Dependency Equilibria and the Causal Structure of Decision and Game Situations

Wolfgang Spohn[*]

Fachgruppe Philosophie
Universität Konstanz
D–78457 Konstanz

Abstract: The paper attempts to rationalize cooperation in the one-shot prisoners' dilemma (PD). The first step consists in introducing (and investigating) a new kind of equilibria (differing from Aumann's correlated equilibria) according to which the players' actions may be correlated (section 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 (section 3). The analysis then raises to a reflective point of view in which the agent integrates his own present and future decision situations into the causal picture of his situation (section 4). This reflective structure is first applied to the toxin puzzle and to Newcomb's problem, showing a way to rationalize drinking the toxin and taking only one box without assuming causal mystery (section 5). The latter result is finally extended to a rationalization of cooperation in PD (section 6).

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 considerable number of astonishingly varied attempts somehow to undermine defection as the rational solution and to establish cooperation as a rational possibility at least in the iterated

[*] This paper was conceived and partially written during my stay at the Center for Interdisciplinary Research of the University of Bielefeld, where I participated at the research group „Making Choices. An Interdisciplinary Approach to Modelling Decision Behavior". 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.

case. But the hard case, I believe, still stands unshaken. Under appropriate conditions backward induction is valid[1]; hence given full rationality (as opposed to 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 the conclusion apalling; it amounts to an outright refutation of the underlying theory of rationality. And I find that all the sophisticated observations that have been made about PD cannot soothen this harsh conclusion.[3] Cooperation *must* be at least a rational possibility in the finitely iterated PD, and under ideal conditions even more so than under less ideal conditions. So, something must be changed in standard rationality theory, i.e. decision and game theory. After a long time of thinking otherwise[4], I have come to the conclusion that it is the single shot case which needs to be reconsidered. This paper tries to do so.

It does so in five steps. Section 2 introduces and discusses a new notion of equilibrium for games in normal form which I call dependency equilibrium. Dependency equilibria may be an object of interest by their own, but they seem to assume an unintelligible causal structure (this may be why they have not been considered in the literature so far). So, the main effort in this paper will be to make causal sense of them; only then they deserve serious interest. To this end, section 3 makes explicit the causal picture that has been more or less implicit or explicit in decision and non-cooperative game theory since their inception. This helps to discover the crucial gap in this picture which I shall propose to fill in section 4 by what I like to call reflective decision. In section 5, the change of view thus brought about is first applied to the so-called toxin puzzle and then to Newcomb's problem where it is shown how drinking the toxin and the one-box solution may be rational *within* causal decision theory. It will be obvious how much this section owes to the conception of resolute choice[5]; the point will be, however, to show how resolute choice may be subsumed

---

[1] Cf. Aumann (1995).

[2] Iterated elimination of weakly dominated strategies is a reasonable procedure in application to the iterated PD, all the more so as the problems it may have as opposed to the elimination of strongly dominated strategies do not obtain in this application. Cf, e.g., Myerson (1991, sect. 2.5 and 3.1).

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

[4] Since Spohn (1978, sect. 5.1) I have been a fervent defender of the two-box solution in Newcomb's problem (see sect. 5 here for more on this), and in Spohn (2000a, sect. 6) I have offered a line of thought for breaking the force of backward induction in the iterated case which I withdraw; I don't see any more how it can be reasonably worked out.

[5] Developed by McClennen (1990).

under the reflective expansion of standard rationality theory. Thus prepared, I can return in section 6 to game theory and to PD in particular and explain how the foregoing confers causal sense to dependency equilibria and how mutual cooperation, which *is* a dependency equilibrium, can be fully rational even in the one shot case.

2. Dependency Equilibria

Let us start with a brief study of the new equilibrium concept. For comparison, it's good to quickly 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. As little as I develop the theory here it is routinely extended to *n*-person games.

So, 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, respectively, of Ann and Bob 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 the player her- or himself. But it is equally appropriate to interpret them as representing the beliefs of the other(s) 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 are 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. Still, it is useful to introduce to set $P_\perp \subseteq P$ of such products. Thus, $p \in P_\perp$ iff there are $s \in S$ and $t \in T$ such that $p = s \otimes t$, i.e. $p_{ik} = s_i \cdot t_k$. Hence, A and B are probabilistically independent just according the distributions in $P_\perp$. This is all the terminology we shall need.

As is well known, $\langle s, t \rangle$ is defined to be 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 no player 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.[6]

Obviously, Ann's and Bob's choices from *A* and *B* are independent in a Nash equilibrium. This is an assumption I would like to give up (for reasons that will become clear later on). Now, Aumann (1974) has introduced an equilibrium concept that allows for dependence between the players. Here is his definition from Aumann (1987) (which is a bit 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 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. Then, one combination is chosen at random according to this distribution, and each player is told only her or his part of the combination. If no player can raise now her or his expected utility by breaking her or his part of the agreed joint distribution and choosing some other pure or mixed strategy instead, then this joint distribution is a correlated equilibrium. Hence, 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 an enlarged game, and it turns out then that all and only the Nash equilibria of the enlarged game correspond to correlated equilibria in the original game.[7] This reflects the fact that correlated equilibria, despite their allowance of dependence, are still Nashian in spirit. The players' standard of comparison is still whether she or he might be better off by independently doing something else,

---

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

[7] For details cf. Myerson (1991, sect. 6.2).

where the expectations about the other player are given by the marginal over his or her strategies.

This standard of comparison is changed in the dependency equilibria to be introduced now. It is not the expected utility given the marginal for the other player, it is 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 her or his conditional expected utility with whatever she or he does 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 and hence no conditional expected utility of a_j is defined. But, of course, the fact that $s_j = 0$ should not render the other figures meaningless. There are three main ways to solve this problem. One may, first, engage into non-standard analysis and probability theory where one can conditionalize with respect to events having infinitesimal probability. This looks circumstantial at the least. One may, secondly, resort to Popper measures which take conditional probabilities as basic and have thus no problem with conditionalizing on null events. This would be the way I prefer.[8] However, the game theoretic community is rather accustomed to the third way which engages 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 probabili-

---

[8] My real preferences, however, go for probabilified ranking functions, a sophistification of Popper measures; cf. Spohn (1988, sect. 7).

ties appearing in this definition are well defined. Though the definition looks a bit more complicated now, the intuitive characterization given above still fits perfectly.

After this correction, the second observation is that dependency equilibria seem to be well in line with decision theory. Most textbooks tell that the general decision rule is to maximize *conditional* expected utility. Savage (1954) still assumed a clear separation of states of the world which have probabilities and consequences which are determined by acts and states, and then the decision rule is simply to maximize expected utility. However, this separation is often not feasible, and the more general picture laid out by Fishburn (1964) is that everything is probabilistically assessed (except perhaps the possible acts themselves), though only conditionally on the possible acts. In this general picture maximizing conditional expected utility is the pertinent decision rule. In view of this it may seem surprising that equilibrium theory did not proceed so far in these terms.

But of course, that's my third point, this is not astonishing at all. The idea behind the general picture is – we will have to look at all this much more carefully in the next section – that conditional probabilities varying from act to act somehow hide causal dependencies which are modeled more generally in a probabilistic way and not in a deterministic way as Savage (1954) did.[9] In the light of this idea, dependency equilibria are a mystery. If Bob chooses knowing what Ann has chosen, then, clearly, his choice causally depends on hers; that's 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 loop, and dependency equilibria seem to assume just this impossibility. So, whoever might have thought of them, he should have dismissed them, it seems, right away as nonsense.[10]

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 honest grounds. For the time being let us look a bit more closely at the properties of dependency equilibria.

This, however, seems to be a messy business. Obviously, the computation of dependency equilibria in two-person games requires to solve quadratic equations, and the more persons, the higher the order of the polynomials we get entangled with. All linear ease is lost. For this reason I cannot offer a well developed theory of depend-

---

[9] 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.

