Australasian Journal of Philosophy
Vol. 59, No. 1; March 1981

CAUSAL DECISION THEORY

David Lewis

Abstract. Newcomb's problem and similar cases show the need to incorporate causal distinctions into the theory of rational decision; the usual noncausal decision theory, though simpler, does not always give the right answers. I give my own version of causal decision theory, compare it with versions offered by several other authors, and suggest that the versions have more in common than meets the eye.

1. Introduction

Decision theory in its best-known form[1] manages to steer clear of the thought that what's best to do is what the agent believes will most tend to cause good results. Causal relations and the like go unmentioned. The theory is simple, elegant, powerful, and conceptually economical. Unfortunately it is not quite right. In a class of somewhat peculiar cases, called Newcomb problems, this noncausal decision theory gives the wrong answer. It commends an irrational policy of managing the news so as to get good news about matters which you have no control over.

I am one of those who have concluded that we need an improved decision theory, more sensitive to causal distinctions. Noncausal decision theory will do when the causal relations are right for it, as they very often are, but even then the full story is causal. Several versions of causal decision theory are on the market in the works of Gibbard and Harper, Skyrms, and Sobel,[2] and I shall put forward a version of my own. But also I shall suggest that we causal decision theorists share one common idea, and differ mainly on matters of emphasis and formulation. The situation is not the chaos of disparate approaches that it may seem.

Of course there are many philosophers who understand the issues very well, and yet disagree with me about which choice in a Newcomb problem is rational. This paper is about a topic that does not arise for them. Noncausal decision theory meets their needs and they want no replacement. I will not enter into debate with them, since that debate is hopelessly deadlocked and I have nothing new to add to it. Rather, I address myself to those who join me in presupposing

[1] As presented, for instance, in Richard C. Jeffrey, *The Logic of Decision* (New York: McGraw-Hill, 1965).

[2] Allan Gibbard and William Harper, 'Counterfactuals and Two Kinds of Expected Utility', in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory*, Volume 1 (Dordrecht, Holland: D. Reidel, 1978); Brian Skyrms, 'The Role of Causal Factors in Rational Decision', in his *Causal Necessity* (New Haven: Yale University Press, 1980); and Jordan Howard Sobel, *Probability, Chance and Choice: A Theory of Rational Agency* (unpublished; presented in part at a workshop on Pragmatics and Conditionals at the University of Western Ontario in May 1978).

that Newcomb problems show the need for some sort of causal decision theory, and in asking what form that theory should take.

2. Preliminaries: Credence, Value, Options

Let us assume that a (more or less) rational agent has, at any moment, a *credence* function and a *value* function. These are defined in the first instance over single possible worlds. Each world W has a credence $C(W)$, which measures the agent's degree of belief that W is the actual world. These credences fall on a scale from zero to one, and they sum to one. Also each world W has a value $V(W)$, which measures how satisfactory it seems to the agent for W to be the actual world. These values fall on a linear scale with arbitrary zero and unit.

We may go on to define credence also for sets of worlds. We call such sets *propositions,* and we say that a proposition *holds* at just those worlds which are its members. I shall not distinguish in notation between a world W and a proposition whose sole member is W, so all that is said of propositions shall apply also to single worlds. We sum credences: for any proposition X,

$$C(X) =^{df} \Sigma_{W \in X} C(W).$$

We define conditional credences as quotients of credences, defined if the denominator is positive:

$$C(X/Y) =^{df} C(XY)/C(Y),$$

where XY is the conjunction (intersection) of the propositions X and Y. If $C(Y)$ is positive, then $C(-/Y)$, the function that assigns to any world W or proposition X the value $C(W/Y)$ or $C(X/Y)$, is itself a credence function. We say that it *comes from* C *by conditionalising on* Y. Conditionalising on one's total evidence is a rational way to learn from experience. I shall proceed on the assumption that it is the only way for a fully rational agent to learn from experience; however, nothing very important will depend on that disputed premise.

