about the other players' strategies. Could we assume that his expectations were given, then his problem of strategy choice would become an ordinary maximization problem: he could simply choose a strategy maximizing his own payoff on the assumption that the other players would act in accordance with his given expectations. But the point is that game theory cannot regard the players' expectations about each other's behavior as given; rather, one of the most important problems for game theory is precisely to decide what expectations intelligent players can rationally entertain about other intelligent players' behavior. This may be called the problem of mutual 'rational expectations'."

In order to solve that problem, he proposes in (1966) not only "postulates of rational behavior in a narrower sense", but also "postulates of rational expectations"; on p. 621 he is then very explicit in stating that these postulates imply "that the *only* variables influencing the players' bargaining behavior will be:

(i) the *payoffs* associated with alternative outcomes for each of the players, and
(ii) the *subjective probabilities* each player assigns to different outcomes being accepted or rejected by the other player(s).

Among these variables, only those mentioned under (i) are *independent* variables while the variables under (ii) are themselves determined by the variables under (i)."

This last claim is all-important to Harsanyi's approach and to standard game theory as well. And I think it is basically wrong. (In fact, if I did not think so, I could have forborne this paper.) However, I cannot argue this strictly, since to this end I had to show for each principle of epistemic rationality one might plausibly entertain that it does not lead from (i) to (ii), and since, with the exception of some basic principles, there is not much agreement as to which principles should be entertained. Epistemic rationality is just much less elucidated than action rationality. No wonder, it is the

time-honoured, but still acute problem of induction in its full philosophical generality. But I shall try to make plausible why I think Harsanyi's claim to be wrong. Let me start by recalling some facts about epistemic rationality.

Firstly, it is clear that one cannot talk absolutely of beliefs being rational or not. A person's belief can be said to be rational only in relation to the evidence this person has. Part of this relation is explicated in deductive logic; whatever follows deductively from the evidence, is rationally to be believed. Inductive logic and statistics as well (which are both more controversial) have tried to clear up more of this relation. Here it has become apparent that the rationality of some epistemic state depends also on the prior epistemic state, i.e. that one should distinguish the problem of rational belief change – how is a prior epistemic state rationally to be modified in the light of new evidence? – from the problem of assessing the rationality of the prior epistemic state – which is the more difficult one. Actually, epistemic rationality is still more complicated; for example, it certainly depends also on the language in which the beliefs are represented. But such profound intricacies are not relevant for our discussion.[14]

Returning now to Harsanyi's claim, let us imagine again that Charlie and Lucy are engaged in some zero-sum game in normal form and assume some RUM (where the "R" still stands only for action rationality). Let us suppose that this does not yet solve the game (i.e. that the game is like our second example in section 4). Now we additionally assume Charlie to be epistemically rational. What does this help? Nothing, I think. We have already seen in section 4 that by deductive logic RUM does not imply anything which would narrow down Charlie's range of possible probabilities about Lucy. And I know of no plausible inductive principle which would do better in this respect. The same holds true of Lucy when we assume her to be epistemically rational. But then it is of no help to Charlie either to believe Lucy to be epistemically rational. And so on. Thus, even if we additionally assume that epistemic rationality is mutual knowledge of some order among Charlie and Lucy, we are not led to infer that they have the subjective probabilities which game theory would like them to have. And this is contrary to Harsanyi's claim that we would be led to infer so, i.e. that the utilities together with all the rationality we might wish (and with mutual knowledge of all this) would determine the subjective probabilities. Of course, this reasoning does not at all exclude that the assumption of epistemic rationality might be quite effective, when Charlie and Lucy are granted other or more evidence than RUM alone.

But instead of criticizing Harsanyi's claim we should perhaps better look at how he supports it. In his (1966) paper, however, from which I have quoted his claim, I have found no such support. There his rationality postulates indeed quite obviously imply that the players' actions depend only on (i) and (ii); but he loses no further word about his stronger claim. Unfortunately, much the same is true of other papers in which he explicates the

program sketched in (1966). (For these papers see the references of Harsanyi (1965) and (1966).)

Perhaps our interest is answered by the theory which he has recently developed together with Reinhard Selten, and which proposes a new two-stage procedure towards solving n-person non-cooperative games (cf. Harsanyi (1975) and (1976)):

> "First, a *prior* subjective *probability distribution* p_i is assigned to the pure strategies of each player i, meant to represent the other players' initial expectations about player i's likely strategy choice. Then, a mathematical procedure, called the *tracing procedure*, is used to define the solution on the basis of these prior distributions p_i. The tracing procedure is meant to provide a mathematical representation for the *solution process* by which rational players manage to coordinate their strategy plans and their expectations, and make them converge to one specific equilibrium point as solution for the game."
>
> (Harsanyi (1976), p. 211.)

