First publ. in: Rationality, rules, and structure / ed. by Julian Nida-Rümelin ... Dordrecht : Kluwer, 2000, pp. 67-84

A Rationalization of Cooperation in the Iterated Prisoner's Dilemma

Wolfgang Spohn
Fachgruppe Philosophie
Universität Konstanz
D–78457 Konstanz

1. Introduction*

This paper consists of two parts. Its first part addresses an obvious and important lacuna in decision theory, our most basic and general theory of practical rationality. For all I can see, this lacuna is only insufficiently addressed in the literature presumably because it is not clear what to do about it and because the attempts to fill it lead onto very shaky grounds. Be this as it may, the first point of the paper will be to show a way how to close the gap.

I believe there are a number of fruitful applications of this amendment to decision theory, and they are the source of my confidence in my proposal.[1] However, in the second part of my paper I would rather like to consider whether the general ideas of the first part help to throw some new light on Prisoner's Dilemma and on Newcomb's Problem, the millstones around the necks of game and decision theory. I am not at all sure of this further application, but I am excited, I admit, by the outlook that it might work.

The ideas of the first part are also formally developed in Spohn (1999, sect. 3). Here, however, I shall try to remain completely informal. This seems feasible for the following reason. Strategic thinking is essentially a recursive matter; one reasons back-

---

* I am indebted to Arthur Merin for improving my English.

[1] See Spohn (1999, sect. 4). This is a larger manuscript the essentials of which are condensed in the first part of this paper.

wards stepwise from the time horizon of one's plans until one reaches the present time of decision. The recursion may become mathematically quite complicated. However, all conceptual difficulties lie in the recursion step. Hence, I shall confine my attention to one such step; doing so allows me to avoid formal matters for the most part.

2. Strategies Generalized

If the lacuna in decision theory is as obvious as I claim, where do we find it? Let us look at the standard account of strategic thinking or sequential decision making. Its starting point is the insight that it would be silly to fix a whole action sequence in advance, come what may. Rather, at the time of decision only the first action has to be decided, and then one sees how the situation develops before one takes further action. However, in finding the optimal first action, one must already anticipate how one's situation may possibly evolve and how one would proceed in each possibility. This thought is the birth of strategic thinking. So, what is a strategy? Standardly, it consists of a first choice and a function assigning a further choice to each possible course of events up to this further choice. But I said I would confine my considerations to one recursive step, and then a strategy s simply says: "Now I do a; and if e_1 happens I do b_1; if e_2 happens, I do b_2; and so forth."

There are many first actions and many strategies. The task is to find the most reasonable first action which requires in general to find the optimal strategy. By which criterion, however, is optimality to be measured? The standard criterion runs as follows, with respect to my above sample strategy: If e_1 obtains and hence the action sequence $<a,b_1>$ is realized, this action sequence is evaluated on the basis of knowledge of e_1; that is, the expected utility of this action sequence is calculated from the given utility function and from the given probabilities conditionalized on e_1. Likewise, if e_2 obtains and the action sequence $<a,b_2>$ is carried out; and so on. The expected utility of the strategy s itself, finally, is the expectation of these expected utilities of the various possible action sequences, i.e. a weighted mixture of these exptected utilities with the present probabilities of e_1, e_2, ... as weights. These simple ideas are the basis for an im-

pressive theory of strategic or sequential decision making with highly sophisticated applications particularly in statistics.[2]

This account rests, however, on the obvious assumption that the agent learns which of the possible events occurs. Otherwise, he could not make his behavior contingent on these events. Of course, the standard theory has always been perfectly clear about this assumption. This suggests, though, that I have just slightly misdescribed what strategies are. A strategy makes the agent's behavior contingent not on various external facts, but rather on various pieces of information he receives or on information states he reaches, where information states are something internal to the agent.

Now the lacuna I have in mind is already evident. For, the conclusion just reached may be expressed a bit more pointedly: The events on which future action is made dependent by strategies are in fact possible decision situations the agent may get into – where I take, here and henceforth, a decision situation to consist not of the external circumstances in which the agent is located, but of his relevant internal make-up. Thus, a decision situation δ consists first of a conceptualization or framing of the external situation – be it in Savage's (1954) terms of states, acts, and consequences or in Jeffrey's (1965) terms of an algebra of propositions and particular action propositions among them, or whatever; and it secondly contains a probability function P and a utility function U defined for the framing from which the relevant expected utilities may be calculated.[3] Whatever the precise construction of P and U, the conceptualization or framing is somehow implicit in them; we may thus represent a decision situation δ by an ordered pair $<P,U>$. Hence, what a strategy s really says, is this: "Now I do a; and if I get into $\delta_1 = <P_1,U_1>$, I do b_1; if I get into $\delta_2 = <P_2,U_2>$, I do b_2; and so forth."

Thus viewed, the standard account looks pretty restricted. It deals only with the case where the decision situations possibly reached by the agent arise from his present decision situation through observation and hence through conditionalizing the present probabilities with respect to the event actually observed. This is perhaps the most important and prevailing case. But obviously there are many other cases; new decision si-

---

[2] See, for instance, Raiffa (1968) as an introductory text and Pratt, Raiffa, Schlaifer (1995) as a more advanced text.