We also define (expected) value for propositions. We take credence-weighted averages of values of worlds: for any proposition X,

$$V(X) =^{df} \Sigma_W C(W/X)V(W) = \Sigma_{W \in X} C(W)V(W)/C(X).$$

A *partition* (or a *partition of* X) is a set of propositions of which exactly one holds at any world (or at any X-world). Let the variable Z range over any partition (in which case the XZ's, for fixed X and varying Z, are a partition of X). Our definitions yield the following *Rules of Additivity* for credence, and for the product of credence and expected value:

(1) $$C(X) = \Sigma_Z C(XZ),$$
$$C(X)V(X) = \Sigma_Z C(XZ)V(XZ).$$

This *Rule of Averaging* for expected values follows:

$$(2) \qquad V(X) = \Sigma_Z C(Z/X)V(XZ).$$

Thence we can get an alternative definition of expected value. For any number v, let [V = v] be the proposition that holds at just those worlds W for which $V(W)$ equals v. Call [V=v] a *value-level proposition.* Since the value-level propositions are a partition,

$$(3) \qquad V(X) = \Sigma_v C([V=v]/X)v.$$

I have idealized and oversimplified in three ways, but I think the dodged complications make no difference to whether, and how, decision theory ought to be causal. First, it seems most unlikely that any real person could store and process anything so rich in information as the C and V functions envisaged. We must perforce make do with summaries. But it is plausible that someone who really did have these functions to guide him would not be so very different from us in his conduct, apart from his supernatural prowess at logic and mathematics and *a priori* knowledge generally. Second, my formulation makes straightforward sense only under the fiction that the number of possible worlds is finite. There are two remedies. We could reformulate everything in the language of standard measure theory, or we could transfer our simpler formulations to the infinite case by invoking nonstandard summations of infinitesimal credences. Either way the technicalities would distract us, and I see little risk that the fiction of finitude will mislead us. Third, a credence function over possible worlds allows for partial beliefs about the way the world is, but not for partial beliefs about who and where and when in the world one is. Beliefs of the second sort are distinct from those of the first sort; it is important that we have them; however they are seldom very partial. To make them partial we need either an agent strangely lacking in self-knowledge, or else one who gives credence to strange worlds in which he has close duplicates. I here ignore the decision problems of such strange agents.[3]

Let us next consider the agent's options. Suppose we have a partition of propositions that distinguish worlds where the agent acts differently (he or his counterpart, as the case may be). Further, he can act at will so as to make any one of these propositions hold; but he cannot act at will so as to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. The partition gives the most detailed specifications of his present action over which he has control. Then this is the partition of the agents' alternative *options.*[4] (Henceforth I reserve the variable A to range over

[3] I consider them in 'Attitudes *De Dicto* and *De Se*', *The Philosophical Review,* 88 (1979): pp. 513-543, especially p. 534. There, however, I ignore the causal aspects of decision theory. I trust there are no further problems that would arise from merging the two topics.

[4] They are his narrowest options. Any proposition implied by one of them might be called an option for him in a broader sense, since he could act at will so as to make it hold. But when I speak of options, I shall always mean the narrowest options.

these options.) Say that the agent *realises* an option iff he acts in such a way as to make it hold. Then the business of decision theory is to say which of the agent's alternative options it would be rational for him to realise.

All this is neutral ground. Credence, value, and options figure both in noncausal and in causal decision theory, though of course they are put to somewhat different uses.

3. Noncausal Decision Theory

Noncausal decision theory needs no further apparatus. It prescribes the rule of V-maximising, according to which a rational choice is one that has the greatest expected value. An option A is *V-maximal* iff $V(A)$ is not exceeded by any $V(A')$, where A' is another option. The theory says that to act rationally is to realise some V-maximal option.