[10] So I did when I first conceived of them in 1982.

ency equilibria. It seems advisable to look at some much discussed simple games in order to get a feeling for the new equilibria, namely Matching Pennies, BoS (Bach or Stravinsky), Hawk and Dove, and PD. This discussion is more vivid when we consider also the other kinds of equilibria for comparison. Afterwards, we can abstract some simple theorems from these examples.

Matching Pennies: This is my paradigm for 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 easily verified that it has exactly one Nash equilibrium and exactly one correlated equilibrium which are characterized by the following distribution:

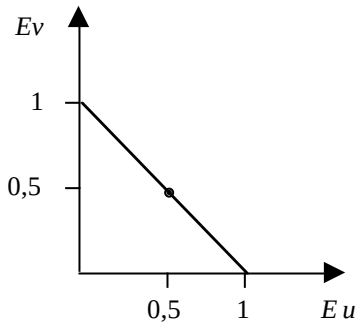
p	b_1	b_2
a_1	¼	¼
a_2	¼	¼

However, it is easily verified as well that the dependency equilibria of this game may be biased towards any of the pure strategy combinations:

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

, where $0 \le x \le \frac{1}{2}$.

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



• : Nash, corr.

– : depend.

Bach or Stravinsky: This game is a paradigmatic mixture of coordination and conflict. Its utilitiy matrix is this:

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

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

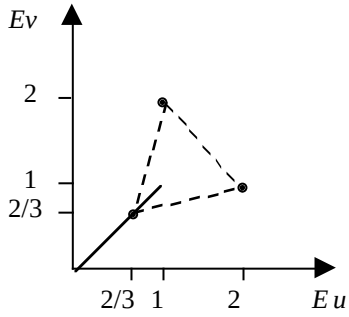
,

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.

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 displays 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:

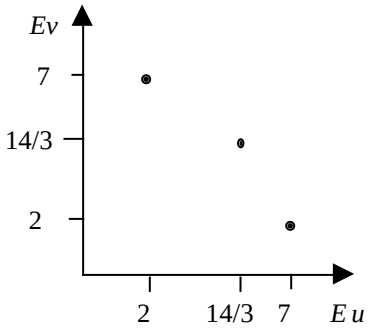
u \ v	b_1	b_2
a_1	6 6	7 2
a_2	2 7	0 0

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	$\frac{4}{9}$	$\frac{2}{9}$
a_2	$\frac{2}{9}$	$\frac{1}{9}$

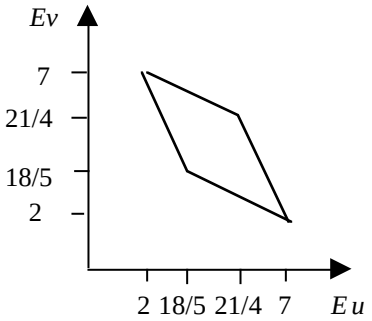


The correlated equilibria reach out further on the diagonal. They are given by

p	b_1	b_2
a_1	x	y
a_2	z	w

, where $x+y+z+w = 1$ and $0 \le \frac{x}{2}, 2w \le y, z$

and they yield the following expected utilities:



Again, we find 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

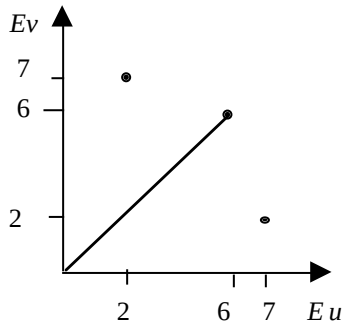
,

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 $y = \frac{1}{18}(2 - 15x + \sqrt{220 - 276x + 225x^2})$

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



Prisoners' Dilemma: This is my final example, the importance of which can hardly be overestimated. 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) 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:

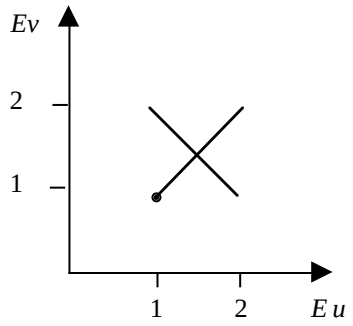
p	b_1	b_2
a_1	$\frac{1}{2}\ x(1+x)$	$\frac{1}{2}\ x(1-x)$
a_2	$\frac{1}{2}\ x(1-x)$	$\frac{1}{2}\ (1-x)\ (2-x)$

, where $0 \le x \le 1$, 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 $-\frac{1}{3} \le x \le \frac{1}{3}$.

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 an old and very simple observation that such dependency between the players may make them cooperate. But now we have found an equilibrium concept which underpins this observation. Moreover, we have seen that correlated equilibria do not provide the right kind of dependency for this purpose, they succumb to defection. Evidently, all this is strong motivation to try to make good sense of these dependency equilibria. This is the task we pursue in the rest of the paper.

For the moment, I shall not further discuss these examples or assess what we have found in them; the purpose of plain illustration is already well served, I think. However, the examples suggest some simple generalizations, all of which can routinely be extended 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 (possibly proper) 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.

Suppose, Ann has to choose first, and Bob chooses second, and both know this. In this case, if Ann does a_i, Bob will choose from $B(a_i) = \{b_k | \ v_{ik} \geq v_{il}$ for all $l = 1,\ldots,n\}$, the set of Bob's best replies. Suppose further that Bob's best reply is always unique so that $B(a_i) = b_{\beta(i)}$ for some function β and all $i \geq 1,\ldots,m$. Hence, Ann will maximize what she receives given the best replies, i.e. she will maximize $u_{i\beta(i)}$. This leads to

Observation 5: If $B(a_i)$ is a singleton for all $i = 1,\ldots,m$ and if $u_{j\beta(j)} \geq u_{i\beta(i)}$ for all $i = 1,\ldots,m$, then p with $p_{j\beta(j)} = 1$ or, for that matter, the pure strategy combination $(a_j, b_{\beta(j)})$ is a dependency equilibrium.

Of course, this holds also with the roles of Ann and Bob reversed. Such sequential decision making (and common knowledge thereof) is the simplest way of conceiving of a dependency between the players. However, it does not help in PD; there, the dependency equilibrium described in Observation 5 is just joint defection. Hence, we have to provide also other ways of understanding these equilibria.

Our observations about PD steer us to the 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 6: 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 $p^r = \frac{r\text{-}1}{r} \cdot p + \frac{1}{r} \cdot q$. 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, p with $p_{ik} = 1$ is a dependency equilibrium.

In *PD*, Hawk and Dove, and BoS this observation fully answers the quest for the Pareto-optima among the dependency equilibria. But it does not generally do this. In Matching Pennies no pure strategy combination is Pareto-better than the Nash equilibrium; still all of them are dependency equilibria.

This accentuates how preliminary my formal investigation of dependency equilibria is. However, it is by far not clear whether dependency equilibria are worth the efforts. This will occupy us for the rest of the paper. If the answer we find is convincing, this may motivate others to deepen the formal investigation.

3. Causal Graphs and Bayesian Nets

I have mentioned that dependency equilibria seem to be a causal mystery. In order to get clear about this, 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 contempted area). This piece of causal theory will clearly confirm some fundamental assumptions of decision and game theory which are causally motivated, but probabilistically expressed. So, it will at first deepen the mystery about dependency

equilibria. At the same time, however, we shall be able to see more clearly where the entry is for gaining a different view.

The standard theory I am alluding to is the theory of causal graphs and Bayesian nets.[11] It deals only with causal dependence and independence between variables. In order to do this, it must consider specific variables and not generic ones. Generic sociological variables would be, e.g., annual income or social status. But it is usually very hard to say anything substantial about causal relations between generic variables. Specific sociological variables would be, e.g., my annual income in 1999 or my social status in 2000, understood as ranges of possible values these variables may take and not as facts consisting in the values these 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 first 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 $\bigtimes V$ 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 carry a temporal order $<$. $A < B$ says that *A precedes B*.[12] I assume $<$ to be a linear order in order to avoid any 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 dependency 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

---

[11] This theory is more or less explicit in the statistical path analysis literature since Wright (1934) and in the linear modeling literature since Haavelmo (1943) and the papers in Simon (1957). The general 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 added the graph theoretical 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).