This – in its details rather complicated – approach would deserve a longer discussion. But let it suffice to indicate why it, too, does not seem to diminish our troubles. If we apply the approach to two-person zero-sum games, then only its second step, the tracing procedure, is relevant (since it drives each prior probability distribution to the same equilibrium point, namely to the only one existing). Let us now consider only one player; assume Charlie to have some prior distribution over Lucy's choice set, which is not an equilibrium distribution. Why should Charlie change these prior probabilities, seemingly without being necessitated by some new evidence and according to the tracing procedure, which can hardly be linked up with any general principle of rational belief change? Why not stick to the prior probabilities which might be well-informed ones (though they would imply that he does not think that Lucy knows them – but why should he think so?)? The only reason Harsanyi gives for indulging in the tracing procedure is just that this prior distribution is not an equilibrium one and that for the reasons retold in our part 2 of the standard story only equilibrium points can be rational solutions of non-cooperative games (cf. Harsanyi (1975), pp. 70–75). Thus he takes for granted what is still in need of clarification for us.

Let us still look at Harsanyi (1973), where he comes nearest to our concerns in trying to overcome the apparent instability of equilibrium points in mixed strategies which we mentioned critically in our commentary to part 4 of the standard story. To this end he presents the following model: Let a non-cooperative n-person game, the "original game", be given in normal form, with $A_1, \ldots, A_n$ being the choice sets of the n players and with $V_1, \ldots, V_n$ being their utility functions on $A_1 \times \ldots \times A_n$. Harsanyi now thinks that

the real game situation may be more realistically described by a slightly different game, the "disturbed game", where the true utilities of each player i are not fixedly given by V_i, but rather oscillate within a small range around the values given by V_i because of "small stochastic fluctuations in his subjective and objective conditions (e.g., in his mood, taste, resources, social situation, etc.)" (Harsanyi (1973), p. 2). The probability laws governing these oscillations may be different for different players, but each player is assumed to know all these distributions. However, each player knows only of his own oscillating utilities how they exactly are at the moment of choice. Thus, in the normal form of the disturbed game, a possible pure strategy of player i is a function which tells him for each possible version of his true utility function which action to choose from A_i. The players' utility functions for the normal form of the disturbed game are then immediately to be inferred from the above description.

Now Harsanyi was able to prove essentially this: the disturbed game has at least one equilibrium point; each equilibrium point of the disturbed game is in pure strategies; if the players choose pure strategies being in equilibrium in the disturbed game, then, according to the probability laws for the utilities, these choices come down to mixed strategies in the original game which are approximately in equilibrium there; and the approximation is the better, the smaller the range of oscillation around the V_i's. This solves the stability problem, since in the disturbed game equilibriums are stable because being in pure strategies, and since choosing a pure strategy in the disturbed game implies choosing a randomized strategy in the original game; moreover, randomization comes about here because of the oscillation of the utilities and need not be carried out intendedly by the players.

It seems as if this model could provide the long sought justification of the epistemic assumptions (3), (6), and (9) in our "theorems". But at what costs does it do so? It has other strong assumptions instead. The conception of oscillating utilities is reminiscent of Thurstone's method of treating psychological variables as random variables (cf. Thurstone (1945)). This method was an important contribution to mathematical psychology, but, roughly, a severe, acknowledged difficulty of this method is to determine the distributions of these random variables (cf. Laming (1973), ch. 2). Thus, in a sense, Harsanyi requires our players to be better Thurstonian psychologists than our able scientists. But one need not interpret the oscillation of utilities as an objective probabilistic indeterminateness of the utilities; one might interpret the probability laws for these oscillations as expressing the subjective uncertainties of the players about one another. Then, however, it would be quite mysterious why the uncertainty about the utilities of, say, player i has exactly the same form for all other players. Now this last objection does not apply to two-person games (because there is only one other player). But even then the reinterpretation will not do, since the utility functions of the disturbed game are assumed there to be known to each player; and this

requires that everyone's probability distribution for the other players' utilities in the original game is known to each player. Thus, however interpreted, one can hardly be happy with the assumptions of Harsanyi's (1973) model. Besides, it still takes for granted that only equilibrium behaviour is rational in games with equilibrium points in pure strategies.

