REVERSING 30 YEARS OF DISCUSSION: WHY CAUSAL DECISION THEORISTS SHOULD ONE-BOX[*]

Wolfgang Spohn

Department of Philosophy
University of Konstanz
78457 Konstanz
Germany

Abstract: The paper will show how one may rationalize one-boxing in Newcomb's problem and drinking the toxin in the Toxin puzzle within the confines of causal decision theory by ascending to so-called reflexive decision models which reflect how actions are caused by decision situations (beliefs, desires, and intentions) represented by ordinary unreflexive decision models.

1. Introduction

Decision theorists have been causal decision theorists (CDTs) all along, I assume, not only since Savage (1954). That there is a position to take has become clear, though, only when an apparent alternative came up, namely Jeffrey's (1965) so called evidential decision theory (EDT), and a problem, namely Newcomb's problem (cf. Nozick 1969), separating the alternatives. Thus, the distinction between CDT and EDT emerged in the late 70's, most conspicuously with Gibbard and Harper (1978).

The present state of discussion is a somewhat acquiesced one, I feel. There is no commonly agreed version of CDT[1], presumably because we well enough understand (subjective) probabilities and utilities, but still not well enough causation and its relation to probabilities. However, the impression that some version of CDT is the right one is overwhelming, and so is the intuition thereby supported that two-boxing is the right thing to do in Newcomb's problem. Some uneasiness remains, even among CDTs. Still, the uneasiness has never led to a general and generally

[*] I dedicate this paper to Karel Lambert, a first version of which I have presented on the conference on occasion of his 75th birthday at UCI in April 2003. And I am indebted to the thoughtful remarks of an anonymous referee.

[1] Cf., e.g., the overview in Joyce (1999, ch. 5) or the papers collected in Campbell, Sowden (1985) and in Sobel (1994).

acceptable version of decision theory that was able to make one-boxing plausible.[2] Hence the acquiescence; the topic seems fought out and further fighting useless.

The uneasiness has a name, the title of Lewis (1981b): if you're so smart, "why ain'cha rich?" The basic answer is: what can we do, if irrationality is rewarded? And CDTs find this answer acceptable in some way or other. Since 1976 when I developed my own account of CDT and Newcomb's problem[3] I, too, was an ardent two-boxer and convinced of that answer. Since a few years, though, the answer sounds self-pitying to me and just wrong; this must be poor rationality that complains about the reward for irrationality.

In this paper I shall explain how we can keep all the insights of CDT and nevertheless rationalize one-boxing. The gist of the paper is presented in section 2; it will make clear what my almost shamefully simple plot will be. In a way I am finished then. However, since my account there will be merely graphical, i.e., in terms of graphs, I have to develop the theory behind those graphs at least as far as required. This will be less simple and the bulk of the paper in the remaining sections 3-6; the final crucial step of my argument will be presented in a precise way only at the end of section 6.

I hope my case regarding Newcomb's problem (NP) will be convincing. It will be even more convincing, I believe, regarding the Toxin puzzle (TP); fully understanding the latter will open the eyes about the former. For this reason I shall deal with both cases in parallel.

2. The Central Idea, in Graphical Terms

In Newcomb's problem (NP) you are standing before two boxes, and you may take the opaque box containing an unknown amount of money or you may take both, the opaque and the transparent box that you see to contain thousand dollars. The unknown amount of money is either nil or a million, depending on an earlier prediction of some being, the predictor, about what you will do; if the prediction is that you take only the opaque box, it contains a million dollar, and if the prediction is that you take both boxes the amount is 0. You know all this, and in particular you

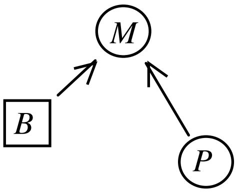
[2] Even Jeffrey changed his mind several times; cf. Jeffrey (1983, 1988, and 1996).

[3] Cf. Spohn (1976/78), ch. 3 and sect. 5.1-2. This German account was neglected; it is very close to Meek and Glymour (1994). For more detailed comparative remarks see Spohn (2001).

know that the predictor is remarkably successful (say, in predicting your actions in other situations and other persons in that situation). What will you do?

There is no point in rehearsing all the arguments for one- and for two-boxing. Standard CDT says you should two-box: your action cannot have any influence on the prediction; the content of the opaque box is fixed, whatever it is; and by two-boxing you end up with thousand dollars more in any case; two-boxing strictly dominates one-boxing. The story is represented by the following *decision graph*:

(NP1)



Here, time always moves from bottom to top, squares represent action nodes, circles represent chance nodes or occurrence nodes, as I shall rather say (since they need not be chancy; unlike the action nodes they are at most under indirect control of the agent and objects of the beliefs of the agent), and the arrows have a causal interpretation. Nodes represent variables; here, B is the action variable of one- or two-boxing, P describes the prediction, and M is the monetary outcome. The standard interpretation of $X \rightarrow Y$ is that Y (directly) causally depends on X, but we shall have to consider this more closely later on. Given this interpretation, (NP1) accurately represents the temporal and causal relations of NP from the point of view of the agent.

“Decision graph” does not seem to be an established term, although its meaning springs to one’s eyes. I shall sketch the theory of decision graphs in section 4; they are precisely what Pearl (2000, p. 23) calls mutilated graphs. Right now, two remarks might suffice.

First, decision graphs should not be confused with the familiar decision trees. In a decision tree the nodes represent events or stats of affairs, and a branch represents an entire possible course of events. Decision trees are temporarily ordered, but their edges have no causal meaning. By contrast, a decision graph is causally structured, and its nodes represent variables. Of course, it is straightforward to construct the associated tree from the graph.

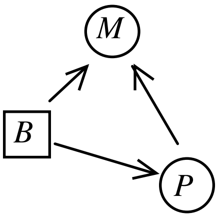
Second, decision graphs should not be confused with influence diagrams (cf. Howard, Matheson 1981). The latter contain also informational arrows that end at an action node and start from all those variables about which the agent is informed at that action node. However, this kind of information flow is not to be represented in decision graphs; it will be taken into account only in what I shall call reflexive decision graphs later on.

The theory of decision graphs makes the obvious assumption that in the situation given by (NP1) the causal independence of *P* from *B* entails its probabilistic independence from *B*. Thus, the dominant action, two-boxing, is also the one maximizing (conditional) expected utility, and hence the one recommended by CDT. Note, by the way, that the temporal relation between *B* and *P* is inessential; all we need to conclude in two-boxing is the causal independence of *P* from *B*. By adding that *P* realizes before *B*, we only make dramatically clear that *B* *cannot* have a causal influence on *P*.

What about the remarkable success of the predictor that suggests that given you one-box it is very likely that she will have predicted that you will one-box, and likewise for two-boxing? How do they enter the picture? They don't. CDTs do not deny them, but they take great pains to explain that they are not the ones to be used in practical deliberation calculating expected utilities; and they diverge in how exactly to conceive of the subjective probabilities to be used instead. I shall return to this issue in a bit more detail at the end of section 4.

Now suppose, just suppose, that the following decision graph would adequately represent NP:

(NP2)



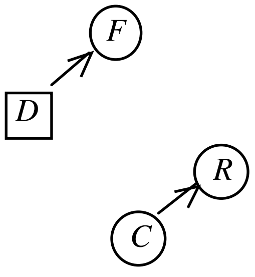
Then, I take it, it would be beyond dispute that the only reasonable thing to do is one-boxing. Nobody has ever doubted this, if NP were a case of backwards causation, as this graph seems to suggest. It is only that we have explicitly excluded backwards causation in NP!

Well, I exclude it, too; anything else would be absurd. Still, I shall defend the claim that (NP2) is an adequate representation of NP. Obviously, this can be so only if the arrows do not quite mean what they were so far told to mean. We shall see what their causal significance exactly is. For the moment I am happy with the conditional conclusion: if (NP2) should be adequate, one-boxing would indeed be rational.