[3] In Spohn (1977, and 1978, ch. 2) I have argued that the decision models of Fishburn (1964) are the most general and appropriate so far available.

tuations may arise in a countless number of ways: due to uncertain observation, as noted by Jeffrey (1965, ch. 11), due to more indirect information, due to make-believe and wishful thinking, due to forgetting, due to endogeneous changes of tastes and preferences, due to drugs, due to the evolution of addictions and phobiae, and so on. Why not take a strategic attitude toward any kind of new decision situation, however it comes about? Decision theory is incomplete unless we generalize to all these cases.

Of course, this incompleteness has not remained unobserved. The first to consider such non-standard cases was, it seems, Strotz (1955/56). It took some time for the second contribution, Pollak (1968), to appear. Further progress was rather sporadic. I mention Peleg, Yaari (1973), Hammond (1976), Yaari (1977), McCain (1979), Becker, Murphy (1988), and at book length Elster (1979) and McClennen (1990). There is more, but not immensely much more. Given the amount of effort spent on developing decision theory in general, my impression that the problem raised by its incompleteness attracts only marginal interest appears justified.

3. The Problem of Optimality

We have already taken the first step towards a completion of decision theory, namely by conceiving of strategies in a more general way; this was the easy part. The second step, though, the evaluation of the generalized strategies, is much more difficult. Indeed, its obscurity sufficiently explains, perhaps, why people have tended to steer clear of these complications. However, in order to attain a full notion of practical rationality, we have to take the second step as well, i.e. to find a general criterion of optimality for strategies thus conceived.

At first, it may seem that we may simply carry over the standard criterion, that is, that we evaluate the action sequence $<a,b_1>$ from the point of view of $\delta_1 = <P_1,U_1>$ and likewise for the other action sequences possibly implemented by the strategy *s*, and that we finally evaluate the strategy *s* itself by the expectation of the values of these action sequences. There is no conceptual difficulty in thus defining the value of strategies.

However, this criterion is certainly inadequate. In order to see this, let us look at a familiar, but still instructive example:

Some friends living in the countryside have invited me to a dinner party, and I am thinking about how to get there. There is no public transport, so I should have to drive there in my own car or I take a taxi. I foresee that the evening will be entertaining and the wine delicious and that I will be pleasantly drunk. If I drive there, I shall have another decision problem at the end of the evening: drive home again or take a taxi? I know already what my sanguine assessment will be: taking a taxi is expensive; fetching the car the other day is annoying; the risk of getting stopped by the police is negligible; the risk of having an accident is as low as ever; etc. Hence, the expected utility of driving myself home is much larger than that of taking a taxi. Moreover, driving there is, considered in itself, much better than going there by a taxi. Applying finally the standard criterion as suggested this leads to the result that the present value of the strategy <driving there, driving home> is much larger than that of the strategy <taking a taxi three, taking a taxi home>.

However, this is clearly *not* what we think to be rational. The rational deliberation goes rather like this: I know that I shall be in a light-hearted mood after having drunk so much and that my assessment will be as described. But that assessment is silly. Now, being sober, I judge that the chance of getting caught by the police is not negligible, that the chance of having an accident is heavily increased, and that the avoidance of these risks is well worth the price of a taxi going there and back. It is this assessment, and not the future silly one, on which I should base my present decision. Hence I better order a taxi right now.

Thus, the deliberational structure appropriate to this case is just opposite to the one for the case of observation, which was adequately handled by the standard theory. In both cases I know what I shall do when I get into this or that decision situation; this follows from the rationality theory presupposed for the recursion step. The cases differ, however, in the following crucial respect: In the case of observation I rely on my future evaluations and use them in order to now assess how desirable it is to get into this or that decision situation; i.e., I base my present decision on my possible future evaluations. In the case of drunkenness, by contrast, I do not do so, finding my future

evaluations inferior; rather, I reevaluate the future situation from my present point of view which I find superior and base my decision on this reevaluation. The upshot is that we have two special cases with convincing, though opposing optimality criteria, each of which bars the other one from serving as a general model. This looks like a serious problem.

Let me repeat the problem in a more artificial and pointed way: Here I am, in my present decision situation $\delta = \langle P,U \rangle$. I expect to be, within a certain span of time, in the situation $\delta_i = \langle P_i,U \rangle$ with a certain subjective probability $P(\delta_i) = p_i$ ($i = 1, ...,n$); the only difference of these future situations to my present one lies in the subjective probabilities I have in them. Now two scenarios are possible: In the one I expect the change from P to one of the P_i to be due to observation[4]; in the other I expect it to be due to forgetting, spontaneous opinionatedness, or whatever. The first scenario is possible only under a certain restriction, namely that P is the mixture of the P_i with the weights p_i[5]; but we may assume that, by chance, the other scenario satisfies this restriction as well. The standard optimality criterion I have sketched above is appropriate in the first scenario; here I can take over the future evaluation of my substrategies in δ_i from the point of view of P_i. But the second scenario resembles the case of the dinner party; here I have a prediction about my behavior in the future situation δ_i, but my intuition tells me clearly that I should base my present evaluation of that future behavior on P and not on P_i. Thus, the two scenarios require different optimality criteria, but there is nothing *in* δ and the δ_i to distinguish the two cases. This is why I think that there cannot be any solution of the problem within the confines of the standard theory of strategic or sequential decision making. What we need is additional structure in order to distinguish between the two scenarios.

---

[4] Which may also be an essentially probabilistic one, as Jeffrey (1965, ch. 11) has proposed, so that P_i need not result from P by conditionalization with respect to an event observed with certainty.

[5] I have stated this condition in Spohn (1978, p. 162); it is equivalent to the reflection principle of van Fraassen (1984).

