Homo Oeconomicus XX(2/3) (ACCEDO Verlagsgesellschaft, München 2003)

Dependency Equilibria and the Causal Structure of Decision and Game Situations

by

Wolfgang Spohn*

Abstract: The paper attempts to rationalize cooperation in the one-shot prisoners' dilemma (PD). It starts by introducing (and preliminarily investigating) a new kind of equilibrium (differing from Aumann's correlated equilibria) according to which the players' actions may be correlated (sect. 2). In PD the Pareto-optimal among these equilibria is joint cooperation. Since these equilibria seem to contradict causal preconceptions, the paper continues with a standard analysis of the causal structure of decision situations (sect. 3). The analysis then raises to a reflexive point of view according to which the agent integrates his own present and future decision situations into the causal picture of his situation (sect. 4). This reflexive structure is first applied to the toxin puzzle and then to Newcomb's problem, showing a way to rationalize drinking the toxin and taking only one box without assuming causal mystery (sect. 5). The latter result is finally extended to a rationalization of cooperation in PD (sect. 6).

Contents:

---

*Prof. Dr. Wolfgang Spohn, Fachbereich Philosophie, Universität Konstanz, D–78457 Konstanz, email: wolfgang.spohn@uni-konstanz.de.

1. Introduction[*]

The driving force behind this paper is, once more, the great riddle posed by the prisoners' dilemma (PD). This has elicited a vast literature and a large number of astonishingly varied attempts to undermine defection as the only rational solution and to establish cooperation as a rational possibility at least in the iterated case. But the hard case, it seems to me, still stands unshaken. Under appropriate conditions backward induction is valid[1]; hence, given full rationality (instead of some form of 'bounded rationality') and sufficient common knowledge, continued defection is the only solution in the finitely iterated PD. The same conclusion is reached via the iterated elimination of weakly dominated strategies.[2] I find this conclusion highly disconcerting; it amounts to an outright refutation of the underlying theory of rationality. Moreover, I find that all the sophisticated observations made so far about

---

[*]This paper was conceived and partially written during my stay at the Center for Interdisciplinary Research of the University of Bielefeld in the summer term 2000, where I participated in the research group „Making Choices. An Interdisciplinary Approach to Modelling Decision Behaviour". I am grateful to the Center and to the organizers of the group, Reinhard Selten, Werner Güth, Hartmut Kliemt, and Joachim Frohn, for the invitation, and I am indebted to all the participants of the research group for the strong stimulation and encouragement I have experienced. Moreover, I am grateful to Max Albert, Robert Aumann, Christopher von Bülow, Werner Güth, Rainer Hegselmann, Arthur Merin, Matthias Risse, Jacob Rosenthal, Thomas Schmidt, and Yanis Varoufakis for comments on preliminary drafts, to Philip Dawid for pointing out to me a serious flaw in section 3.2 (which is now repaired in a satisfying way, I hope), to Adelheid Baker for meticulously checking my English, and to Hartmut Kliemt for being indulgent with this overly long paper. In particular the rich comments by Max Albert, Matthias Risse, and Yanis Varoufakis have shown me the insufficiency of my paper. The arrangement of the subject matter may be infelicitous in various respects, the comparative remarks in section 6.3 and throughout the text are too scarce in almost all respects. But clearly, the topic is inexhaustible, and the question is: how to live up to an inexhaustible topic? Only by coming to a preliminary end, none the less.

[1]Cf. Aumann (1995).

[2]Iterated elimination of weakly dominated strategies is a reasonable procedure when applied to the iterated PD, all the more so as the criticism this may meet compared to the elimination of strongly dominated strategies do not obtain in this application. Cf., e.g., Myerson (1991, sects. 2.5 and 3.1).

PD have failed to tone down this harsh conclusion.[3] Cooperation *must* remain at least a rational possibility in the finitely iterated PD, and under ideal conditions even more so than under less ideal ones.[4] Thus, something needs to be changed in standard rationality theory, i.e., decision and game theory. After a long time of thinking otherwise[5], I have come to the conclusion that it is the single-shot case that needs to be reconsidered, which this paper tries to do.

This is how my plot is supposed to go.[6] Section 2 introduces and discusses a new notion of equilibrium for games in normal form which I call dependency equilibrium. In particular, mutual cooperation will be a dependency equilibrium in the single-shot PD. These equilibria may be an object of interest in their own right, but they seem to assume an unintelligible causal loop, namely, that the players' actions causally influence each other (this may be why they have not been considered in the literature). Thus, my main effort in this paper will be to make causal sense of them; then and only then they deserve interest.

To this end, section 3 introduces some basics of the theory of causation that has become standard and has been implicitly or explicitly assumed in decision and non-cooperative game theory since their inception. The crucial observation will be this (section 3.2): Whenever we reduce a richer causal model to a coarser one by deleting some variables, each (apparent) causal dependence in the coarser model expresses either a possibly indirect causal dependence or the relation of having a common cause in the richer model (or a still more complicated relation which will turn out, though, to be irrelevant for our concerns). Hence, the apparent mutual causal dependence of the actions in a dependency equilibrium may only signify that they have a common cause.

This will lead us to the question of how an action is caused. In rational decision theory the answer can only be that it is caused by the decision situation (= the agent's subjective state). Hence, if the causes of actions are to enter the causal picture of the

---

[3]As I have more fully explained in Spohn (2000, sect. 5).

[4]Aumann's proof of backward induction assumes common knowledge of resilient rationality (CKR), i.e., the mutual belief that the players will behave rationally even after arbitrary deviations from the rational equilibrium path. Aumann (1995, p. 18) grants that "CKR is an ideal condition that is rarely met in practice" ("ideal" in the sense of "idealized"), and since this assumption looks implausible even as a rationally entertainable belief (cf. Rabinowicz 1998), one may hope to find a loophole here. I am skeptical, however. In actual life, I have certainly to reckon with the irrationality of my fellow humans, but this should not be the only possibility how my repeated cooperation may turn out as rational. It should all the more prove as rational, given the most unshakeable mutual beliefs in common rationality.

[5]Since Spohn (1978, sect. 5.1) I have been a fervent defender of the two-box solution of Newcomb's problem, but I have changed my mind (see sect. 5.2 below). In Spohn (2000, sect. 6) I have offered a line of thought for breaking the force of backward induction in the iterated case, but I am withdrawing it since I do not see anymore how it can be reasonably worked out.

[6]In case this summary is too abstract, it may at least serve as a reference.

agent, we have to develop standard decision theory to become what I call reflexive decision theory (because the possible decision situations themselves are now reflected in decision models). This is the task addressed in section 4, where the crucial observation will be this: A decision situation (= the agent's subjective state) may have other effects besides the action. If we now reduce the reflexive model to a standard non-reflexive model, the action will appear to cause these other effects in the reduced model, though reflexion shows that they actually have only a common cause.

In section 5, this observation is applied to the toxin puzzle and to Newcomb's problem. This is suggested by the fact that these cases allow (though they do not force) us to conceive the agent's decision situation as having such side effects, i.e., as somehow causing the prediction of the relevant predictor. Thereby we can rationalize drinking the toxin or taking only one box *within causal decision theory* in a perfectly straightforward way. Having got thus far, I transfer this kind of analysis to the two-person case of PD where we may similarly conceive the decision situations of the players as being causally entangled. Thus, the dependency equilibria in PD with their apparent causal loop in the actions acquire causal sense and may thus be rationalized. In this way, cooperation emerges at least as a rational possibility, and backward induction cannot even start. This is the task dealt with in section 6.1.

I will be mainly occupied with developing this line of thought in detail. However, a fictitious dialogue in section 6.2 and a number of comparative remarks in section 6.3 will, I hope, further clarify the nature of my approach.

The paper borrows from many sources. But I should emphasize that, with regard to section 4, my main debt is to Eells (1982) who was the first to take the reflexive perspective in decision theory. In section 5, it will become obvious how strongly I am influenced by the theory of resolute choice developed by McClennen (1990). I hope to advance it by showing how resolute choice may be subsumed under the reflexive extension of standard rationality theory.

2. Dependency Equilibria

2.1 Nash, correlated and dependency equilibria

Let me start with an outline of the new equilibrium concept. For comparison, it is useful to rehearse Nash equilibria and Aumann's correlated equilibria. We shall deal only with normal form games. Hence, the refinements of Nash equilibria relating to the extensive form are out of our focus. It suffices to consider two-person games. While I hardly develop the theory here, it may be routinely extended, it seems, to n-person games.

Thus, let $A = \{a_1,\ldots,a_m\}$ be the set of pure strategies of Ann (row chooser) and $B = \{b_1,\ldots,b_n\}$ the set of pure strategies of Bob (column chooser). Let u and v be the utility functions of Ann and Bob, respectively, from $A \times B$ into $\boldsymbol{R}$; we abbreviate $u_{ik} = u(a_i, b_k)$ and $v_{ik} = v(a_i, b_k)$.

Moreover, let S be the set of mixed strategies of Ann, i.e., the set of probability distributions over A. Hence, $s = \langle s_1,\ldots,s_m\rangle = (s_i) \in S$ iff $s_i \geq 0$ for $i = 1,\ldots,m$ and $\sum_{i=1}^{m} s_i = 1$. Likewise, let T be the set of mixed strategies of Bob. Mixed strategies have an ambiguous interpretation. Usually, the probabilities are thought to be intentional mixtures by each player. But it is equally appropriate to interpret them as representing the beliefs of others about the player. Indeed, in relation to dependency equilibria, this will be the only meaningful interpretation.

We shall envisage the possibility that the actions in a game may be governed by any probability distribution whatsoever. Let P be the set of distributions over $A \times B$. Thus, $p = (p_{ik}) \in P$ iff $p_{ik} \geq 0$ for all $i = 1,\ldots,m$ and $k = 1,\ldots,n$ and $\sum_{i,k} p_{ik} = 1$. Each $p \in P$ has a marginal s over A and a marginal t over B. But since p may contain arbitrary dependencies between A and B, it is usually not the product of the marginals s and t. This is all the terminology we shall need.

As is well known, $\langle s, t\rangle$ is defined as a *Nash equilibrium* iff for all $j = 1,\ldots,m$ $\sum_{i,k} s_i t_k u_{ik} \geq \sum_{k} t_k u_{jk}$ (or, equivalently, for all $s^* \in S$ $\sum_{i,k} s_i t_k u_{ik} \geq \sum_{i,k} s_i^* t_k u_{ik}$) and if the corresponding condition holds for the other player. Hence, in a Nash equilibrium neither Ann nor Bob can raise her or his expected utility by changing from her or his equilibrium strategy to some other pure or mixed strategy, given the other player sticks to his or her equilibrium strategy. There is no need here to rehearse the standard rationale for Nash equilibria, and there is no time to discuss its strengths and weaknesses.[7]

Obviously Ann and Bob's choices from A and B are independent in a Nash equilibrium. This is an assumption I would like to abandon (for reasons that will become clear later on). Aumann (1974) has introduced an equilibrium concept that allows for dependence between the players. Here is his definition from Aumann (1987) (which is a little simpler and less general than his original definition which would require us to introduce additional structure):

Let $p \in P$ have marginals $s \in S$ and $t \in T$. Then p is a *correlated equilibrium* iff for all $j = 1,\ldots,m$ $\sum_{i,k} p_{ik} u_{ik} \geq \sum_{k} t_k u_{jk}$ (or, equivalently, for all $s^* \in S$ $\sum_{i,k} p_{ik} u_{ik} \geq \sum_{i,k} s_i^* t_k u_{ik}$) and if the corresponding condition holds for the other player. The most

[7]This has been done many times, also by myself in Spohn (1982).

straightforward way to understand this, which is offered by Aumann himself (1987, pp. 3f.), is the following: Somehow, Ann and Bob agree on a joint distribution over the strategy combinations or outcomes of their game. One combination is chosen at random according to this distribution, and each player is told only their part of the combination. If no player can raise their expected utility by breaking their part of the agreed joint distribution and choosing some other pure or mixed strategy instead, then this joint distribution is a correlated equilibrium. Thus, correlated equilibria are self-enforcing, they do not need external help from sanctions or agreements.

Correlated equilibria appear to fall outside non-cooperative game theory. However, one can model the selection of a joint distribution for the original game as an additional move in a game enlarged by preplay communication, and it then turns out that all and only the Nash equilibria of the enlarged game correspond to correlated equilibria in the original game.[8] This reflects the fact that correlated equilibria, despite their allowance of dependence, are still non-cooperative in essence. The players' standard of comparison is still whether they might be better off by independently doing something else, where the expectations about the other player are given by the marginal over their strategies.

This standard of comparison is changed in the dependency equilibria introduced below. It is not the expected utility given the marginal for the other player, but rather the *conditional expected utility* given the conditional probabilities determined by the joint distribution.

Here is a first attempt to formalize this idea: Let $p \in P$ have marginals $s \in S$ and $t \in T$. Let $p_{k|i}$ be the probability of b_k given a_i, i.e., $p_{k|i} = p_{ik} / s_i$, and $p_{i|k} = p_{ik} / t_k$ the probability for a_i given b_k. Now, p is a *dependency equilibrium* iff for all i with $s_i > 0$ and all $j = 1, \ldots m$ $\sum_k p_{k|i} u_{ik} \geq \sum_k p_{k|j} u_{jk}$ and if the corresponding condition holds for the other player. Thus, in a dependency equilibrium each player maximizes their conditional expected utility with whatever they do with positive probability according to the joint equilibrium distribution.

This provokes at least three immediate remarks.

The first point to be taken up is a technical flaw in the above definition. If some a_j has probability 0 in the joint distribution p, i.e., if $s_j = 0$, then no conditional probability given a_j is defined. Yet, the fact that $s_j = 0$ should not render the other figures meaningless. There are three ways to solve this problem. One may, first, engage in non-standard probability theory where one can conditionalize with respect to events having infinitesimal probability. This looks circumstantial at the least. One may, second, resort to Popper measures that take conditional probabilities as basic and have

[8]For details, cf. Myerson (1991, pp. 255-257).

thus no problem with conditionalizing on null events. This would be the way I prefer.[9] However, the game theoretic community is rather accustomed to the third way, engaging in epsilontics, i.e., in approaching probability 0 by ever smaller positive probabilities. This strategy is easily applied to our present problem.

Let us call a distribution $p \in P$ *strictly positive* iff $p_{ik} > 0$ for all i and k. Now we correct my flawed definition by an approximating sequence of strictly positive distributions; this is my official definition: $p \in P$ is a *dependency equilibrium* iff there is a sequence $(p^r)_{r \in N}$ of strictly positive distributions such that $\lim_{r \to \infty} p^r = p$ and for all i with $s_i > 0$ and $j = 1,\ldots,m$ $\lim \sum_k p^r_{k|i} u_{ik} \geq \lim \sum_k p^r_{k|j} u_{jk}$ and for all k with $t_k > 0$ and all $l = 1,\ldots,n$ $\lim \sum_i p^r_{i|k} v_{ik} \geq \lim \sum_i p^r_{i|l} v_{il}$. All the conditional probabilities appearing in this definition are well-defined. Though the definition looks more complicated now, the intuitive characterization given above still fits perfectly.

After this correction, the second point is that dependency equilibria seem to be well in line with decision theory. Most textbooks state that the general decision rule is to maximize *conditional* expected utility. Savage (1954) still assumed a clear separation of states of the world having probabilities and consequences carrying utilities; consequences are then determined by acts and states. The pertinent decision rule is simply maximizing expected utility. However, this separation is often not feasible, and the more general picture put forward by Fishburn (1964) is that everything is probabilistically assessed (except perhaps the acts themselves), though only conditionally on the possible acts. In this general picture, maximizing *conditional* expected utility is the appropriate decision rule. It may seem surprising that this situation in decision theory has so far not been reflected in equilibrium theory.

But of course, and that is my third point, this is not astonishing at all. The idea behind the general picture is (we shall have to look at all this much more carefully in the next section) that the conditional probabilities somehow hide causal dependencies which are more generally modeled in a probabilistic and not in a deterministic way (as Savage 1954 did).[10] In the light of this idea, dependency equilibria are a mystery. If Bob chooses after observing what Ann has chosen, then, clearly, his choice causally depends on hers; that is the simplest case of a one-way causal dependence. But how can Ann's choice at the same time depend on Bob's? That would amount to a causal

---

[9]My real preferences, though, are for probabilified ranking functions, a sophistication of Popper measures; cf. Spohn (1988, sect. 7).

[10]States of the world may then be distinguished by their probabilistic and causal independence from the acts; but they do no longer play the special role of contributing to the deterministic causation of the consequences.

loop, and dependency equilibria seem to assume just this impossibility. So, whoever might have thought of them, he should have dismissed them right away as nonsense.[11]

However, the case is not as hopeless as it seems, and the main effort of this paper will be to make causal sense of dependency equilibria. I am not sure whether I shall fully succeed, but I hope to prepare the ground, at least. For the time being, let us look a little more closely at the properties of dependency equilibria.