Here is the guiding intuition. How would you like to find out that A holds? Your estimate of the value of the actual world would then be $V(A)$, if you learn by conditionalising on the news that A. So you would like best to find out that the V-maximal one of the A's holds (or one of the V-maximal ones, in case of a tie). But it's in your power to find out that whichever one you like holds, by realising it. So go ahead — find out whichever you'd like best to find out! You make the news, so make the news you like best.

This seeking of good news may not seem so sensible, however, if it turns out to get in the way of seeking good results. And it does.

4. Newcomb Problems

Suppose you are offered some small good, take it or leave it. Also you may suffer some great evil, but you are convinced that whether you suffer it or not is entirely outside your control. In no way does it depend causally on what you do now. No other significant payoffs are at stake. Is it rational to take the small good? Of course, say I.

I think enough has been said already to settle that question, but there is some more to say. Suppose further that you think that some prior state, which may or may not obtain and which also is entirely outside your control, would be conducive both to your deciding to take the good and to your suffering the evil. So if you take the good, that will be evidence that the prior state does obtain and hence that you stand more chance than you might have hoped of suffering the evil. Bad news! But is that any reason not to take the good? I say not, since if the prior state obtains, there's nothing you can do about it now. In particular, you cannot make it go away by declining the good, thus acting as you would have been more likely to act if the prior state had been absent. All you accomplish is to shield yourself from the bad news. That is useless. (*Ex hypothesi,* dismay caused by the bad news is not a significant extra payoff in its own right. Neither is the exhilaration or merit of boldly facing the worst.) To decline the good lest taking it bring bad news is to play the ostrich.

The trouble with noncausal decision theory is that it commends the ostrich as rational. Let G and $-G$ respectively be the propositions that you take the small

good and that you decline it; suppose for simplicity that just these are your options. Let E and $-E$ respectively be the propositions that you suffer the evil and that you do not. Let the good contribute g to the value of a world and let the evil contribute $-e$; suppose the two to be additive, and set an arbitrary zero where both are absent. Then by Averaging,

$$(4) \quad V(-G) = C(E/-G)V(E-G) + C(-E/-G)V(-E-G) = -eC(E/-G)$$
$$V(G) = C(E/G)V(EG) + C(-E/G)V(-EG) = -eC(E/G) + g$$

That means that $-G$, declining the good, is the V-maximal option iff the difference $(C(E/G) - C(E/-G))$, which may serve as a measure of the extent to which taking the good brings bad news, exceeds the fraction g/e. And that may well be so under the circumstances considered. If it is, noncausal decision theory endorses the ostrich's useless policy of managing the news. It tells you to decline the good, though doing so does not at all tend to prevent the evil. If a theory tells you that, it stands refuted.

In Newcomb's original problem,[5] verisimilitude was sacrificed for extremity. $C(E/G)$ was close to one and $C(E/-G)$ was close to zero, so that declining the good turned out to be V-maximal by an overwhelming margin. To make it so, we have to imagine someone with the mind-boggling power to detect the entire vast combination of causal factors at some earlier time that would cause you to decline the good, in order to inflict the evil if any such combination is present. Some philosophers have refused to learn anything from such a tall story.

If our aim is to show the need for causal decision theory, however, a more moderate version of Newcomb's problem will serve as well. Even if the difference of $C(E/G)$ and $C(E/-G)$ is quite small, provided that it exceeds g/e, we have a counterexample. More moderate versions can also be more down-to-earth, as witness the medical Newcomb problems.[6] Suppose you like eating eggs, or smoking, or loafing when you might go out and run. You are convinced, contrary to popular belief, that these pleasures will do you no harm at all. (Whether you are right about this is irrelevant.) But also you think you might have some dread medical condition: a lesion of an artery, or nascent cancer, or a weak heart. If you have it, there's nothing you can do about it now and it will probably do you a lot of harm eventually. In its earlier stages, this condition is hard to detect. But you are convinced that it has some tendency, perhaps slight, to cause you to eat eggs, smoke, or loaf. So if you find yourself indulging, that is at least some evidence that you have the condition and are in

[5] Presented in Robert Nozick, 'Newcomb's Problem and Two Principles of Choice', in N. Rescher *et al.*, eds., *Essays in Honor of Carl G. Hempel* (Dordrecht, Holland: D. Reidel, 1970).