4. A General Solution

The difference between the scenarios was intuitively clear; in consisted in how I arrive at the future situations, through observation, through drinking alcohol, through forgetting, or whatever. Thus the additional structure should somehow provide this kind of information. This can be done as follows.

The general problem, to repeat, is this: I am now in the decision situation δ, having probabilities P and utilities U, pondering about which action a to take next, and anticipating that I shall get into one of the possible decision situations $\delta_1 = \langle P_1,U_1\rangle$, $\delta_2 = \langle P_2,U_2\rangle$, ... Where I get to is usually act-dependent; hence P is to provide also all probabilities of the form $P(\delta_1 \mid a)$ telling how likely I am to get into δ_1 given that I do a. There are a lot of strategies of the general form as explained above, but only one of them is optimal[6], and it tells what to do right now. Two principles, I think, answer the quest for optimality:

There is, first, the *principle of consistency*.[7] It says that a strategy s can be optimal in δ only if for each decision situation δ_i possibly reached it prescribes a substrategy s_i which is optimal by the lights of δ_i – where it is precisely the criterion of optimality I am about to develop which recursively says what is optimal in δ_i. This principle seems unassailable to me. I simply assume that I shall behave rationally also in the future decision situations I shall reach. This assumption is part and parcel of normative rationality theory; insofar I expect non-rational behavior of myself, it does not occur in a genuine decision situation towards which I take a strategic stance and which is dealt with by normative theory of rationality.

However, the principle of consistency does not yet provide any kind of evaluation of the strategies for the original situation δ. This is afforded by the second principle I want to propose, the *principle of reevaluation*. It says that each future decision situation δ_i has to be reevaluated from the relevant superior decision situation. What do I mean thereby? What is relevant? What is superior?[8]

---

[6] Or some of them are; but I shall simply proceed in the singular.

[7] Stated already in Strotz (1955/56).

[8] For a more detailed and precise description see Spohn (1999, sect. 3).

To take up the latter first, superiority is a primitive notion of my account. I simply assume that a decision maker has a notion of one possible decision situation being superior to another and that this relation of superiority among the possible or even counterfactual decision situations is an element of the present decision situation. Formally, we may assume that this relation is a partial ordering. Substantively, there are many plausible examples. The more informed situation is superior to the less informed one, even if the surplus information consists of bad news. The reflection principle of van Fraassen (1984) is, I think, the basic principle guiding purely doxastic changes to superior situations. Forgetting or brain-washing leads to an inferior situation. Maturing and aging and the accompanying shift of interests and utilities presumably lead to an incomparable situation. Becoming more sophisticated is usually superior to remaining primitive; acquiring higher virtues, if such there are, is presumably superior to retaining lower virtues. Quite generally, the autonomy of the evolution of one's desires and utilities is an important criterion of superiority; hence, addictions set the addict on the road to inferiority. And so on. Certainly, there are many examples on which opinions diverge. We need not decide upon them, however, because only the subjective opinion of the decision maker matters in his situation. The examples I gave should only make clear that we all have a rich notion of the relation I call the relation of superiority.

Secondly, what is relevance supposed to mean here? When considering the possibility of moving from δ to δ_1, the principle of consistency requires me to evaluate the options open in δ_1 from the point of view of δ_1 itself. But the example of the dinner party has shown that these options have to be reevaluated. The general suggestion is, of course, that this reevaluation is done from a superior point of view. In the case of the dinner party, the present sober state provides the superior point of view. In the case of observation, the observationally informed state is the superior point of view by itself; hence, the reevaluation of δ_1, being identical with the evaluation in δ_1, is not a genuine one in this case. In this way, the relation of superiority is supposed to afford the required discriminations.

However, there are usually many, and many fancyful, situations superior to δ_1. So, from which one is δ_1 to be reevaluated? Not just anyone of them will do; otherwise, we would get lost in fancy and ambiguity. Hence, we have to restrict considerations to the

relevant superior situation. In the two examples above, I already referred to the relevant situation at hand. And in general I propose to explain that the superior situation which is relevant to the move from δ to δ_1 is that situation δ_1' superior to or identical with δ_1 which is feasible in the sense that I would reach δ_1' if, in going from δ to δ_1, I could cancel all moves into inferior direction and use instead the opportunities to move into superior direction. Again, the precise content of this explanation should be studied by looking at various examples.[9] Yet I hope that the general idea is already clear enough.

Now, I can state my principle of reevaluation. It says that the evaluation of a strategy *s* in the original situation δ has to proceed in the following way: For each situation δ_i possibly reached, determine first the relevant superior situation δ_i', then reevaluate the substrategy s_i of *s* for δ_i according to δ_i', and finally define the value of *s* as the expectation of the reevaluations of the substrategies – where the expectation is, of course, conditional on the first action prescribed by *s*.

The required optimality criterion can now be stated in full generality as well: it says that a strategy is *optimal* if and only if it is consistent and has a maximal value among all consistent strategies according to the principle of reevaluation.

By way of comparison, I should mention that my principle of reevaluation, combined with the principle of consistency, is so far a specific proposal for what has been called sophisticated choice. However, other accounts of sophisticated choice like Strotz (1955/56), Pollak (1968), Peleg, Hammond (1976), and Yaari (1977) do not employ anything similar to my superiority relation; insofar, my proposal seems to offer a substantial refinement of these accounts.