2.2 Some examples

The computation of dependency equilibria seems to be a messy business. Obviously it requires to solve quadratic equations in two-person games, and the more persons, the higher the order of the polynomials we become entangled with. All linear ease is lost. Therefore, I cannot offer a well-developed theory of dependency equilibria. Thus, it seems advisable to look at some much discussed simple games in order to develop a feeling for the new equilibria, namely, Matching Pennies, BoS (Bach or Stravinsky), Hawk and Dove, and PD. This discussion becomes more vivid when we consider the other kinds of equilibria for comparison. Afterwards, we can infer some simple theorems from these examples.

Matching Pennies: This is the paradigm for a pure conflict, i.e., a zero- or constant-sum game. It is characterized by the following utility matrix:

u \ v	b_1	b_2
a_1	0, 1	1, 0
a_2	1, 0	0, 1

It is clear that it has exactly one Nash equilibrium and exactly one correlated equilibrium. It is characterized by the following distribution:

p	b_1	b_2
a_1	$1/4$	$1/4$
a_2	$1/4$	$1/4$

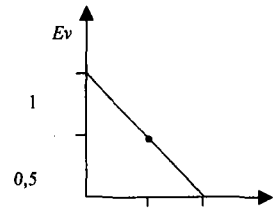
[11] Which I did when I first conceived of them in 1982.

By contrast, it is easily verified that the dependency equilibria of this game may be biased toward the diagonal or toward the counter-diagonal:

p	b_1	b_2
a_1	x	$\frac{1}{2}-x$
a_2	$\frac{1}{2}-x$	x

where $0 \leq x \leq \frac{1}{2}$.

It is instructive to represent the players' expected utilities in the various equilibria by a joint diagram:



• : Nash, corr.
– : depend.

Bach or Stravinsky: This game is a paradigmatic coordination game superimposed by a conflict. Its utility matrix is:

u \ v	b_1	b_2
a_1	2, 1	0, 0
a_2	0, 0	1, 2

As is well known, this game has three Nash equilibria, two in pure strategies (the players can meet on the diagonal) and a mixed one:

p	b_1	b_2
a_1	1	0
a_2	0	0

p	b_1	b_2
a_1	0	0
a_2	0	1

p	b_1	b_2
a_1	$\frac{2}{9}$	$\frac{4}{9}$
a_2	$\frac{1}{9}$	$\frac{2}{9}$

The correlated equilibria of this game form just the convex closure of the Nash equilibria:

p	b_1	b_2
a_1	x	$\leq 2x, 2y$
a_2	$\leq \frac{x}{2}, \frac{y}{2}$	y

The dependency equilibria are again of three kinds:

p	b_1	b_2
a_1	1	0
a_2	0	0

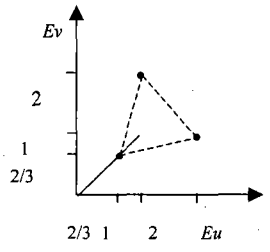
p	b_1	b_2
a_1	0	0
a_2	0	1

provided the zero rows and columns are approximated in an appropriate way, and

p	b_1	b_2
a_1	x	$\frac{2}{3}-x$
a_2	$\frac{1}{3}-x$	x

, *where* $0 \leq x \leq \frac{1}{3}$.

The players' expected utilities in these equilibria come to this:



• : Nash
-- : corr.
•, – : depend.

Quite similar observations can be made about pure coordination games without conflict like meeting at one of two places.

Hawk and Dove: This game represents another very frequent type of social situation. It will show even more incongruity among the equilibrium concepts. So far, one may have thought that the correlated equilibria are the convex closure of the Nash equilibria. But this is not true. I shall consider the utility matrix preferred by Aumann because it illustrates that there are correlated equilibria which Pareto-dominate mixtures of Nash equilibria; hence, both players may improve by turning to correlated equilibria. However, they may improve even more by looking at dependency equilibria. Here is the utility matrix:

u \ v	b_1	b_2
a_1	6, 6	7, 2
a_2	2, 7	0, 0

There are again three Nash equilibria with the following expected utilities:

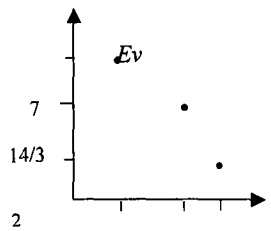
p	b_1	b_2
a_1	0	1
a_2	0	0

p	b_1	b_2
a_1	0	0
a_2	1	0

p	b_1	b_2
a_1	4/9	2/9
a_2	2/9	1/9

The correlated equilibria reach out further on the diagonal. They are given by

1.1.1



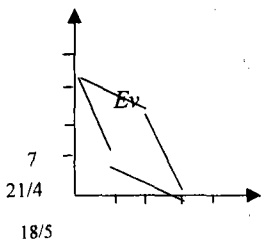
p	b_1	b_2
a_1	x	y
a_2	z	w

, *where* $x+y+z+w = 1$ *and* $0 \leq x, 2w \leq y, z$

and they yield the following expected utilities:

1.1.2

1.1.3



Again, we have three kinds of dependency equilibria:

p	b_1	b_2
a_1	0	1
a_2	0	0

p	b_1	b_2
a_1	0	0
a_2	1	0

provided the zero rows and columns are approximated in an appropriate way, and

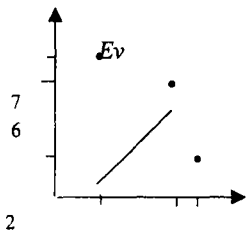
p	b_1	b_2
a_1	x	y
a_2	y	$1-x-2y$

, *where* y =

$\frac{1}{18}(2-15x+\sqrt{4+156x+9x^2})$.

This makes evident that we slip into quadratic equations. The corresponding expected utilities reach out still further on the diagonal:

1.1.4



Clearly, 6 > 21/4 > 14/3, the maximal values reached on the diagonals of the three diagrams.

Prisoners' Dilemma: This is my final and perhaps most important example. Its utility matrix is:

u \ v	b_1	b_2
a_1	2 | 2	3 | 0
a_2	0 | 3	1 | 1

There is only one Nash equilibrium:

p	b_1	b_2
a_1	0	0
a_2	0	1

Indeed, defection (= a_2 or, respectively, b_2) strictly dominates cooperation (= a_1 or b_1); hence, there can be no other Nash equilibrium. For the same reason, this is also the only correlated equilibrium.

The dependency equilibria, by contrast, have a much richer structure. They come in two kinds:

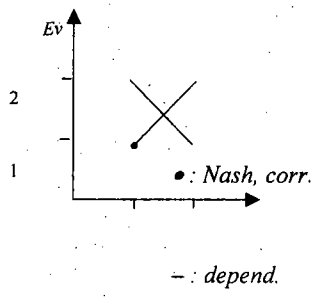
p	b_1	b_2
a_1	$\frac{1}{2}x(1+x)$	$\frac{1}{2}x(1-x)$
a_2	$\frac{1}{2}x(1-x)$	$\frac{1}{2}(1-x)(2-x)$

, where $0 \leq x \leq 1$, *and*

p	b_1	b_2
a_1	$\frac{3}{8}(1-x)(1+x)$	$\frac{1}{8}(1-x)(1-3x)$
a_2	$\frac{1}{8}(1+x)(1+3x)$	$\frac{3}{8}(1-x)(1+x)$

, where $-\frac{1}{3} \leq x \leq \frac{1}{3}$.

The expected utilities in all these equilibria look very simple:



It is of particular interest here that joint cooperation is among the dependency equilibria; indeed it weakly Pareto-dominates all other such equilibria. Of course, it is a well-worn and very simple observation that such dependence between the players may make them cooperate. But now we have found an equilibrium concept that underpins this observation. Moreover, we have seen that correlated equilibria do not provide the right kind of dependence for this purpose, they succumb to defection. Evidently, all this is strong motivation to try to make good sense of dependency equilibria. This is the task I shall pursue in the rest of the paper.

2.3 Some observations

For the moment, I shall not further discuss or assess these examples beyond the illustrations given. However, the examples suggest some simple generalizations, all of which can be extended, it seems, to the n-person case.

Observation 1: Each Nash equilibrium of a two-person game is a correlated equilibrium.
Proof: Just look at the definitions.

Observation 2: The set of correlated equilibria of a two-person game is convex.

Again, the proof is evident from the definition. Of course, we find both observations already in Aumann (1974, sect. 4). They entail that the convex closure of the Nash equilibria of a game is a subset of the set of correlated equilibria.

The next observations are closer to our concerns:

Observation 3: Each Nash equilibrium of a two-person game is a dependency equilibrium.
Proof: Again, just look at the definitions.

Observation 4: Generally, dependency equilibria are not included among the correlated equilibria, and vice versa.
Proof: Just look at the examples above.

In BoS we saw that there are also very bad dependency equilibria, and in PD we luckily found one dependency equilibrium weakly Pareto-dominating all the others. This suggests the following question: Which dependency equilibria are Pareto-optimal within the set of dependency equilibria? Clearly, these are the most interesting or attractive ones. Here is a partial answer:

Observation 5: Let $q = s \otimes t$ be a Nash equilibrium and suppose that the pure strategy combination (a_i,b_k) is at least as good as this equilibrium, i.e., that $u_{ik} \geq \sum_{j,l} s_j\, t_l\, u_{jl}$ and $v_{ik} \geq \sum_{j,l} s_j\, t_l\, v_{jl}$. Then this combination, or p with $p_{ik} = 1$, is a dependency equilibrium.

Proof: Define $p^r = \frac{r-1}{r} \cdot p + \frac{1}{r} \cdot q$, and assume that p^r is strictly positive. Obviously $\lim_{r\to\infty} p^r = p$. Moreover, $\lim_{r\to\infty} \sum_l p^r_{l|i} \cdot u_{il} = u_{ik}$, and for all $j \neq i$ and all r $\sum_l p^r_{l|j} \cdot u_{jl} = \sum_l t_l \cdot u_{jl}$.

But now we have $u_{ik} \geq \sum_{j,l} s_j t_l u_{jl} \geq \sum_l t_l u_{jl}$: the first inequality holds by assumption and the second because $\langle s, t \rangle$ is a Nash equilibrium. The same considerations apply to the other player. Hence, given our assumption, p with $p_{ik} = 1$ is a dependency equilibrium.

If p^r should not be strictly positive, modify q such that those a_j with $j \neq i$ and $s(a_j) = 0$ receive some positive probability by q and such that $q(b_l \mid a_j) = t_l$, and correspondingly for those b_l with $l \neq k$ and $t(b_l) = 0$. Then the modified p^r is strictly positive, and the same proof goes through.

In PD, Hawk and Dove, and BoS this observation fully satisfies the quest for the Pareto-optima among the dependency equilibria. But it does not generally do so. In Matching Pennies no pure strategy combination is Pareto-better than the Nash equilibrium; yet mixtures of them in which equivalent strategy combinations have equal weight are dependency equilibria.

This accentuates how preliminary my formal investigation of dependency equilibria is. However, it is not yet clear whether dependency equilibria are at all worth the efforts. If the answer we shall find is convincing, this may be sufficient motivation to deepen the formal investigation.

3. Causal Graphs, Bayesian Nets, Reductions, and Truncations

I have mentioned that dependency equilibria seem to be a causal mystery. For the sake of clarity, it is helpful to look at some basics of the probabilistic theory of causation which has become sort of a standard (if there is any in this contended area). This piece of causal theory will clearly confirm some fundamental assumptions of decision and game theory that are causally motivated, but probabilistically expressed. Thus, it will at first deepen the mystery about dependency equilibria. At the same time, however, we shall be able to see more clearly how to gain a different view.

3.1 Causal graphs and Bayesian nets

The standard theory I am alluding to is the theory of causal graphs and Bayesian nets.[12] It deals only with causal dependence and independence between variables. In order to

[12]This theory has been discussed more or less explicitly in the statistical path analysis literature since Wright (1934) and in the linear modeling literature since Haavelmo (1943) and the papers collected in Simon (1957). The structure and the crucial role of the general properties of conditional probabilistic independence seem to have been recognized not before Spohn (1976) and Dawid (1979). Pearl and his collaborators rediscovered these properties and added the graph theoretic methods as summarized in Pearl (1988). Since then an impressive theoretical edifice has emerged, best exemplified by Spirtes et al. (1993), Shafer (1996) and Pearl (2000).

do so, it must consider specific variables and not generic ones. Generic variables, say, of a sociological kind, would be annual income or social status. But it is usually very hard to say anything substantial about causal relations between generic variables. Specific variables of a sociological kind would be, e.g., my income in 2001 or my social status in 2002, insofar they are understood as ranges of possible values the variables may take, and not as facts consisting in the values the variables actually take. Hence, the realization of specific variables is always *located* at a specific time and usually also at a specific place or in a specific object or person.

The basic ingredient of the causal standard theory is thus a non-empty set U of variables which we assume to be finite; U is also called a *frame*. We may represent each variable by the (finite) set of the possible values it may take (this presupposes that the variables are mutually disjoint sets). For $V \subseteq U$, each member of the Cartesian product $\times V$ of all the variables or sets in V is a *possible course of events within V*, a possible way how all the variables in V may realize.

Due to their specificity, the variables in U have a temporal order $<$. $A < B$ says that *A precedes B.*[13] I assume $<$ to be a linear (and not a weak) order, thus avoiding questions about simultaneous causation. Moreover, due to their specificity the variables in U also display causal structure; their causal order is a partial order agreeing with the temporal order. That is, if $A \Rightarrow B$ expresses that *A influences B* or *B causally depends* on *A*, then $\Rightarrow$ is a transitive and asymmetric relation in U, and $A \Rightarrow B$ entails $A < B$.

Since U is finite, we can break up each causal dependence into a finite chain of direct causal dependencies. This simplifies our description. If $A \rightarrow B$ expresses that *A directly influences B*, or *B* directly *causally depends* on *A*, then $\rightarrow$ is an acyclic relation in U agreeing with the temporal order, and $\Rightarrow$ is the transitive closure of $\rightarrow$. Of course, directness and indirectness is relative here to the frame U; a direct causal dependence in U may well become indirect or, as we shall see, even spurious in refinements of U.

Graphs are relations visualized. Thus, we may say as well that $\langle U, \rightarrow \rangle$ is a directed acyclic graph agreeing with the temporal order[14] or, as we define it, a *causal graph.* Let me introduce some terminology we shall need:

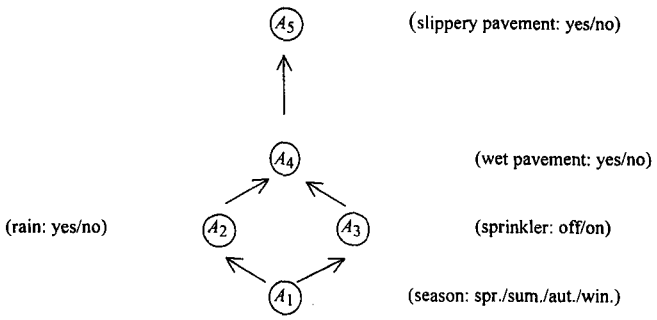
$Pa(B)$ = the set of parents of $B = \{A \mid A \rightarrow B\}$,
$Pr(B)$ = the set of variables preceding $B = \{A \mid A < B\}$, and
$Nd(B)$ = the set of non-descendants of $B = \{A \mid A \neq B$ and not $B \Rightarrow A\}$.

---

[13]This is not the A and B from sect. 2. From now on A, B, C, etc. are used to denote any single variables whatsoever. Of course, A and B from sect. 2 are also variables.

[14]The temporal order is often left implicit or neglected, presumably because the statistical literature is more interested in generic variables. However, as long as one is not engaged in the project of a causal theory of time, one must presuppose temporal order when talking about causation.

A small example may be instructive. It is Pearl's favorite. The indices indicate the temporal order:



This is a very simple causal graph showing how the season influences the wetness of the pavement via two different channels, and the wetness in turn directly influences the slipperiness.

So far, we have just structure. However, the causal structure must somehow relate to how the variables realize, and since we shall consider realization probabilities here, this means that the causal structure must somehow relate to these probabilities. I should emphasize that these probabilities may be objective ones (whatever this means precisely), in which case they relate to the objective causal situation, or they may be some person's subjective probabilities, in which case they reflect the causal beliefs of that person.[15] The latter perspective will be the relevant one for us.

But what exactly is the relation between causation and probability? Spirtes et al. (1993) state two crucial conditions, the causal Markov condition and the minimality condition. In order to explain them, we need the all-important notion of conditional independence:

Let p be a probability measure for U (i.e., $p(v) \geq 0$ for each $v \in \times U$ and $\sum_{v \in \times U} p(v) = 1$). Then, for any mutually disjoint sets of variables $X, Y, Z \subseteq U$ X is said to be *conditionally independent* of Y *given* Z w.r.t. p – in symbols: $X \perp Y / Z$ – iff for all $x \in \times X$, $y \in \times Y$ and $z \in \times Z$ $p(x \mid y,z) = p(x \mid z)$, i.e., if, given any complete information about Z, no information about Y teaches us anything about X.