In order to see what I may be up to with (NP2), let us look at the Toxin puzzle (TP) invented by Kavka (1983); it is more suitable for making my point. The story is as follows: At this evening you are approached by a weird scientist. She requests you to form the intention to drink a glass of toxin tomorrow noon. If you drink it, you will feel awful for a few hours, but then you will recover without any aftereffects. If and only if you have formed the intention by midnight, you will be rewarded with ten thousand dollars. Whether you have formed the intention can be verified by a cerebroscope the scientist has developed. The reward only depends on your intention or rather the verdict of the cerebroscope; what you actually do tomorrow noon is of no further relevance. Of course, you think that ten thousand dollars by far outweigh a few hours of sickness. But how do you get them? It is clear in advance that once you stand before the glass of toxin you have no incentive whatsoever to drink it; by then the cerebroscope has made its verdict, whatever it is. Hence, it seems you cannot honestly form that intention. You may pretend as well as you can; but this is no way to deceive the cerebroscope.

Let us again represent the situation by a decision graph. Note that a decision graph only contains action and occurrence nodes, and the only action and occurrence variables involved in the toxin story are these:

(TP1)



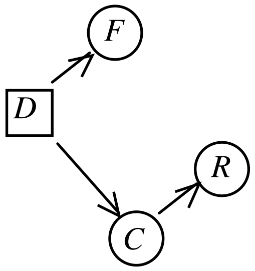
D is the action variable of drinking the toxin or not, *F* tells how you feel tomorrow afternoon, *C* is the variable for the cerebroscope reading, and *R* says whether or not you get the reward. The causal influences in the story run only in the way indicated by the arrows.

Given this representation and given again that the causal independence of *C* from *D* implies its probabilistic independence, it is clear that at *D* only not drinking the toxin maximizes (conditional) expected utility and hence that it is difficult or impossible to have the contrary intention.

One will object that (TP1) forgets about the most important variable, the intention to be formed. Yes, certainly. I shall undo this neglect in a moment. The neglect is due to the fact that a decision graph contains all the action and occurrence nodes to be considered in the represented situation for making a decision or forming an intention; but it does not contain the intention itself as a separate variable.

However, let me first make the same move as in the case of NP. Suppose, just suppose, that the decision graph adequately representing TP would be this:

(TP2)

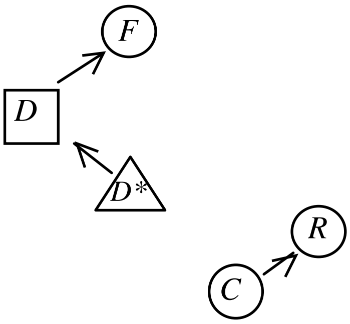


Then, again, drinking the toxin would obviously and uncontestedly be the rational action maximizing conditional expected utility. The only mystery, again, is the arrow from *D* to *C*, since I have not yet explained how to avoid the absurd and unwanted interpretation of this arrow as backwards causation.

The mystery dissolves when we undo the neglect already observed and explicitly introduce the intention as a separate variable. That is, we shall now distinguish a third kind of node, intentional nodes or rather decision nodes or variables that realize in entire decision situations, mental complexes of beliefs and desires, focusing or concluding in a decision or intention to act in a certain way; I shall represent such decision nodes by triangles. Most of the literature is quite sloppy at this point and refer to the square action nodes also as choice or decision nodes. Thus the present distinction, that is crucial for this paper, is blurred right from the beginning and cannot be reimported into the picture. So, how should we represent the causal situation of the toxin story with these richer means?

This is not entirely clear. One idea is that, willy-nilly, you take the final decision only tomorrow noon when you stand before the glass of toxin. This yields the following causal diagram or reflexive decision graph, as I shall call it for reasons to become clear soon:

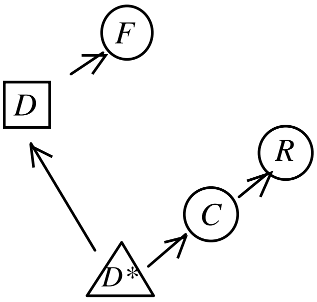
(TP1*)



Here, *D** is the decision variable realizing as this or that mental complex deciding about drinking or not drinking the toxin. Again, as in (TP1), it seems that the only thing you can rationally decide in *D** is not to drink the toxin; the causal picture shows that you did not manage to be decided before midnight, and the late decision can only be to choose the dominant act, not drinking, whether or not the cerebroscope has read your mind correctly.

However, (TP2*) does not represent what the scientist asked you for. He asked you for being decided before midnight. Thus, the relevant causal diagram should be rather this:

(TP2*)



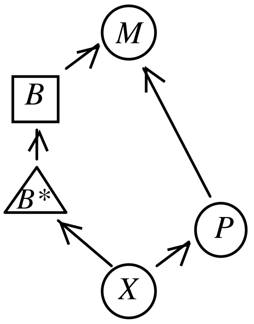
Here, you are decided or resolved in some way or other before midnight. Being decided includes the intention to maintain the decision till the time of acting has come and to refuse reconsideration, it even forbids the thought that one might perhaps reconsider the case. (In section 6 I shall say a bit more about this strong, but certainly not overly strong reading of being decided.) So, what is now your decision in D^* before midnight? Drinking the toxin, of course, since ten thousand dollars outweigh feeling sick for a few hours and since you rely on the cerebroscope. And what do you do tomorrow noon? You march into the room, take the glass, and empty it, or at least try to do so as hard as you can (your throat may revolt), without wasting any further thought.

The crucial theoretical point now is this: The starred and the unstarred decision graphs apparently represent the same decision situation. The unstarred graphs are the ordinary ones simply reflecting on the relevant action and occurrence nodes. The starred graphs additionally reflect on the ordinary decision situations deliberating on and causing the actions; this is why I shall call them reflexive decision graphs. Still, they are, reflexively and unreflexively, about the same decision. That is, we must somehow understand the unstarred, unreflexive graphs as *reduced* versions of the starred, reflexive ones. Obviously, (TP1) is the reduced version of (TP1*). Similarly, we must understand (TP2) as a reduced version of (TP2*). (TP2*) adequately represents the causal relations given an early decision; there D^* is a common cause both of C and D. What remains of that causal configuration when D^* is reduced away to yield (T2)? A shadow arrow, as it were, from D to C, which does not stand for backwards causation, as it first appeared, but is the residue of an implicit common cause relation. This is the truth about (TP2).

Of course, I am speaking only figuratively so far. I have not explained the theory behind the starred graphs, and I have not explained the exact relation between the starred and the unstarred graphs. This is the task of the next sections; in particular, section 6 will make precise sense of what I just called a "shadow arrow". Still, my suggestion should be clear without these explanations. It is that (TP2*) is causally entirely straight, that (TP2) only mirrors (TP2*), that both rationalize drinking the glass of toxin, and that this is how you can gain ten thousand dollars.

Let us return to Newcomb's problem and take the same reflexive move. In fact, that move is already found in Eells (1982, ch. 6-8). Like everybody else, Eells was wondering about the strong correlation between the predictor's prediction and your or everybody's behavior. If neither the prediction causes the behavior nor the behavior the prediction, as it was constitutive of the story, then the only explanation of the correlation is a common cause; this is what Reichenbach's common cause principle tells us. It may be unclear what the common cause is, but it must exist. So, let us call it variable X. How does X causally influence your behavior? Since your intention, or your entire decision situation or cognitive-motivational complex concluding in your intention, is a complete cause of your action, *X*'s influence can directly affect only your decision situation and only indirectly your behavior. Thus, the causal picture Eells draws is this:

(NP1*)

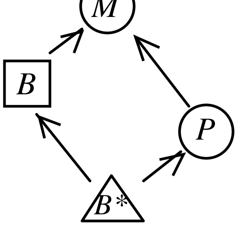


Here, *B** is the variable for the intention to one- or two-box, or rather for the decision situation concluding in such an intention.

What is rational according to (NP1*)? Only two-boxing, as Eells convincingly argues. He argues that whatever the indirect influence of *X* on *B* might be, it is screened off by your self-awareness of *B**. Given you know what decision situation you are in there is no longer any correlation between *B* and *X* and thus none between *B* and *P*, and thus only two-boxing maximizes (conditional) expected utility. In other words, (NP1*) is just a reflexive version of (NP1); if we reduce (NP1*) by *B** (and *X*), we return to (NP1).