[6] Discussed in Skyrms, and Nozick, *opera cit.*; in Richard C. Jeffrey, 'Choice, Chance, and Credence', in G. H. von Wright and G. Fløistad, eds., *Philosophy of Logic* (Dordrecht, Holland: M. Nijhoff, 1980); and in Richard C. Jeffrey, 'How is it Reasonable to Base Preferences on Estimates of Chance?' in D. H. Mellor, ed., *Science Belief and Behaviour: Essays in Honour of R. B. Braithwaite* (Cambridge: Cambridge University Press, 1980). I discuss another sort of moderate and down-to-earth Newcomb problem in 'Prisoners' Dilemma is a Newcomb Problem', *Philosophy and Public Affairs*, 8 (1979): pp. 235-240.

for big trouble. But is that any reason not to indulge in harmless pleasures? The V-maximising rule says yes, if the numbers are right. I say no.

So far, I have considered pure Newcomb problems. There are also mixed problems. You may think that taking the good has some tendency to produce (or prevent) the evil, but also is a manifestation of some prior state which tends to produce the evil. Or you may be uncertain whether your situation is a Newcomb problem or not, dividing your credence between alternative hypotheses about the causal relations that prevail. These mixed cases are still more realistic, yet even they can refute noncausal decision theory.

However, no Newcomb problem, pure or mixed, can refute anything if it is not possible. The Tickle Defence of noncausal decision theory[7] questions whether Newcomb problems really can arise. It runs as follows: 'Supposedly the prior state that tends to cause the evil also tends to cause you to take the good. The dangerous lesion causes you to choose to eat eggs, or whatever. How can it do that? If you are fully rational your choices are governed entirely by your beliefs and desires so nothing can influence your choices except by influencing your beliefs and desires. But if you are fully rational, you know your own mind. If the lesion produces beliefs and desires favourable to eating eggs, you will be aware of those beliefs and desires at the outset of deliberation. So you won't have to wait until you find yourself eating eggs to get the bad news. You will have it already when you feel that tickle in the tastebuds — or whatever introspectible state it might be — that manifests your desire for eggs. Your consequent choice tells you nothing more. By the time you decide whether to eat eggs, your credence function already has been modified by the evidence of the tickle. Then $C(E/G)$ does not exceed $C(E/-G)$, their difference is zero and so does not exceed g/e, $-G$ is not V-maximal, and noncausal decision theory does not make the mistake of telling you not to eat the eggs.'

I reply that the Tickle Defence does establish that a Newcomb problem cannot arise for a fully rational agent, but that decision theory should not be limited to apply only to the fully rational agent.[8] Not so, at least, if rationality is taken to include self-knowledge. May we not ask what choice would be rational for the partly rational agent, and whether or not his partly rational methods of decision will steer him correctly? A partly rational agent may very well be in a moderate Newcomb problem, either because his choices are influenced by something besides his beliefs and desires or because he cannot quite tell the strengths of his beliefs and desires before he acts. ('How can I tell what I think

[7] Discussed in Skyrms, *op. cit.*; and most fully presented in Ellery Eells, "Causality, Utility and Decision", forthcoming in *Synthese*. Eells, argues that Newcomb problems are stopped by assumptions of rationality and self-knowledge somewhat weaker than those of the simple Tickle Defence considered here, but even those weaker assumptions seem to me unduly restrictive.