Conditional probabilistic dependence is closely tied up with causal dependence according to a causal graph $\langle U, \rightarrow \rangle$. The *causal Markov condition* says that, for all

[15]This assertion sounds nice, and I do not think it is really wrong, but it deserves a most careful explanation. In fact, it is the most profound philosophical problem with causation what to say here precisely.

$A \in U$, given the parents of A, A is irrelevant to all other variables preceding it, or indeed to all other non-descendants – formally: that for all $A \in U$

$$A \perp Pr(A) \setminus Pa(A) \,/\, Pa(A) \text{ (where } \setminus \text{ denotes set theoretic subtraction),}$$

or equivalently (though the proof is not entirely trivial – cf. Verma, Pearl 1990 and theorem 9 in Pearl 1988, p. 119):

$$A \perp Nd(A) \setminus Pa(A) \,/\, Pa(A).$$

And the *minimality condition* says that, for all $A \in U$, the set $Pa(A)$ of parents of A is indeed the smallest set of variables preceding A or of non-descendants of A, respectively, for which these conditional independencies hold w.r.t. p.

We say that p *agrees with* the causal graph $\langle U, \rightarrow \rangle$ or that $\langle U, \rightarrow, p \rangle$ is *a Bayesian net* iff p satisfies the causal Markov and the minimality condition w.r.t. $\langle U, \rightarrow \rangle$.[16] In fact, in such a Bayesian net $\langle U, \rightarrow, p \rangle$ we can infer from p alone the set of parents of each variable and thus the whole causal graph agreeing with p.[17]

Let me illustrate these definitions with the above example: p satisfies the causal Markov condition w.r.t. the graph concerning (obviously Californian) pavements iff

$$A_3 \perp A_2 \,/\, A_1, \quad A_4 \perp A_1 \,/\, \{A_2, A_3\}, \quad A_5 \perp \{A_1, A_2, A_3\} \,/\, A_4,$$

or, equivalently, iff for all $a_i \in A_i$ $(i = 1,\ldots,5)$

$$p(a_1,a_2,a_3,a_4,a_5) = p(a_1) \cdot p(a_2 \mid a_1) \cdot p(a_3 \mid a_1) \cdot p(a_4 \mid a_2,a_3) \cdot p(a_5 \mid a_4).$$

The latter equation, by the way, makes clear how information about causal structure allows for a vast reduction of probabilistic information, an observation computer scientists are eagerly exploiting for implementing probability measures.[18] p satisfies the minimality condition iff moreover *none* of the following holds:

$$A_2 \perp A_1, \quad A_3 \perp A_1 \,/\, A_2, \quad A_4 \perp A_2 \,/\, A_3, \quad A_4 \perp A_3 \,/\, A_2, \quad A_5 \perp A_4.$$

The conditional independencies and dependencies characteristic of the causal Markov and the minimality condition are the basic ones entailed by the causal structure. But there is a very useful and graphic way to discover all conditional

---

[16] This definition is due to Pearl (1988, p.119).

[17] This was precisely my explication of direct causal dependence in probabilistic terms in Spohn (1976/78, sect. 3.3) and (1980).

[18] For a good introduction into the computational aspects of Bayesian nets, see Jensen (1996).

dependencies and independencies implied by the basic ones. This is delivered by the so-called criterion of d-separation.[19] Let us say that a path in the graph $\langle U, \rightarrow\rangle$[20] is *blocked* or *d-separated* by a set $Z \subseteq U$ of nodes (or variables) iff

(a) the path contains some chain $A \rightarrow B \rightarrow C$ or fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z, or
(b) the path contains some collider $A \rightarrow B \leftarrow C$ such that neither B nor any descendant of B is in Z.

We continue to define for any mutually disjoint $X, Y, Z \subseteq U$ that Z *d-separates* X and Y iff Z blocks every path from a node in X to a node in Y.

The notion may look complicated at first, but one becomes quickly acquainted with it. In our sample graph, for instance, A_2 and A_3 are d-separated only by $\{A_1\}$, but neither by $\varnothing$ nor by any set containing A_4 or A_5.

The importance of d-separation is revealed by the following *theorem*: For all $X, Y, Z \subseteq U$, if X and Y are d-separated by Z, then $X \perp Y / Z$ according to all measures p agreeing with $\langle U, \rightarrow\rangle$; and conversely, if X and Y are not d-separated by Z, then not $X \perp Y / Z$ according to almost all p agreeing with $\langle U, \rightarrow\rangle$.[21] This shows that d-separation is indeed a reliable guide for discovering conditional independencies entailed by the causal structure, and in fact all of them for almost all measures. We shall make use of this fact later on.

Spirtes et al. (1993) define a causal graph $\langle U, \rightarrow\rangle$ and a probability measure p for U to be *faithful* to one another iff indeed for all mutually disjoint $X, Y, Z \subseteq U$ $X \perp Y / Z$ w.r.t. p if and only if X and Y are d-separated by Z.[22] Thus, the second part of the theorem just stated says that almost all p agreeing with $\langle U, \rightarrow\rangle$ are faithful to $\langle U, \rightarrow\rangle$. But sometimes it is useful to exclude the exceptional cases by outright assuming faithfulness.

3.2 Reductions of causal graphs and Bayesian nets

An important issue in the theory of causation is how causal graphs and Bayesian nets change with changing frames. If the frame is extended there is no determinate answer

---

[19]Invented by Thomas Verma; see Verma, Pearl (1990), and also Pearl (1988, p. 117).

[20]A path is just any connection between two nodes disregarding the directions of the arrows, i.e., any sequence $\langle A1,...,An\rangle$ of nodes such that for each i = 1,...,n-1 either Ai $\rightarrow$ Ai+1 or Ai $\leftarrow$ Ai+1.

[21]Cf. Pearl (2000, p. 18). The proof is involved; see Spirtes et al. (1993, theorems 3.2 and 3.3). "Almost all" is here understood relative to the uniform distribution over the compact space of all probability measures for U.

[22]This is not quite faithful to Spirtes et al. (1993). Their definition of faithfulness on p.56 is a different one, and in their theorem 3.3 they prove it to be equivalent with the definition given here.

because probabilities can be arbitrarily extended to the richer frame. But if we start with a Bayesian net on a large frame and reduce it, then the Bayesian net on the reduced frame must have a definite shape. The question is which one.

Let us simplify matters by focussing on reductions by a single variable only. Larger reductions can then be generated by iterating such minimal reductions. So, how does a causal graph change when a node, C, is deleted from the frame U? The answer, which is not entirely obvious, is prepared by the following definition:

The causal graph $\langle U^*, \rightarrow^* \rangle$ is called the *reduction* of the causal graph $\langle U, \rightarrow \rangle$ by the node C iff:

1. $U^* = U \setminus \{C\}$,
2. for all $A, B \in U^*$ $A \rightarrow^* B$ iff either $A \rightarrow B$, or not $A \rightarrow B$ and one of the following three conditions holds:
   (i) $A \rightarrow C \rightarrow B$ (let us call this the *IC-case*), or
   (ii) $A < B$ and $A \leftarrow C \rightarrow B$ (let us call this the *CC-case*), or
   (iii) $A < B$ and there is a variable $D < B$ such that $A \rightarrow D \leftarrow C \rightarrow B$ (let us call this the *N-case*).

Thus, the reduced graph contains all the arrows of the unreduced graph not involving the deleted variable C. And it contains an arrow $A \rightarrow^* B$ where the unreduced graph contains none exactly when B is rendered *i*ndirectly *c*ausally dependent on A by the deleted C (the IC-case), or when the deleted C is a *c*ommon *c*ause of A and B (the CC-case), or when A is the *n*eighbor of such a CC-case involving B (the N-case).

The N-case may look a bit complicated. Let us make it more graphic (as another mnemonic aid for its label):

The N-case is always accompanied by a CC-case. Note the importance of the temporal relation $D < B$. If $B < D$, we only have a CC-case involving B and D, where

The justification for this definition is provided by the following *theorem*: Let $\langle U, \rightarrow, p\rangle$ be a Bayesian net, let $\langle U^*, \rightarrow^*\rangle$ be the reduction of $\langle U, \rightarrow\rangle$ by C, and let p^* be the marginalization or restriction of p to $U^* = U \setminus \{C\}$. Then the causal graph agreeing with p^* is a (proper or improper) subgraph of $\langle U^*, \rightarrow^*\rangle$, and if p is faithful to $\langle U, \rightarrow\rangle$, then it is $\langle U^*, \rightarrow^*\rangle$ itself which agrees with p^*.

Proof: Suppose that B directly causally depends on A w.r.t. p^*, i.e., that not $A \perp B / X$, where $X = Pr(B) \setminus \{A, C\}$. We need not specify here relative to which probability measure the conditional (in)dependence statements are to be taken, because p and p^* completely agree on them within U^*. According to the above theorem about d-separation not $A \perp B / X$ entails that A and B are not d-separated by X in $\langle U, \rightarrow\rangle$. This may be so because $A \rightarrow B$ or because the IC-, the CC-, or the N-case obtains, but in no other case; in all other cases each path between A and B must be blocked by X, as is easily checked. Hence, not $A \perp B / X$ entails $A \rightarrow^* B$. This proves the first part of the theorem.

Now, suppose that B does not directly causally depend on A w.r.t. p^*, i.e., that $A \perp B / X$, and suppose that p is faithful to $\langle U, \rightarrow\rangle$. Hence, A and B must be d-separated by X in $\langle U, \rightarrow\rangle$. Therefore, as we have just seen, neither $A \rightarrow B$ nor one of the IC-, CC-, or N-case can obtain. That is, $A \rightarrow B$ does not hold. This proves the second assertion of the theorem.

In the case that p is not faithful to $\langle U, \rightarrow\rangle$, the theorem cannot be strengthened, because in that case there may hold a lot of conditional independencies not foreseen by the criterion of d-separation. Hence, d-separation may tell us $A \rightarrow^* B$, even though $A \perp B / X$, which excludes a direct causal dependence of B on A w.r.t. p^*.

However, if p is faithful to $\langle U, \rightarrow\rangle$, this situation cannot arise, and we have a complete answer about the behavior of reductions.[23] Indeed, it is more illuminating to reverse the perspective again and to read the theorem not as one about reductions, but as one about extensions. Our picture of the world is always limited, we always move within a small frame U^*. So, whenever we construct a causal graph $\langle U^*, \rightarrow^*\rangle$ agreeing to our probabilities p^*, we should consider this graph as the reduction of a yet unknown, more embracive graph $\langle U, \rightarrow\rangle$. And the theorem then tells us (i) that *where there is no direct causal dependence according to the small graph, there is none in the extended graph*, and (ii) that *what appears to be a direct causal dependence in the small graph may be either confirmed as such in the extended graph, or it may turn out to be spurious and to resolve into one of the IC-, CC-, or N-case*. This observation will acquire crucial importance in sections 4 and 5. To be precise, the observation is

[23]One should note, though, that even if p is faithful to $\langle U, \rightarrow\rangle$, p* need not be faithful to $\langle U^*, \rightarrow^*\rangle$. Indeed, p* cannot be faithful if the N-case applies, since in that case we have $A \perp B$, though A and B are not d-separated by $\varnothing$ in $\langle U^*, \rightarrow^*\rangle$.

guaranteed only if the extended probabilities p are faithful to the extended graph $\langle U, \rightarrow\rangle$. But since almost all probability measures agreeing with $\langle U, \rightarrow\rangle$ are faithful to it, we may reasonably hope to end up with such a p.

So much for the standard theory of probabilistic causation. Calling it standard is perhaps justified in view of the impressive list of its predecessors and defenders. It is also more or less explicit in a great deal of applied work and in particular in large parts of decision and game theory. But it is still contested, most critically perhaps by Cartwright (1989, 1999), who splits up the causal Markov condition into two parts, a proper Markov condition relating only to the past of the parents of the relevant node, and a screening-off condition (as in Reichenbach's principle of the common cause) relating to the other non-descendants. Cartwright accepts the proper Markov condition, but vigorously rejects the screening-off condition. This is tantamount to the assertion of interactive forks, as introduced and defended by Salmon (1980, 1984).

But even if the theory is not contested, the underlying conceptions may be quite different. In Spohn (2001) I have elaborated, for instance, on the differences between my picture and that of Spirtes et al. (1993). It is important to know of these divergences; in philosophy no opinion is really standard in the end. In the following, though, I shall neglect these debates and proceed with the theory to which I refer as the standard one.

3.3 Actions, truncations, and basic decision models

So far, actions and agents have not entered the picture. A Bayesian net describes either some small part of the world or some person's partial view of the world. But this person might be a detached observer having only beliefs and no interests whatsoever about that part. This, however, is not the agent's view as it is modeled in decision theory. In order to accommodate it, we have to enrich our picture by adding two ingredients.

The first ingredient consists in desires or interests that are represented by a utility function. Each course of events is more or less valued, and accordingly a *utility function* u is a function from $\times U$ into $\boldsymbol{R}$.

So far, we still might have a mere observer, though an interested one. But an agent wants to take influence, to shape the world according to his interests. Hence, we must assume that some variables are action variables that are under direct control of the agent and take the value set by him. Thus, the second ingredient is a partitioning of the frame U into a set H of *action variables* and a set W of *occurrence variables*, as I call them for want of a better name.

Are we done now? No. The next important step is to see that not any structure $\langle U, \rightarrow, H, p, u\rangle$ (where $W = U \setminus H$) will do as a decision model; we must impose some restrictions.

A minor point to be observed here is that H does not contain all the variables in U which represent actions of the agent. Rather, H contains only the action variables still open from the agent's point of view. That is, the decision model is to capture the agent's decision situation at a given time t. Thus, H contains only the action variables later than t, whereas the earlier variables representing acts of the agent are already past, no longer the object of choice, and thus part of W.

Given this understanding of H, the basic restriction is that the decision model must not impute to the agent any cognitive or doxastic assessment of his own actions, i.e., of the variables in H. The agent does not have beliefs or probabilities about H. In the first place, he has an intention about H, formed rationally according to his beliefs and desires or probabilities and utilities, and then he may as well have a derivative belief about H, namely, that he will conform to his intention about H. But this derivative belief does not play any role whatsoever in forming the intention. I have stated this *"no probabilities for acts" principle* in Spohn (1977, sect. 2) since it seemed to me to be more or less explicit in all of the decision theoretic literature (cf., e.g., Fishburn 1964, pp. 36ff.) except Jeffrey's evidential decision theory (1965); the principle was also meant as a criticism of Jeffrey's theory. The arguments I have adduced in its favor have been critically examined by Rabinowicz (2002). My present attitude toward the principle will become clear in the next section.

It finds preliminary support, though, in the fact that it entails another widely observed principle, namely, that the action variables in H are exogenous in the graph $\langle U, \rightarrow\rangle$, i.e., uncaused or parentless. Why does this *"acts are exogenous" principle*, as I call it here, follow? If the decision model is not to contain probabilities for actions, it must not assume a probability measure p for the whole of U. Only probabilities for the occurrence variables in W can be retained, but they may, and should, be conditional on the various possible courses of action $h \in \times H$; the actions may, of course, matter to what occurs in W. Hence, we must replace the measure p for U by a family $(p_h)_{h \in \times H}$ of probability measures for W. Relative to such a family, Bayesian net theory still makes perfect sense; such a family may also satisfy the causal Markov and the minimality condition and may agree with, and be faithful to, a given causal graph.[24] However, it can do so only when action variables are parentless. For a variable to have parents in agreement with the probabilities, conditional probabilities for it must be explained, but

[24]My definitions and theorems concerning conditional independence in Spohn (1978, sect. 3.2) dealt with the general case relating to such a family of probability measures. The graph theoretic material may be supplemented in a straightforward way.

this is just what the above family of measures must not do concerning action variables. Therefore, these variables cannot have parents.

Pearl (2000, ch. 3) thinks along very similar lines when he describes what he calls the *truncation* of a Bayesian net: He starts from a Bayesian net $\langle U, \rightarrow, p\rangle$. U contains a subset H of action variables. p is a measure for the whole of U and thus represents rather an external observer's point of view. Therefore, the action variables in H have so far no special role and may have any place in the causal graph $\langle U, \rightarrow\rangle$. Now Pearl imagines that the observer turns into an agent by becoming empowered to set the values of the variables in H according to his will so that the variables in H do not evolve naturally, as it were, but are determined through the intervention of the agent. Then Pearl asks which probabilities should guide this intervention. Not the whole of p. Rather, the intervention cuts off all the causal dependencies the variables in H have according to $\langle U, \rightarrow\rangle$ and puts itself into place. Hence, the agent should rather consider the *truncated* causal graph $\langle U, \rightarrow^*\rangle$ which is defined by deleting all arrows leading to action variables, i.e., $A \rightarrow^* B$ iff $A \rightarrow B$ and $B \notin H$. Thereby the action variables turn exogenous, in accordance with our principle above.