Do we have to despair in finding some kind of justification for (3), (6), and (9)? We indeed have to, I think, if we are looking for it in the field defined by what I have called Harsanyi's claim, i.e. only in the game situation at hand. In fact, this section has now led us exactly into the predicament where we ended up in section 4. And the way out was already hinted at there: we need not confine the evidence on which the epistemic rationality of the players is operating to the game situation at hand. After all, we might as well ask for some support or evidence for the assumptions (4), (5), (7), and (8), which are epistemic ones, too (by assuming Charlie to believe something). Here it is very clear that a player's belief of his fellows being rational and having such and such utilities cannot be evidenced in the given game situation alone; rather it can only be acquired through long and rich human experience (the details of which are obscure). Thus, this might also be the appropriate field of evidence for (3), (6), and (9); in particular, a player may have been in game situations a great many times, and he may thereby have formed the beliefs we would like him to have. In fact, Brown (1951) has already brought up this idea in connection with his iterative process of approximating equilibrium points of two-person zero-sum games by fictitious playing, which is also called the Brown-Robinson process[15]. Let me adapt this process to a rather simple story about Charlie and Lucy.

Suppose that Charlie and Lucy play a certain zero-sum game in normal form where their choice sets and utility functions are given, respectively, by A_1 and A_2, and $U_1 = U$ and $U_2 = -U$. They play it not once, but many, possibly infinitely many times. But they are simple-minded, they do not conceive the situation as a supergame, they even do not think about the other being rational and having such and such utilities. In each play they only maximize their expected utility as determined by their utility functions and their momentary subjective probabilities for the other's actions. Still, they are epistemically rational in adjusting their probabilities in the light of past experiences:

However, we do not want to be so restrictive as to assume that both conform to what is called the straight rule[16], i.e. that after the n-th play their probabilities for the other's actions in the $n+1$-st play are identical with the relative frequencies of the other's actions in the first n plays; by assuming this we would exactly copy the original Brown-Robinson process. We want to be a bit more liberal, in order to connect the process at issue with established principles of epistemic rationality.

We first assume that they follow the rule of conditionalization, which says that someone's probability $P_{t'}(C)$ for some event C at some time t' is to be

equal to his conditional probability $P_t(C \mid E)$ for C at some earlier time t, where E is the experience he has gathered between t and t'. This is the most basic rule of rational belief change.[17] For Charlie, e.g. this means that after n plays his probabilities for Lucy's actions in the $n+1$-st play are his prior probabilities for these conditionalized by what she has done in the first n plays.

Secondly, in order to retain the merits of the straight rule, we assume that they satisfy the so-called axiom of convergence or Reichenbach axiom, which says for Charlie, e.g., that the difference between his probabilities for Lucy's actions in the $n+1$-st play and the relative frequencies of these actions in the first n plays, whatever they are, converges to zero (for $n \to \infty$). Thus, one might say that the Reichenbach axiom ensures that in the end experience gets the upper hand of prior conceptions; it is therefore generally considered as a further minimal requirement of epistemic rationality.[18]

Now, if Charlie and Lucy have this much epistemic and action rationality, and if the original game has exactly one equilibrium point consisting of Charlie's (mixed or pure) strategy s_1 and Lucy's s_2, then we have: for each $a \in A_1$, the relative frequency with which Charlie chooses a in the described game process converges to the probability with which a shows up in s_1, and the corresponding is true of Lucy. Because of this, Charlie also tends to develop the appropriate belief (3) about Lucy, and vice versa for Lucy.[19] Thus, in the given special case this story meets all the demands arisen in our discussion above.

My point in telling this (mathematically trivially) liberalized version of the Brown-Robinson process is, to repeat it, not to remind us of something like the intuitive attractiveness of the Brown-Robinson idea; this would be superfluous. Rather, it is that *some such story must be told*, if we want to have reasonable theoretical grounds for such epistemic premises as (3), (6), and (9), which in turn must be included in game theoretic theorizing, if it is to be unassailable. And this is so because only such stories about game learning processes can give a theoretical account of the evidence leading epistemically rational players to the beliefs (3), (6), and (9) – an evidence which, as I have argued, cannot be found in the given game situation alone.

One might object that there are many ways of coming to the beliefs (3), (6), and (9) – the simplest one being that an advisory game theorist tells the players what to do and to believe (perhaps by telling the standard story of section 3) and that the players believe him. Of course, it may and often does go this way. But this is of no help to the game theorist: Firstly, he would not want to restrict his theory to people enlightened by him, and secondly, he certainly has no theory at all about the communicative exchange between him and the players, i.e. no theory about this way of coming to the beliefs (3), (6), and (9).