[8] In fact, it may not apply to the fully rational agent. It is hard to see how such an agent can be uncertain what he is going to choose, hence hard to see how he can be in a position to deliberate. See Richard C. Jeffrey, "A Note on the Kinematics of Preference", *Erkenntnis,* **11** (1977): 135-141. Further, the "fully rational agent" required by the Tickle Defence is, in one way, not so very rational after all. Self-knowledge is an aspect of rationality, but so is willingness to learn from experience. If the agent's introspective data make him absolutely certain of his own credences and values, as they must if the Defence is to work, then no amount of evidence that those data are untrustworthy will ever persuade him not to trust them.

till I see what I say?' — E. M. Forster.) For the dithery and the self-deceptive, no amount of *Gedankenexperimente* in decision can provide as much self-knowledge as the real thing. So even if the Tickle Defence shows that noncausal decision theory gives the right answer under powerful assumptions of rationality (whether or not for the right reasons), Newcomb problems still show that a general decision theory must be causal.

5. Utility and Dependency Hypotheses

Suppose someone knows all there is to know about how the things he cares about do and do not depend causally on his present actions. If something is beyond his control, so that it will obtain — or have a certain chance of obtaining — no matter what he does, then he knows that for certain. And if something is within his control, he knows that for certain; further, he knows the extent of his influence over it and he knows what he must do to influence it one way or another. Then there can be no Newcomb problems for him. Whatever news his actions may bring, they cannot change his mind about the likely outcomes of his alternative actions. He knew it all before.

Let us call the sort of proposition that this agent knows — a maximally specific proposition about how the things he cares about do and do not depend causally on his present actions — a *dependency hypothesis* (for that agent at that time). Since there must be some truth or other on the subject, and since the dependency hypotheses are maximally specific and cannot differ without conflicting, they comprise a partition. Exactly one of them holds at any world, and it specifies the relevant relations of causal dependence that prevail there.

It would make no difference if our know-it-all didn't really know. If he concentrates all his credence on a single dependency hypothesis, whether rightly or wrongly, then there can be no Newcomb problems for him. His actions cannot bring him news about which dependency hypothesis holds if he already is quite certain which one it is.

Within a single dependency hypothesis, so to speak, V-maximising is right. It is rational to seek good news by doing that which, according to the dependency hypothesis you believe, most tends to produce good results. That is the same as seeking good results. Failures of V-maximising appear only if, first, you are sensible enough to spread your credence over several dependency hypotheses, and second, your actions might be evidence for some dependency hypotheses and against others. That is what may enable the agent to seek good news not in the proper way, by seeking good results, but rather by doing what would be evidence for a good dependency hypothesis. That is the recipe for Newcomb problems.

What should you do if you spread your credence over several dependency hypotheses? You should consider the expected value of your options under the several hypotheses; you should weight these by the credences you attach to the hypotheses; and you should maximise the weighted average. Henceforth I reserve the variable K to range over dependency hypotheses (or over members of partitions that play a parallel role in other versions of causal decision theory). Let us define the *(expected) utility* of an option A by:

$$U(A) =^{df} \Sigma_K C(K)V(AK).$$

My version of causal decision theory prescribes the rule of U-*maximising* according to which a rational choice is one that has the greatest expected utility. Option A is U-maximal iff $U(A)$ is not exceeded by any $U(A')$, and to act rationally is to realise some U-maximal option.

In putting this forward as the rule of rational decision, of course I speak for myself; but I hope I have found a neutral formulation which fits not only my version of causal decision theory but also the versions proposed by Gibbard and Harper, Skyrms, and Sobel. There are certainly differences about the nature of dependency hypotheses; but if I am right, these are small matters compared to our common advocacy of utility maximising as just defined.

In distinguishing as I have between V and U — value and utility — I have followed the notation of Gibbard and Harper. But also I think I have followed the lead of ordinary language, in which 'utility' means much the same as 'usefulness'. Certainly the latter term is causal. Which would you call the useful action: the one that tends to produce good results? Or the one that does no good at all (or even a little harm) and yet is equally welcome because it is a sign of something else that does produce good results? (Assume again that the news is not valued for its own sake.) Surely the first — and that is the one with greater utility in my terminology, though both may have equal value.