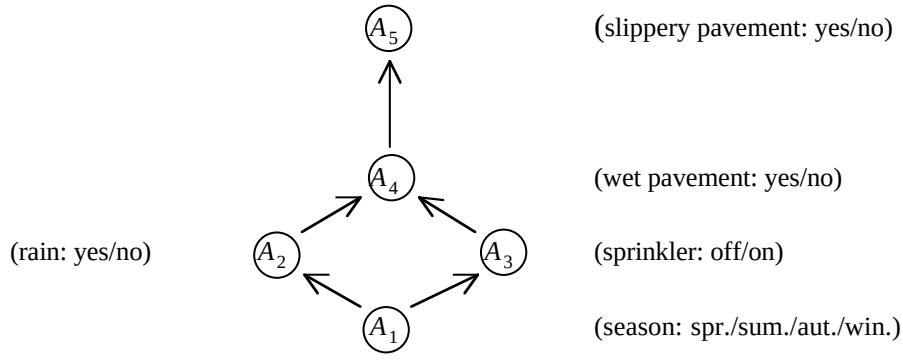
[12] This is not the *A* and *B* from section 2. From now on *A*, *B*, *C* etc. are used to denote any single variables whatsoever. Of course, the *A* and *B* from section 2 are also variables.

relation in U agreeing with the temporal order, and $\Rightarrow$ is the transitive closure of $\rightarrow$. Of course, directness and indirectness is here relative to the frame U; a direct causal dependence in U may well turn indirect in refinements of U.

Graphs are relations visualized. So, we may say as well that $\langle U, \rightarrow \rangle$ is a directed acyclic graph agreeing with the temporal order[13] or, as we define, 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 $B \Rightarrow A\}$.

A small example may be instructive. It's Pearl's favorite (though it is difficult to conceive of its season variable A_1 as a specific variable – 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 here realization probabilities, this means that the causal structure must somehow relate to these probabilities. I should emphasize that these probabilities maybe objective ones (whatever this means precisely), in which case they relate to the objective causal situation, or they may be

[13] The temporal order is often left implicit or even neglected, presumably because the statistical literature is rather interested in generic variables. I find the temporal order essential.

some person's subjective probabilities, in which case they reflect the causal beliefs of that person.[14] The latter perspective will be the relevant one for us later on.

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 (that is, $p(w) \geq 0$ for each course of events $w \in \times U$ and $\sum_{w \in \times U} p(w) = 1$). 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 anything about X.

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

$$A \perp Pr(A) \setminus Pa(A) / Pa(A),$$[15]

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, respectively, of non-descendants of A 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 Markov and the minimality condition w.r.t. to $\langle U, \rightarrow \rangle$.[16] In

---

[14] This assertion sounds nice, and I don't think it is really wrong, but it would deserve most careful explanation; the most profound philosophical problem with causation is just what to say here precisely.

[15] \ denotes set theoretic subtraction.

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

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.[17]

Let me illustrate these definitions with the above example: p satisfies the 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 eagerly exploited for implementing probability measures.[18]) And 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 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 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 $\langle U, \rightarrow\rangle$ is *blocked* or *d-separated* by a set $Z \subseteq U$ of nodes (or variables) iff

(1) 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
(2) the path contains some collider $A \rightarrow B \leftarrow C$ such that neither B nor any descendant of B is in Z.

---

[17] This was precisely my explication of direct causal dependence in probabilistic terms in Spohn (1978, sect. 3.3) and (1980).

[18] For an elementary introduction into the computational aspects of Bayesian nets see Jensen (1996).

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

Then 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.

This looks a bit complicated at first, but one gets quickly acquainted with this notion. 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 this notion is revealed by the following *theorem*: For all $X, Y, Z \subseteq U$, if X and Y are d-sparated 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 $\langle U, \rightarrow\rangle$.[20]

The proof is quite involved[21], but it 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] So, the second part of the theorem says that almost all p agreeing with $\langle U, \rightarrow\rangle$ are faithful to $\langle U, \rightarrow\rangle$. But sometimes it is useful to explicitly exclude the exceptional cases and to assume faithfulness outright.

One point where this is useful is the reduction of causal graphs. One might wonder how a faithful graph looks like after the deletion of a node. The answer is straightforward: Define the causal graph $\langle U^*, \rightarrow^*\rangle$ to be the *reduction* of the causal graph $\langle U, \rightarrow\rangle$ by the node A iff:

(1) $U^* = U \setminus \{A\}$,
(2) $B \rightarrow^* C$ iff either (i) $B \neq A \neq C$ and $B \rightarrow C$, or (ii) $B \rightarrow A \rightarrow C$, or (iii) $B < C$ and $B \leftarrow A \rightarrow C$.

Larger reductions can be generated successively and are obviously independent of the order of the succession.

---

[20] Cf. Pearl (2000, p. 18).

[21] See theorems 3.2 and 3.3 in Spirtes et al. (1993).

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

The notion of d-separation then helps to a straightforward proof of another *theorem*: if $\langle U, \rightarrow\rangle$ is faithful to p, then the reduction $\langle U, \rightarrow\rangle$ by A is faithful to the marginalization or restriction of p to $U \setminus \{A\}$.

The reverse perspective is more interesting, though: Our picture of the world is always limited, we always move within a small frame U. So, whenever we construct a causal graph $\langle U, \rightarrow\rangle$ faithful to our beliefs, we should consider that graph as the reduction of a yet unknown, more embracive graph. Each apparent direct causal dependence $B \rightarrow C$ in the graph $\langle U, \rightarrow\rangle$ may thus be the result of a reduction and unfold into an indirect causal dependence or a conjunctive fork with a common cause rendering the causal dependence of C on B spurious. This is what Reichenbach's principle of the common cause taught us in effect. This observation will turn out as crucial later on.

So much for the standard theory of probabilistic causation. Calling it the standard is perhaps justified in view of the impressive list of its predecessors and defenders. It is also more or less explicit in a lot of applied work and in particular in most of decision and game theory. But it is still contested, most critically perhaps by Cartwright (1989, 1999), who splits up the above Markov condition into two parts, a proper Markov condition relating only to the past of the parents of the 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 (2000b) 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 sequel, though, I shall neglect these debates and proceed on the standard theory introduced above.

So far, actions and agents did not enter the picture. A Bayesian net described either some small part of the world or some person's partial view of the world. But this person could be a wholly detached observer having merely beliefs and no interests whatsoever about that part of the world. 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 two ingredients.

The first ingredient consists in desires or interests which 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. H rather contains only the action variables still open from the agent's point of view to be modeled. That is, the decision model is to capture the agent's decision situation at a given time t, and then 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 implicit in all of the decision theoretic literature 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 (1999). 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 without parents. 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 $h \in \bigtimes H$; the actions may, of course, matter to what occurs in W. Hence we must replace the measure p for U by a family $(p_h)_{h \in \bigtimes H}$ of probability measures for W. Relative to such a family, Bayesian net theory still makes perfect sense; such a family can also satisfy the Markov and the minimality condition and can agree with, and be faithful to, a given causal graph.[23] However, it can do so only if the action variables are parentless. For a variable to have parents in agreement with the probabilities, conditional probabilities for this variable must be explained, but 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 $\langle U, \rightarrow, p \rangle$ with U containing 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 getting 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. And Pearl's question is 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 $\langle U, \rightarrow^* \rangle$ in which all arrows leading to action variables are deleted; that is, $A \rightarrow^* B$ 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 $p_h(w) = p(w \mid h)$ ($h \in \bigtimes H$, $w \in \bigtimes W$); this would reestablish the deleted dependencies. We rather 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):

---

[23] My definitions and theorems concerning conditional independence in Spohn (1978, sect. 3.2) were already so general to relate to such a family of probability measures. The graph theoretic material may be supplemented in a straightforward way.

> If $z \in \bigtimes U$ is a course of events in U[24] and if for each $A \in U$ a is the value A takes according to z and $pa(a)$ the values the variables in $Pa(A)$ take according to z, then $p(z) = \prod_{A \in U} p(a \mid pa(a))$.