However, the common cause need not be so obscure. What, if the decision situation itself is the common cause? This is my shamefully simple suggestion. Then we get the following reflexive decision graph:

(NP2*)



It is obvious that in the situation thus presented one-boxing is rational. If my decision determines or strongly influences the prediction, then I rationally decide to one-box, and when standing before the boxes I just do this. The situation is not at all different from the Toxin case. Again, (NP2) is simply the reduced version of (NP2*), and its arrow from *B* to *P* does not express backwards causation, but only mirrors that *B**, which is only implicit (or a latent variable) in (NP2), is a common cause of both, *B* and *P*. This is how you get the million, in full agreement with CDT!

My point may be reinforced with the much simpler smoker's story. The true story is, of course, that smoking causes, or raises the risk of lung cancer; this is why you should not smoke. There is a well-known alternative story. Smoking as such is harmless. There is, however, a gene, the smoker's gene, that probabilistically causes both, lung cancer and the desire to smoke. Should you smoke, then, if you have the desire? Even though you are likely then to have the gene and to develop lung cancer? Yes, of course, CDT says; there is no use whatsoever in abstaining from smoking in that scenario. It corresponds to (NP1*). Now consider a third variant. There, perversely, the desire to smoke itself causes lung cancer *not* via smoking, but via some other causal route; thus, the desire is a common cause of smoking and cancer. The scenario corresponds to (NP2*). Knowing these causal relations, should you smoke? Certainly not. Given these side effects, you should not even have the desire to smoke and hence not smoke; and if you are rational, this is what you do and what your desires are, judged again from the stance of CDT.

Let me emphasize why I claim still to move within the confines of CDT. The borderline between CDT and EDT is not entirely clear; each version of CDT and of EDT would draw it in a slightly different way. In any case, two-boxing in NP cannot be a defining characteristic of CDT; otherwise, Eells (1982) could not have been classified as a version of EDT. Roughly, one might say that EDTs take evidential relations between actions and other occurrences to be relevant for rational decisions (where these evidential relations may be or, as many and I think, must be due to common causes). (NP2) and (TP2) do so, too; there seems no other way to interpret their strange backward arrows. Hence, when I endorse (NP2) and (TP2), I seem to adhere to EDT.

Not so! Like CDTs, I deny the general decision relevance of purely evidential relations; I do not advise the person with the smoker's gene to abstain from smoking. I grant the relevance of evidential relations only in the very special case of (TP2) and (NP2), where these relations derive from the intention itself being a common cause, i.e., are backed by reflexive decision models like (TP2*) and (NP2*). These reflexive models are in full agreement with CDT; all their arrows are ordinary causal arrows. I claimed it to be intuitively clear what the rational action is in the reflexive models, namely drinking the toxin and one-boxing (but I shall be more precise in section 6). And I said that the unreflexive (TP2) and (NP2) are not to be taken at face value, but can only be understood as reductions of the reflexive models (something to be unfolded in the subsequent sections). This is why I say that even as a CDT you should one-box.

One may object that there is a relevant difference between NP on the one hand and TP or the smoker's case on the other. In TP you were explicitly asked for an early decision, and then you may perhaps have difficulties with being really decided so early. By contrast, NP is told in such a way that you are standing before the boxes, deliberating and deciding only then; there can only be a late decision. So, (NP2*) cannot be an adequate reflexive representation of the situation.

I believe this is the misleading part of the NP story. Perhaps, you start deliberating on the matter only when standing before the boxes, because you are informed about the plot only then. This does not necessarily mean, though, that you are deciding only then. Perhaps – indeed, this is what I am suggesting – you were committed to one-box all along, and by deliberating you discover to be so committed all along. In any case, this is the only way how (NP2*) makes sense: You are decided early enough to one-box, simply by being rational, and this influences the predictor's prediction, presumably simply by his observation of your consistent and continuous rationality. (Of course, as in TP, the predictor could distinguish between your being really or only apparently committed.)

Being committed or decided all along without ever having reflected on the matter? This sounds strange, and this may be the weak part of my account of NP, but, as I would insist then, the only weak part; and it is *not* a weak part of my account of TP. On the other hand, it is not so strange perhaps. You will grant that you have many beliefs without ever having reflected on them, for instance, about the absence of ice bears in Africa. You have never posed the question to yourself; but for a long time your mind is fixed how to respond. Similarly, I trust, your introspection will reveal that often your reflection does not issue in a decision, but rather finds that you were already decided or committed. This is what I plead for in the artificial and concocted NP case.

In a way, the rest of the paper is an extended appendix; I have to explain how all these graphs exactly work: causal graphs and Bayesian nets without action nodes (in section 3), decision graphs with action nodes (in section 4), and reflexive decision graphs with decision nodes (in section 6), after motivating the reflexive move (in section 5). We shall thus move from the familiar to the less familiar. In the end, though, the relation between reflexive and unreflexive decision models should be fully clear.

3. Causal Graphs, Bayesian Nets and Their Reduction

Before entering all the decision business, we must briefly look at pure causal theorizing that has been neatly codified in the theory of so-called causal graphs and Bayesian nets.[4] It deals only with causal dependence and independence between variables. In order to do so, it must consider specific variables and not generic ones. Generic variables, say, of a sociological kind, would be annual income or social status. But it is usually very hard to say anything substantial about causal relations between generic variables. Specific variables of a sociological kind would be, e.g., my income in 2007 or my social status in 2008 (insofar they are understood as ranges of possible values the variables may take, and not as facts consisting in the values the variables actually take). Hence, the realization of specific variables is always *located* at a specific time and usually also at a specific place or in a specific object or person. Therefore, there can be causal order among specific variables.

Thus, the basic ingredient is a non-empty set U of variables that we assume to be finite; U is also called a *frame*. I shall use A, B, C, etc. for denoting single variables in U, and V, X, Y, Z, etc. for denoting sets of variables, i.e. subsets of U. We may represent each variable by the (finite) set of the possible values it may take (this presupposes that the variables are mutually disjoint sets). For $V \subseteq U$, each member of the Cartesian product $\times V$ of all the variables or sets in V is a *possible course of events within* V, a possible way how all the variables in V may realize.

Due to their specificity, the variables in U have a temporal order $<$. $A < B$ says that *A precedes B*. I assume $<$ to be a linear (not a weak) order, in order to avoid issues of simultaneous causation. Due to their specificity the variables in U also display causal order which is a partial order agreeing with the temporal order. That

[4] This theory has been started in the statistical path analysis literature since Wright (1934). In the meantime, an impressive theoretical edifice has emerged, best exemplified by Pearl (1988, 2000), Spirtes et al. (1993), and Shafer (1996). Jensen (2001) is perhaps the best readable introduction into this theory.

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 $A < B$.[5]

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

Graphs are relations visualized. Thus, we may as well say that $\langle U, \rightarrow\rangle$ is a directed acyclic graph agreeing with the temporal order[6] or, as we define it, a *causal graph*. If we neglect the various types of nodes or variables, the previous section provided various examples for causal graphs. Let me introduce some terminology we shall need:

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

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

But what exactly is the relation between causation and probability? Spirtes et al. (1993, sect. 3.4) state two crucial conditions, the causal Markov condition and the

[5] We may be content here with the usual assumption of the transitivity of causal dependence. I think, though, that the matter is more complicated. In Spohn (forthcoming, sect. 14.11-12) I argue that causation between facts or events is transitive (this is contested in the literature) and that this entails that causal dependence between variables is transitive only under special, though widely applying conditions.

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

minimality condition. In order to explain them, we need the all-important notion of conditional independence:

Let p be a probability measure for U, i.e., over the power set of $\times U$. 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: $X \perp Y \mid Z$ – iff for all $x \in \times X$, $y \in \times Y$ and $z \in \times Z$ $p(x \mid y, z) = p(x \mid z)$, i.e., if, given any complete information about Z, no information about Y teaches us anything about X (and vice versa).