It is essential to define utility as we did using the unconditional credences $C(K)$ of dependency hypotheses, not their conditional credence $C(K/A)$. If the two differ, any difference expresses exactly that news-bearing aspect of the options that we meant to suppress. Had we used the conditional credences, we would have arrived at nothing different from V. For the Rule of Averaging applies to any partition; and hence to the partition of dependency hypotheses, giving

$$(5) \qquad V(A) = \Sigma_K C(K/A)V(AK).$$

Let us give noncausal decision theory its due before we take leave of it. It works whenever the dependency hypotheses are probabilistically independent of the options, so that all the $C(K/A)$'s equal the corresponding $C(K)$'s. Then by (5) and the definition of U, the corresponding $V(A)$'s and $U(A)$'s also are equal. V-maximising gives the same right answers as U-maximising. The Tickle Defence seems to show that the K's must be independent of the A's for any fully rational agent. Even for partly rational agents, it seems plausible that they are at least close to independent in most realistic cases. Then indeed V-maximising works. But it works because the agent's beliefs about causal dependence are such as to make it work. It does not work for reasons which leave causal relations out of the story.

I am suggesting that we ought to undo a seeming advance in the development of decision theory. Everyone agrees that it would be ridiculous to maximise the 'expected utility' defined by

$$\Sigma_Z C(Z)V(AZ)$$

where Z ranges over just any old partition. It would lead to different answers for different partitions. For the partition of value-level propositions, for instance, it would tell us fatalistically that all options are equally good! What to do? Savage suggested, in effect, that we make the calculation with unconditional credences, but make sure to use only the right sort of partition.[9] But what sort is that? Jeffrey responded that we would do better to make the calculation with conditional credences, as in the right hand side of (2). Then we need not be selective about partitions, since we get the same answer, namely $V(A)$, for all of them. In a way, Jeffrey himself was making decision theory causal. But he did it by using probabilistic dependence as a mark of causal dependence, and unfortunately the two need not always go together. So I have thought it better to return to unconditional credences and say what sort of partition is right.

As I have formulated it, causal decision theory is causal in two different ways. The dependency hypotheses are causal in their content: they class worlds together on the basis of likenesses of causal dependence. But also the dependency hypotheses themselves are causally independent of the agent's actions. They specify his influence over other things, but over them he has no influence. (Suppose he did. Consider the dependency hypothesis which we get by taking account of the ways the agent can manipulate dependency hypotheses to enhance his control over other things. This hypothesis seems to be right no matter what he does. Then he has no influence over whether this hypothesis or another is right, contrary to our supposition that the dependency hypotheses are within his influence.) Dependency hypotheses are 'act-independent states' in a causal sense, though not necessarily in the probabilistic sense. If we say that the right sort of partition for calculating expected utility is a causally act-independent one, then the partition of dependency hypotheses qualifies. But I think it is better to say just that the right partition is the partition of dependency hypotheses, in which case the emphasis is on their causal content rather than their act-independence.

If any of the credences $C(AK)$ is zero, the rule of U-maximising falls silent. For in that case $V(AK)$ becomes an undefined sum of quotients with denominator zero, so $U(A)$ in turn is undefined and A cannot be compared in utility with the other options. Should that silence worry us? I think not, for the case ought never to arise. It may seem that it arises in the most extreme sort of Newcomb problem: suppose that taking the good is thought to make it absolutely certain that the prior state obtains and the evil will follow. Then if A is the option of taking the good and K says that the agent stands a chance of escaping the evil, $C(AK)$ is indeed zero and $U(A)$ is indeed undefined. What should you do in such an extreme Newcomb problem? V-maximise after all?