The next task is to find the probabilities that agree with the truncated graph. We must not simply put $p_h(w) = p(w \mid h)$ ($h \in \times H$, $w \in \times W$); this would reestablish the deleted dependencies. Rather, we have to observe the factorization of the whole of p provided by the causal graph $\langle U, \rightarrow\rangle$ (which I have already exemplified above with the Californian pavements):

If $v \in \times U$ is a course of events in U[25] and if for each $A \in U$ a is the value A takes according to v and $pa(a)$ the values the variables in $Pa(A)$ take according to v, then $p(v) = \prod_{A \in U} p(a \mid pa(a))$.

Then we have to use the truncated factorization[26] that deletes all factors concerning the variables in H from the full factorization:

If $h \in \times H$ and $w \in \times W$ and if for each $A \in W$ a is the value A takes according to w and $pa(a)$ the values the variables in $Pa(A)$ take according to h and w, then $p_h(w) = \prod_{A \in W} p(a \mid pa(a))$.

For the family (p_h) thus defined, we say that $\langle U, \rightarrow^*, (p_h)\rangle$ is the *truncation* of $\langle U, \rightarrow, p\rangle$ with respect to H, and we can easily prove that (p_h) agrees with $\langle U, \rightarrow^*\rangle$ if p agrees with $\langle U, \rightarrow\rangle$; this is built in into the truncated factorization. Thus, as Pearl and I agree, it is this family (p_h) that yields the probabilities to be used by the agent. Hence,

---

[25] I use „v" since „u" is already reserved for the utility function.

[26] Cf. Pearl (2000, p. 72).

Pearl also subscribes to the two principles above.[27] The notion of truncation will receive a crucial role in sections 4 and 5.

We may resume this discussion by defining a *basic decision model*. This is a structure $\langle U, \rightarrow, H, (p_h), u\rangle$, where $\langle U, \rightarrow\rangle$ is a causal graph, H is a set of exogenous variables, (p_h) is a family of probability measures for W agreeing with $\langle U, \rightarrow\rangle$, and u a utility function from $\times U$ into $\boldsymbol{R}$.

What is the associated decision rule? Maximize conditional expected utility, i.e., choose a course of action $h \in \times H$ for which $\sum_{w \in \times W} u(h,w) \cdot p_h(w)$ is maximized. However, this decision rule is naïve insofar as it neglects the fact that the agent need not decide for a whole course of action; rather, he needs to choose only from the (temporarily) first action variable and may wait to decide about the later ones. Thus the naïve decision rule has not taken into account strategic thinking. We shall have several reasons for undoing this neglect below.

So far, I have not really argued for the two principles and thus for the given basic form of decision models. I have only claimed that it is more or less what we find in most of the decision theoretic literature. I find it very natural to read Savage (1954) and Fishburn (1964) in this way, and I have referred to the more recent literature about causal graphs such as Spirtes et al. (1993) and Pearl (2000). This is not an argument, but it carries authority. We shall continue the topic in the next section.

Let me point out an important consequence, though. In a basic decision model all non-descendants of an action variable are probabilistically independent of it. This is entailed by the exogeneity of action variables, as is easily verified with the help of d-separation. In other words: what is causally independent from actions is also probabilistically independent from them.

This observation provides an immediate solution of Newcomb's problem.[28] According to Nozick (1969), the initial paper on the problem, Newcomb's problem is constituted by the fact that there may be events (such as the prediction of the mysterious predictor) which are causally independent from my actions, but nevertheless probabilistically relevant. According to the observation just made, this alleged fact is spurious; there are no such events, and hence there is no Newcomb's problem, as I have explained in Spohn (1978, sect. 5.1). Of course, there is more to say about Newcomb's problem, and I shall say more below. But I believe that thereby the

---

[27]In this paragraph I have slightly assimilated Pearl's conception to mine, though in a responsible way, I believe. In principle, the truncation procedure is already described in Spohn (1978, pp. 187ff.), though without graph-theoretic means. It should also be noted that Spirtes et al. (1993, pp. 75ff.) make essential use of the transition from unmanipulated to manipulated graphs, as they call it. This transition closely corresponds to Pearl's truncation.

[28]For a presentation of Newcomb's problem, see sect. 5.

stubborn intuition of two-boxers, which I have espoused for more than 20 years, is well explained: if Newcomb's problem is modeled by a basic decision model, two-boxing is the only rational action.

The observation also explains a constitutive feature of non-cooperative game theory, namely, that the actions of the players are causally independent; they do not communicate or interact in any way. And the players have to be aware of this causal independence. Hence, if this observation is correct, the players' actions are probabilistically independent as well (also from their own point of view). This is what has been assumed all along in non-cooperative game theory, and this is why we seem to be forced to adopt something like Nash equilibria, which are the only equilibria conforming to this probabilistic independence.

All this shows that basic decision models as defined above are deeply entrenched in decision and game theoretic thinking. The last point, in particular, underscores the suspicion, raised in section 2, that dependency equilibria do not make causal sense. Thus, our search for causal sense can only take one direction: we have to scrutinize the assumptions underlying basic decision models. This is our next task.

4. Reflexive Decision Theory

How can one doubt the "no probabilities for acts" and the "acts are exogenous" principle? I see essentially two ways. On the one hand, the agent himself may *make* his actions dependent on the behavior of other variables and thus turn the action variables into endogenous ones; this is what is called strategic behavior. By deciding for a certain strategy the agent obviously accepts certain probabilities for the actions covered by the strategy, in contradiction to the two principles. On the other hand, it is hard to see why the agent should not be able to reflect on the causes of his own actions, just as he does concerning the actions of others. This reflexion should clearly enable him to have (probabilistic) predictions about his future actions, again in contradiction to the principles. We shall see that both approaches come to the same thing; but let us dwell upon them separately and more carefully.

4.1 Dependency schemes and strategies

Let us take up strategies first. Concerning basic decision models, I have already mentioned that it would be a naïve decision rule simply to choose a course of action with maximal expected utility. Usually it is better to wait and see what happens and act accordingly. How can this be accounted for in our graph theoretic framework?

The most general way is this: According to a given basic decision model $\langle U, \rightarrow, H, (p_h), u\rangle$ all action variables in H are exogenous. What the agent does in thinking

about strategies is to enrich the causal graph $\langle U, \rightarrow \rangle$ by some edges each of which ends at some action variable and starts at some preceding occurrence variable; this means to reverse the truncation described in the previous section. Of course, the agent does not only create such dependencies, he considers to create them in a specific way expressed by specific probabilities. This is captured in the following definition: A *dependency scheme* q for a given basic decision model is a function which specifies for each action variable $A \in H$ a probability distribution for A conditional on each realization of $Pr(A)$, i.e., of all the variables preceding A.

On the basis of the probability family (p_h) each dependency scheme q determines a probability measure p_q for the whole of U defined as follows: for $w \in \times W$ and $h \in \times H$ $p_q(h,w) = p_h(w) \cdot q(h \mid w)$ – where $q(h \mid w)$ denotes the probability that the action sequence h realizes according to q given w. That is: if, for $A \in H$, a denotes the value A takes according to h and $pr(a)$ the values the variables in $Pr(A)$ take according to h and w, then $q(h \mid w) = \prod_{A \in H} q(a|pr(a))$. These are just the factors we need in order to fill up a truncated factorization to yield a complete one.

This, in turn, enables us to define the *expected utility* of each dependency scheme q: $Eu(q) = \sum_{h \in \times H} \sum_{w \in \times W} u(h,w) \cdot p_q(h,w)$. This suggests a more general and reasonable decision rule: If your situation is represented by the given model, choose a dependency scheme with maximal expected utility! Is this rule a good one?

No, the problem is that not every dependency scheme represents a feasible strategy. I have lost my glasses, for instance. What to do? Clearly, the optimal dependency scheme would be to search in my office if I have forgotten them in my office, to look into the fridge if I have put them into the fridge, etc. This would clearly be the fastest way to find my glasses. But it is obviously not feasible; my problem is just that I do not know where I have put them. Hence, dependency schemes maximizing expected utility tell only how the agent and his actions would be optimally embedded into the causal graph according to his subjective view. Whether he is able to embed himself in such a way is another question.

This raises the following question: Which of the dependency schemes are feasible strategies that the agent is able to realize by himself? Generally, one can only say that the latter form a convex subset of the former. The reason is the frame-relativity of dependency schemes. One should think at first that there is no need to consider probabilistic dependency schemes because it is always better to establish good deterministic dependencies. However, there is no guarantee that the frame contains the variables which the agent is able to connect up with in a deterministic way. Perhaps the agent at best receives incomplete information about the variables included in the frame. In this case only a probabilistic dependency is within his power. Hence, as long

as we do not make special assumptions about which variables are in the frame U, no more can be said about the feasibility of dependency schemes.

So, we should perhaps include in the frame those variables to which the agent can establish a deterministic dependence. Which are they? The answer seems clear. The agent can make his action depend only on those variables whose state he learns before the time of action. Maybe his behavior is correlated with the state of certain variables, though he does not notice it. But if so, the behavior is not intentional. Thus, for the dependence to be intentional, the agent has to know the states he wants to correlate with.

Again, no general statement seems available concerning the kind of variables the agent learns about. They must be observable, for sure; but the decline of empiricism has shown that this characterization is vague and loose. Still, there *is* a general statement: Whatever the external events the agent does, or does not, notice, he knows his own state before the time of action, he knows the decision situation he is in (i.e., his subjective view of it), which is generated, among other things, by the external events he has noticed.

This seems generally true. Hence, a general procedure for discovering the feasible strategies among the dependency schemes would be to extend the causal graph of the given basic decision model by a number of *decision nodes*, as I call them, such that each action node is preceded by a decision node, and then to define a *strategy* as a dependency scheme which makes each action node depend only on its associated decision node. Here, a decision node is quite a complex variable consisting of all the decision situations the agent might be in concerning the associated action node. Obviously a decision node causally depends, in turn, on many other variables; thereby, the action node's direct intentional dependence on the decision node ramifies into various indirect dependencies (where "direct" and "indirect" is relative to the extended causal graph). Moreover, it is obvious which deterministic shape the action node's intentional dependence on the decision node should take: the relevant decision rule, say, maximizing conditional expected utility, states which action to perform in which decision situation.

It should be clear that we have to elaborate the content of the previous paragraphs in detail. This is what we shall do in section 4.3. But one point should be stated right away. What I have explained so far entails that as soon as the agent has decided for a certain strategy or dependency scheme, he can, on the basis of this decision, predict with which probability he will perform which action supported by his strategy. This is the first way for apparently rebutting the "no probabilities for acts" principle.

4.2 Reflexion

I have announced a second way, at which we should look next before further developing the above ideas. This way is even more straightforward: Why should the agent be unable to take doxastic attitudes like predicting, explaining, etc. toward his own actions, if he can very well do so toward the actions of others? One should indeed think that he is particularly well endowed in his own case because he has so much more data about himself than about anybody else.

Hence, the question is rather: How should the agent predict his own future behavior? There seem countless ways. The agent knows his habits ("sure, I'll brush my teeth this evening when I go to bed; that's what I always do!") or the conventions ("of course, I'll drive on the right tomorrow; everybody does!"), he knows his anxieties and the resulting behavior ("I won't hike through Devil's gorge!"), and so on. All these pieces of behavior may, or may not, be under the rational control of the agent. If they are, as is likely in the case of habits and conventions (at least in the examples given), the prediction is incomplete unless it mentions that the particular instantiation of a habit or convention is confirmed by rational control. This means, in turn, that the prediction of a piece of behavior is really based on the prediction of the (tacit or explicit) rational deliberation leading to it. If a piece of behavior is not under rational control, as it may be in the case of anxiety, then, it seems to me, it cannot be the object of a practical deliberation and does not deserve the status of an action node in a decision model; from the point of view of a practical deliberation, it is just an occurrence to reckon with, not an action to be intentionally chosen.[29]

To conclude, the agent should predict and explain his actions at future action nodes as intentional and rational actions with the help of decision theory, just as he explains and predicts the actions of others. Hence, if we want to make explicit these means for predicting and explaining actions within the decision model, we should extend it by decision nodes, as we have envisaged them in our discussion of strategies. The agent has (probabilistic) predictions about the decision situations he will face, and accordingly he has (probabilistic) predictions about the future actions, again just as in the case of strategies.

It may seem surprising how the active mode of considering which feasible strategy to choose and the passive mode of predicting future actions can come to the same thing. But it is not so surprising, after all; the two modes melt into each other in this special case. If I predict my likely future actions from my likely future decision situations, this is like forming a conditional intention. And conversely, if I choose among feasible strategies that make future actions dependent on future decision

[29]Psychology and self-observation teaches that this distinction is not clear-cut at all. However, for the sake of theoprizing we sometimes have to paint black and white.

situations, the chosen dependence is not really subject to my present evaluation and intention. Rather, all the parameters on which the evaluation and intention is based, i.e., the relevant subjective probabilities and utilities, are already specified in the future decision situation on which the action depends; the decision is deferred to that situation. One description is as good as the other; and so the active mode of decision and the passive mode of prediction merge.

Thus, it seems that we have a convincing double safe argument against our principles. Did we succeed to refute them? It is not clear whether this conclusion would help with the task set at the end of section 3. And it would be premature, in any case. Before jumping to conclusions, we should rather scrutinize how decision models that include decision nodes really look like.

4.3 Reflexive decision models and their truncated reductions

In section 4.1 I have already sketched such *reflexive decision models*, as I would like to call them, since they model how the agent reflects on his own attitudes. These models need to be worked out. The resulting structure, however, will be *very* complex; we cannot and need not fathom it here in full depth. I shall render precisely only those aspects required for my argument; the others will be left sketchy.

Here is my proposal in the form of an extensively annotated partial definition: $\bar{\delta} = \langle U, \rightarrow, H, D, p, u \rangle$ is a *reflexive decision model* iff the following conditions (1) – (8) are satisfied:

(1) H, the set of *action variables*, and D, the set of *decision variables*, are disjoint subsets of U; as before, $W = U \setminus (H \cup D)$ is the set of *occurrence variables*.

This simply introduces the decision variables as new ingredients.

(2) $\langle U, \rightarrow \rangle$ is a causal graph such that each action node has exactly one decision node as the only parent, i.e., for each $A \in H$ there is a $\Delta \in D$ with $Pa(A) = \{\Delta\}$, and each decision node has at least one action node as a child, i.e., for each $\Delta \in D$ there is an $A \in H$ with $\Delta \in Pa(A)$.

This was the upshot of our preceding discussion. For each action node it is just the parental decision node that provides the intentional or explanatory or predictive determinants of which element of the action node is performed. It is thus obvious that only decision nodes can be parents of action nodes, and indeed that each action node can have only one parental decision node. Thus, (2) is the minimum required.

The question is rather whether (2) should be strengthened. One might require that no two action nodes have the same parental decision node, or that each action node be

immediately preceded in time by its parental decision node.[30] One might also wonder how a decision node can have other children than action nodes. It will be crucial for my argument in the next section to reject all such strengthenings of (2). Therefore, I shall defer further discussion of this point.

(3) u is a utility function from $\times(U \setminus D)$ into $\boldsymbol{R}$.

The point of this condition is to exclude the decision nodes from the utility function; in my view, being in, or getting into, this or that decision situation does not hold any utility in itself. I have argued for the point in Spohn (1999, pp. 49ff.). But since it does not play any role here, I shall not dwell upon it.

(4) p is a probability measure for U.

This reflects the point that there seems to be no restriction on the domain of the agent's probability function under the present perspective. Later conditions, though, will restrict the values p may take.

The next condition is concerns the self-localization of the agent in the reflexive decision model $\bar{\delta}$. Such a model is to represent the agent's own practical point of view resulting in a decision, not that of an external observer. The point is reflected in the fact that H represents his *own* possible future actions and p and u his *own* cognitive and conative attitudes. But at which time? The answer is immediate: the agent is to decide about the first of his action nodes (and possibly later ones as well) and hence finds himself, as it were, *in* the first decision node. That is, the time when the agent takes the attitudes p and u is the time of the first decision node.

At that very time the agent knows in which decision situation he presently finds himself. He may not have foreseen it, and he may have forgotten it later on; but at the time of decision he knows his subjective view of his situation; and the model represents only this view. This knowledge is captured in the next condition:

(5) If $\Delta_0 \in D$ is the temporally first decision node, there is a particular $\delta_0 \in \Delta_0$ such that $p(\delta_0) = 1$.[31]

This is embarrassing, though. δ_0 is obviously to represent the present decision situation of the agent of which he is aware; on the other hand, the reflexive model $\bar{\delta}$,

[30]Since we have assumed the variables to be linearly ordered in time, the second strengthening implies the first.