Conditional probabilistic dependence is closely tied up with causal dependence according to a causal graph $\langle U, \rightarrow \rangle$. The *causal Markov condition* says that, for all $A \in U$, given the parents of A, A is independent of all other variables preceding it, or indeed of all other non-descendants – formally: that for all $A \in U$

$$A \perp Pr(A) - Pa(A) \mid Pa(A) ,$$

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

$$A \perp Nd(A) - Pa(A) \mid Pa(A).$$

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

We say that p *agrees with* the causal graph $\langle U, \rightarrow \rangle$ or that $\langle U, \rightarrow, p \rangle$ is *a Bayesian net* iff p satisfies the causal Markov and the minimality condition w.r.t. $\langle U, \rightarrow \rangle$ (cf. Pearl 1988, p. 119). In fact, in such a Bayesian net $\langle U, \rightarrow, p \rangle$ we can infer from p alone the set of parents of each variable and thus the whole causal graph agreeing with p. This was indeed my explication of direct causal dependence in probabilistic terms in Spohn (1976/78, sect. 3.3) and (1980).

The conditional independencies and dependencies characteristic of the causal Markov and the minimality condition are the basic ones entailed by the causal structure. There is, however, 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 (cf. Verma and Pearl 1990, and Pearl 1988, p. 117). Let a *path* be any connection between two nodes disregarding the directions of the arrows, i.e., any sequence $\langle A_1, ..., A_n \rangle$ of nodes such that for each $i = 1, ..., n -$

1 either $A_i \to A_{i+1}$ or $A_i \leftarrow A_{i+1}$. And let us say that a path in the graph $\langle U, \to \rangle$ is *blocked* by a set $Z \subseteq U$ of nodes (or variables) iff

(a) the path contains some chain $A \to B \to C$ or fork $A \leftarrow B \to C$ such that the middle node B is in Z, *or*

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

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

The importance of d-separation is revealed by the following *theorem*: For all X, Y, $Z \subseteq U$, if X and Y are d-separated by Z, then $X \perp Y \mid Z$ according to all measures p agreeing with $\langle U, \to \rangle$; and conversely, if X and Y are not d-separated by Z, then not $X \perp Y \mid Z$ according to almost all p agreeing with $\langle U, \to \rangle$.[7] This shows that d-separation is indeed a reliable guide for discovering conditional independencies entailed by the causal structure, and in fact all of them for almost all measures. We shall make use of this fact later on.

Spirtes et al. (1993, sect. 3.4.3) define a causal graph $\langle U, \to \rangle$ and a probability measure p for U to be *faithful* to one another iff indeed for all mutually disjoint X, Y, $Z \subseteq U$ $X \perp Y \mid Z$ w.r.t. p if and only if X and Y are d-separated by Z.[8] Thus, the second part of the theorem just stated says that almost all p agreeing with $\langle U, \to \rangle$ are faithful to $\langle U, \to \rangle$. But sometimes it is useful to exclude the exceptional cases by outright assuming faithfulness.

Spirtes et al. (1993) take their conditions connecting causality and probability only as assumptions that widely apply and then help inferring causal from probabilistic relations. In particular, they are guaranteed to apply to causally sufficient graphs that are closed under the common cause relation; i.e., if each common cause of two or more variables in the graph is represented in the graph, too. For a causally insufficient set of variables there may always be latent variables confounding the manifest causal picture. One may wonder, then, how one can ever be sure or confident that a given set is causally sufficient.

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

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

By contrast, my idea has been, as mentioned, to *define* direct causal dependence via the causal Markov and the minimality condition. I discuss this disagreement in Spohn (2001 and forthcoming, sect. 14.9); there is no point in going through this intricate issue here. Let me explain, though, which shape the problem of latent variables and causal sufficiency takes within my approach.

To begin with, it is clear that when I use Bayesian nets to define causal dependence, I can thereby define only a frame-relative notion of causal dependence. This relativization can be undone for sure only in the universal frame containing all variables whatsoever. However, this universal frame is entirely unmanageable, presumably even fictitious. So, we should rather study the causal relations relative to smaller and larger frames. This will reveal the extent to which the causal relations in the smaller frame are indicative of those in the larger frame (and thus eventually of those in the universal frame). If we extend a given frame, it is, however, difficult to say how the smaller Bayesian net develops into the larger one, simply because the probabilities can be arbitrarily extended to the richer frame. Hence, we should rather look at the reverse process; we should start with a Bayesian net on a large frame and reduce it. Then the question about the shape of the reduced Bayesian net must have a definite answer.[9] Here it is:

Let us simplify matters by focusing on minimal reductions by a single variable. Larger reductions are generated by iterating minimal reductions. So, how does a causal graph change when a node, C, is deleted from the frame U? The answer is prepared by the following definition:

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

(1) $U^r = U - \{C\}$,

(2) for all $A, B \in U^r$ $A \rightarrow^r B$ iff either $A \rightarrow B$, or not $A \rightarrow B$ and one of the following three conditions holds:

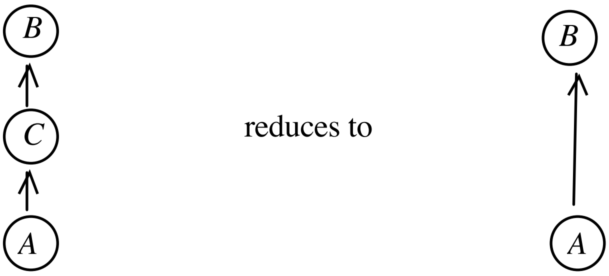
  (i) $A \rightarrow C \rightarrow B$ (call this the *IC-case*), or

  (ii) $A < B$ and $A \leftarrow C \rightarrow B$ (call this the *CC-case*), or

[9] Richardson, Spirtes (2002, 2003) have studied a similar question and developed an elaborate theory of how directed acyclic graphs with latent variables reduce to so-called maximal ancestral graphs when the latent variables are eliminated. Their study is much more ambitious and difficult insofar as they do not presuppose, as I do here, any temporal order for their graphs. Since I take arrows to express also temporal order, I am able to pursue my question within the simpler framework of directed acyclic graphs.

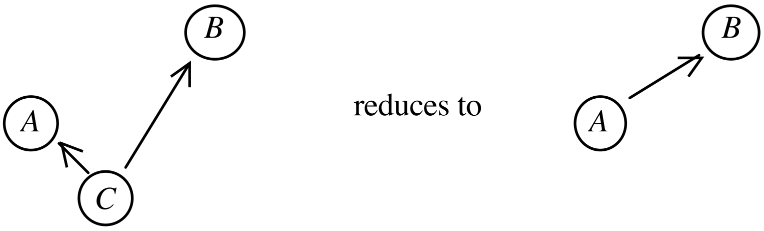
(iii) $A < B$ and there is a variable $D < B$ such that $A \rightarrow D \leftarrow C \rightarrow B$ (call this the *N-case*).

Thus, the reduced graph contains all the arrows of the unreduced graph not involving the deleted variable C. And it contains an arrow $A \rightarrow^{r} B$ where the unreduced graph contains none exactly when B is rendered *i*ndirectly *c*ausally dependent on A by the deleted C:



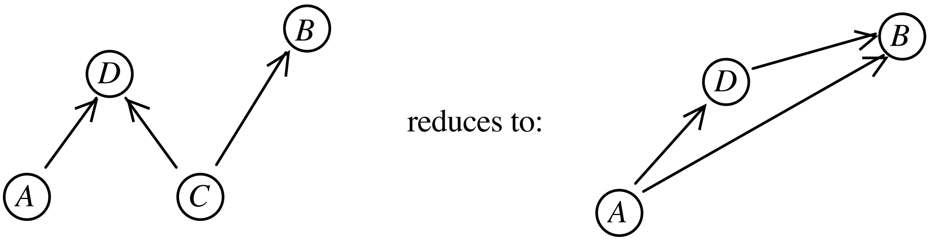
(the IC-case),

or when the deleted C is a *c*ommon *c*ause of A and B:



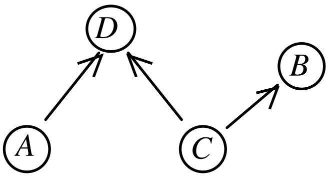
(the CC-case),

or when A is the *n*eighbor of such a CC-case involving B:

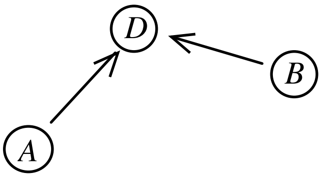


(the N-case).