[9] Leonard J. Savage, *The Foundations of Statistics* (New York: Wiley, 1954): p. 15. The suggestion is discussed by Richard C. Jeffrey in 'Savage's Omelet', in P. Suppe and P. D. Asquith, eds., *PSA 1976,* Volume 2 (East Lansing, Michigan: Philosophy of Science Association, 1977).

No; what you should do is not be in that problem in the first place. Nothing should ever be held as certain as all that, with the possible exception of the testimony of the senses. Absolute certainty is tantamount to a firm resolve never to change your mind no matter what, and that is objectionable. However much reason you may get to think that option A will not be realised if K holds, you will not if you are rational lower $C(AK)$ quite to zero. Let it by all means get very, very small; but very, very small denominators do not make utilities go undefined.

What of the partly rational agent, whom I have no wish to ignore? Might he not rashly lower some credence $C(AK)$ all the way to zero? I am inclined to think not. What makes it so that someone has a certain credence is that its ascription to him is part of a systematic pattern of ascriptions, both to him and to others like him, both as they are and as they would have been had events gone a bit differently, that does the best job overall of rationalising behaviour.[10] I find it hard to see how the ascription of rash zeros could be part of such a best pattern. It seems that a pattern that ascribes very small positive values instead always could do just a bit better, rationalising the same behaviour without gratuitously ascribing the objectionable zeros. If I am right about this, rash zeros are one sort of irrationality that is downright impossible.[11]

6. Reformulations

The causal decision theory proposed above can be reformulated in various equivalent ways. These will give us some further understanding of the theory, and will help us in comparing it with other proposed versions of causal decision theory.

Expansions: We can apply the Rule of Averaging to expand the $V(AK)$'s that appear in our definition of expected utility. Let Z range over any partition. Then we have

$$(6) \qquad U(A) = \Sigma_K \Sigma_Z C(K)C(Z/AK)V(AKZ).$$

(If any $C(AKZ)$ is zero we may take the term for K and Z as zero, despite the fact that $V(AKZ)$ is undefined.) This seems only to make a simple thing complicated; but if the partition is well chosen, (6) may serve to express the utility of an option in terms of quantities that we find it comparatively easy to judge.

Let us call a partition *rich* iff, for every member S of that partition and for

[10] See my 'Radical Interpretation', *Synthese*, 23 (1974): pp. 331-344. I now think that discussion is too individualistic, however, in that it neglects the possibility that one might have a belief or desire entirely because the ascription of it to him is part of a systematic pattern that best rationalises the behaviour of *other* people. On this point, see my discussion of the madman in 'Mad Pain and Martian Pain', in Ned Block, ed., *Readings in Philosophy of Psychology*, Volume 1 (Cambridge, Massachusetts: Harvard University Press, 1980).

[11] Those who think that credences can easily fall to zero often seem to have in mind credences conditional on some background theory of the world which is accepted, albeit tentatively, in an all-or-nothing fashion. While I don't object to this notion, it is not what I mean by credence. As I understand the term, what is open to reconsideration does not have a credence of zero or one; these extremes are not to be embraced lightly.

every option A and dependency hypothesis K, $V(AKS)$ equals $V(AS)$. That means that the AS's describe outcomes of options so fully that the addition of a dependency hypothesis tells us no more about the features of the outcome that matter to the agent. Henceforth I reserve the variable S to range over rich partitions. Given richness of the partition, we can factor the value terms in (6) part way out, to obtain

$$(7) \qquad U(A) = \Sigma_S (\Sigma_K C(K)C(S/AK))V(AS).$$

Equation (7) for expected utility resembles equation (2) for expected value, except that the inner sum in (7) replaces the conditional credence $C(S/A)$ in the corresponding instance of (2). As we shall see, the analogy can be pushed further. Two examples of rich partitions to which (7) applies are the partition of possible worlds and the partition of value-level propositions [V=v].