[31]This condition of consciousness or self-knowledge has first been stated by Eells (1982, p. 176).

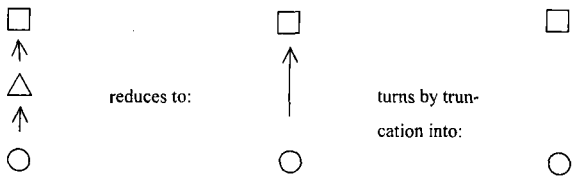
which we are about to define, does so as well. But δ_0 is only a part and not the whole of $\hat{\delta}$. How can this be?

The first response is that two different decision models, in the present case δ_0 and $\hat{\delta}$, may well represent the same situation; the representation relation is rarely one-one in model construction. Indeed, if one decision model is a reduction of another, they may be said to represent the same situation.[32] The second response is that we face a general difficulty here. Whenever one models states of reflexion, the object of reflexion cannot be understood as the whole reflexive state itself.[33] The embarrassment is thus a common one.

Here is an account of what δ_0 is, if not the whole of $\hat{\delta}$. It is not the basic submodel resulting from the full reflexive model by eliminating all decision nodes; it is only the first decision node Δ_0 itself that needs to be eliminated. This elimination results, more precisely, in the *truncated reduction* of $\hat{\delta}$ by $\{\Delta_0\}$ defined as follows:

For any decision node $\Delta \in D$, let $Ac(\Delta)$ denote the set of *action children* of Δ (which must not be empty according to condition (2)) and $Oc(\Delta)$ denote the set of *other* (occurrence or decision) *children* of Δ (which may, but need not be empty). Then, the *truncated reduction* of $\hat{\delta}$ by $\{\Delta_0\}$ is obtained by first reducing $\hat{\delta}$ by $\{\Delta_0\}$ not precisely in the way described in section 3.2, but in a slightly modified way and then truncating this reduction with respect to $Ac(\Delta_0)$. What is the slightly modified way?

Arrows in which Δ_0 is not involved are simply maintained in the reduction as defined in section 3.2. Likewise, the reduction contains arrows from the parents of Δ_0 to the children of Δ_0; this is the IC-case. The arrows arriving at action children will then fall victim to the truncation. (In the following diagrams triangles stand for decision nodes, squares for action nodes, and circles for occurrence nodes.)

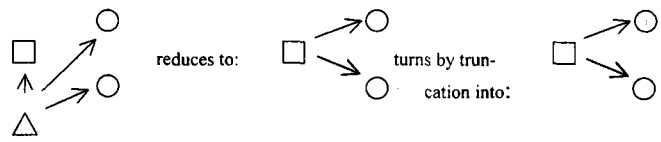


(the IC-case)

[32] I have not explicitly defined the reduction of basic decision models; but our definition of the reduction of Bayesian nets is easily extended. Such reductions are at the heart of the theory of small worlds of Savage (1954, sect. 5.5). In Spohn (1978, sects. 2.3 and 3.6) I have elaborated on their theoretical importance.

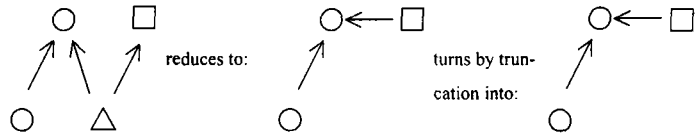
[33] This is so at least if we stick to standard ways and do not resort to the model theoretic means devised by Barwise (1990), which attempt to accommodate such circular phenomena in a straightforward way.

The slight modification occurs in the CC-case. Here, we have to stipulate, for reasons to be immediately explained, that all arrows between $Ac(\Delta_0)$ and $Oc(\Delta_0)$ created by the reduction run from $Ac(\Delta_0)$ to $Oc(\Delta_0)$ *irrespective* of the temporal order, i.e., even in the case where the arrows thus run backwards in time.



(*two CC-cases; the fat arrows show the modification*)

This modification entails that the N-case cannot obtain. The modification of the CC-case treats all occurrence children of Δ_0 as if they were later than the action children of Δ_0, and thus the "N" can take only the form of a simple CC-case:



(no genuine N-case)

Hence, I propose:

(6) $\delta_0 = \langle U \setminus \{\Delta_0\}, \rightarrow^*, H, D \setminus \{\Delta_0\}, (p_g), u\rangle$ is the truncated reduction of $\tilde{\delta}$ by $\{\Delta_0\}$ in the sense just defined (where g runs through $\times Ac(\Delta_0)$).

We have to make clear to ourselves what this amounts to: The causal graph $\langle U \setminus \{\Delta_0\}, \rightarrow^*\rangle$ of δ_0 is obtained from $\langle U, \rightarrow\rangle$ by deleting, together with Δ_0, all arrows ending or starting at Δ_0 and, provided $Oc(\Delta_0)$ is not empty, by adding arrows from all $A \in Ac(\Delta_0)$ and all $B \in Pa(\Delta_0)$ to all $C \in Oc(\Delta_0)$. The action nodes in Ac (Δ_0) are thereby turned into exogenous variables, making the other children of Δ_0, if any, directly causally dependent on all the parents *and* all the action children of Δ_0.

This may appear not entirely intelligible. The first mystery may be how a decision node may at all have any causal influence that is not mediated by action nodes. But let

us grant this point for the moment; it will become clear when we consider specific examples with non-empty $Oc(\Delta_0)$ in the next section.[34]

The second mystery is the modification of the CC-case which creates arrows running backwards from $Ac(\Delta_0)$ to earlier members of $Oc(\Delta_0)$. Do we thereby allow for backward causation? No. Recall our observation in section 3.2 that in a reduced causal graph an arrow $A \to B$ generally signifies only that B directly causally depends on A *or* that the IC-, the CC- or the N-case applies to A and B. Here, it can only be the CC-case which applies to the arrows from $Ac(\Delta_0)$ to $Oc(\Delta_0)$. Note that we had no choice here but to assume the anomalous backward arrows. If we had added only forward arrows in the reduction, i.e., arrows from the earlier members of $Oc(\Delta_0)$ to $Ac(\Delta_0)$ and from $Ac(\Delta_0)$ to the later members of $Oc(\Delta_0)$, then only the latter, but not the former, would have survived the truncation. But there is no reason whatsoever to treat the former and the latter arrows in the truncation in a different way; the temporal location of the members of $Oc(\Delta_0)$ is irrelevant to the causal structure of the situation and should not make any difference. Hence, the information about the common cause of $Ac(\Delta_0)$ and $Oc(\Delta_0)$ has to be completely, and not only partially, retained in the truncated graph. The only way to do this is by adding in the reduction only arrows starting from, and not leading to, $Ac(\Delta_0)$, as it is shown in the above diagram of the CC-case. This point will be a crucial step of my whole argument because it allows us to interpret critical action dependencies of probabilities as indicating merely a common cause relation.

Concerning the rest of δ_0, it is clear that δ_0 contains the same utility function as $\hat{\delta}$ since decision nodes do not carry utilities anyway. The probability family (p_g) of δ_0, finally, is derived from the measure p of $\hat{\delta}$ by eliminating the reflexive probability of condition (5) and all probabilities entailed by it, in particular the probabilities for the actions in $Ac(\Delta_0)$. The procedures described in section 3 then guarantee that the remaining family (p_g) agrees with the reduced and truncated graph.[35]

The upshot of all this is that δ_0 contains the same decision relevant items as the reflexive model $\hat{\delta}$ and indeed all of them; the surplus of the reflexive model is only the agent's firm belief that he *is* in δ_0 and what follows from this belief. In this way, the circularity problem that plagued our modeling of reflexive states is solved.

---

[34]Let me point out, though, that the presence of this kind of causal influence is a most crucial point behind the insightful considerations of Frank (1988). The decision situations in which one finds oneself are accompanied by emotions which show in other ways than merely in the actions optimal in these situations.

[35]There is a familiar problem about conditionalization here. The relation between decision nodes and the appertaining action nodes will be essentially deterministic. Hence, if $p(\delta 0) = 1$, then $p(g^*) = 1$ for the action (course) g^* that is optimal in $\delta 0$ and $p(g) = 0$ for all other actions g so that p_g remains undefined for them. But the problem is familiar from sect. 1, and it may be corrected with similar means. I neglect it here in order to avoid further complications.

However, I should emphasize that thereby the "no probabilities for acts" principle has reentered the picture, if only in relation to the variables in $Ac(\Delta_0)$; the other action variables are taken care of by the later decision nodes. The reason is that δ_0, which observes this principle, contains precisely what is needed for determining or causing the optimal action. The surplus of the reflexive model has no effect in this respect – though, of course, we can positively assert this only after the decision rule determining optimal actions has been introduced.[36]

This is all we need for our present purposes. However, the definition of the reflexive decision model $\hat{\delta}$ is not yet completed. Let me indicate what is still missing.

For one thing, we require

(7) a condition concerning the shape of all the decision situations δ in all the decision nodes $\Delta \in D$.

This condition would not differ so much from our detailed condition (6) on δ_0. The most important difference will be that the agent may envisage arbitrary changes of probabilities and utilities in all the possible decision situations, whereas the probabilities and utilities of δ_0 had, of course, to agree with those of $\hat{\delta}$. To be sure, theoretical work only becomes substantial by considering various specific forms of change. For example, a case that is treated extensively in decision theory is the one where the agent deliberates whether first to collect (possibly costly) evidence and then to decide on the basis of probabilities changed accordingly. But many more changes may be conceived[37], and conceptually one should allow for all kinds of changes.

Another, final, ingredient is missing, indeed the most difficult and important one: the decision rule that specifies for each possible decision situation what is rationally or optimally done in it. In our reflexive context, this also determines the agent's beliefs about the relation between decision and action nodes, since he believes to be and to stay rational (otherwise, it would be wrong from the outset to consider only decision nodes as parents of action nodes). Hence, the final condition is:

(8) For any decision node Δ and any situation $\delta \in \Delta$, if $g \in \times Ac(\Delta)$ is irrational in δ according to the relevant decision rule, then $p(g \mid \delta) = 0$.

This condition is stated only negatively. In case there are several optimal actions in δ the general model should not ordain specific probabilities for those actions.

---

[36]This has been one of my two arguments for this principle in Spohn (1977, sect. 2), the one which Rabinowicz (2001) considers to be the stronger one.

[37]In Spohn (1999, sect. 4) I tried to start a systematic treatment of the possible kinds of changes.

But what is the relevant decision rule? I am not sure, and I think nobody knows. It is obvious that it will be a recursive one. Each situation δ in the last decision node Δ_n is free of further decision nodes; the strategic horizon does not extend further. Hence, each such δ is a basic decision model, and the rule of maximizing conditional expected utility is good enough for it. Having determined optimal choices for all situations in the decision nodes $\Delta_{k+1},\ldots,\Delta_n$, we may then continue considering all the feasible strategies for the situations in Δ_k and maximize expected utility, as defined in section 4.1, and so on until we reach Δ_0. This procedure may be strictly defined. Indeed, it is what is usually called *sophisticated choice*: predict what you will find rational, and thus will do, in the future situations you might reach, and do right now what is rational from your present point of view given these predictions![38] However, I do not think that sophisticated choice is adequate for reflexive decision models in general. Since I have made no assumptions concerning the probabilities and utilities in later situations and their evolution from earlier ones, the agent's future points of view and his present one may mesh in more complicated ways than is conceived in sophisticated choice. I have argued the point and proposed an improved decision rule sketchily in Spohn (2000, sect. 4) and in formal detail in Spohn (1999, sect. 3). The issue is extremely difficult, though. Fortunately, it is irrelevant for our present purpose, and these meager hints may suffice. The only decision rule we will need in the next two sections will be the maximization of conditional expected utility in basic decision models, which forms the unproblematic recursion base.

This finishes my (partial) construction of the reflexive decision model $\bar{\delta}$. I admit, it has become quite involved. But decision nodes have a rich internal structure and are richly connected within the graph. So these are unavoidable costs of reflexion. Every item of the construction is significant and consequential, and I recommend the whole structure for further investigation.

4.4 Where do we stand?

Before going on, let us take stock where we presently stand with respect to our concern for understanding dependency equilibria. Our proximate goal was to undermine the exogeneity of action nodes in basic decision models. This we have achieved, indeed in the only feasible way, I think, by introducing decision nodes; it would have been inconsistent to proclaim decision theory as a theory of rational action and then to resort to other theories when reflecting on the causation of actions.

---

[38]Sophisticated choice was first developed by Strotz (1955/56) and further elaborated and discussed, among others, by Pollak (1968), Yaari (1977), and Hammond (1976, 1988). See also the thorough discussions in McClennen (1990).

What does it help, though, to understand action nodes in this way as endogenous? Nothing, it seems. One may have feared that, by allowing for the endogeneity of action nodes, we plunge right into the absurdities of evidential decision theory, by recommending, for instance, to refrain from smoking in Fisher's scenario of the smoking gene (even if one would like to smoke) or, generally, to perform an act which is symptomatic of the more favorable circumstances, even though it does nothing to promote these more favorable circumstances. But this does not follow. If we look at δ_0, the present situation of the agent of which he has self-knowledge, we find again the nodes in $Ac(\Delta_0)$ to be exogenous; the decision rule for δ_0 does not recommend to choose the symptomatic act. The same is true for the reflexive model $\widehat{\delta}$ that is governed by the same decision rule. But here the fatal consequence is blocked by self-knowledge; since the agent knows he is in δ_0 and in no other decision situation, the chosen act cannot be symptomatic of the circumstances; it would be so only via alternative decision situations. In fact, this is precisely how Eells (1982, ch. 8) blocks the consequence and thus establishes two-boxing as the rational solution of Newcomb's problem within the confines of evidential decision theory. Hence, we have generally maintained the very argument which we attempted to circumvent in order to make causal sense of dependency equilibria: namely the argument that causal independence from actions implies probabilistic independence.

Still, our progression into reflexive decision theory with its introduction of decision nodes provides us with the key to a solution of our problem. But the key lies elsewhere; it does not lie in the endogeneity of action nodes, but rather in the efficacy of the decision nodes extending beyond action nodes. The above weak version of condition (2) leaves room for the possibility that a decision has a causal relevance that the appertaining action has not. And the decision rule for the decision model that results from the truncated reduction of a reflexive model states what is rational if that possibility should obtain. I want to develop this idea in the next two sections. For this purpose, let us leave these (too) abstract fields and apply the structure to two apparently simple, albeit most significant examples, namely, the toxin puzzle and Newcomb's problem. This application will turn out to be highly instructive.

5. The Toxin Puzzle and Newcomb's Problem

5.1 The toxin puzzle

What is the *toxin puzzle*? It was invented by Kavka (1983) and goes as follows: This noon someone who seems to be a rich neuroscientist, let us call her the predictor, approaches you and makes you an offer. She shows you a glass filled with a fluid

called toxin. It tastes disgusting and makes you feel terribly kind of seasick for about three hours, but has no after-effects; afterwards everything is fine again. The offer the predictor makes is this: If you have the firm intention by midnight to drink the glass of toxin tomorrow noon, she will reward you by transferring $ 10,000 to your bank account briefly after midnight. You can trust on this, even if you cannot confirm by tomorrow noon whether the money has arrived. She has a kind of cerebroscope that reliably establishes whether you have the intention by midnight. What you do the next day is irrelevant; you are rewarded for having the intention, not for actually drinking the toxin.

Suppose you clearly prefer $ 10,000 over three hours of terrible sickness. Otherwise, you would reject the offer. So this seems to be easily earned money. Where is the problem? It is that it seems impossible for you to form the relevant intention and to earn $ 10,000. For you know already now that tomorrow noon, when you will be standing in front of the glass of toxin, you will have no reason to drink it. The case will be decided, the predictor has transferred the money or not, according to what the cerebroscope told her, and then it would be clearly less pleasant to drink the toxin than to abstain from doing so. Thus, how could you ever form the required intention in the full awareness of all this?

Well, is it really impossible? Two things should be clear. On the one hand, intention and action are not analytically tied together. One may drink the toxin, or do anything, without intending to do so. The popular view among action theorists is that actions analytically presuppose at least some intention or have to be intended under some description; only mere behavior may occur without intention.[39] However, in this respect our action nodes should rather be understood as behavior nodes. How an action node realizes depends on nothing but the behavior shown by the agent. Of course, he intends it to be an action, and if the behavior is actually caused by a decision node, it *is* an action. But this is not how an action node in itself is characterized. Conversely, one may have the firmest intention possible to drink the toxin without actually drinking it. Philosophers call this weakness of will and have their problems with this possibility. But it *is* one. Perhaps, after smelling the toxin the disgust is too great to be overcome even with the best intention. Perhaps you even start drinking it, but your throat automatically revolts, and you spit it out without control. However, even if the predictor foresees that this will happen, she should not deny you the intention.