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



reduces to:



The justification for this definition is provided by the following *theorem*: Let $\langle U, \rightarrow, p\rangle$ be a Bayesian net, let $\langle U^r, \rightarrow^r\rangle$ be the reduction of $\langle U, \rightarrow\rangle$ by C, and let p^r be the marginalization or restriction of p to $U^r = U - \{C\}$. Then the causal graph agreeing with p^r is a (proper or improper) subgraph of $\langle U^r, \rightarrow^r\rangle$, and if p is faithful to $\langle U, \rightarrow\rangle$, then it is $\langle U^r, \rightarrow^r\rangle$ itself which agrees with p^r. (For a proof see Spohn 2003, pp. 214f.)

In the case that p is not faithful to $\langle U, \rightarrow\rangle$, the theorem cannot be strengthened, because in that case there may hold conditional independencies not foreseen by the criterion of d-separation. Hence, d-separation may tell us $A \rightarrow^r B$, even though $A \perp B \mid Pr(B) - \{A, C\}$, which excludes a direct causal dependence of B on A relative to the reduced frame and p^r. However, if p is faithful to $\langle U, \rightarrow\rangle$, this situation cannot arise, and we have a complete answer about the behavior of reductions.[10]

As envisaged above, we should again reverse the perspective and read the theorem not as one about reductions, but as one about extensions. Our picture of the world is always limited, we always move within a small frame U^r. So, as I said above, whenever we construct a causal graph $\langle U^r, \rightarrow^r\rangle$ agreeing to our probabilities p^r, we should consider this graph as the reduction of a yet unknown, more embracive graph $\langle U, \rightarrow\rangle$ (and in the final analysis as the reduction of the universal graph). The theorem then tells us (i) that *where there is no direct causal dependence according to the small graph, there is none in the extended graph*, and (ii) that *what appears to be a direct causal dependence in the small graph may be either confirmed as such in the extended graph, or it unfolds into the IC- or the CC-case; and what appears to be causal triangle in the small graph, may resolve into the N-case*. To be precise, this is guaranteed only if the extended probabilities p are faithful to the

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

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

This observation is already half of the truth about the relation between the unstarred, unreflexive and the starred, reflexive decision graphs in section 2; for the full truth we have to dwell on the action and the decision nodes.

4. Decision Graphs and Basic Decision Models

So far, a Bayesian net describes either some small part of the world or some person's partial view of the world, a view of a detached observer having only beliefs and no interests whatsoever about that part. However, this is not the agent's view as we have to model it now. In order to accommodate it, we have to enrich our picture by adding two ingredients.

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

We might still have a mere observer, though an interested one. However, 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 the agent's direct control and take the values 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 have called them.

This is so far only the formal frame. The next important step is to see that not any structure $\langle U, \rightarrow, H, p, u\rangle$ (where $\langle U, \rightarrow, p\rangle$ is a Bayesian net and $W = U - H$) will do as a decision model; we must impose some restrictions.

A minor point to be observed here is that H does not contain all the variables in U which represent actions of the agent. Rather, H contains only the action variables still under consideration from the agent's point of view. That is, the decision model is to capture the agent's decision situation at a given time, and H contains only the action variables later than this time, 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, because the agent has no or no decision relevant beliefs or probabilities about H and because the model should contain only what is decision relevant. 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; however, this derivative belief does not play any role in forming the intention and should hence not be part of the decision model. I have stated this *"no probabilities for acts" principle* in Spohn (1977, sect. 2) since it seemed to me to be more or less explicit in all of the decision theoretic literature (cf., e.g., Fishburn 1964, pp. 36ff.) except Jeffrey's EDT (1965); Levi (1986) prominently supports this principle under the slogan "deliberation crowds out prediction". The arguments in its favor were critically examined by Rabinowicz (2002) in a most careful way. My present attitude toward the principle is a bit more sophisticated and will become clear in section 6.

It finds preliminary support, though, in the fact that it entails another widely observed principle, namely, that the action variables in H are exogenous in the graph $\langle U, \rightarrow \rangle$, i.e., uncaused or parentless. Why does this *"acts are exogenous" principle*, as I call it, follow? If the decision model is denied to contain probabilities for actions, it must not assume a probability measure p for the whole of U. Only probabilities for the occurrence variables in W can be retained, but they may, and should, be conditional on the various possible courses of action $h \in \times H$; the actions will usually matter to at least part of what occurs in W. Hence, we must replace the measure p for U by a family $(p_h)_{h \in \times H}$ of probability measures for W. Relative to such a family, Bayesian net theory still makes perfect sense; such a family may also satisfy the causal Markov and the minimality condition and may agree with, and be faithful to, a given causal graph.[11] However, it can do so only when action variables are parentless. For a variable to have parents in agreement with the probabilities, conditional probabilities for it must be explained, but this is just what the above family of measures must not do concerning action variables. Therefore, these variables cannot have parents.

Pearl (2000, ch. 3) thinks along very similar lines when he describes what he calls the *truncation* of a Bayesian net: He starts from a Bayesian net $\langle U, \rightarrow, p \rangle$. U contains a subset H of action variables. p is a measure for the whole of U and thus

[11] My definitions and theorems concerning conditional independence in Spohn (1976/78, sect. 3.2) dealt with the general case relating to such a family of probability measures. The graph theoretic material may be supplemented in a straightforward way.

represents rather an external observer's point of view. Therefore, the action variables in H so far have no special role and may have any place in the causal graph $\langle U, \rightarrow\rangle$. Now, Pearl imagines that the observer turns into an agent by becoming empowered to set the values of the variables in H according to his will so that the variables in H do not evolve naturally, as it were, but are determined through the intervention of the agent. Then Pearl asks which probabilities should guide this intervention. Not the whole of p. Rather, the intervening agent cuts off all the causal dependencies the variables in H have according to $\langle U, \rightarrow\rangle$ and puts himself into place. Hence, the agent should rather consider the *truncated* causal graph $\langle U, \rightarrow^t\rangle$ which is defined by deleting all arrows leading to action variables, i.e., $A \rightarrow^t B$ iff $A \rightarrow B$ and $B \notin H$. Thereby the action variables turn exogenous, in accordance with our principle above.

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

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

What we have to use then in deciding about a course of action in $\times H$ is the *truncated factorization* (cf. Pearl 2000, p. 72) that deletes all factors concerning the variables in H from the full factorization:

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

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

Hence, Pearl also subscribes to the two principles above.[12] The notion of truncation will be the other ingredient in understanding the relation between the starred and the unstarred decision graphs of section 2.

Let us resume our discussion so far. We may define a *decision graph* $\langle U, \rightarrow, H\rangle$ to be a causal graph $\langle U, \rightarrow\rangle$ together with a set H of exogenous nodes of U. And we may define a *basic decision model* to be a structure $\langle U, \rightarrow, H, (p_h), u\rangle$, where $\langle U, \rightarrow, H\rangle$ is a decision graph, (p_h) is a family of probability measures for $W = U - H$ agreeing with $\langle U, \rightarrow\rangle$, and u is a utility function from $\times U$ into $\boldsymbol{R}$.

What is the associated decision rule? Maximize conditional expected utility, i.e., choose a course of action $h \in \times H$ for which $\sum_{w \in \times W} u(h,w) \cdot p_h(w)$ is maximal! However, this decision rule is naive insofar as it neglects the fact that the agent need not decide for a whole course of action; rather, he needs to choose only from the (temporarily) first action variable and may wait to decide about the later ones. In other words, the naive decision rule has not taken into account strategic thinking. We shall have several reasons for undoing this neglect in section 5.

An important consequence is that 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 is a tenet characteristic of CDT. Thus, if we are to account for NP within CDT, we have to represent it, it seems, by the decision graph (NP1) and the accompanying basic decision model.[13]

I have not really argued for the two principles and thus for defining basic decision models in the above way. I have implied, though, that it is more or less what we find in most of the decision theoretic literature, and this observation may perhaps suffice. Indeed, I take the two principles as one way of saying what is constitutive of CDT. Or to put the point more cautiously:

The common aim of CDTs is to find a representation in which states or variables causally independent from the actions are also probabilistically independent and which use such probabilities for calculating the agent's expected utilities. If there are

[12] 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. The notion of intervention is crucial also for Spirtes et al. (1993, pp. 75ff.); they model it through their transition from unmanipulated to manipulated graphs, as they call it. Their procedure closely corresponds to Pearl's truncation.