Sophisticated choice has been penetratingly criticized by McClennen (1990), sect. 11.3; it may seem, hence, that my proposal also falls victim to this criticism. This would be a false impression, however. In section 6 I shall sketch an extension of my proposal, and I shall indicate (see footnote 24) how this extension is able to encompass resolute choice, the alternative offered and defended by McClennen (1990), ch. 12. If I am right in this, my proposal may indeed be used to unify various acounts in the field.

---

[9] They may be found in Spohn (1999, sect. 4). There, in sect. 3, I also address the worry whether *the* relevant superior situation is always unique. It is not, but the problem may be overcome.

My principle of reevalutation looks complicated, and if one properly works out the full recursion, it begins to look even more imperspicuous.[10] Still, I am quite confident that my proposal is reasonable. It is perfectly general, it agrees with the restricted standard account, in simple cases it boils down to something simple, and then its power emerges in dealing with a great variety of simple cases in a way which seems intuitively very plausible.[11] But I am not going here to defend these claims by looking at all these cases and studying more closely the basic relation of superiority. I rather want to turn to an application which is too shaky to serve as support of my general account, but which I would like to put up for discussion all the more because it is without doubt important.

5. Prisoner's Dilemma: An Endless Story

The application I have in mind is the prisoner's dilemma (PD) the relevance of which to almost all areas of practical philosophy is unsurpassed. Let me briefly resume my point of departure:

In the one-shot PD I am prepared, for the time being[12], to accept that the only rational solution is defection; analogously, the only rational thing to do in the one-shot Newcomb problem is to take both boxes. In both cases, if you had cooperated or taken only one box, you may rightfully regret not having chosen to get more. Thus I am firmly on the side of causal as opposed to evidential decision theory. The decision maker has only an absolute, act-independent subjective probability for the other player's cooperation or the predictor's prediction, and, whatever the probability is, it is rational for the decision maker to take the dominating action, i.e. defecting or two-boxing. If that probability were act-dependent, that would express the decision maker's thought to have a causal influence on the other player's behavior or on the predictor's prediction – which, ex hypothesi, he denies and excludes.[13]

---

[10] See Spohn (1978, sect. 4.4, and 1999, sect. 3).

[11] See the cases discussed in Spohn (1999, sect. 4).

[12] But see footnote 27.

[13] This is roughly how I argued in Spohn (1978, sect. 5.1). The basis of the argument is a probabilistic theory of causation which entails that exactly one of the four combinations of probabilistic and

Problems arise, however, with the iterated PD; indeed, I take it to be a scandal of normative rationality theory that there still does not seem to be a fully rational account in favor of cooperation. Let me briefly sketch ten different views on the iterated PD. Each view contributes a highly illuminating idea; their collection shows the tremendous intellectual challenge the iterated PD continues to present; however, even their collection is not sufficiently revealing, I think.

(1) Intuitively, the case appears quite clear. If two subjects play PD many times and do not manage to set up cooperation, but are caught in defection, they are terribly silly. They are not only collectively silly, due to a tragic conflict between individual and collective rationality. Rather, they seem to be individually accountable for their failure. At least one of them must have been individually irrational; the other one may have been irrational, too, or she may have been rational, though unable to do better than defect against her silly opponent. Intuition equally clearly tells us what is rational in the iterated PD: namely to start and maintain a pattern of mutual, conditional trust and kindness which secures long and stable cooperation, relative to which disintegration in the final plays would be annoying, but negligible. This intuition is supported by the computer tournaments of Axelrod (1984) in which the tit-for-tat strategy, which instantiates this intuition in an exemplary way, was most successful.

However, attempting to back up this intuition by a theory of rationality is utterly frustrating. The central cause of all frustrations is, of course, the famous backward induction argument purporting to show that in the finitely iterated PD the only equilibrium strategy for both players is always to defect. Let us see what people have done about it.

(2) One idea is to move into the context of evolutionary game theory. Here, whole populations are occupied with playing PD, and one can choose plausible set-ups in which the cooperative parts of the population turn out to be much more successful then

---

causal dependence and independence of the prediction (or the other player's action) on/from the decision maker's action is impossible, namely the combination of probabilistic dependence and causal independence – which is just the case which Nozick (1969) and many following him have found troublesome. Within the framework of directed acyclic graphs, Meek and Glymour (1994) give an account of intervening or deciding, as opposed to prediction, which embodies the very same conclusion.

the defecting ones so that society evolves to consist mainly of cooperative individuals.[14] However, this move, illuminating as it is for empirical theory construction, simply changes the topic; we wanted to learn about individual rationality, but evolutionary game theory does not teach us anything in this respect.

(3) In practice, there is a simple and usually effective method to make cooperation individually rational: we empower an authority to offer rewards for cooperation and to punish defection. This often helps, but it is sometimes difficult to implement and sometimes hardly feasible. More importantly, from a theoretical point of view, this means changing the utility functions of the players until there is no longer any PD; this is a way to avoid PD, not to solve it.

(4) So, let us look at the twin of PD, Newcomb's problem. Here, evidential decision theory seems to offer a viable rationalization of taking only one box.[15] The dominance argument is thereby invalidated and backward induction deprived of its basis, and thus one may think of carrying over this rationalization to PD. I cannot engage now into the ramified argument between evidential and causal decision theory[16]; let me only express my conviction that this move is of no avail: either, evidential decision theory gets the causal relations right as in Eells (1982) and recommends two-boxing; or it neglects causal relations, or it gets them wrong, and is therefore inadequate.[17]

(5) Davis (1977) argues that the players should decide in PD according to a mirror principle saying: "If two rational agents have the same evidence and preferences, they will make the same (nonrandom) choice." (The name and the phrasing of the principle are due to Sorensen 1985, p. 158.) We may assume that both players firmly believe that this principle holds and that they satisfy its premise. Thus they are certain to do the same, and then cooperation emerges as the only rational alternative even in the one-shot case.