On the other hand, if you leave open the slightest possibility to reconsider the case the next day (with predictable outcome), or even if you only hope that your body will unintentionally refuse the toxin, this defeats your present intention to drink the toxin. If the cerebroscope detects anything of this kind, any false thought, the predictor is right to deny you the reward.

---

[39] This view was forcefully introduced by Davidson (1971).

How, then, should we model this decision problem with the means introduced before? A preliminary point is that we have to translate the action theoretic talk of intentions into decision theoretic talk in which the term "intention" does not occur. My proposal is to translate "intending to do a" as "being decided to do a" and the latter as "being in a decision situation in which it is optimal to do a". These translations enter delicate terrain. Intention talk and decision talk do not square easily.[40] Thus, I propose simply to accept the first translation. Concerning the second translation, one should note two things: that we are dealing only with rational persons who do not err in their decision and always opt for the optimal or rational action, and that decision theory does not take decision making as a process in time which starts with analyzing one's situation and eventually results in a decision; being in a decision situation that is represented by some model automatically entails knowing the optimal action according to this model.[41] With these two observations, the second translation seems justified.

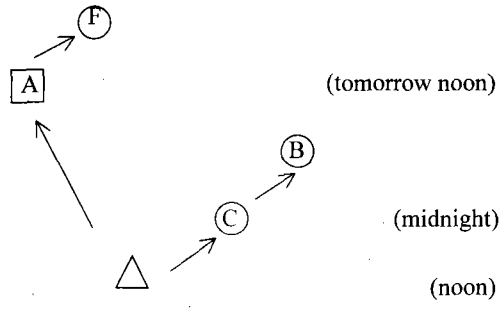
Given these translations, it is clear that the model for the toxin puzzle should contain five nodes or variables: an action node A = {drink, do not drink}, three occurrence nodes F = {feeling seasick for three hours, not feeling so}, C = {cerebroscope is positive, it is negative} and B = {\$ 10,000 are on your bank account, they are not}, and a decision node Δ = {decide to drink, decide not to drink}. These nodes are, I find, properly arranged in the following causal graph:

---

[40]For instance, action theorists are occupied with the locution „doing a with the intention i", a locution that appears very crude from the decision theoretic point of view because actions are usually guided not by one goal, but by many values as they are represented in a utility function.

[41]What lurks in the background here is the long discussion about the so-called deduction problem in doxastic logic, i.e., about the question of whether or not one may assume belief to be consistent and deductively closed (cf., e.g., Stalnaker 1999, ch. 13 and 14). The fundamental opposition here is that between semantic or regulative and computational theories. Regulative theories neglect the computational aspect and operate only on the semantic level by setting there the normative standards for correct computations, and thus assume belief to be deductively closed. In computational theories this is considered illegitimate and one tries to do without this assumption. I observe that theories about propositional attitudes naturally move on the regulative level, that they are substantial theories within their idealized confines, and that computational theories are poor or more or less equally bad idealizations.

These observations hold for decision theory as well. For instance, Pollock (1995) is one of the few exceptions having tried to develop a decision theory at the computational level. But already its epistemic part is to be criticized, as I argue in Spohn (2002). Hence, in my view, the only successful decision theories move on the regulative level – hence my remark about the „automatic" knowledge of the optimal action.



The positions of *A*, *B*, *C*, and *F* are beyond dispute, but the place of Δ needs discussion. Let us first complete the reflexive model, though. The utilities of the various courses of events are clear. The causal relations may be assumed to be more or less deterministic; that is, given Δ is positive (decide to drink), the probability for the other variables being positive is roughly 1, and given Δ is negative, the other variables are almost certainly negative, too. Moreover, the probability for Δ being positive is also 1: its conditional expected utility is roughly the utility of all variables being positive, the conditional expected utility of the negative decision is roughly the utility of all variables being negative, the first value is clearly larger than the second, and hence the positive decision is the optimal one (which means being determined to drink the toxin, actually drinking it, feeling seasick, and receiving the reward).

The crucial point in achieving this result is, of course, the introduction of the decision node *and* its side effect upon the cerebroscope (which is, to be sure, the fantastic part of the story). If we looked only at the action node, we would see only its causal influence on the sickness and thus conclude that abstaining from the toxin is the dominant action, as is maintained by those who find it impossible to have the positive intention. We should therefore inspect this crucial point more closely.

Why assume one decision node as the temporally first one? One alternative seems to be to assume two decision nodes, one at noon today and the other at tomorrow noon, both governing the same action node *A*; that is, you decide now to drink the toxin, and later you reconsider and decide not to drink it. Well, this may be what actually happens; we should not consider it impossible that a person is firmly decided, but nevertheless reconsiders and arrives at a different decision later on. This may even be reasonable; perhaps some new fact has emerged which one would never have dreamt of and which puts the previous decision into an entirely new light. However, the decision model models the perspective of the agent and not what may actually happen. This perspective cannot accommodate two decision nodes; to envisage a second decision node at which to decide about the toxin would cancel out the first decision node; it would mean that the case is, after all, *not* decided at the first node. Therefore, I

think condition (2) is right in insisting that each action node is governed by exactly one decision node.

Are there any further alternatives? I see only one, namely, to place the decision node Δ immediately before the action node A. My condition (2) above allows for any temporal distance between an action node and the relevant decision node. This is perhaps too liberal. It might be tempting to think that as long as there is time to decide about an action, it is not yet decided; this would imply that the decision to do a is located immediately before a, and hence Δ immediately before A. But I am not tempted by this idea. I do not see why one should exclude the possibility to decide a matter very early, to maintain one's decidedness until the time of action has come, and then to do what one has decided. One may object that maintaining one's decidedness is a matter of choice, something one may or may not do. Perhaps. But then we are back at the first alternative, which we have already rejected. Again, one has to distinguish between an external, objective perspective and the agent's intentional perspective. Objectively, the agent always has a choice – I do not see why I should deny this – so that the matter is decided only at the last possible moment. By contrast, from the agent's present point of view, there can be only one choice: either right now, and then it is a matter of intention to stick to that choice; or later on, and then it is a matter of prediction what the situation and the choice will be. And it is the agent's point of view which is to be captured by our modeling.

Part of the present intention or decision to do a some time later is the intention not to reopen the issue in the meantime. This intention and the capability to realize it may be subsumed under the label "commitment power".[42] More or less of this power may be needed. We shall soon see that the same point will be relevant to Newcomb's problem and that the commitment power needed there is minimal. In the case of Ulysses and the Sirens the required commitment power is superhuman; intention and will is powerless against the seduction of the Sirens; and knowing this, Ulysses is well-advised to have the sailors' ears blocked and let himself be bound to the mast. The toxin case is between the extremes; it takes some courage to resist the temptation and to overcome the disgust. Of course, to some extent it is vague and, I suspect, even a matter of convention when the intention is (too) feeble and the temptation (too)

[42]One must be careful here. „Commitment" is rather a family concept in game theory. Each specific application is clear, but what unites the various applications is not entirely clear. However, one prevailing idea seems to be that in order to be committed to do a, I have to commit myself to do a, where the latter is an action by which I establish external forces, as it were, that will compel me to do a. In this sense, commitment power is the power to establish such external forces. In my view, a commitment can exist without such externalization (even though this is sometimes characterized as magic or self-hypnosis), and in this sense, commitment power is, as explained, the capability to maintain one's decidedness (which may, of course, be supported by external means). McClennen (1990, sects. 9.7 and 12.9) has best explained the meaning of „commitment" as I am using it.

strong, that is, when the weakness of the will is or is not exculpable. And this vagueness can, of course, not be resolved by the cerebroscope. It is an unavoidable, not a central aspect of the above example.

It should be clear by now that the distinction I am belaboring here is tantamount to the distinction between sophisticated and resolute choice by McClennen (1990). Sophisticated choice takes a predictive attitude toward future actions and optimizes present actions given the predictions. Resolute choice, by contrast, takes an intentional or commissive attitude toward future actions and optimizes present and future actions at once. Like McClennen I find nothing miraculous in that intentional attitude of resolute choice. Hence, it is a welcome aspect that the distinction is reflected in my formal framework. But the way in which it is reflected is perhaps more important, namely simply by the temporal location of the decision node(s) in the reflexive model. Reflexive decision theory is thus able to reconcile these two methods of choice; they are both instantiations of one kind of model and one decision rule.

Indeed, there is a choice here between sophisticated and resolute choice. Obviously it is not fixed in advance at which time the decision node Δ is to be located. You may decide the case right now, or you may defer the decision to tomorrow. And you choose when to decide according to where you have the higher conditional expected utility. In the model presented above it is higher than in the model where Δ is later. Thus, in the above reflexive model of the toxin puzzle I have already specified the result of this optimization. Hence, whether the agent proceeds according to sophisticated or to resolute choice, is itself a matter of optimal choice. Reflexive decision models do not yet represent this choice since they assume a fixed causal graph. But they show a way to extend the theory so as to cover this choice as well. This extension would be most interesting, but it is beyond the scope of this paper, all the more so as we have left the general decision rule for reflexive decision models aside.

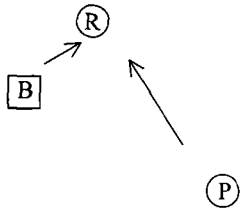
5.2 Newcomb's problem

With all these points in mind let us turn now to *Newcomb's problem* which is much closer to our ultimate concern since it bears often observed similarities to the prisoners' dilemma. The problem is quickly explained:

You will be led into a room with two boxes on the table. One is transparent and visibly contains \$ 1,000; the other is opaque, but you are told that it contains either \$ 0 or \$ 1,000.000: You have a choice: you may take either the opaque box only (one-boxing) or *both* boxes (two-boxing). This seems an easy choice – of course, you do not want to miss out on the additional thousand dollars in the transparent box – but there is a highly irritating detail. A predictor has already predicted what you will do and has

filled the opaque box accordingly. If her prediction is that you will take only one box, she fills it with a million dollars; if she predicts you will be two-boxing, nothing is put into the opaque box. As I mentioned, the prediction and the filling of the boxes are already completed; any cheatful change is excluded. The predictor is surprisingly knowledgeable, and you know this. In, say, 95 percent of the cases her prediction is correct, concerning one-boxers and two-boxers alike; and very likely she will be correct in your case as well. Does this story behind the opaque box change your mind? Will you now take only one box?

The majority tends to stick to two-boxing, and there is a considerable variety of theories justifying this.[43] I find the simplest theory emerges when we look at the natural formalization of the situation in a basic decision model, which contains three variables; the action variable B = {one-boxing, two-boxing}, and two occurrence variables P = {prediction of one-boxing, prediction of two-boxing}, R (= return) = {receiving \$ 0, \$ 1,000, \$ 1,000,000, \$ 1,001,000}. The temporal, and thus the causal, relations are clear:



You cannot influence the prediction because it has already taken place, and even in a reflexive context the prediction cannot influence your decision because you have no information whatsoever about it. The point now is that the agent's probabilities have to agree with this causal graph. Hence, the prediction is not only causally, but

[43]See, e.g., the papers collected in Campbell, Sowden (1985) and Sobel (1994), and, more recently, Joyce (1999).

also probabilistically independent of your choice[44], and so two-boxing turns out as the dominating action and as the only rational choice.[45]

For more than twenty years I was firmly convinced by this consideration. But nagging doubts have kept returning. It is not only that the minority of one-boxers is the personified bad conscience of the majority of two-boxers.[46] Rather, it is the question of Gibbard and Harper (1978): "If you're so smart, why ain'cha rich?" The answer of Lewis (1981) is that rational persons have no chance if irrationality is (pre-)rewarded. But this answer is feeble-minded; somehow rationality should show in the adaptability to the conditions of success.

The doubts are reinforced if we look at the iteration of Newcomb's problem. Suppose you are confronted with this situation a hundred times, say, once a day for a hundred days. So, if all goes well, you will have earned a hundred million dollars! But it seems you cannot earn them; backward induction stands against it: The above model applies to the last round; hence you will decide then to take both boxes. If the predictor does not have a false theory about you, she will predict exactly this, and you will receive just one thousand dollars. Thus, it is clear already now what will happen in the last round, and since this is clear, the only rational thing for you to do in the penultimate round is again to take both boxes, and so forth. On the whole, you will thus end up with a hundred thousand dollars instead of a hundred million. This seems absurd.

Of course, one should scrutinize this use of backward induction. Perhaps it is more problematic here than in the genuine game theoretic application (though I do not see why). However, despite the doubts that have been raised against backward induction, it is a forceful argument.[47] Therefore, I think it would be too weak a strategy only to try to undermine backward induction. Indeed, the absurdity is also created by the claim that only two-boxing is rational in the one-shot case. This claim was doubtful as well. Hence let us look more carefully at the original problem.

---

[44]This entails that you cannot transfer the predictor's observed success rate of 95 percent to your own case and form the corresponding probabilities conditional on your own options. That this divergence between the statistical and the subjective probabilities in your own case lies at the heart of Newcomb's problem was forcefully argued by Kyburg (1980).

[45]This is the view of Newcomb's problem which I have presented in Spohn (1978, sect. 5.1). It is explained also in Meek, Glymour (1994) with the explicit use of the theory of Bayesian nets.

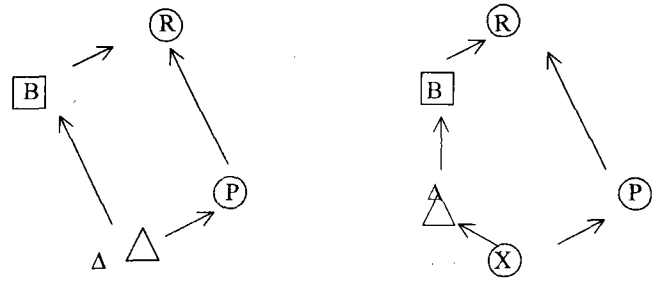
[46]Jeffrey, for instance, has changed his mind several times, in (1983), (1988), and (1996).

[47]The doubts were initiated by Binmore (1987), Reny (1989), Bicchieri (1989) and Pettit, Sugden (1989). Aumann (1995) proves backward induction to be valid under certain assumptions. This result may be used as a defense of backward induction (against Aumann's explicit advise). However, Rabinowicz (1998) in turn makes clear the respect in which these assumptions are unreasonably strong. Rabinowicz is right, I think, but the gap he opens seems too small to accommodate all the positive conclusions about cooperation I am after.

Let us start by asking what you might or should think about the predictor, if you were in this situation. No doubt you would be wondering how she manages to be so successful and what her theory about human agents might be. On the other hand, it is clear what your own theory about yourself is. You firmly believe to be rational and to maximize conditional expected utility. This is how you fix your intentions concerning your present actions, and this is how you predict your future actions in future situations. That this is so is, of course, presupposed in this paper. If your self-theory is entirely different, this paper simply does not apply to you.

These remarks do not yet fully determine your self-theory since maximizing conditional expected utility may mean various things explicated in a number of theories that differ in certain details. We will come to this in a moment. What I want to focus on right now is the tension between your assuredness regarding your self-theory and your bewilderment about the predictor. Perhaps the predictor has simply been lucky to have been successful so many times in the past; but this is too improbable. Perhaps her prediction is guided, say, by the form of your nose and thus by an entirely wrong theory; but again success with such a theory appears to be sheer luck. Thus, in view of your self-theory, you should rather assume that her means of prediction are correlated not only with your action, but also with your ways of deciding and thus

theory about you should be the same as your self-theory. This ideal is not so difficult to attain; after all, we are here publicly discussing your self-theory. And certainly, you would like this ideal to obtain; you cannot believe that there is any better theory about you than your self-theory.



(a) decision causes prediction

(b) decision and prediction have a common cause

Case (b) shows how the situation was conceived by Eells (1982, p. 211), who was the first to introduce the reflexive point of view, and how thereafter the correlation between prediction and action was usually understood. If we look here at the reduced and truncated model $\delta_0 \in \Delta$, which is reflected by the full model $\bar{\delta}$, we find that B and P, action and prediction variable, are causally and probabilistically independent. Hence, two-boxing is the dominant and thus optimal action in δ_0 and in $\bar{\delta}$ as well.

Fisher's smoking scenario has the same structure as case (b) provided the arrow from P to R is deleted. Here, X would be the variable of the smoking gene that is responsible for the desire to smoke as well as for lung cancer, P would represent getting lung cancer, B would be the action variable about smoking, R would be a variable about the pleasure derived from smoking, and Δ would be the decision variable containing the various possible desirabilities for smoking caused by X. Then, as before, if I find myself in the actual decision situation δ_0 with its preference for smoking, it would be rational to smoke since it is causally and probabilistically independent from P and thus dominates nonsmoking.