[13] Meek and Glymour (1994) have further elaborated on this consequence.

probabilities correlating a causally independent variable with the actions, as there plausibly are in NP, basing expected utilities on them yields bad advice. However, CDTs then disagree about the precise form for suitably representing the agent's probabilities. Gibbard and Harper (1978) think that probabilities about conditionals (that express causal relations) are to be used instead of conditional probabilities; Lewis (1981a) endorses this idea, too. Skyrms (1984, ch. 4) sticks to ordinary conditional probabilities, but enriches the conditions by causal hypotheses. And so forth. I am not starting to argue about these ideas. Let me only say that I always had a strong preference for solutions that get along with ordinary probabilities about ordinary propositions (i.e., not containing modalities), as do Kyburg (1980), Meek and Glymour (1994), Pearl (2000), and all those working in this tradition. What I think, then, is that the two principles are constitutive for versions of CDT sharing this preference. In any case, they lead to the account of NP just sketched and to the idea that the truncated factorization instead of the full probability measure has to be used for calculating expected utilities.

However, somehow I want to stick to these principles and still to reverse the conclusions. I have intimated in section 2 that this is the consequence of ascending to the reflexive perspective. Let us see how this works in detail.

5. Strategies and Reflexion

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

Let us take up strategies first. Concerning basic decision models, I have already mentioned that it would be a naive 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 means to reverse the truncation described in the previous section. Of course, the agent does not only create such dependencies, he considers to create them in a specific way expressed by specific probabilities. This is captured in the following definition: A *dependency scheme q* for a given basic decision model is a function which specifies for each action variable $A \in H$ a probability distribution for A conditional on each realization of $Pr(A)$, i.e., of all the variables preceding A.

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

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

Again no! The problem is that not every dependency scheme represents a strategy that is feasible to the agent. 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 obviously be the fastest way to find my glasses. But it is not feasible to me; my problem is just that I do not know where I have put them. Hence, dependency schemes maximizing expected utility tell only how the agent and his actions would be optimally embedded into the causal graph according to his subjective view. Whether he is able to embed himself in such a way is another question.

This raises the following question: Which of the dependency schemes are feasible strategies that the agent is able to realize by himself? Generally, one can only say that the latter form a convex subset of the former. It is convex because any mixture of feasible strategies is feasible in turn. The reason why we cannot say more lies in the frame-relativity of dependency schemes. Prima facie one may think that all deterministic dependency schemes are feasible and that the probabilistic strategies simply result from mixing the deterministic ones (and then one should indeed think that there always are deterministic strategies among the optimal ones so that it would suffice to consider only deterministic strategies). However, there is no guarantee that the frame contains the variables that the agent is able to connect up with in a deterministic way. Perhaps the agent at best receives incomplete information about the variables included in the frame. In this case only a probabilistic dependency is within his power. Hence, as long as we do not make special assumptions about which variables are in the frame U, no more can be generally said about the feasibility of dependency schemes.

So, we should perhaps include in the frame those variables to which the agent can establish a deterministic dependence. Which are they? The answer seems clear. The agent can intentionally make his action depend only on those variables whose state he learns and thus knows with certainty before the time of action. However, no general statement seems available concerning the kind of variables the agent learns about. They must be observable, for sure; but the decline of empiricism has shown that this characterization is vague and loose. Still, there *is* a general statement: Whatever the external events are 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 is as Eells (1982, ch. 6) has conceived it.

This observation provides us with a general procedure for discovering the feasible strategies among the dependency schemes, namely by extending the causal graph of the given basic decision model by *decision nodes* as already introduced in section 2, such that each action node is preceded by a decision node, and then to define a feasible *strategy* as a dependency scheme which makes each action node depend only on its associated decision node. Obviously a decision node causally depends, in turn, on many other variables; thereby, the action node's direct intentional dependence on the decision node ramifies into various indirect dependencies (where "direct" and "indirect" is relative to the extended causal graph). Moreover, it is obvious which deterministic shape the action node's intentional dependence on the decision node should take: the relevant decision rule, say, maximizing conditional expected utility, states which action to perform in which decision situation.

It should be clear that we have to elaborate the content of the previous paragraphs in more detail. This is what we shall do in the next 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 for each action supported by his strategy with which probability he will have the opportunity to perform it. 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. toward his own actions, if he can very well do so toward the actions of others? One should indeed think that he is particularly well endowed in his own case because he has so much more data about himself than about anybody else.

Hence, the question is rather: How should the agent predict his own future behavior? There seem countless ways. The agent knows his habits ("sure, I'll brush my teeth this evening when I go to bed; that's what I always do!") or the conventions ("of course, I'll drive on the right tomorrow; everybody does!"), he knows his anxieties and the resulting behavior ("I definitely won't hike through Devil's gorge!"), and so on. These pieces of behavior may or may not be under the rational control of the agent. If they are, as is likely in the case of habits and conventions (at least in the examples given), the prediction is incomplete unless it mentions that the particular instantiation of a habit or convention is confirmed by rational control. This means, in turn, that the prediction of a piece of behavior is really based on the prediction of the (tacit or explicit) rational deliberation leading to it. If a piece of behavior is not under rational control, as it may be in the case of anxieties, 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. (Psychology and self-observation teaches that this distinction is not so clear-cut. However, for the sake of theorizing we sometimes have to paint black and white.)

To conclude, the agent should predict and explain his actions at future action nodes as intentional and rational actions with the help of decision theory, just as he explains and predicts the actions of others. Hence, if we want to make explicit these means for predicting and explaining actions within the decision model, we should extend it by decision nodes, as already envisaged. 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 double safe argument against our principles. Did we succeed to refute them? And thus to refute CDT as based on them? It would be premature to jump to conclusions. We should rather scrutinize how reflexive decision graphs and models that include decision nodes really look like.

6. Reflexive Decision Graphs and Models and their Truncated Reductions

Our discussion so far has provided all the ingredients for coping with our final task. We only have to put them together. First, the causal relations should obviously be summarized thus: $\langle U, \rightarrow, H, D\rangle$ is a *reflexive decision graph* iff (i) H, the set of *action variables*, and D, the set of *decision variables*, are disjoint subsets of U, (ii), as before, $W = U - (H \cup D)$ is the set of *occurrence variables*, and (iii) $\langle U, \rightarrow\rangle$ is a causal graph such that (iv) each action node has exactly one decision node as the only parent, i.e., for each $A \in H$ there is a $\Delta \in D$ with $Pa(A) = \{\Delta\}$, and (v) 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 discussion in section 5. For each action node it is just the parental decision node that provides the intentional or explanatory or predictive determinants of the action performed at that node. 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.

The latter condition may seem too strong. Is it not possible that one decides twice, and in different ways, about one and the same decision node? The fiancée has firmly promised and honestly intends to marry the fiancé, but after a fortnight she equally firmly does no longer want to marry that man, and we may grant her some good reasons. Should we deny her honesty and say that at least on one occasion she had no real intention? I do not think so. We may well admit that in actual fact an agent may be decided twice about one and the same decision node. However, reflexive decision graphs do not represent the perspective of the external observer; they capture the agent's subjective perspective. And in that perspective one cannot envisage to decide twice; if one envisages a second decision for the same action node, one thereby grants that the first decision is none. This minimal condition is part of our explication.

The question is rather whether the explication should be strengthened. Three further conditions suggest themselves; but it is crucial to reject all of them.

First, one might require that no two action nodes have the same parental decision node. However, this would exclude that one decides about a whole course of actions at once, and there is no need of doing so. It may be wise to fix only the first action and then to see how the situation develops; but sometimes it may be wise to fix a whole course of actions at once. Anyway, doing so should not be conceptually excluded.

Second, one might require that each action node be immediately preceded in time by its parental decision node. However, our discussion of the Toxin puzzle in section 2 clearly shows the inadequacy of this condition. Of course, I can be decided early, and in TP I get the reward only by being decided early. This is also what our everyday experience tells us. Often I go to bed with a plan in mind that I simply execute the next morning. And where to make the next vacation is something to be decided many months before. Ever so often we must, can, and do fix our plans far in advance, and thereby we take real decisions.[14]

Third, our explication allows a decision node to have other children than action nodes, but one might wonder how this is possible. Well, in TP it was explicitly assumed to be possible via the cerebroscope. This also corresponds to our everyday experience. We do not infer the attitudes of our fellows, their beliefs, desires, and intentions, merely from their actions; they are closely connected to their emotions and are thus revealed also by their mimics, gestures, and other emotional responses.[15] This is crucial for human intercourse; extremely controlled persons who allow a glimpse into their inner life only through their actions are somehow eerie. Thus, we should reckon with side effects of decision nodes. It is clear from section 2 that my entire argument stands and falls with this possibility.