On the other hand, it must be admitted that the theory of game learning processes is not in a too promising shape. The Brown-Robinson process and

its liberalization are nice examples, but it hardly extends beyond the domain of two-person zero-sum games (cf. Rosenmüller (1971)). However, the assumptions of the Brown-Robinson process are rather poor; our Charlie and Lucy are there not even treated as genuine game theoretic subjects, because according to these assumptions each of them has to see the other as some irregular die whose propensity of landing with this or that side up has to be found out. Thus, the natural idea would be to enrich the assumptions of the game learning process by treating Charlie and Lucy as game theoretic subjects, i.e. by letting them know the other's utilities and by letting them theorize about the other's epistemic states. Whether such assumptions would make game learning processes move to the desired result in more general than only two-person zero-sum games, is, however, a very open question.

The long and the short of all this: In the absence of more concrete results, at least a general moral may be drawn from the previous discussion. Distinguish strictly between action rationality and epistemic rationality. If your concern is action rationality, then design full decision models for your subjects and determine rational action by the rule of maximizing expected utility; and if this alone does not satisfy you, if you search for some account for the epistemic assumptions written into the decision models, then keep strictly to some rules of epistemic rationality as basic and as widely acceptable as possible. Otherwise, theoretical and foundational confusion threatens.[20]

Notes

1 The problems of the normal form exhibited by Selten (1975) have no bearing on my considerations, which therefore apply as well to his improved conception.

2 For this action theoretic topic cf. e.g. Churchland (1970).

3 Though it is not the only and in my view even not the best one; cf. Spohn (1978), ch. 2.

4 Cf. e.g. Stegmüller (1970), ch. III–V, and (1973a), ch. VIII, or Putnam (1975), ch. 11, 12, and 22.

5 Similarly, a good and unified view of a strategically thinking and acting person is as one which theorizes decision-theoretically about his or her own future action; cf. Spohn (1978), ch. 4.

6 Cf. e.g. Lewis (1969) and Schiffer (1972).

7 For all details see Savage (1954), sect. 5.5, and Spohn (1978), sect. 2.3 and 3.6.

8 Since in Savage (1954) the decision rule is to maximize expected utility, it is no wonder that his reduction method likewise operates with expected utilities.

9 Cf. e.g. Luce, Raiffa (1957), p. 75.

10 It may be worth noting here that the attempts of explaining single-case probabilities in terms of long-run considerations have also proved to be inconclusive; cf. Hacking (1965), ch. IV.

11 In fact, I am a bit ashamed of the triviality of our "theorems". I had hoped to present something more informative; and indeed, there are many assumptions, maybe weaker, maybe more plausible, which one might try instead of RUE.

However, I have found no assumption as effective as RUE. But, after all, mathematical novelty is not my aim here.

12 Let me note by the way, that (1), (2), (4), (5), (7), and (8) are Charlies's half of no stronger RUM than RUM_2 and that (1)–(9) are Charlie's half of no stronger RUE than RUE_2. This is welcome, I think, since it seems that the higher we climb up the RUM- or the RUE-hierarchy, the more we lose ourselves in oddities.

13 Cf. e.g. Hempel (1961/62) or Churchland (1970) and other literature on rational explanation and explanation of actions.

14 This implies a trivial, but, I think, pertinent side remark: namely that rational belief and true belief must be strictly distinguished. Though probably most rational beliefs are true, most truths could only irrationally believed today (because our evidence is so poor), and many rational beliefs are false (because our evidence often is misleading). This is not to say that rational and true belief would not be interrelated, but the nature of that connection is a deep and open philosophical problem (cf. Peirce (1960), vol. V, §§384–385+405–408, or Putnam (1978)). Now the game theorist assumes his players to have many true beliefs, e.g. when he thinks the players to know the objective probabilities of chance moves, or when he assumes some RUM (all the $2n$ beliefs imputed, say, to Lucy by RUM_n are true according to RUM_n); and the point is that, whenever he does so, he introduces a genuine, new assumption which can in no way be accounted for by the epistemic rationality of the players alone. It seems to me that the standard story was not always quite clear about this point; for instance, when assuming not more than first order beliefs about rationality etc. (cf. part 1 of our standard story) the (wrong) idea might have been that the higher order beliefs somehow fall in by the rationality assumed.

15 Because Robinson (1951) has proved that Brown's idea works – cf. also Luce, Raiffa (1957), pp. 442ff.

16 Cf. e.g. Carnap (1952), §14.

17 The straight rule is not compatible with the rule of conditionalization, i.e. there is in general no prior probability measure with respect to which conditionalization yields the posterior probabilities dictated by the straight rule. In fact, this is the strongest theoretical ground for rejecting the straight rule. Cf. Carnap (1952), §14.