And then we have to use the truncated factorization[25] which deletes all factors concerning the variables in H from the full factorization:

> If $h \in \bigtimes H$ and $w \in \bigtimes W$ 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 prove then that (p_h) agrees with $\langle U, \rightarrow^*\rangle$ if p agrees with $\langle U, \rightarrow\rangle$. So, as Pearl and I perfectly agree, it is the family (p_h) thus defined that yields the probabilities to be used by the agent. In this way, Pearl also subscribes to the two principles above.[26] The notion of truncation will become important in the next section.

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 $\bigtimes U$ into $\boldsymbol{R}$.

What is the associated decision rule? Maximize conditional expected utility, i.e., choose a course of action $h \in \bigtimes H$ for which $\sum_{w \in \bigtimes W} u(h,w) \cdot p_h(w)$ is maximized. However, this decision rule is naïve insofar it neglects the fact that the agent need not have to decide for a whole course of action; rather, he has 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 soon.

---

[24] „u“ is already reserved for the utility function.

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

[26] 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.

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 entirely 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 is authority, without doubt. We shall continue the topic in the next section.

Let me point out an important consequence, though, right away: 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.[27] 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 stubborn intuition of two-boxers, which has been mine 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 non-cooperation game theory. It is constitutive of non-cooperative game theory that the choices of the players are causally independent; they do not communicate or interact in any way. And the players are, of course, aware of this causal independence. Hence, if the observation is correct, the players' choices are probabilistically independent as well (also from their own point of view). This is what game theorists have assumed all along about non-cooperative game theory, and this is why we seem to be forced to something like Nash equilibiria 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.

---

[27] For a presentation of Newcomb's problem see section 5.

So our search for this causal sense can only take one direction: we have to scrutinize the assumptions underlying basic decision models. This is our next task.

4. Reflective 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 knows obviously with which probability he will take which action, 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 other persons. This reflection should clearly enable him to have (probabilistic) predictions about his future actions, again in contradiction to the principles. We shall see that both thoughts come out the same; but let us dwell upon them separately and a bit more carefully.

I 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 to 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 is the reversal of 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 $w \in \bigtimes W$ and $h \in \bigtimes H$ $p_q(h,w) = p_h(w) \cdot q(h \mid w)$ – where $q(h \mid w)$ denotes the probability that the action sequence h realizes according to q and the course of events w. More explic-

itly: if for each $A \in H$ a is the value A takes according to h and $pr(a)$ the values that the variables in $Pr(A)$ take according to h and w, then $q(h \mid w) = \prod_{A \in H} q(a \mid pr(a))$. These are just the factors we need in order to fill up a truncated factorization to a complete one.

This, in turn, enables us to define an *expected utility* for each dependency scheme q: $Eu(q) = \sum_h \sum_w u(h,w) \cdot p_q(h,w)$ (where h ranges over $\times H$ and w over $\times W$). 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 just is that I don't know where I have put them. Hence, dependency schemes maximizing expected utility rather 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.

Still, we would like to know, of course, 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 at all that the frame contains the variables which the agent is able to connect up with in a deterministic way. Perhaps the agent receives at best incomplete information about the variables in the frame preceding the relevant action variable, in which case only a probabilistic dependency is within his power. Hence, as long as we don't make special assumptions about which variables are in the frame U, no more can be said about the feasibility of dependency schemes.

So, maybe we should 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 the state of which he knows, or learns to know, before the time of action. Maybe his behavior is correlated with the state of certain variables without the agent noticing it. But if so, that behav-

ior 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 to know. 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 enrich 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 consider only dependency schemes which make 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 enriched causal graph). Moreover, it is obvious which direction the action node's intentional dependence on the decision node should take: the relevant decision rule, say, maximizing conditional expected utility, tells which action to perform in which decision situation.

It should be clear that we have to elaborate the content of the previous paragraphs in careful detail. This is what we shall do later on in this section. 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 do which action entertained by his strategy. This is the first way for apparently rebutting the "no probabilities for acts" principle.

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. towards his own actions, if he can very well do so towards the actions of other persons? One

should rather think that he is particularly endowed in his own case because he has so much more data about himself than about anybody else.

As long as no good response to this question is in sight, we might rather try to say how the agent should 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 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, and so on. All this behavior may, or may not, be under the rational control of the agent. If it is, as it 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 the habit or convention is confirmed by rational control. This means, in turn, that the prediction of the behavior is really based on the prediction of the (tacit or explicit) rational deliberation leading to it. If the 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.

To conclude, the agent should predict and explain his actions at the future action nodes as intentional and rational action 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 in the decision model, we should enrich it with 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 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 doubly knit 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 the previous section. And it would be premature, by all means. Before jumping to conclusions, we should rather scrutinize how decision models that include decision nodes do really look like. I have already sketched such *reflective decision models*, as I would like to call them since they model how the agent reflects on his own future attitudes; but the sketch needs to be worked out.

Here is my proposal in form of an extended, annotated definition: $\hat{\delta} = \langle U, \rightarrow, H, D, p, u \rangle$ is a *reflective decision model*, if 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 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 action 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, So, (2) is the minimum to be 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 is temporarily immediately preceded by its parental decision node.[28] One might also wonder how a decision node may have other than action nodes as chil-

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

dren. It will be crucial for my argument in the next section to reject all such strengthenings of (2). Therefore I shall defer the 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 receive a utility by its own. I have argued for the point in Spohn (1999, pp. 49ff.). But since it does not play any role here, I shall not further dwell upon it.

(4) *p* is a probability measure for *U*.

This reflects the point that there seems to be no restriction to 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 about the self-localization of the agent in the reflective decision model $\hat{\delta}$. 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 the agent is, as it were, *in* the first decision node. That is, the time at which the agent has 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 is. 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 $p(\delta_0) = 1$.

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 reflective model $\hat{\delta}$ 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.[29] The second response is that we face here a general difficulty. Whenever one models states of reflection, the object of reflection cannot be understood as the whole reflective state itself.[30] The embarrassment is thus a common one.

Nevertheless, I owe an account of what δ_0 is, if not the whole of $\hat{\delta}$. It is not the basic submodel resulting from the full reflective 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:

Let first, for any decision node $\Delta \in D$, $Ac(\Delta)$ denote the set of *action children* of Δ (which is not empty according to condition (2)) and $Oc(\Delta)$ denote the set of *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 $\{\Delta_0\}$ and then truncating this reduction with respect to $Ac(\Delta_0)$, where – this is important – all members of $Ac(\Delta_0)$ are treated in the reduction as temporally preceding all members of $Oc(\Delta_0)$. So 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 $\hat{\delta}$ by $\{\Delta_0\}$ (where g runs through $\times Ac(\Delta_0)$).

We should make clear to us what this really 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 by adding arrows from all $A \in Ac(\Delta_0)$ and all $B \in Pa(\Delta_0)$ to all $C \in Oc(\Delta_0)$, provided $Oc(\Delta_0)$ is not empty. The action nodes in $Ac(\Delta_0)$ are thereby turned into exogenous variables, and the other children of Δ_0, if there are any, are thereby made directly causally dependent on all the parents *and* all the action children of Δ_0. This may appear not entirely intelligible, and the first

---

[29] 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) and in Spohn (1978, sect. 2.3 and 3.6) I have elaborated on their theoretical importance.

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

mystery may be how at all a decision node may have any causal influence not mediated by action nodes. These things will become clearer when we shall consider specific examples with non-empty $Oc(\Delta_0)$ in the next section. But recall our observation in the previous section that in a reduced causal graph an arrow $A \rightarrow B$ generally signifies only that B causally depends on A *or* that A and B have a common cause; and this holds as well for all the arrows from $Ac(\Delta_0)$ to $Oc(\Delta_0)$. This also explains the strange condition that the members of $Ac(\Delta_0)$ are to be treated as preceding the members of $Oc(\Delta_0)$: if the reduction would create arrows running the other way from $Oc(\Delta_0)$ to $Ac(\Delta_0)$, these arrows would be lost in the subsequent truncation, and so the existence of the common cause of $Ac(\Delta_0)$ and $Oc(\Delta_0)$ would not show up any more in the truncated graph.