So, am I wedded to familiar views? No. Case (a) is a possible scenario as well, making even more sense and allowing very different conclusions; this is what I would like to defend in the following. Case (a) presupposes that the decision variable Δ temporally precedes the prediction variable P. This seems to contradict the very description of Newcomb's problem; here you are standing before the two boxes, and now is the time to decide. However, we have observed in the toxin puzzle that the temporal position of the decision variable is something that is within your choice, too. Thus, when you are standing before the boxes, you may as well consider the decision as having been taken a long time ago and as having only to be executed right now. Indeed, it would be rational to place the decision variable at such an early time because this yields the higher conditional expected utility. And, although it may be unclear how

the predictor actually arrives at her prediction, it follows that, if she has observed your rationality in the past and thus uses the correct theory to predict your behavior, she should take you to be decided to take only one box and hence put the million dollars into the opaque box. This leaves open the exact temporal position of Δ. You can place it as early as you want, in principle as early as the inception of your rationality (but, of course, this, too, is not an event with a well-defined temporal location), and in any case early enough for the predictor to observe, and to become convinced of, your rationality.

What I would like to claim is that case (a) at least presents an intentional picture you *may* have. I cannot say you should have it because it is too unclear how the predictor operates. Maybe case (b) actually applies, however early Δ is placed, even though the predictor is much more likely to observe the consequences of your rationality rather than its causal preconditions. Maybe case (a) applies, but your trust in the efficiency of your decidedness regarding the predictor is not firm enough to justify one-boxing. Maybe the predictor thinks that despite being decided to take only one box you will succumb to the temptation to take the additional thousand dollars. Or maybe the predictor has indeed been nothing but lucky. All this is unclear. Still, case (a) is a way to rationalize one-boxing *without* causal mystery, and if you could be sufficiently sure that the predictor has developed a correct picture of you, then case (a) provides the optimal representation of your situation and you should be decided to take only one box.

According to this point of view, the presentation of Newcomb's problem has been utterly deceptive. It was presented as if you had to decide now so that the causal situation could only be as in case (b). But actually you can also consider yourself to be decided so that now, when standing before the boxes, you simply have to carry out your previous commitment without being in a new decision situation.

All this shows, by the way, the difference to the smoking gene example. There is no way to conceive that example as in (a) and thus to rationalize nonsmoking.

The same caveat applies here as in the toxin puzzle. Even if case (a) represents your intentional picture of the situation (as it should, given you have enough trust into the predictor), it is not a picture of what actually happens. Standing before the boxes with the best of your intentions, you may still reopen the issue, decide anew, and conclude to take both boxes. Or you might simply succumb to the temptation to take two boxes (although there is no aversion to overcome in this case in order to stick to your commitment). Or you might believe you can outsmart the predictor, and so on. Either way, you may end up taking both boxes against your rational intention.

However, let us look again at the iterated case where you have the relevant choice a fixed number of times. (An informal look will suffice, it would be too cumbersome to formalize it.) According to the old picture, case (b), backward induction showed that

you should rationally take both boxes every time, that the predictor, if she has a correct theory about you, should predict just this, and that your only way to circumvent this result is to somehow convince the predictor of your *irrationality* (say, by starting to take only one box). The new picture, case (a), is completely reversed. Now the predictor should rationally expect you to take only one box; and if you are as rational as always to take only one box, everything is fine. If, however, she observes you taking two boxes, this will undermine her belief in your rationality, and she will presumably adjust her predictions. The crucial point is here that the backward induction has lost its base because we have shown two-boxing to be not the only rational solution of the single-shot case. On the contrary, if the predictor has the right theory about you, one-boxing is the *only* rational solution of the single-shot case. And backward induction then shows that this is so in the iterated case as well. At any rate, whatever your temptation to deviate from your rational intention in the single-shot case may be, it should be effectively suppressed in the iterated case.

Let me make sure that the crucial step in my argument is crystal clear: It is my stipulation in section 4.3 that in the truncated reduction of the reflexive model the causal arrows have to go from the action children of the deleted decision node to its other children, even if the latter precede the former. This anomaly can be observed in our examples. There, the truncated reduction δ_0 contains an edge from B to the earlier P (Newcomb) or from A to the earlier C (toxin), respectively, and the probabilities in δ_0 agree with this. But this does not amount to assuming backward causation. Rather, the arrow $B \rightarrow P$ in the truncated reduction δ_0 means only that P causally depends on B, *or* B and P have a common cause (since the N-case cannot apply); and the latter is indeed true of our cases. Hence, rather than creating a mystery, we have an *explanation* how the probabilities can be as they are in δ_0 without assuming a causal anomaly, namely, by understanding these probabilities and the dependencies they embody as the result from reducing and truncating a richer, entirely normal causal graph.

Finally, only because the probabilities are as they are in δ_0, the rationality is reversed, and drinking the toxin and taking one box turn out to maximize conditional expected utility. In a nutshell, the point is this: (1) that we may maximize the conditional expected utility of our decision or that of our action, (2) that the two maximizations may fall apart, (3) that in such cases only the former, i.e., the maximization of the conditional expected utility of the decision, is rational, (4) that this maximization can be made explicit only in a reflexive model, and (5) that it is only the truncated reduction of the reflexive model as here defined which preserves this maximization on the level of unreflexive models.

6. *Prisoners' Dilemma*

6.1 How cooperation may be rational in prisoners' dilemma

Let us return to our starting point, the game theoretic dependency equilibria and in particular to the single-shot PD which we found to have a single Pareto-optimal dependency equilibrium and which we need to attend for avoiding the absurdities emerging in the finite iteration. Does the foregoing shed new light on PD?

Hope is nourished by the long-observed kinship between Newcomb's problem and PD. Newcomb's problem is a one-sided decision problem because the predictor is not modeled as an agent or player. It is left unclear what drives the predictor. She appears to be a wealthy philanthropist, or maybe she is a goddess, and her only interest is to predict correctly.[49]

By contrast, PD is a two-sided Newcomb's problem. Each player has well-defined utilities and acts according to them, and each player takes both the role of the agent and that of the other agent's predictor. Since defection dominates cooperation, it is clear for both how to act and what to predict: joint defection. At least, this is the unsatisfactory standard view, just as two-boxing is the majority view regarding Newcomb's problem. Hence, the foregoing change of view should also open up a new perspective on PD.

However, because of the double role of the players, the situation is more complicated. If the roles were separated, the picture would be clearer. If, say, Ann were in the role of the predictor and Bob in the role of the agent, Bob could not hope to have an influence on Ann's decision situation by being decided early enough. Because Ann has interests of her own and not the predictor's philanthropic ones, she would in any case choose defection as her "prediction".

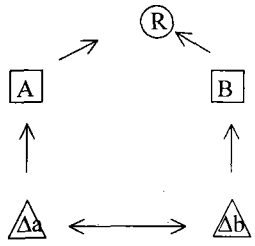
But in fact the roles are not separated in this way. With her decision, Ann does not only predict Bob's action, but also tries to influence his decision; likewise, by forming a certain intention or decision, Bob does not only try to impress Ann, but also responds to her intention or decision. Thus, we have to conceive their decision situations as entangled, as shown in the following causal graph:

---

[49]Selten, Leopold (1982) have analyzed the situation as a game. With bi = i-boxing and pi = predicting i-boxing (i = 1,2), the most plausible utility matrix is given by

u \ v	p_1	p_2
b_1	1 2	0 0
b_2	0 3	1 1

Here (b2, p2) is the only Nash (and the only correlated) equilibrium, whereas (b1, p1), (b2, p1), and (b2, p2) are dependency equilibria.



Here, Δ_b and B are occurrence nodes from Ann's perspective; her decision node is only Δ_a and her action node is only A. The converse holds for Bob.

By drawing a double arrow between Δ_a and Δ_b, I have enriched causal graphs in a way initially not allowed. How are we understand this double arrow? Not as a causal loop, which would be impossible. I think we should rather conceive the variables Δ_a and Δ_b as temporally extended so that there is enough time for causal influence to flow back and forth and thus to establish the interdependence indicated by the double arrow. Hence, there is no causal mystery here.

However, by assuming this interdependence, have I not violated the independence constitutive of PD outright? No, I do not think so. The independence in PD applied to the actions, in the first place. And this holds here as well; there is no causal influence running from A to B or from B to A.

That is a lame argument, one will object; of course, the causal independence was also meant to hold between the players' decision situations. There the criminals sit in their prison cells, pondering what to do and being firmly precluded from any communication. Well, one *may* represent their situation in this way; then one would have to delete the double arrow in the diagram above, and joint defection would ensue as the only rational solution.

But one *need not* represent their situation in this way. My point here is the same as in Newcomb's problem, where one could take one's decision as having been determined a long time before one is actually standing before the boxes. The temporal location of the decision was not fixed there; rather, it was rational in the sense of maximizing conditional expected utility to locate it as early as possible.

Likewise in PD. I am not simply referring to the trivial fact that the criminals may have communicated and coordinated long before they were imprisoned and that later on they may stick to their old intentions or decisions. I rather think that the players should see their decision situations, even in the absence of communication, as being entangled in the way described above, and that this entanglement can rationally take only the form of the Pareto-optimal dependency equilibrium since this is the only way how they can individually(!) maximize conditional expected utility. This is what full

rationality demands for all PD-like situations; anything else but cooperation would be a deviation from full rationality, i.e., irrational.

However, it should be clear again that all this only describes how the players should rationally view and decide their decision situations. What actually goes on may be different. A player may doubt that the co-player(s) are fully rational in the sense explained (perhaps because it is likely that they have not fully grasped the theory of rationality developed here), and he may thus have different conditional probabilities and different expected utilities. Or he may be tempted, against his or her intention or decision, to reconsider the issue and conclude to defect (or not to drink the toxin, or to take two boxes), and there may be strong incentives to do so. And so on. But this does not disprove the rationality of cooperation.

I grant that such factors may make cooperation unlikely in the one-shot PD. The situation changes, however, in the iterated PD. Here, the rationality of the players becomes more clearly observable. In particular, if one player defects, this displays his deficient rationality, with all the detrimental effects on the conditional probabilities and future cooperation. Of course, this is far from a theoretical treatment of the iterated PD. But it completely reverses the standard rationality account of the iterated PD in, I find, a plausible and desirable way. The cornerstone of this reversal is, of course, the different view of the one-shot case, which undermines backward induction in favor of continued defection right at the beginning. Moreover, the reversal seems to open up the possibility of rationalizing the intuitively compelling and experimentally confirmed tit-for-tat strategy and similar strategies embodying kindness and responsiveness[50], a rationalization that still seems to be wanting.

6.2 A dialogue

Incredulity and rejection was the most frequent response to my accounts of the toxin puzzle, Newcomb's problem, and PD. It may be helpful to dramatize the central doubts in a little dialogue between the opposing wisdom (*OW*), which I myself still feel very strongly, and the daring proponent (*DP*), whose role I have taken here.

OW: To be honest, I don't believe a word of what you said.

DP: I can imagine.

OW: Let's return to the toxin puzzle. Here you stand the other day at noon, about to drink the toxin. Do you really want to deny that you would be better off not drinking it?

DP: Yes.

OW: You're kidding.

---

[50]Cf., e.g., Axelrod, Dion (1988).

DP: No. Look! It simply does not occur to me when standing before the toxin that I'd prefer to refrain from it. This has been clear all along. But I prefer $ 10,000 even more. Therefore, I decided at midnight to drink the toxin, and it was part of my decision not to decide anew; otherwise, my decision would not have been a genuine one, and I would not have got the $ 10,000.

OW: But you *can* decide anew, and you would be better off.

DP: Yes, I can, but I shouldn't. I don't deny that you are maximizing conditional expected utility when you refrain from drinking the toxin. But I have given you a perfectly reasonable explanation of how I am maximizing conditional expected utility, too. I would already be satisfied if you could grant this. And these are not two different senses of maximization. We apply the same theory, we only differ in structuring the decision situation. Of course, I have claimed more. I have also claimed that what maximizes expected utility depends on *when* you decide, and that if you can choose when to decide, you are better off in the toxin case if you decide early.

OW: But the final decision is taken only immediately before the action.

DP: What do you mean by "final"? Is there a decision and a final decision? No, I think to assume that decisions always immediately precede actions would be an unsupported and restrictive assumption. You *can* choose when to decide in the toxin case. By the way, often you just carry out old decisions, too.

OW: But only because my present deliberation would confirm my old decision. Or because the decision costs are too high for always deliberating anew.

DP: Are you sure these are the only reasons?

OW: Well, let me ask you another question: How do you manage to maintain your decidedness? Somehow, you have to commit yourself.

DP: What do you mean?

OW: I mean you have to take some measure ensuring that you will really stick to your decision in the end, say, by instructing a friend to return the $ 10,000 if you don't drink the toxin.

DP: That's a point where I do not understand you at all. You present the case as if I had to guard against my rationality and had to force myself into doing something that I will judge to be irrational and that I would never do in the absence of force. This is a very distorted picture of commitment. I don't have to guard against my rationality; that's almost analytic. Of course, others may not trust in my sense of commitment, and then I may have to prove it to them by some such measure. But I don't have to prove anything to myself. No, I am committed merely by the insight into the rationality of being committed. This is part and parcel of what it means to be rational. It is not a mysterious or superhuman additional capacity. Whether it is rational to be committed depends on the situation, but in the toxin case it is. Of course, there are also intricate intertemporal situations that are difficult to assess, but the toxin case is a very simple

one in which the basic point stands out very clearly. What I have to guard against is my weakness that I may not follow my insight. I admit this is often necessary, but that's another story.

OW: Hm, we could be moving in circles here. Let me turn to Newcomb's problem. You have added another mystery here. In the toxin case, you actually decide by midnight and think you can maintain this decision. In the Newcomb case, you stand before the boxes, you reflect on the case for the first time in your life, and now you say you discover, as it were, that there is nothing to decide, that you have been decided for a long time, and that you only have to carry out your old commitment. How can you be decided, when you never reflected on the case?

DP: I agree, this sounds odd, at least if you conceive of a decision as an occurring mental event. But if you don't, you may well discover that you are committed. Since I have not pondered about Newcomb's problem before, the idea cannot be that the predictor somehow observed me pondering and deciding to take only one box and that she filled the opaque box accordingly. The idea is rather that the predictor has had many opportunities to observe my rationality in general, that she inductively infers that I will behave rationally on other occasions as well, that rationality, i.e., maximizing conditional expected utility as explained, requires me to be committed in this very special type of situation, and that knowing this both of us behave accordingly. The same applies to the toxin case, by the way. If I am rational, the predictor's cerebroscope should see this, and then we both can deduce that I have the intention to drink the toxin. If you accept my toxin story, it's really a small step to accept my Newcomb story.

OW: Hm. But it is not a small step to PD.

DP: I agree. I fear my presentation has diminished the differences. On the other hand, if what I said so far is acceptable, this *must* have consequences for PD. Do you have something particular in mind?

OW: Several things! First, you have extensively explained why the probabilities as they are in your dependency equilibria need not imply a causal interdependence between the players' actions. The interdependence may as well obtain only between their decision situations. Now I can imagine how the latter interdependence comes about by all kinds of signaling, commitment moves in my sense, or what have you. But if I understand you correctly, you allow for such things, but you say they are not required.

DP: Yes, precisely.

OW: But how am I then to conceive of that interdependence? That's definitely a step beyond your *intra*personal "discovery to be committed".

DP: Yes, certainly. But if photons can be entangled, why should persons not manage to be so as well?

OW: Come on, that's a bit farfetched, isn't it?

DP: Ah, you know well enough that I would never make any quick inferences from the quantum to the human domain, say, from uncertainty to freedom. No, purely from the perspective of a theory of causation the comparison is not so farfetched. But you are right, just as I said that rationality requires you to be committed in the toxin and the Newcomb case, I now say that rationality requires both players to be entangled in the Pareto-optimal dependency equilibrium and to be committed to cooperation. This entanglement and commitment emerges with the situation without requiring binding agreements or whatever. That's what maximizing conditional expected utility dictates.

OW: This sounds as if the players should choose their beliefs so as to maximize conditional expected utility.

DP: No, certainly not. They do believe in common rationality, and this belief has consequences.

OW: All this is hard to believe. But let's grant your entanglement of the players' original decision situations. Now, however, my very first doubt returns. Suppose you are one of the prisoners, you are in prison, and you find yourself entangled with your partner, as you describe it. But you are free to break the entanglement, and you should notice that you are better off by doing so. So, you should break it and leave your co-player in her imaginary entanglement. Or rather, you should realize that the same holds for your co-player, and that there is no way to save the entanglement, anyhow.

DP: I feared that you would come up with the hardest question of all. I see the difference to the toxin and the Newcomb case. There I am committed to myself. I would not maximize expected utility, if I were not fully decided early enough. In PD, by contrast, I am committed to my co-player (as she is to me), and by breaking my commitment, I am harming her, but not, it appears, myself. That makes a difference. On the other hand, if there are commitments in the sense I have described above, why shouldn't there be joint commitments in the same sense, whose violation would be individually irrational?

OW: ???