In order to extend reflexive decision graphs to decision models, we only have, it seems, to add probabilities and utilities: $\hat{\delta} = \langle U, \rightarrow, H, D, p, u\rangle$ is a *reflexive decision model* iff $\langle U, \rightarrow, H, D\rangle$ is a reflexive decision graph, p is a probability measure for U, i.e., over the algebra of all subsets of $\times U$ agreeing with $\langle U, \rightarrow\rangle$, and u is a utility function from $\times U$ into $\boldsymbol{R}$.

The fact that p unrestrictedly distributes over the whole of U, in contrast to the probabilities in basic decision models, reflects the point that in the reflexive perspective there is no such restriction; the agent now has beliefs about his own actions and even about his own decision situations. However, the utility function u might be restricted to map only $\times(U - D)$ into $\boldsymbol{R}$. Arguably, decision nodes should be excluded from the utility function; being in, or getting into, this or that decision situation does not meaningfully hold any utility in itself. Let us ignore the point, though; it will not play any role here.

Still, our characterization of reflexive decision models is incomplete. We have to impose *four* further conditions. Only the first two are relevant for fully understanding what was going on in section 2; therefore I shall carefully discuss them. The other two will be only mentioned at the end; we shall see why it is not important in the present context, and indeed too difficult, to fully state them.

The first of these additional conditions concerns the self-localization of the agent in the reflexive decision model $\hat{\delta}$. A basic decision model does not contain the agent

[14] The temporal decoupling of decision and action is also an essential ingredient of Bratman's account of intention, planning, and agency; see Bratman (1999, chs. 1 - 4).

[15] Frank (1988) profoundly elaborates on this aspect of rational action.

himself; it only represents his field of deliberation. This is different with the reflexive model $\hat{\delta}$. It also contains the agent's possible or actual decision situations, and hence the question is: at which decision node is he presently situated? The answer is obvious: the agent is to decide about the first of his action nodes (and possibly later ones as well) and thus finds himself, as it were, *in* the first decision node. That is, the time at which the reflexive model characterizes the agent is the time of its first decision node.

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

(SK) 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 condition of consciousness or self-knowledge has first been stated and accepted by Eells (1982, p. 176).

This raises a problem. δ_0 is obviously to represent the present decision situation of the agent of which he is aware; on the other hand, the reflexive model $\hat{\delta}$, which we are about to characterize, does so as well. But δ_0 is only a part and not the whole of $\hat{\delta}$. How can this be? Now we have finally returned to the crucial issue of section 2, the relation between the unstarred and the starred graphs, the unreflexive basic and the reflexive models.

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.[16] The second response is that it is indeed a general difficulty that we face here. Whenever one models states of

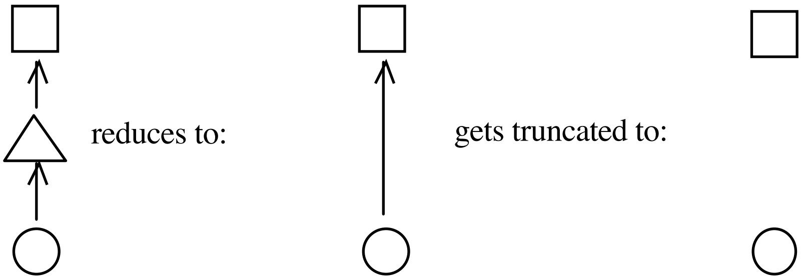
[16] I have not explicitly defined the reduction of basic decision models; but our definition of the reduction of Bayesian nets is easily extended. Such reductions are at the heart of the theory of small worlds of Savage (1954, sect. 5.5). In Spohn (1976/78, sects. 2.3 and 3.6) I have elaborated on their theoretical importance.

reflexion, the object of reflexion cannot be understood as the whole reflexive state itself.[17] Thus, the problem is a common one.

Here is an account of what δ_0 is, if it cannot be the whole of $\hat{\delta}$. δ_0 is not the basic submodel resulting from the full reflexive model by eliminating all decision nodes; it is only the first decision node Δ_0 itself that needs to be eliminated. Eliminating it from the reflexive graph means reducing the graph by it. However, this reduction need not produce a decision graph; we have to additionally truncate the reduction result with respect to the action nodes decided in δ_0. Thus, the elimination of Δ_0 results in what we might call the *truncated reduction* of $\hat{\delta}$ by Δ_0. It is fully described as follows:

For any decision node $\Delta \in D$, let $Ac(\Delta)$ denote the set of *action children* of Δ (that must not be empty according to our definition of reflexive decision graphs) and $Oc(\Delta)$ the set of *other* (occurrence or decision) *children* of Δ (that may, but need not be empty). Then, the *truncated reduction* of $\hat{\delta}$ by Δ_0 is obtained by first reducing $\hat{\delta}$ by Δ_0 not precisely in the way described in section 3, but in a slightly modified way I am about to explain and then truncating this reduction with respect to $Ac(\Delta_0)$ as described in section 4. The slightly modified way is this:

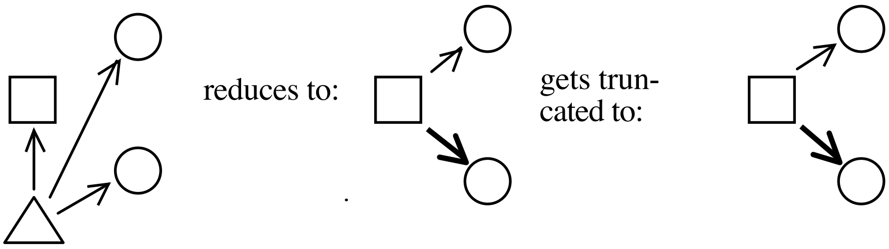
Arrows in which Δ_0 is not involved are simply maintained in the reduction as defined in section 3. Next, the reduction contains arrows from all parents of Δ_0 to all children of Δ_0; this is the IC-case. The arrows arriving at occurrence or other children will be preserved after truncation, whereas the arrows arriving at action children will fall victim to truncation:



(the IC-case)

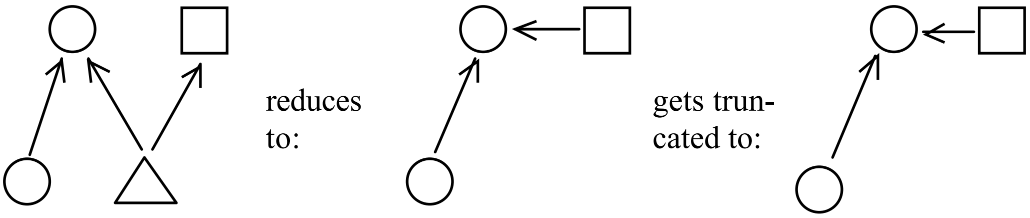
[17] This is so at least if we stick to standard ways and not resort to the model theoretic means devised by Barwise (1990) who attempts to accommodate such circular phenomena within set theory without the foundation axiom.

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



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

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



(no genuine N-case)

Hence, I propose as a second additional condition on reflexive decision models:

(TR) The basic decision model $\delta_0 = \langle U - \{\Delta_0\}, \rightarrow^{rt}, H, D - \{\Delta_0\}, (p_g), u\rangle$ is the truncated reduction of $\hat{\delta}$ by Δ_0 in the sense just defined (where g runs through $\times Ac(\Delta_0)$).