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 reflective 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 the previous section guarantee then that the remaining family (p_g) agrees with the reduced and truncated graph.

There is a familiar problem about conditionalization here. The relation between decision nodes and the appertaining action nodes will be esentially deterministic. Hence, if $p(\delta_0) = 1$, then $p(g^*) = 1$ for the action (course) g^* that is optimal in δ_0 and $p(g) = 0$ for all other actions g so that p_g remains undefined for them. But the problem is familiar from section 1, and it may be corrected with similar means. I neglect it here in order to avoid further complications.

The upshot of all this is that δ_0 contains the same decision relevant items as the reflective model $\hat{\delta}$ we are about to define, and indeed all of them; the surplus of the reflective model is only the agent's firm belief that he *is* in δ_0 and its consequences. In this way, the circularity problem that plagued our modeling of reflective states is solved.

However, I would like to 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 reflective model does not do any work in this respect – though, of course, we can positively assert this only after the

decision rule determining optimal actions has been introduced.[31] I shall discuss the consequences of this observation at the end of this section.

This lengthy discussion of the precise structure of δ_0 considerably facilitates the next point. So far, I have said nothing about the values of all the other decision variables. But surely, these values must observe structural restrictions; they must be suitable decision models. So, what can we say about an arbitrary value δ of an arbitrary decision variable $\Delta \in D$? More or less the same as about $\delta_0 \in \Delta_0$:

Again, I assume that $\delta \in \Delta$ contains the truncated reduction of $\langle U, \rightarrow \rangle$ by Δ as a causal graph. This means that the agent believes that he will have the same causal picture in δ as he has now. One might here conceive of even more generality, but I shall not do so. The action nodes in δ are only those in $H_\delta = \bigcup\{Ac(\Delta') \mid \Delta' = \Delta \text{ or } \Delta \in Pr(\Delta')\}$, i.e. the children of Δ or of later decision situations. This is due to the time-dependency of the notion of an action node; action variables are only action variables still to be decided upon and thus always in the future of the decision situation they are part of. The beliefs and desires of δ_0 had, of course, to agree with those of $\hat{\delta}$. But nothing of this sort can and should be assumed about the future decision situation δ; the agent can envisage arbitrary changes of probabilities and utilities. Of course, theoretical work only becomes substantial, when various specific forms of change are considered.[32] Conceptually, however, one should allow for all kinds of changes. The only restriction is that the probabilities in δ have to agree with the causal graph in δ. All this is summarized in condition (7):

(7) For each $\Delta \in D$ and $\delta \in \Delta$ $\delta = \langle U \setminus \{\Delta\}, \rightarrow^*, H_\delta, D_\delta, (p_g^\delta), u_\delta \rangle$, where:
$\langle U \setminus \{\Delta\}, \rightarrow^* \rangle$ is the truncated reduction of $\langle U, \rightarrow \rangle$ by $\{\Delta\}$,
$H_\delta = \bigcup\{Ac(\Delta') \mid \Delta' = \Delta \text{ or } \Delta \in Pr(\Delta')\}$ as specified above,
$D_\delta = \{\Delta' \in D \mid \Delta \in Pr(\Delta')\}$,
$(p_g^\delta)_{g \in \times Ac(\Delta)}$ is a family of probability measures for $U \setminus (\{\Delta\} \cup Ac(\Delta))$ agreeing with the causal graph $\langle U \setminus \{\Delta\}, \rightarrow^* \rangle$, and
u_δ is a utility function from $\times (U \setminus D)$ into $\boldsymbol{R}$.

This also holds for the other values of Δ_0 besides δ_0. One should note, however, that these other values of Δ_0 are not decision relevant, since they are only counter-

---

[31] This has been one of my two arguments for this principle in Spohn (1977, sect. 2) the one which Rabinowicz (1999) considers to be the stronger one.

[32] In Spohn (1999, sect. 4) I have tried to start this work.

factual possibilities for the agent; he knows that he is in δ_0. Their relevance is only a causal one; they demonstrate that being in δ_0 causes the agent to do what is optimal in δ_0; if he had been in another decision situation, he would have done something else. And this is, of course, something he himself believes.

All this is pretty complicated, but we have still to take the most difficult step; we have still to specify the relation between decision and action nodes to be believed by the agent. I shall, however, simplify things because the intricacies involved in this step do not really matter here. In order to specify this relation we have to say what the rational or optimal actions are in all the situations contained in the decision nodes. Since the agent believes to be and to stay rational, he believes that these are the actions he will do in these situations (otherwise it would be wrong, in the first place, to consider decision nodes as parents of action nodes).

So, what is the decision rule for all the situations appearing in the reflective model $\hat{\delta}$? One difficulty is that we cannot say this for all situations at once; we have rather to apply a recursive procedure, i.e. backward induction. Look first at the last decision node Δ_n. Each situation $\delta \in \Delta_n$ is free of further decision nodes; the strategic horizon does not extend further. Hence, each such δ is a basic decision model, and it is rational to choose an action that maximizes conditional expected utility in δ. In other words, if $g \in \bigtimes Ac(\Delta_n)$ is irrational, i.e., does not maximize conditional expected utility in δ, then $p(g \mid \delta) = 0$. This does not include a specific recomendation for a tie breaker, for what to believe if there are several equally optimal actions. Rightly so, I think.

Next, assume we already know what is rational at the $n - k$ latest decision nodes $\Delta_{k+1},\ldots,\Delta_n$, and consider any situation $\delta \in \Delta_k$. The agent has a clear prediction what he will do in any of the situations in $\Delta_{k+1},\ldots,\Delta_n$. Together with any choice g from $\bigtimes Ac(\Delta_k)$ this yields a dependency scheme, indeed a feasible strategy, s_g for δ, which has an expected utility $Eu_\delta(s_g)$ from the point of view of δ. The rational thing to do in δ is then to realize some g for which this value is maximal among all such feasible strategies. In this way, we can work backwards until we have reached Δ_0 and are thus able to say which choice from $\bigtimes Ac(\Delta_0)$ is rational in δ_0 or, what comes to the same, in $\hat{\delta}$. Let me summarize this by the condition (8), which is my last one:

(8) For any decision node δ and any situation $\delta \in \Delta$, if $g \in \bigtimes Ac(\Delta)$ is irrational in δ according to the explanation just given, then $p(g \mid \delta) = 0$.

This sounds good, and for our present purposes it is more than good enough (because we can neglect the recursive complications in the next sections). But really it is not good enough. What I have explained so far 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![33] However, sophisticated choice needs more sophistication. Since we made no restrictive 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 mesh in much more complicated ways than I have just explained. At least I have argued so and proposed an improved decision rule sketchily in Spohn (2000, sect. 4) and in formal detail in Spohn (1999, sect. 3).

This finishes my construction of the reflective decision model $\hat{\delta}$. It has become quite involved, I admit. But decision nodes have a rich internal structure and are richly connected within the graph. So these are unavoidable costs of reflection. Every item of the construction is significant and consequential; let me therefore briefly recapitulate which constructive choices we have faced:

I have mentioned several possible strengthenings of condition (2) and have promised to discuss them in the next section. One may consider a more embracive domain for the utility function in condition (3) (though I think one should not). One may doubt the reflective knowledge embodied in condition (5). One may also doubt the particular construction of the situation in the decision nodes in conditions (6) and (7) with its reintroduction of the "no probabilities for acts" principle. As I have pointed put, however, this is related to the particular shape of the decision rule in condition (8) which I have stated only preliminarily and which would require much more scrutiny. So I recommend the whole structure for further investigation which I can carry out here only very incompletely.

Before going on, let us take stock where we presently stand with respect to our concern of 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, it seems to me, by introducing decision nodes; it would have been inconsistent to proclaim decision theory as a theory of

---

[33] 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).

rational action and then to resort to other theories when reflecting on the causation of actions.