---

[14] Axelrod, Dion (1988) briefly present the intricacies and ramifications of the evolutionary treatment of PD.

[15] See, for instance, Gibbard, Harper (1978).

[16] See, for instance, the papers collected in Campbell, Sowden (1985).

[17] Indeed, I am surprised how small the impact of the heated philosophical discussion has been in economics and in game theory; evidentialism seems to be a philosophical, but only weakly contagious disease.

Is there something wrong with the mirror principle? No, I think one should seek ways to maintain it. However, if its acceptance by the players is taken with Davis (1977) as entailing their neglect or denial of the causal independence of their actions, then we are back at the evidentialism just discarded. If it is not so taken, then it becomes clear, I think, that the mirror principle is still incomplete; it does not say anything about the rational mechanism leading the agents from given evidence and preferences to a certain choice. If that mechanism is standard game theory, both players will expect one another to defect. Whether there is another account of rationality entailing cooperation remains the crucial question which is not answered by the mirror principle and which I want to tackle in the next section.

(6) In Spohn (1982, pp. 254-6) I argued that the standard argument for equilibrium behavior proceeding from the common knowledge of the game situation is incomplete in a very similar way. And indeed there are many reasons for finding fault with the standard Nash equilibrium concept.[18] Do these reasons open an escape from the backward induction argument? No; however serious the doubts are about equilibria in general, they seem inappropriate in the finitely iterated PD since the backward induction argument shows that always defecting is not only an equilibrium, but indeed the unique (weakly) rationalizable strategy in the sense that no other strategy survives the iterated elimination of weakly dominated strategies. Thus, mutual knowledge of the utility functions and the Bayesian rationality of the players is sufficient for establishing continued defection as the only rational option.

(7) So, perhaps the crucial fault lies in the logic of the backward induction argument itself? This doubt has been raised by Binmore (1987), Bicchieri (1989), and Pettit, Sugden (1989) and denied, for instance, by Aumann (1995). I side with Aumann, but surely one should scrutinize this discussion much more carefully than I can do here. Let me only add that as soon as one grants backward induction to hold under certain idealizations, however strong, the problem remains. If cooperation in the iterated PD can be rational for reasonable and well-informed players like us, it should be so all the more

---

[18] See, for instance, the diagrams of van Damme (1991, pp. 335f.) showing the impressive ramifications of the equilibrium concept.

for perfectly rational and perfectly informed players; one cannot be satisfied with allowing an exception and prescribing defection in the strongly idealized case.[19]

(8) The most substantial game theoretic contributions are still to be mentioned. One line of thought is to assimilate the very often iterated PD to the infinitely iterated PD. Then one may adduce the rich battery of so-called folk theorems[20] showing that in the infinitely iterated case there are infinitely many more or less cooperative equilibria. This is an ingenious and very sophisticated observation. But it is obviously not fully satisfying, since it imputes to the players the clearly and knowably false assumption of infinite repetition.

One may, however, interpret the folk theorems in a different way: In one variant of these theorems the utilities in the future plays are discounted by some factor $\alpha < 1$, and this discount factor may also be understood as expressing the players' subjective probability in each play that there will be a next play at all, so that the probability for an infinity of plays is in effect 0. However, I find even this interpretation implausible because there is still a positive probability for any finite number of plays. The plausible assumption would be that we all are sure to play PD at most, say, a million times in our life and very likely much less; and this assumption turns the strategic situation into a finitely iterated PD. Hence, cooperation should be rationally possible also in this case.

(9) In conversation, Teddy Seidenfeld proposed to me another variant which drastically changes the picture: make the continuation of the game in some way dependent on past cooperation; there may or may not be an upper limit to the number of plays. This idea has already been fruitfully applied by Feldman and Thomas (1987) in the context of evolutionary game theory. Its point in the context of individual decision making is obvious: this variant set-up provides for a simple and theoretically sound rationalization of cooperation in standard decision theoretic terms and avoids the devastating backward induction. This is a beautiful idea, but it provides only a partial solution and dissolves neither the original problem nor the desire to solve it as well.

(10) The idealizations required for backward induction, i.e. the relevant assumptions of mutual knowledge, may well fail, of course. This is the entry of the perhaps most

---

[19] As does Sobel (1993, sect. 6).

[20] Cf., for instance, van Damme (1991, ch. 8) or Osborne, Rubinstein (1994, ch. 8).

interesting attempts to establish a cooperative solution. The catch notions are the (trembling hand) perfect equilibria of Selten (1975) and the sequential equilibria of Kreps, Wilson (1982). Very roughly, the idea pursued here is that a rational strategy must define rational behavior even for situations which can only be reached if some players behave irrationally and that one should always expect with a small probability that irrational behavior occurs intentionally or unintentionally. Taking these things into account may promote cooperation in the following way: I cooperate in the first play, because my hand trembles or maybe because I follow a sophisticated plan. My partner is surprised, but then starts thinking how my perceived irrationality can be explained and maintained, and perhaps he reaches the conclusion that he should cooperate as well in the second play, and so on; cooperation may thus be the perfectly rational continuation of a somehow irregular or irrational beginning.[21]

This picture may be realistic, but from the point of view of normative rationality theory it seems distorted. The normative intuition which demands compliance is that cooperation is rationally possible and indeed rational, and that it must be so without any help from direct or indirect gaps or deficiencies in rationality; it seems not good enough to show how cooperation can emerge as a form of bounded rationality.