18 Cf. Stegmüller (1973b), pp. 502ff. – One might find it objectionable that the Reichenbach axiom expresses a limit property of subjective probabilities and says as such nothing about their actual form. However, there are "actual" properties of probabilities, most notably symmetry properties, which are known to imply the Reichenbach axiom. Cf. Carnap, Jeffrey (1971), parts 4 and 5.

19 All this is easily proved; Robinson's (1951) proof concerning the Brown-Robinson process simply extends to our somewhat liberalized version. If there should be more than one equilibrium point, a more complicated, but equally satisfying proposition holds true; cf. Robinson (1951). However, in contrast to the Brown-Robinson process, nothing can here be said about the convergence rate, because the Reichenbach axiom assumes nothing about convergence rates.

20 I warmly thank Prof. Reinhard Selten for encouragement and healthy scepticism, Ulrike Haas and Andreas Kemmerling for advice in putting and arranging things, Clara Seneca for checking my English, and the staff of *Theory and Decision* for showing me that there are some people for which this paper might be worth reading.

Bibliography

Brown, G.W. (1951), "Iterative Solution of Games by Fictitious Play", in: T.C. Koopmans (ed.), *Activity Analysis of Production and Allocation*, New York: Wiley & Sons, pp. 374–376

Carnap, R. (1952), *The Continuum of Inductive Methods*, Chicago: University Press

Carnap, R., Jeffrey, R.C. (eds.) (1971), *Studies in Inductive Logic and Probability*, vol. I, Berkeley: University of California Press

Chernoff, H. (1954), "Rational Selection of Decision Functions", *Econometrica* 22, 422–443

Churchland, P.M. (1970), "The Logical Character of Action-Explanations", *Philosophical Review* 79, 214–236

Hacking, I. (1965), *Logic of Statistical Inference*, Cambridge: University Press

Harsanyi, J.C. (1965), "Bargaining and Conflict Situations in the Light of a New Approach to Game Theory", *The American Economic Review* 55, 447–457

Harsanyi, J.C. (1966), "A General Theory of Rational Behavior in Game Situations", *Econometrica* 34, 613–634

Harsanyi, J.C. (1973), "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points", *International Journal of Game Theory* 2, 1–23

Harsanyi, J.C. (1975), "The Tracing Procedure: A Bayesian Approach to Defining a Solution for *n*-Person Noncooperative Games", *International Journal of Game Theory* 4, 61–94

Harsanyi, J.C. (1976), "A Solution Concept for *n*-Person Noncooperative Games", *International Journal of Game Theory* 5, 211–225

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

Laming, D. (1973), *Mathematical Psychology*, London: Academic Press

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

Luce, R.D., Raiffa, H. (1957), *Games and Decisions*, New York: Wiley & Sons

McClennen, E.F. (1976), "Some Formal Problems with the von Neumann and Morgenstern Theory of Two-Person Zero-Sum Games, I: The Direct Proof", *Theory and Decision* 7, 1–28

von Neumann, J., Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton: University Press, ^{2}1947

Peirce, C.S. (1960), *Collected Papers*, vol. I–VI (ed. by C. Hartshorne and P. Weiss), Cambridge, Mass.: Harvard University Press

Putnam, H. (1975), *Mind, Language, and Reality. Philosophical Papers*, vol. 2, Cambridge: University Press

Putnam, H. (1978), "Realism and Reason", in: H. Putnam, *Meaning and the Moral Sciences*, London: Routledge & Kegan Paul, pp. 121–140

Rawls, J. (1971), *A Theory of Justice*, Cambridge, Mass.: Harvard University Press

Robinson, J. (1951), "An Iterative Method of Solving a Game", *Annals of Mathematics* 54, 296–301

Rosenmüller, J. (1971), "Über Periodizitätseigenschaften spieltheoretischer Lernprozesse", *Zeitschrift für Wahrscheinlichkeitstheorie* 17, 259–308

Savage, L.J. (1954), *The Foundations of Statistics*, New York: Wiley & Sons
Schiffer, S.R. (1972), *Meaning*, Oxford: University Press
Selten, R. (1975), "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games", *International Journal of Game Theory* 4, 25–55.
Spohn, W. (1978), *Grundlagen der Entscheidungstheorie*, Kronberg/Ts.: Scriptor
Stegmüller, W. (1970, 1973a), *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie*, Band II, *Theorie und Erfahrung*. 1. Halbband 1970, 2. Halbband 1973, Berlin, Heidelberg: Springer
Stegmüller, W. (1973b), *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie*, Band IV, *Personelle und Statistische Wahrscheinlichkeit*, Berlin, Heidelberg: Springer
Thurstone, L.L. (1945), "The Prediction of Choice", *Psychometrika* 10, 237–253