What does it help, though, to understand action nodes in this way as endogenous? Nothing so far. 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 to smoke in Fisher's scenario of the smoking gene (even if one would like to smoke) or, generally, to do the 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 reflective model $\hat{\delta}$ that is governed by the same decision rule. But here the fatal consequence is blocked by self-knowledge; since the agent knows to be 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 still maintained the feature that causal independence from actions implies probabilistic independence, which we attempted to circumvent somehow in order to make causal sense of dependency equilibria.

This does not mean, however, that our progression into reflective decision theory was in vain. On the contrary, the introduction of decision nodes was crucial because it will provide us with the key to the 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. This is the basic idea for which I left room in the above weak version of condition (2) and which I want to develop in the next two sections. For this purpose, let us leave the (too) abstract fields and let us apply the structure to two apparently simple, but most significant examples, namely the toxin puzzle and Newcomb's problem. The application will turn out to be highly instructive.

5. The Toxin Puzzle and Newcomb's Problem

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, call her the predictor, approaches you and makes you an offer. She shows you a glass filled with a fluid called toxin. It tastes terrible and makes one feel kind of seasick for about three hours, but has no further 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 $ 1.000 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 which reliably establishes whether you have that intention by midnight. What you do the other day is irrelevant; you are rewarded for having the intention, not for actually drinking the toxin.

Suppose you think this might be easily earned money; three hours of sickness are well paid by $ 1.000. Otherwise, you would reject the offer. Where is the problem then? It is that it seems impossible for you to form the relevant intention and to earn $ 1.000. 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 it. So, how could you 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. Among action theorists the view is popular that actions analytically presuppose at least some intention; only mere behavior may be without intention. 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 the agent shows. 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 action nodes are characterized by themselves. Conversely, one may have the intention, as firm as possible, to drink the toxin without actually drinking it. Philosophers call that weakness of will and have their problems with this possibility. But it *is* one. Perhaps, after smelling the toxin the disgust is too big to be overcome even with the

best intention. Perhaps, one even starts drinking it, but one's throat automatically revolts, and one spits it out without control. However, even if the predictor foresees that this will happen to you, she should not deny you the intention.

On the other hand, if you leave open the slightest possibility to reconsider the case the other day (with predictable outcome), or if you even only hope that your body will unintentionally refuse the toxin, this destroys 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.

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.[34] Thus, I simply beg 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 represented by some model automatically entails knowing the optimal action according to this model.[35] With these two observations, the second translation seems justified.

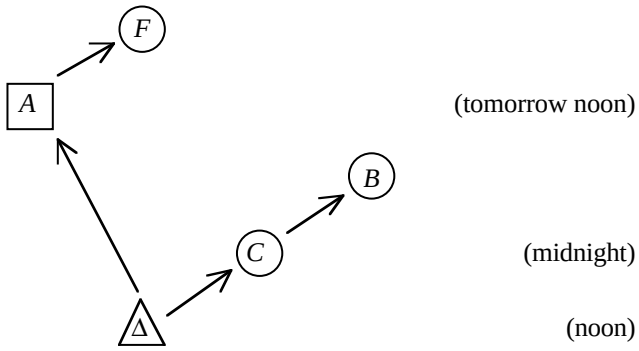
Given these translation, it is clear that the model for the toxin puzzle should contain five nodes or variables: an action node A = {drink, do not drink}, three occurrence nodes F = {feeling seasick for three hours, not feeling so}, C = {cerebro-

---

[34] For instance, action theorists are occupied with the locution „doing *a* with the intention *i*" 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.

[35] What lurks in the background here is the long discussion about the so-called deduction problem in doxastic logic, i.e. about the question 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 and computational theories. Semantic theories neglect the computational aspect, set rather only the normative standards for correct computations on the semantic level, and thus assume belief to be deductively closed. Computational theories find this illegitimate and try to do without this assumption. I observe that theories about propositional attitudes naturally move on the semantic level, that they are substantial theories within their idealized confines, and that computational theories are poor or equally bad idealizations. These observations hold for decision theory as well. Hence, in my view, the only successful decision theories move on the semantic level – whence my remark about the „automatic" knowledge of the optimal action.

scope is positive, it is negative} and *B* = {$ 1.000 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:



The positions of *A*, *B*, *C* and *F* are beyond dispute, but the place of Δ needs discussion. Let us first complete the reflective 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 also almost certainly negative. And the probability for being Δ being positive is also 1: its conditional expected utility is the utility of all variables being positive; the conditional expected utility of the negative decision is the utility of all variables being negative; the first value is 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 efficacy upon the cerebroscope (which is, to be sure, the phantastic part of the story). If we would look only at the action node, we would only see its causal influence on the sickness and thus conclude that not drinking is the dominant action, as is maintained by those who take it to be impossible to have the positive intention. We should therefore inspect this crucial point a bit 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 on you reconsider and decide not to drink. Well, this may be what actually happens; we should not consider it to be impossible that a person is

most firmly decided, and nevertheless reconsiders and arrives at a different decision later on. This may even be reasonable; perhaps some new fact has turned up which one would never have dreamt of and which puts the old decision into an entirely new light. However, the decision model models the perspective of the agent and not what may actually happen. In this perspective, there cannot be two decision nodes; to envisage a second decision node at which to decide about the toxin would annihilate 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 further alternatives? I see only one, namely to place the decision node Δ immediately before the action node *A*. My condition (2) above allowed 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 so tempted. I don't see why one should exclude the possibility to decide a matter very early, to maintain the decidedness till the time of action has come, and then to do what one decided to do. One may object that maintaining the decidedness is up to choice, something one may or may not do. Yes, 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 don't see why I should deny this –, so that the matter is decided only at the latest possible moment. From the agent's present point of view, by contrast, 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 this 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 between. This intention and the capability to realize it is is subsumed in game theoretic literature under the label "committing power".[36] 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 no committing power whatsoever will be needed there. In the case of Ulysses and the Sirens the required committing power

---

[36] Committing power is quite a colorful phenomenon in game theory. The relevant sense is best explained in McClennen (1990, sect. 9.7 and 12.9).

is superhuman; intention and will is powerless against the seduction of the Sirens; and knowing this Ulysses is well-advised to close the sailors' ears and to let bind himself 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 when the temptation is (too) strong, that is, when, and when not, the weakness of will is exculpable. And this vagueness can, of course, not be resolved by the cerebroscope. It is an unavoidable, not a central aspect of the example.

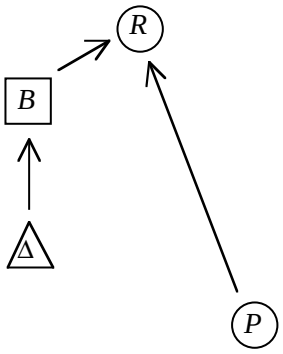
It should be clear by now that the distinction I am belaboring here is tantamount to the distinction between sophisticated and resolute choice by McClennen (1990). Sophisticated choice takes a predictive attitude towards future actions and attempts to optimize present actions given the predictions. Resolute choice, by contrast, takes an intentional or commissive attitude towards future actions and thus optimizes present and future actions at once. And like McClennen I find nothing miraculous or incredulous in that intentional attitude of resolute choice. It is most welcome that this distinction is thus reflected in my formal framework. But the way how it is reflected is even more important, namely simply by the temporal location of the decision node(s) in the reflective model. The reflective 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.

There is a final moral to be drawn from our modeling of the toxin puzzle. Obviously, it is not fixed in advance when the decision node Δ is to be located. You rather have a choice; you may decide the case right now, or you may defer the decision to tomorrow. And, of course, you choose according to where you have the higher conditional expected utility. This is in the model given above and not in the model where Δ is placed after *C*, because we assumed that you prefer \$ 1.000 to the avoidance of sickness. Thus, in reflectively modeling the toxin puzzle as above I have already presented the result of this optimization. Generally, whether the agent proceeds according to sophisticated or to resolute choice is by itself a matter of optimal choice that can be accounted for by reflective decision models. There is certainly a general theory about this kind of choice, and an utterly interesting one. But it is beyond the scope of this paper, all the more so as we have left the general decision rule for reflective decision models in the dark.