So the suggestion in particular from (7) and (10) is that it is the normative rationality theory itself which needs to be reformed, and my brief summary was, I hope, not unfair in suggesting that no working idea for this reform seems available.

6. A Way Out?

However, in the first part of this paper I have already proposed a reform of rationality theory. Does it help, perhaps, to illuminate the present problem as well? Yes, I think it does in a certain way – a way which may seem cheap or miraculous; but it would be surprising, on the other hand, if the solution would have to be very complicated or sophisticated. So, here is the line of thought I want to propose:

---

[21] A precise story is told by Kreps et al. (1982).

Intuitively, we would reason as follows in the finitely iterated PD: if the first play is considered in itself, I should defect there because I have no influence whatsoever on what you do in the first play. However, when I cooperate now this may have the effect of raising the probability of your cooperation in later plays. That is how cooperation in the first play may have maximal expected utility. So far, so good. But how could I raise the probability of your cooperation in the second play? By the same consideration, your cooperation in the second play can only get a positive probability if you think that it raises the probability of my cooperation in the third play. And so the hope of raising the probability of cooperation is deferred to later and later plays until the final play where we know already that it will be badly disappointed. Hence, there is no rational hope in making cooperation likelier, and thus the intuitive reasoning fails.[22] Of course, this is again nothing but a form of the old backward induction argument.

Let me put this impossibility in a somewhat different way: It is constitutive of PD that I do not believe in a correlation between our choices in the first play; your choice is just an independent, i.e. causally and (hence[23]) probabilistically independent state of the world for me. But somehow we would like to believe in a probabilistic correlation between our actions in later plays. How could we have this belief?

The first difficulty is that as a decision maker pondering about my future actions I do not have any subjective probabilities for these future actions of mine; I determine the best or rational course of action, and then we may or may not add the epiphenomenal belief in that course of action. Indeed, I still think that this is an important point which Savage (1954) got right and Jeffrey (1965) wrong. From the point of view of rationality theory it is only the above generalized strategic thinking which allows us to have a probabilistic assessment of the actions considered, namely by assuming subjective probabilities for getting into various future decision situations which rationality theory must view as complete deterministic causes for the actions taken in them.

Hence, we can believe in a correlation between our actions in the later plays of PD only if we believe our decision situations in the later plays to be correlated. So, I have to

---

[22] By contrast, however, this reasoning would be perfect in the scenario considered in sect. 5, (9) above.

[23] See footnote 13 above.

imagine myself being in various possible decision situations later on, say, in the second play. How could they vary? Apparently only by containing varying beliefs, i.e. varying subjective probability functions. But how then could the possible decision situations in the second play also vary in their optimal actions, as they must when a correlation between these actions is to emerge as well? Only by containing varying assumptions about the correlations in the third or in later plays. They cannot contain such assumptions concerning the second play itself, because then I can no longer believe our choices in the second play to be correlated; I have to believe then into the causal and (hence) probabilistic independence of these choices. In this way a present belief in future correlation can only derive from a future belief in still more distant correlation, and again the castle in the air collapses in the last play. Thus, if belief in correlated future action presupposes belief in correlated future decision situations, and if these future situations differ only in their subjective probabilities, the backward induction argument strikes again, and there is no rational way to entertain such a belief.

However, the last conclusion suggests to consider a further possibility: namely that the future decision situations we would like to believe to be correlated differ also in their utility functions. How might this come about? This possibility seems to violate the very set-up of the iterated PD. But no, there is, I think, a way of rationalizing this suggestion.

Let us start from the supposition that I believe that our actions are correlated, not in the first play, but in the second and later plays. As a consequence, I realize that we are caught in a continued perspectival trap by the iterated PD which consists in the fact that the actions in later plays which I now believe to be correlated cannot seem correlated to me at the later time of choice; at that later time I can view your action only as a causally and hence probabilistically independent state of the world.

Now, the first part of this paper enters the argument. For, if I perceive these later plays as a trap, I judge these later decision situations as we have conceived them so far to be inferior in the specific sense discussed earlier. The relevant superior situation from which to assess this inferior situation is the one in which my cooperation and my defection in the later plays are reevaluated by receiving an additional utility or disutility which makes their overall utility correspond to the expected utility they would have under the presently assumed correlation with your actions in the later plays.

It is not implausible, I think, to apply here the superiority/inferiority distinction in this way. On reflection, such a trap-like structure turns out to be not unusual; quite often the formation of our decision situations is in some way negatively or counterproductively correlated with the external circumstances. For instance, a strong desire to eat often arises due to some more or less subconscious frustration which, unlike hunger, does not vanish by eating; hence, a situation in which such a desire for food is present is inferior to the same situation without this desire. Similarly, there is a not uncommon tendency of men to fall in love, i.e. to assign a high utility to having relations, with women who maintain an utter reserve and to ignore women who are obliging. Sometimes, men have this tendency because they do not really want to get close to women. But others severely suffer from this tendency; for them the situation of being in love with an unapproachable woman and not being interested in a responsive one is inferior to the reverse situation. These examples differ, however, from iterated PD, and iterated Newcomb as well, because they embody a trappy desire formation, whereas it is belief formation which is trappy in iterated PD and Newcomb. Indeed, it is so in a perfectly schematic way simply due to my moving in time; at any time the other player's next choice, or the predictor's next prediction, is probabilistically independent from my next choice only from the present, but not from any earlier point of view. This remarkable feature should, however, not distract from the similarity to the other cases.