DP: Look, I need not defend the view that cooperation is the only rational option in the single-shot PD. The only point I want to insist upon is that one *may* represent the decision situations of the players as originally entangled in the weakly Pareto-optimal dependency equilibrium – an entanglement, which is not broken by later reconsideration and by a new decision and in which it is rational to cooperate even in the single-shot case. That's all that is required for breaking the force of backward induction. As soon as you look at iterations the joint commitment will unfold its momentum.

OW: Hm, I would like to see this worked out. But what does commitment and entanglement help in case of several Pareto-optimal dependency equilibria?

DP: Not much. PD is particularly simple insofar as entanglement alone already yields a unique solution. But the problem of selecting from multiple equilibria exists here as elsewhere. The rich literature on bargaining theory should apply in my setting as well as in the usual one.

OW: Let me ask you a final question. By calling your dependency equilibria equilibria, you use a venerable name. This seems to be a false pretense. Equilibria are so called because they are *self-stabilizing*. This is true of correlated equilibria as well. But I don't see how this characterization should apply to dependency equilibria.

PD: I don't find much of a difference. First of all, you know well enough that mixed Nash equilibria are not self-stabilizing as such.

OW: Yes, of course, they have to be backed up by Harsanyi's (1973) account, say, according to which they are approximations of equilibria in perturbed games.

PD: I admit I have not thought about a similar account for mixed dependency equilibria. But a pure dependency equilibrium *is* self-stabilizing. Both players have an incentive to stick to it since each would lower their conditional expected utility by deviating from it. But even in a mixed dependency equilibrium, just as in a mixed Nash equilibrium, the players don't have an incentive to deviate, in the sense that each will do only what has positive probability in the relevant marginal of the joint probability distribution. Otherwise, they would again lower the conditional expected utility.

OW: But this is just the argument why *correlated* equilibria are self-stabilizing.

DP: No, in correlated equilibria deviation includes the break of the dependency in the equilibrium distribution. That's not the sense appropriate for dependency equilibria. There, deviation has to keep *within* the entanglement of the players' decision situations, as embodied in the equilibrium distribution. Hence my consistent talk of maximizing *conditional* expected utility. Of course, this is to say that this explanation of self-stabilization presupposes all my stories about joint commitment and entanglement. If you don't swallow the latter, you can't accept the former.

OW: You seem to have, or want to have, an answer to every question.

PD: This is my role here, isn't it?

OW: But roles aside, do you really believe every word you say?

PD: How do you expect me to respond? I have grown up in the same tradition as you. It is a good and strong tradition, but presumably not one with which one can be fully content, and certainly not one from which one can easily free oneself. How can I know then what I really believe?

6.3 Some comparative remarks

So many pages have been filled with PD, indeed so many pages at least resembling mine in spirit, that I cannot aim here for any completeness in my comparative remarks. In fact, I do not have an argument with similar approaches. The question is not really who is wrong and who is right, but rather to which extent the various approaches are at least illuminating. Hence, my concluding comparative remarks rather aim at clarifying in which way I believe my ideas to differ from similar ones.

(1) Some philosophers may say that they have offered a much simpler rationalization of cooperation in the one-shot PD than the one I have put forward, namely, via the so-called *mirror principle*[51] which says that whenever Ann and Bob are in the same decision situation, they act in the same way. In PD they are in the same situation because of the symmetry of the story. Hence, only joint cooperation and joint defection are possible outcomes. If they believe in the principle, they also believe that these are the only possible outcomes. Hence, since for both joint cooperation is better than joint defection, it is rational for both to cooperate.

This argument is entailed by my account, so I agree with its conclusion. But it is too quick. It avoids causal considerations, and it does not present a theory of rationality that would entail the rationality of cooperation for Ann and Bob in their particular situation. Therefore, it does not exclude that mutual defection satisfies the mirror principle as well, as the standard theory has it. In a way, it takes its conclusion for granted. By contrast, my account tries to back up the mirror argument by specifying a theory of rationality in which the rationality of cooperation emerges as a conclusion. Moreover, it is hard to see how to apply the mirror principle when the situation is not symmetric; but there is no such presupposition in my account.

(2) In a similar spirit, some game theorists may say that my account really belongs to *cooperative game theory* within which the cooperative solution of PD is no mystery, anyway. However, if we follow Osborne, Rubinstein (1994, pp. 255f.) and take cooperative game theory to refer to groups of players without considering "the details of how groups of players function internally", this is not true. I have at least attempted to offer a theory of *individual* rationality rationalizing cooperation. If instead we follow Harsanyi, Selten (1988, p. 1) and "define cooperative games as those permitting enforceable agreements", it is again not true. Clearly, my rationalization of cooperation refrained from alluding to enforceable agreements. Whether it succeeded is, of course, another question. In any case, I have tried to tell a story about individual rationality backing up cooperative game theory to some extent.

(3) To continue on the issue, Harsanyi, Selten (1988, pp. 4ff.) showed how cooperation can emerge as an ordinary Nash equilibrium in a non-cooperative PD by

[51]Cf. Davis (1977) and Sorensen (1985).

adding a preplay of commitment moves. Ann starts making her conditional commitment move ("I commit myself to cooperate, if Bob does so as well"), Bob follows with his unconditional commitment move ("I commit myself to cooperate"), and then both cooperate. Thereby we do not even have to assume a causal *inter*dependence of the players' decision situations; the dependence is successively generated by the commitment moves. We also find the relevant conditional beliefs in their original decision situations, which turn into unconditional beliefs in cooperation after the commitment moves have been taken. But, of course, the idea of Harsanyi and Selten is that there is some external mechanism sufficiently lowering the pay-offs in case the commitments are violated. As I have explained, this is not my conception of commitment, I hoped to do without such a mechanism. Only in this way can I hope to extend my argument to the iterated PD in which cooperation should emerge without external help, an argument that has not been pursued by Harsanyi and Selten.

(4) My proposal closely resembles the old theory of conjectural variations.[52] Friedman (1983, p. 107) concludes that "at the level of simultaneous decisions in a single-period model, conjectural variation is not meaningful" and hence interprets the old single-period models of conjectural variation as being implicitly about a dynamical process that has been more recently treated in multiperiod models. So, does this verdict apply as well to the account offered here? Yes, at least according to opposing wisdom, as I called it. However, the effective objection has been that single-period conjectural variations assume a causal dependence that does simply not exist. This objection does not apply here. At least I have taken pains with explaining why dependency equilibria do not assume the causal dependence they seem to assume. Whether my resort helps making sense of the theory of conjectural variations is, however, beyond my judgment. The answer may very well be negative. After all, duopoly and PD are quite different chapters of game theory.

(5) Without doubt, my closest allies are Albert, Heiner (2001) and Heiner, Albert, Schmidtchen (2000). They also seek a causally unassailable rationalization of one-boxing in Newcomb's problem, and they also generalize their account to a treatment of the single-shot PD. It is surprising to see, however, how intentions can be so close and how their realizations can still differ in so many details. Hence, some comparative remarks are called for.

There is no doubt that their approach is more detailed and makes sense more immediately. This is so because they choose quite a different setting, namely, that of evolutionary game theory. Their idea is that solvers of Newcomb's problem come in two types, the greedy type and the temptation resistant type. These labels are used just for illustration, the real difference between the types lies in their decision rules. The greedy type bases their calculation of expected utilities on the disconnection principle,

[52]As Max Albert has pointed out to me. This theory is well reviewed in Friedman (1983, sect. 5.1-2).

as they call it, which favors two-boxing, and the temptation resistant type on the backtracking principle favoring one-boxing. Moreover, the types are somehow predictable; they send out some kind of signal which displays their type (more or less reliably) and which the predictor can use for her prediction. It is plausible then, and Albert and Heiner explain this more rigorously, that only the temptation resistant type will survive if evolution rewards high pay-offs. A similar, though more complex account applies to PD. The players come in various types, exchanging signals displaying their type, and behaving according to their type and the signal received. Then, again, evolution will favor the (conditionally) cooperative type.

At face value, at least, this is not a story about individual rationality. And it seems it cannot be turned into one. I am lucky, or not lucky, in the evolutionary lottery that assigns a type to me, but the question of which type I should rationally *adopt* does not arise. This is certainly a major difference to my attempt, and I do not see how it could be bridged.[53]

A number of further differences are entailed by this. Also considering the causes of actions, Albert and Heiner introduce what they call motivation packages, which resemble my decision nodes. However, they do not engage in what I have called reflexive decision theory. The causal picture they develop is that of the external theorist rather than that of the agent himself, a fact that shows up also in their rejection of my (or rather Eells') reflexive condition (5) in section 4.3. Indeed, a motivation package in their sense does not only consist of beliefs and desires or probabilities and utilities, but also of a decision criterion, and their point is that the types may differ only in the latter component of their motivation packages. By contrast, I have no place for variation here. For me, there is only one decision rule, the rational one (at least as long as I have not given up and acquiesced in many rationalities). And since I am dealing only with rational subjects, there is only one decision rule for the subjects as well. Still, I can mimic the types of Albert and Heiner. The temptation resistant (or, as I would say, the commissive) type is decided early, whereas the greedy type is decided late (though both types apply the same decision rule). And because I can decide when to be decided, I can choose my type. However, this is not to say that my commissive type follows their backtracking principle. My account rather proceeds in terms of truncated reductions as explained, which agrees with the backtracking principle in the Newcomb case, but perhaps not in general. Moreover, since my idea is that there is a unique rationality, I may hope to dispense with signals (though, of course, they are helpful in actual life as well as in the presentation of models). The common insight of the predictee and the predictor (in the Newcomb case) or the players (in PD) into what

[53]In private communication Max Albert tells me they need not stick to the evolutionary approach and could turn their story into one about individual rationality. This would indeed shift the discussion lines between us.

is rational in their kind of situation should suffice. But I grant that this may be a mysterious part of my account, a part Albert and Heiner cannot share, all the more so, as they would certainly reject my conception of commitment.

Thus, I have listed seven or eight differences, and there may be more. This strikes me as rather fruitful. It would be palliative to say that these differences confirm our common intention. But they certainly span a fascinating field of further investigation that may produce still better rationalizations of cooperation.

References

Albert, M. and R.A. Heiner (2001), "An Indirect-Evolution Approach to Newcomb's Problem", *Homo Oeconomicus*, forthcoming.

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

Aumann, R. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55, 1-18.

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

Axelrod, R. and D. Dion (1988), "The Further Evolution of Cooperation", *Science* 242, 1385-1390.

Barwise, J. (1990), "On the Model Theory of Common Knowledge", in: J. Barwise, *The Situation in Logic*, CSLI Lecture Notes 17, Cambridge 1990.

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

Binmore, K. (1987), "Modelling Rational Players, Part I", *Economics and Philosophy* 3, 179-213.

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

Cartwright, N. (1989), *Nature's Capacities and Their Measurement*, Oxford: Clarendon Press.

Cartwright, N. (1999), "Causal Diversity and the Markov Condition", *Synthese* 121, 3-27.

Davidson, D. (1971), "Agency", in: R. Binkley et al. (eds.), *Agent, Action, and Reason*, Toronto: University of Toronto Press, 3-25.

Davis, L. (1977), "Prisoners, Paradox, and Rationality", *American Philosophical Quarterly* 114, 319-327.

Dawid, A.P. (1979), "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society*, Series B 41, 1-31.

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

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

Frank, R.H. (1988), *Passions Within Reasons*, New York: Norton & Company.

Friedman, J.W. (1983), *Oligopoly Theory*, Cambridge: Cambridge University Press.

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, Dordrecht: Reidel, 125-162.

Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations", *Econometrica* 11, 1-12.

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

Hammond, P. (1988), "Consequentialist Foundations for Expected Utility", *Theory and Decision* 25, 25-78.

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

Harsanyi, J.C., R. Selten (1988), *A General Theory of Equilibrium Selection in Games*, Cambridge, Mass.: MIT Press.

Heiner, R.A., M. Albert and D. Schmidtchen (2000), "Rational Contingent Cooperation in the One-Shot Prisoner's Dilemma", unpublished manuscript.

Jeffrey, R.C. (1965/83), *The Logic of Decision*, Chicago: Chicago University Press.

Jeffrey, R.C. (1988), "How to Probabilize a Newcomb Problem", in: J.H. Fetzer (ed.), *Probability and Causality*, Dordrecht: Reidel, 241-251.

Jeffrey, R.C. (1996), "Decision Kinematics", in: K.J. Arrow et al. (eds.), *The Rational Foundations of Economic Behaviour*, Basingstoke: Macmillan, 3-19.

Jensen, F.V. (1996), *An Introduction to Bayesian Networks*, London: UCL Press.

Joyce, J.M. (1999), *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press.

Kavka, G. (1983), "The Toxin Puzzle", *Analysis* 43, 33-36.

Kyburg jr., H.E. (1980), "Acts and Conditional Probabilities", *Theory and Decision* 12, 149-171.

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

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

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

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

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

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

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, Ca.: Morgan Kaufmann.

Pearl, J. (2000), *Causality. Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

Pettit, P., 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.

Pollock, J.L. (1995), *Cognitive Carpentry*, Cambridge, Mass.: MIT Press.

Rabinowicz, W. (1998), "Grappling With the Centipede", *Economics and Philosophy* 14, 95-125.

Rabinowicz, W. (2002), "Does Deliberation Crowd Out Self-Prediction?", to appear in *Erkenntnis*.

Reny, P. (1989), "Common Knowledge and Games With Perfect Information", *Proceedings of the Philosophy of Science Association* 2, 363-393.

Salmon, W.C. (1980), "Probabilistic Causality", *Pacific Philosophical Quarterly* 61, 50-74.

Salmon, W.C. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton, N.J.: Princeton University Press.

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

Selten, R. and U. Leopold (1982), "Subjunctive Conditionals in Decision and Game Theory", in: W. Stegmüller, W. Balzer and W. Spohn (eds.), *Philosophy of Economics*, Berlin: Springer, 191-200.

Shafer, G. (1996), *The Art of Causal Conjecture*, Cambridge, Mass.: MIT Press.

Simon, H.A. (1957), *Models of Man*, New York: Wiley.
Sobel, J.H. (1994), *Taking Chances: Essays on Rational Choice*, Cambridge: Cambridge University Press.
Sorensen, R.A. (1985), "The Iterated Versions of Newcomb's Problem and the Prisoner's Dilemma", *Synthese* 63, 157-166.
Spirtes, P., C. Glymour and R. Scheines (1993), *Causation, Prediction, and Search*, Berlin: Springer.
Spohn, W. (1976/78), *Grundlagen der Entscheidungstheorie*, Ph.D. Thesis Munich 1976, published at Scriptor, Kronberg/Ts. 1978.
Spohn, W. (1977), "Where Luce and Krantz Do Really Generalize Savage's Decision Model", *Erkenntnis* 11, 113-134.
Spohn, W. (1980), "Stochastic Independence, Causal Independence, and Shieldability", *Journal of Philosophical Logic* 9, 73-99.
Spohn, W. (1982), "How to Make Sense of Game Theory", in: W. Stegmüller, W. Balzer and W. Spohn (eds.), *Philosophy of Economics*, Berlin: Springer, 239-270; reprinted in: Y. Varoufakis and A. Housego (eds.), *Game Theory: Critical Concepts, Vol. 4, Discontents*, London: Routledge, 2001, 213-241.
Spohn, W. (1988), "Ordinal Conditional Functions: A Dynamic Theory of Epistemic States", in: W.L. Harper and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, Dordrecht: Kluwer, 105-134.
Spohn, W. (1999), "Strategic Rationality", Forschungsberichte der DFG-Forschergruppe "Logik in der Philosophie", Nr. 24, Konstanz.
Spohn, W. (2000), "A Rationalization of Cooperation in the Iterated Prisoner's Dilemma", in: J. Nida-Rümelin and W. Spohn (eds.), *Rationality, Rules, and Structure*, Dordrecht: Kluwer, 67-84.
Spohn, W. (2001), "Bayesian Nets Are All There Is to Causal Dependence", in: D. Costantini, M.C. Galavotti and P. Suppes (eds.), *Stochastic Dependence and Causality*, Stanford: CSLI Publications, forthcoming.
Spohn, W. (2002), "A Brief Comparison of Pollock's Defeasible Reasoning and Ranking Functions", *Synthese* 131, 39-56.
Stalnaker, R.C. (1999), *Context and Content*, Oxford: Oxford University Press.
Strotz, R.H. (1955/56), "Myopia and Inconsistency in Dynamic Utility Maximization", *Review of Economic Studies* 23, 165-180.
Verma, T. and J. Pearl (1990), "Causal Networks: Semantics and Expressiveness", in: R.D. Shachter et al. (eds.), *Uncertainty in Artificial Intelligence*, vol. 4, North-Holland, Amsterdam, 69-76.
Wright, S. (1934), "The Method of Path Coefficients", *Annals of Mathematical Statistics* 5, 161-215.
Yaari, M.E. (1977), "Endogeneous Changes in Tastes: A Philosophical Discussion", *Erkenntnis* 11, 157-196.