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

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 don't let go 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 take only one box she fills it with a million dollars; if she predicts your two-boxing, nothing is put into the opaque box. As I said, the prediction and the filling is already completed; any cheatful correction is excluded. The predictor is surprisingly able, and you know this. In 95 percent of the cases, say, her prediction is correct, concerning one-boxers and two-boxers alike; and very likely she is correct in your case as well. Does this story behind the opaque box change your mind? Do you now take only one box?

The majority tends to stick to two-boxing, and there is a considerable variety of theories justifying this.[37] I find the simplest emerges when we look at the natural formalization of the situation in a reflective decision model, which contains four variables; the action variable B = {one-boxing, two-boxing} two occurrence variables P = {prediction of one-boxing, prediction of two-boxing} , R (= return) = {receiving $ 0, $ 1.000, $ 1.000.000, $ 1.001.000}, and of course, the decision variable Δ governing B. The temporal, and thus the causal, relations seem clear:



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

You cannot influence the prediction because it has already taken place, and 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 also probabilistically independent of the decision and the action, and so two-boxing turns out as the dominating action and as the only rational choice.[38]

For more than twenty years I have found this consideration to be cogent. But nagging doubts have returned. It is not only that the minority of one-boxers is the personified bad conscience of the majority of two-boxers.[39] It's rather the question of Gibbard and Harper (1978): "If you're so smart, why ain't you rich?" The answer of Lewis (1981) is that the rational ones 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 can't 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 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 millions. 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 don't think so). However, despite the doubts that have been raised against backward induction it still seems to me to be a forceful argument.[40] Therefore I think it would be too weak a strategy only to try to undermine backward induction. But the ab-

---

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

[39] Jeffrey, for instance, has changed his mind several times from (1983) over (1988) to (1996).

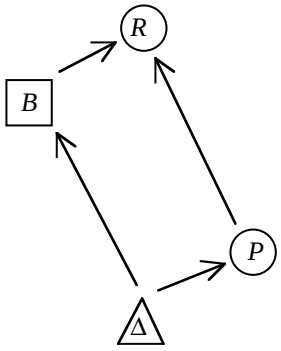
[40] The doubts were initiated by Binmore (1987), Reny (1989), Bicchieri (1989) and Pettit, Sugden (1989). Aumann (1995) defends backward induction by proving it to be valid under certain assumptions. Rabinowicz (1998) in turn makes clear the respect in which these assumptions are unreasonably strong. Rabinowicz is right, I think, but the gap he opens seems too small to harbor all the positive conclusions about cooperation I am after.

surdity is also created by the claim that only two-boxing is rational in the one-shot case. So, since this was doubtful as well, let us look a bit more carefully at the original problem.

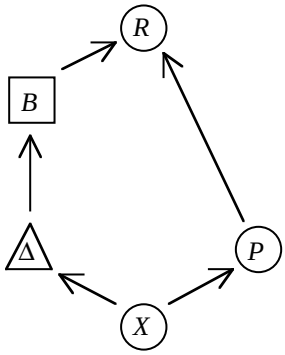
Let us start by asking what you might or should think about the predictor, if you are in this situation. No doubt you will be wondering how she manages to be so successful, and what her theory about human agents might be. On the other hand, it is clear what your own theory about yourself is. You firmly believe to be rational and to maximize conditional expected utility. That is how you fix your intentions concerning the actions to be decided now, and that 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.

Now, being rational in the sense of maximizing conditional expected utility can mean various things; there are a number of theories about this differing in details (due to Newcomb's problem, to be sure). I shall return to this. What is more important for the moment is the tension between your assuredness of your self-theory and your bewilderment about the predictor. Perhaps, the predictor has simply been lucky to be 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 success with such a theory amounts again to sheer luck. So, 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 causing the action. And there is no big mystery about this. Ideally, the predictor's 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.

Now, whatever her precise thoughts, there can only be a correlation between her prediction and your decision either if one causally depends on the other or if both have a common cause; that's what Reichenbach's principle of the common cause tells us. The case that your decision causally depends on the predictor's prediction may certainly be excluded from your point of view. Hence, two possibilities remain which are depicted in the following graphs:



(a) (decision causes prediction)



(b) (decision and prediction have a common cause)

Case (b) is how the situation has been conceived by Eells (1982, p. 211), who was the first to explicitly introduce the reflective 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 $\hat{\delta}$, 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 $\hat{\delta}$ as well.

Fisher's smoking scenario has the same structure as case (b), provided we delete the arrow from *P* to *R*. Here, *X* would be the variable to have, or not to have, the smoking gene which creates the desire to smoke and causes lung cancer, *P* would stand for having or getting lung cancer, or not, *B* would be the action variable to smoke or not to smoke, *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 non–smoking.

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 sequel. Case (a) presupposes that the decision variable Δ temporally precedes the prediction variable *P*. However, 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. But 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 can 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 for a long time and accordingly 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 getting convinced of, your rationality.

What I would like to claim is at least that case (a) presents an intentional picture you *may* have. I can't 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 upon 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) gives 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 take yourself to be decided so that now, when standing before the boxes, you simply have to carry out your old 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 hence no way to rationalize non-smoking.

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 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 that you better 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 your commitment). Or you might believe you can outsmart the predictor. And so on. Either way, you can 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). But now the picture (= case (a)) is completely reversed. Now the predictor should rationally expect you to take only one box, and if you are so rational 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 than tells that it is so in the iterated case as well. By all means, 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.

A final point: I have already abstractly explained my stipulation that in the truncated reduction of the reflective model the action childs of the deleted decision node should be treated as preceding its other childs. This creates an anomaly, as becomes obvious in our examples. There, the truncated reduction δ_0 contains an edge from *B* to the earlier *P* (Newcomb) or, respectively, from *A* to to the earlier *C* (toxin), and the probabilities in δ_0 agree with this. Strictly speaking, this requires a modification of our original definition of a causal graph. And it seems tantamount to assuming (or having the agent to assume) backwards causation. So, haven't we lapsed into causal mystery, after all? No, the mystery is only apparent and dissolves when we recognize that an arrow $B \to P$, say, in the truncated reduction δ_0 means only that *P* causally depends on *B or B* and *P* have a common cause; and of course, in our cases

only the latter interpretation can apply. Hence, rather than creating a mystery by ourselves we have an explanation for 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.

6. Prisoners' Dilemma

Let us return, after all these decision theoretic considerations, to where we started: to the game theoretic dependency equilibria and in particular to the single-shot prisoners' dilemma which we found to have a single Pareto-optimal dependency equilibrium and which we should attend for avoiding the absurdities showing in the finite iteration. Does the foregoing shed new light on prisoners' dilemma?

Hope is nourished by the long observed kinship between Newcomb's problem and prisoners' dilemma. 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 philanthrop, maybe she is a goddess, and her only interest appears to be to predict correctly.[41]

By contrast, prisoners' dilemma is a two-sided Newcomb's problem. Each player has well-defined utilities and acts according to them, and each player is both in the role of the agent and in the role 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 in Newcomb's problem. Hence, the foregoing change of view should open also a new perspective on PD.

---

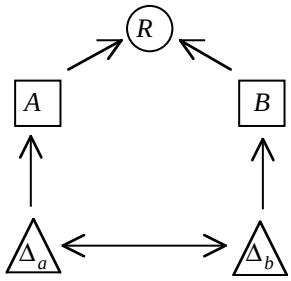
[41] Selten, Leopold (1982) have analyzed the situation as a game. With b_i = i-boxing and p_i = predicting i-boxing (i = 1,2) the most plausible utility matrix is given by

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

Here (b_2, p_2) is the only Nash equilibrium, whereas (b_1, p_1), (b_2, p_1), and (b_2, p_2) are dependency equilibria.

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 philanthropic ones of the predictor in Newcomb's problem, she would in any case choose defection as her "prediction".