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

This may not appear intelligible because the modification of the CC-case may generate arrows running backwards from $Ac(\Delta_0)$ to earlier members of $Oc(\Delta_0)$. We certainly do not want to allow for backward causation. However, the modification does not do so. Recall our observation in section 3 that in a reduced causal graph, relative to a restricted frame, an arrow $A \rightarrow^r B$ generally signifies only that B directly causally depends on A *or* that the IC-, the CC- or the N-case applies to A and B. Here, the only case that can apply to the arrows between $Ac(\Delta_0)$ and $Oc(\Delta_0)$ is the CC-case. It is important to see that we had no choice here but to assume the anomalous backward arrows. If we had added only forward arrows in the reduction, i.e., arrows from the earlier members of $Oc(\Delta_0)$ to $Ac(\Delta_0)$ and from $Ac(\Delta_0)$ to the later members of $Oc(\Delta_0)$, then only the latter, but not the former, would have survived the truncation. But there is no reason whatsoever to treat the former and the latter arrows in the truncation in a different way; the temporal location of the members of $Oc(\Delta_0)$ is irrelevant to the causal structure of the situation and should not make any difference. Hence, the only way to uniformly and adequately retain the information about the common cause of $Ac(\Delta_0)$ and $Oc(\Delta_0)$ in the truncated graph is by adding in the reduction only arrows starting from, and not leading to, $Ac(\Delta_0)$, as it is shown in the above diagram of the CC-case. This point is the final crucial step in my explanation how CDT can and should represent NP and TP first by the reflexive decision graphs (NP2*) and (TP2*) and thus as well by the unreflexive decision graphs (NP2) and (TP2) that are the formers' truncated reductions.

(TR) does not yet fully specify the basic decision model δ_0 to which $\hat{\delta}$ reduces; it does not yet fix the probabilities and utilities of δ_0. It is clear, though, that δ_0 should contain the same utility function as $\hat{\delta}$, at least if we follow my above suggestion and assume that decision nodes do not carry utilities by themselves. And the probability family (p_g) of δ_0 is derived from the measure p of $\hat{\delta}$ by eliminating the reflexive probability of condition (SK) and all probabilities entailed by it, in particular the probabilities for the actions in $Ac(\Delta_0)$. The procedures described in sections 3 and 4 then guarantee that the remaining family (p_g) agrees with the reduced and truncated graph.

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

However, I should emphasize that thereby the "no probabilities for acts" principle has reentered the picture, if only in relation to the variables in $Ac(\Delta_0)$; the other action variables are taken care of by the later decision nodes. The reason is that δ_0, which observes this principle, contains precisely what is needed for determining or causing the optimal action. The surplus of the reflexive model has no effect in this respect.[18] In other words, the reflexive model allows for completely defined beliefs or probabilities and does therefore justice to the intuitions of those not accepting restrictions in the domain of probabilities. At the same time, it keeps the essence of the "no probabilities for acts" and the derived "acts are exogenous" principle through its reducibility to the associated basic decision model. This is how I still maintain the principles and thereby the core of CDT. Nevertheless, the conclusions as to what is rational in NP and TP have reverted in the reflexive model as well as in its truncated reduction. This finishes my main argument.

Still, I should at least indicate how my definition of reflexive decision models may and should be completed. I had announced that besides the conditions (SK) and (TR) two further conditions would be needed to completely characterize reflexive decision models. Let me sketch what is missing.

For one thing, we require a condition concerning the shape of all the decision situations δ in all the decision nodes $\Delta \in D$. This condition would not differ so much from our detailed condition (TR) on δ_0. The most important difference will be that the agent may envisage arbitrary changes of probabilities and utilities in all the possible decision situations, whereas the probabilities and utilities of δ_0 had, of course, to agree with those of $\hat{\delta}$. To be sure, theoretical work only becomes substantial by considering various specific forms of change. For example, a case that is treated extensively in decision theory is the one where the agent deliberates whether first to collect (possibly costly) evidence and then to decide on the basis of probabilities changed accordingly. But many different kinds of change may be conceived, forgetting, for instance, or what economists call endogenous preference change, and conceptually one should allow for all types of change.

A final condition is missing, indeed the most difficult and important one: the decision rule that specifies for each possible decision situation which action(s) should

[18] This has been one of my two arguments for this principle in Spohn (1977, sect. 2), the one which Rabinowicz (2002) considers to be the stronger one.

be rationally or optimally performed in it. In the reflexive context, this also determines the agent's beliefs about the relation between decision and action nodes, since he believes to be and to stay rational (otherwise, it would be wrong from the outset to consider later action nodes really as actions nodes governed by decision nodes). Hence, the final condition is:

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

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

But what is the relevant decision rule? I don't know, and I think nobody knows. It is obvious that it will be a recursive one. Each situation δ in the last decision node Δ_n is free of further decision nodes; the strategic horizon does not extend further. Hence, each such δ is a basic decision model, and the rule of maximizing conditional expected utility as envisaged in section 4 is good enough for it. Having determined optimal choices for all situations in the decision nodes $\Delta_{k+1}, \ldots, \Delta_n$, we may then continue considering all the feasible strategies for the situations in Δ_k and maximize expected utility as envisaged in section 5, and so on until we reach Δ_0. This procedure may be strictly defined. It is indeed familiar from what is called *sophisticated choice*. In fact, though, it is more general insofar as it allows deciding about shorter or longer courses of actions at once and thus contains elements of so-called *resolute choice*. Hence, this procedure seems to offer a combination of these two choice rules.[19]

However, we obviously enter here a very different and very difficult topic that is beyond the issues of CDT. Let me only point out that the problems concerning an adequate general decision rule have no impact on my treatment of NP and TP, because the reflexive models representing them contain only one decision and one action node. Then the truncated reduction generates a basic decision model with just one action node as defined in section 4, and for this model the standard rule of maximizing conditional expected utility is adequate – and indeed recommends drinking the toxin and one-boxing according (TP2) and (NP2).

[19] Sophisticated choice was first developed by Strotz (1955/56) and further elaborated and discussed, among others, by Yaari (1977), and Hammond (1976, 1988). See also the thorough discussions in McClennen (1990) who champions resolute choice besides sophisticated choice. In Spohn (2009) I argue that the issue of an adequate general decision rule is still more complicated.

References

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

Bratman, Michael E. (1999), *Faces of Intention. Selected Essays on Intention and Agency*, Cambridge, Cambridge University Press.

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

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

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

Frank, Robert H. (1988); *Passions Within Reason. The Strategic Role of the Emoitions*, New York, W. W. Norton & Company.

Gibbard, A., and 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.

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

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

Howard, R. A., J. E. Matheson (1981), "Influence Diagrams", in: R. A. Howard, J. E. Matheson (eds.), *Readings on the Principles and Applications of Decision Analysis, vol. 2*, Menlo Park, Ca.: Strategic Decisions Group, pp. 719-762.

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. (2001), *Bayesian Networks and Decision Graphs*, Springer, Berlin.

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

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

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

Levi, I. (1986), *Hard Choices. Decision Making Under Unresolved Conflict*, Cambridge University Press, Cambridge.

Leiws, D. (1981a), "Causal Decision Theory", *Australasian Journal of Philosophy* 59, 5-30.

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

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

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

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

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.

Rabinowicz, W. (2002), “Does Practical Deliberation Crowd Out Self-Prediction?”, *Erkenntnis* 57, 91-122.

Richardson, T., and P. Spirtes (2002), “Ancestral Markov Graphical Models”, *Annals of Statistics* 30, 962-1030.

Richardson, T., and P. Spirtes (2003), “Causal Inference Via Ancestral Graph Models”, in: P. Green, N. Hjort, S. Richardson (eds.), *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, pp. 83-105.

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

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

Skyrms, B. (1984), *Pragmatics and Empiricism*, Yale University Press, New haven.

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

Spirtes, P., C. Glymour, and R. Scheines (1993), *Causation, Prediction, and Search*, Springer, Berlin, 2nd ed. 2000.

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

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

Spohn, W. (1980), “Stochastic Independence, Causal Independence, and Shieldability”, *Journal of Philosophical Logic* 9, 73-99.

Spohn, W. (2001), “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, pp. 157-172.

Spohn, W. (2003), “Dependency Equilibria and the Causal Structure of Decision and Game Situations”, *Homo Oeconomicus* 20, 195-255.

Spohn, W. (2009), „Why the Received Models of Considering Preference Change Must Fail”, in: T. Grüne-Yanoff, S. O. Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, Springer, Dordrecht.

Spohn, W. (forthcoming), *Ranking Theory. A Tool for Epistemology.*

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

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