So, let us return to PD. What is the point of thus applying the superiority/inferiority distinction? It does not yet seem to help: As I have explained in the first part, the relevant superior situation is so far only a hypothetical situation, and if the decision situations I actually expect to reach are those in the trap, the only consistent strategy is still always defecting; hypothetical reevaluation alone cannot change this. So a further step is required. It says that if my superiority assessment is as explained, *I should actually move into the superior situation.* This may seem strange, but it is the core of the solution I want to propose. In the examples discussed in the first part, I was moved into new decision situations by external forces like observation, alcohol, and other things in a way which was not under my immediate control. In the present case, by contrast, I want to suggest that it is the pure insight into the trap-like structure of the whole set-up which should rationally move me into the superior situation with its adjusted utility function.

What I thus propose, in effect, is a *law of rational utility change* which is not a mere change of expected utilities due to changes of subjective probabilities. This resembles the usual practical solution of PD mentioned in section 5, (3), which consists in changing the utility functions of the players from outside; my suggestion is that internal rationality alone should have the same effect as external punishment.

In fact, I am inclined to think that such a law of rational utility change has a much wider application beyond PD. Whenever I seem to be stuck in an inferior position, I should rationally change my evaluation so as to reach the so far only hypothetical superior point of view. This is so in the case of addictions; if I realize I am (getting) addicted, I should assign an evaluative malus to future addictive behavior. Likewise, if I realize I am overeating out of frustration, doing so should rationally receive a negative evaluation. If I become conscious of my self-frustrating longing for unapproachable women, this should dampen my longing in future cases. And so on. However, such a law of rational utility change refers only to my evaluations. It is not an automatic consequence that I manage to anchor the changed evaluation in my motivation and thus in my behavior. This anchoring seems indeed easy in the PD case, but it may be difficult or virtually impossible in the case of addictions, uncontrolled eating, etc. However, this problem does not directly affect the question of what my evaluations should rationally be.[24]

With such a law of rational utility change my argument may now be brought to an end. The full theory of rationality is now the one amended by these laws and mechanisms. Hence, if I think that you are rational I think that you conform to this amended theory as well. Moreover, I do not firmly believe into a fixed correlation between our actions in the later plays. Correspondingly, there is no fixed superior decision situation into which I should move. Rather, how strong a correlation I assume depends on our present actions which may or may not intensify the assumed correlation. In this way, cooperating may or may not become the rational thing to choose in the later plays, and it

---

[24] The background of this remark is the observation that the notion of utility has at least three different aspects, namely what Kusser (1989) calls evaluation, motivation, and satisfaction, which are ideally and perhaps also usually congruent (this is why they have hardly been distinguished in the received decision theory), but which may, and often do, diverge, thus making room for many interesting phenomena. See also Kusser, Spohn (1992).

may do so for me and for you in a correlated way.[25] Rationally, however, you and I should start with believing in a strong correlation which will then be confirmed so that long-standing cooperation will indeed emerge. The crucial point of this reasoning is that thereby the present belief in the correlation in later plays does not derive from the later belief in still more distant correlation. In this way, the reasoning breaks the force of the backward induction argument.

To summarize, my argument is that the subjective assumption that the chosen actions will be correlated puts in force an enriched theory of rationality employing superiority assessments and a new law for changing utilities, and this enriched theory in turn makes the assumption of correlation rationally entertainable. By this kind of bootstrapping, correlation is rationally believable and cooperation thus rationally possible in a full sense, even in the finitely iterated PD.[26]

Mutatis mutandis, these considerations should apply to the iterated Newcomb problem. This then would be my offer as a causal decision theorist to the evidentialist whose intuition I share that it cannot be rational to stay poor if one has the chance to get rich. It is not true that "the reason why we", the causal decision theorists, "are not rich is that the riches were reserved for the irrational" (Lewis 1981, p. 377); the reason is that we were caught in too narrow a notion of rationality. The truly rational man does not pity himself because only allegedly irrational men are consistently (pre-)rewarded; he should be able to adapt, and my proposal shows how to do so in a rational way.[27]

References

---

[25] It would be interesting at this point to study the relation to the theory of correlated equilibria initiated by Aumann (1974). Indeed, my account may help to explain how a correlated equilibrium may arise without explicit agreement or external arbiter.

[26] This enriched account of sophisticated choice is also able to incorporate resolute choice as favored by McClennen (1990). This is most easily explained with the toxin puzzle invented by Kavka (1983) which is also a trap-like decision situation preventing one from getting the large reward for having formed the firm intention to drink the painful toxin. McClennen recommends in this situation to resolve first to form the intention and then to actually drink the toxin without reconsideration. In my account, one realizes the trap, changes the utilities accordingly, forms thus the required intention, and sticks to it by drinking the toxin even after reconsideration. In this way, utility change may yield the same result as resolute choice.

[27] It has been suggested to me several times in conversation that my reasoning might be applied to the one-shot case as well. This may be so, though the causal independence of your action (in PD) or the predictor's prediction (in Newcomb) from my action looks unshakeable. It is obvious that such suggestions can only be checked after my proposal in the last section is carried out with formal rigor – something which needs to be done some time soon.

Aumann, R.: 1974, 'Subjectivity and Correlation in Randomized Strategies', *Journal of Mathematical Economics* 1, 67-96.