But in fact the roles are not so separated. 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 to be entangled as shown in the following causal graph:



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 to be 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, I seem to have outright violated the independence constitutive of PD. But no, I think otherwise. The independence in PD was an independence between 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's 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 is here the same as in Newcomb's problem where one could take one's decision as being 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. Thereby, I do not simply refer to the trivial fact that the criminals may have communicated and coordinated long before they were imprisoned and that later on they may simply 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, because 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.

It should again be clear, however, that all this only describes how the players should rationally view their decision situations and how they should rationally decide. What goes on actually may be different. A player may doubt that the co-player(s) are fully rational in the sense explained (perhaps because 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 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 story about the iterated PD in, I find, a highly plausible and desirable way. The cornerstone of the reversal is, of course, the different view of the one-shot case which demolishes backward induction in favor of continued defection right at the beginning. Moreover, the

reversal seems to open the possibility of rationalizing the intuitively convincing and experimentally confirmed tit-for-tat strategy and similar strategies embodying kindness and responsiveness[42], a rationalization that still seems to be wanting.

I should mention that the single-shot PD is a very simple game for the theory developed here insofar as it has only one Pareto-optimal dependency equilibrium. This uniqueness was of course essential for the determinateness of my rationality claims above. Usually, the uniqueness fails, and then we get involved into the difficult problem of selecting among multiple equilibria. I have nothing to say about this problem. I only want to remark that the selection problem seems somewhat restricted insofar as it refers only to Pareto-optimal strategy combinations and that bargaining theory has to say a lot about this selection, though it remains to be checked to which extent bargaining theory can be applied to dependency equilibria.

There is a much simpler and more direct rationalization of cooperation in the one-shot PD than the one I have offered, namely via the so-called mirror principle.[43] The principle 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 complete 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 joint cooperation is for both 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 much too quick. It avoids causal considerations, and it does not really present a theory of rationality that would entail the rationality of cooperation for Ann and Bob in their particular situation. Therefore it does not really exclude that mutual defection satisfies the mirror principle as well, as the standard theory has it; it rather takes its conclusion for granted. By contrast, my account tries to back up the mirror argument by specifying the rationality theory in such a way that 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.

A final consequence of the theory put forward here seems worth to be emphasized: In standard game theory the assumption that has been called Harsanyi's doc-

---

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

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

trine plays a crucial role.[44] It says that any difference in opinion must be due to a difference in the information received; this entails in particular the common prior assumption[45] that all subjects have the same uninformed prior probabilities. At times, Harsanyi makes an even stronger claim, e.g., when he says in (1966, p. 621), that his postulates of rational expectations imply "that the *only* variables influencing the players' bargaining behavior will be: (i) the *payoffs* associated with alternative outcomes for each of the players, and (ii) the *subjective probabilities* each player assigns to different outcomes being accepted or rejected by the other player(s). Among these variables, only those mentioned under (i) are *independent* variables while the variables under (ii) are themselves determined by the variables under (i)." The effect of the doctrine is that as long as information gathering is public each player knows what the other players believe. According to the stronger claim this common knowledge of belief is indeed a consequence of the common knowledge of the game alone, i.e., of its structure and of the utility functions and the rationality of the players. This enriched common knowledge in turn yields an immediate justification of Nash equilibrium behavior. Thus, Harsanyi's doctrine may be said to mark the special status of game theory within the general theory of practical rationality.

I was always critical of Harsanyi's doctrine. As I have argued at length in Spohn (1982, sect. 5 and 6) the rationality of actions is one matter and the rationality of beliefs another, independent matter required for determining rational action, in game situations just as elsewhere. The theory of rationalizability, as I foreshadowed it there and as it was developed in detail by Bernheim (1984) and Pearce (1984), is all one can deduce about the beliefs of the players from the common knowledge of the game.

Now, however, it seems I have to revise my opinion, though in a way unforeseen by Harsanyi (and by myself). For, if my rationalization of cooperation in *PD* is sensible, it is not only the cooperative strategy, but also the special form of the players' conditional probabilities which is determined as rational by the structure of the game; they are the only probabilities which can rationally emerge from the causal interdependence of their decision situations as depicted in the above causal graph. Sometimes, rationalizability also achieves the determination of beliefs, namely whenever only the unique Nash equilibrium is rationalizable. But now I have to say that it may determine the wrong beliefs, for instance in PD where the Pareto-optimal de-

---

[44] Cf. Hargreaves Heap, Varoufakis (1995, pp. 23ff. and 75ff.).

[45] Cf. Aumann (1987, sect. 5).

pendency equilibrium is not rationalizable, as PD. Moreover, the determination of probabilities goes much farther here; at least when there is a unique Pareto-optimal dependency equilibrium my theory would prescribe rational choices *and* beliefs.

So let me resume: In section 2 I have introduced a new and possibly interesting equilibrium concept. In section 3 we have elaborated on the fact that the new equilibria seem to assume a causal dependency between the players' actions which is explicitly denied in non-cooperative game theory. Then, in a long argument, I have explained this apparent causal dependency by rising to a reflective point of view and conceiving the actions as causes of interdependent decision situations. The upshot is that dependency equilibria do not offend causal preconceptions and should thus be taken seriously in game theory.

I confess to feel a remaining uncertainty. Up to section 5, I find my account to be completely straightforward and to show a clear way how to rationally intend to drink the toxin and how to rationally take only one box. By comparison, the interdependence of the decision situation of the players in *PD* may have seemed less clear. But this relates, I think, to the reestablishment of Harsany's doctrine and the ensuing peculiarity of game situations, which, doubtlessly, deserves further scrutiny.

References

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

Aumann, R. (1987), "Corelated 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., 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.

Bernheim, B.D. (1984), "Rationalizable Strategic Behavior", *Econometrica* 52, 1007-1028.

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

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

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

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 University Press, Cambridge.

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

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

Hargreaves Heap, S.P., Y. Varoufakis (1995), *Game Theory. A Critical Introduction*, Routledge, London.

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

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

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

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

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

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

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

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

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

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*, Harvard University Press, Cambridge, Mass.

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

Pearce, D.G. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection", *Econometrica* 52, 1029-1050.

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

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

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.

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

Rabinowicz, W. (1999), "Does Deliberation Crowd Out Prediction?", Manuscript, to appear.

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 University Press, Princeton, N.J.

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

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

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

Simon, H.A. (1957), *Models of Man*, Wiley, New York.

#Sobel, J.H. (1993), “Backward-Induction Arguments: A Paradox Regained”, *Philosophy of Science* 60, 114-133.

Sobel, J.H. (1994), *Taking Chances: Essays on Rational Choice*, Cambridge University Press, Cambridge.

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

Spirtes, P., C. Glymour, R. Scheines (1993), *Causation, Prediction, and Search*, Springer, Berlin.

Spohn, W. (1977), “Where Luce and Krantz Do Really Generalize Savage’s Decision Model”, *Erkenntnis* 11, 113-134.

Spohn, W. (1976/78), *Grundlagen der Entscheidungstheorie*, Ph.D. Thesis Munich 1976, published at Scriptor, Kronberg/Ts. 1978.

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, W. Spohn (eds.), *Philosophy of Economics*, Springer, Berlin, pp. 239-270.

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

Spohn, W. (1999), "Strategic Rationality", Forschungsberichte der DFG-Forschergruppe "Logik in der Philosophie", Nr. 24, Konstanz.

Spohn, W. (2000a), “A Rationalization of Cooperation in the Iterated Prisoner’s Dilemma”, in: J. Nida-Rümelin, W. Spohn (eds.), *Rationality, Rules, and Structure*, Kluwer, Dordrecht, pp. 67-84.

Spohn, W. (2000b), “Bayesian Nets Are All There Is to Causal Dependence”, in: D. Costantini, M.C. Galavotti, P. Suppes (eds.), *Stochastic Dependence and Causality*, CSLI Publications, Stanford, forthcoming.

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

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

Verma, T., J. Pearl (1990), “Causal Networks: Semantics and Expressiveness”, in: R.D. Shachter et al. (eds.), *Uncertainty in Artificial Intelligence*, vol. 4, North-Holland, Amsterdam, pp. 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.