Aumann, R.: 1995, 'Backward Induction and Common Knowledge of Rationality', *Games and Economic Behavior* 8, 6-19.

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

Axelrod, R., and D. Dion: 1988, 'The Further Evolution of Cooperation', *Science* 242, 1385-1390.

Becker, G.S., and K.M. Murphy,: 1988, 'A Theory of Rational Addiction', *Journal of Political Economics* 96, 675-700.

Bicchieri, C.: 1989, 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge', *Erkenntnis* 30, 69-85.

Binmore, K.: 1987, 'Modeling Rational Players: Part I', *Economics and Philosophy* 3, 179-214.

Campbell, R., and L. Sowden (eds.): 1985, *Paradoxes of Rationality and Cooperation*, The University of British Columbia Press, Vancouver.

Davis, L.: 1977, 'Prisoners, Paradox, and Rationality', *American Philosophical Quarterly* 114, 319-327.

Eells, E.: 1982, *Rational Decision and Causality*, Cambridge University Press, Cambridge.

Elster, J.: 1979, *Ulysses and the Sirens*, Cambridge University Press, Cambridge.

Feldman, M.W., and E.A.C. Thomas: 1987, 'Behavior-Dependent Contexts for Repeated Plays of the Prisoner's Dilemma II: Dynamical Aspects of the Evolution of Cooperation', *Journal of Theoretical Biology* 128, 297-315.

Fishburn, P.C.: 1964, *Decision and Value Theory*, Wiley, New York.

Gibbard, A., and W.L. Harper: 1978, 'Counterfactuals and Two Kinds of Expected Utility', in: C.A. Hooker, J.J. Leach, and E.F. McClennen (eds.), *Foundations and Applications of Decision Theory, vol. 1*, Reidel, Dordrecht, pp. 125-162.

Hammond, P.: 1976, 'Changing Tastes and Coherent Dynamic Choice', *Review of Economic Studies* 43, 159-173.

Jeffrey, R.C.: 1965, *The Logic of Decision*, Chicago University Press, Chicago, 2nd. ed. 1983.

Kavka, G.: 1983, 'The Toxin Puzzle', *Analysis* 43, 33-36.

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

Kreps, D.M., and R. Wilson: 1982, 'Sequential Equilibria', *Econometrica* 50, 863-894.

Kusser, A.: 1989, *Dimensionen der Kritik von Wünschen*, Athenäum, Frankfurt a.M.

Kusser, A., and W. Spohn: 1992, 'The Utility of Pleasure is a Pain for Decision Theory', *Journal of Philosophy* 89, 10-29.

Lewis, D.: 1981, "Why Ain'cha Rich?", *Noûs* 15, 377-380.

McCain, R.A.: 1979, 'Reflections on the Cultivation of Taste', *Journal of Cultural Economics* 3, 30-52.

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

Meek, C., and C. Glymour: 1994, 'Conditioning and Intervening', *British Journal for the Philosophy of Science* 45, 1001-1021.

Nozick, R.: 1969, 'Newcomb's Problem and Two Principles of Choice', in: N. Rescher et al. (eds.), *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht, pp. 114-146.

Osborne, M.J., and A. Rubinstein: 1994, *A Course in Game Theory*, MIT Press, Cambridge, Mass.

Peleg, B., and M.E. Yaari: 1973, 'On the Existence of a Consistent Course of Action When Tastes are Changing', *Review of Economic Studies* 40, 391-401.

Pettit, P., and R. Sugden: 1989, 'The Backward Induction Paradox', *Journal of Philosophy* 86, 169-182.

Pollak, R.A.: 1968, 'Consistent Planning', *Review of Economic Studies* 35, 201-208.

Pratt, J.W., H. Raiffa, and R. Schlaifer: 1995, *Introduction to Statistical Decision Theory*, MIT Press, Cambridge, Mass.

Raiffa, H.: 1968, *Decision Analysis*, Addison-Wesley, Reading, Mass.

Savage, L.J.: 1954, *The Foundations of Statistics*, Dover, New York, 2nd. ed. 1972.

Selten, R.: 1975, 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games', *International Journal of Game Theory* 4, 25-55.

Sobel, J.H.: 1993, 'Backward-Induction Arguments: A Paradox Regained', *Philosophy of Science* 60, 114-133.

Sorensen, R.A.: 1985, 'The Iterated Versions of Newcomb's Problem and the Prisoner's Dilemma', *Synthese* 63, 157-166.

Spohn, W.: 1977, 'Where Luce and Krantz Do Really Generalize Savage's Decision Model', *Erkenntnis* 11, 113-134.

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

Spohn, W.: 1982, 'How to Make Sense of Game Theory', in: W. Stegmüller, W. Balzer, and W. Spohn (eds.), *Philosophy of Economics*, Springer, Berlin, pp. 239-270.

Spohn, W.: 1999, "Strategic Rationality", Forschungsberichte der DFG-Forschergruppe "Logik in der Philosophie", Nr. #, Konstanz.

Strotz, R.H.: 1955/56, 'Myopia and Inconsistency in Dynamic Utility Maximization', *Review of Economic Studies* 23, 165-180.

van Damme, E.: 1991, *Stability and Perfection of Nash Equilibria*, Springer, Berlin, 2nd ed.

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

Yaari, M.E.: 1977, 'Endogeneous Changes in Tastes: A Philosophical Discussion', *Erkenntnis* 11, 157-196.