Imaging: Suppose we have a function that selects, for any pair of a world W and a suitable proposition X, a probability distribution W_X. Suppose further that W_X assigns probability only to X-worlds, so that $W_X(X)$ equals one. (Hence at least the empty proposition must not be 'suitable'.) Call the function an *imaging function,* and call W_X the *image of W on X.* The image might be sharp, if W_X puts all its probability on a single world; or it might be blurred, with the probability spread over more than one world.

Given an imaging function, we can apply it to form images also of probability distributions. We sum the superimposed images of all the worlds, weighting the images by the original probabilities of their source worlds. For any pair of a probability distribution C and a suitable proposition X, we define C_X, the *image of* C *on X,* as follows. First, for any world W',

$$C_X(W') =^{df} \Sigma_W C(W) W_X(W');$$

think of $C(W) W_X(W')$ as the amount of probability that is moved from W to W' in making the image. We sum as usual: for any proposition Y,

$$C_X(Y) =^{df} \Sigma_{W \in Y} C_X(W).$$

It is easy to check that C_X also is a probability distribution; and that it assigns probability only to X-worlds, so that $C_X(X)$ equals one. Imaging is one way — conditionalising is another — to revise a given probability distribution so that all the probability is concentrated on a given proposition.[12]

[12] Sharp imaging by means of a Stalnaker selection function is discussed in my 'Probabilities of Conditionals and Conditional Probabilities', *The Philosophical Review,* **85** (1976): pp. 297-315, especially pp. 309-311. This generalisation to cover blurred imaging as well is due to Peter Gärdenfors, 'Imaging and Conditionalisation' (unpublished, 1979); a similar treatment appears in Donald Nute, *Topics in Conditional Logic* (Dordrecht, Holland: D. Reidel, 1980), Chapter 6. What is technically the same idea, otherwise motivated and under other names, appears in my 'Counterfactuals and Comparative Possibility', *Journal of Philosophical Logic,* **2** (1973): pp. 418-446, Section 8; in John L. Pollock, *Subjunctive Reasoning* (Dordrecht, Holland: D. Reidel, 1976): pp. 219-236; and in Sobel, *op. cit.* The possibility of deriving an imaging function from a partition was suggested by Brian Skyrms in discussion of a paper by Robert Stalnaker at the 1979 annual meeting of the American Philosophical Association, Eastern Division.

For our present purposes, what we want are images of the agent's credence function on his various options. The needed imaging function can be defined in terms of the partition of dependency hypotheses: let

$$W_A(W') =^{df} C(W'/AK_W)$$

for any option A and worlds W and W', where K_W is the dependency hypothesis that holds at W. In words: move the credence of world W over to the A-worlds in the same dependency hypothesis, and distribute it among those worlds in proportion to their original credence. (Here again we would be in trouble if any of the $C(AK)$'s were zero, but I think we needn't worry.) It follows from the several definitions just given that for any option A and proposition Y,

$$(8) \qquad C_A(Y) = \Sigma_K C(K)C(Y/AK).$$

The inner sum in (7) therefore turns out to be the credence, imaged on A, of S. So by (7) and (8) together,

$$(9) \qquad U(A) = \Sigma_S C_A(S)V(AS).$$

Now we have something like the Rule of Averaging for expected value, except that the partition must be rich and we must image rather than conditionalising. For the rich partition of possible worlds we have

$$(10) \qquad U(A) = \Sigma_W C_A(W)V(W).$$

which resembles the definition of expected value. For the rich partition of value-level propositions we have something resembling (3):

$$(11) \qquad U(A) = \Sigma_v C_A([V=v])v.$$

7. Primitive Imaging: Sobel

To reformulate causal decision theory in terms of imaging, I proceeded in two steps. I began with the dependency hypotheses and used them to define an imaging function; then I redefined the expected utility of an option in terms of imaging. We could omit the first step and leave the dependency hypotheses out of it. We could take the imaging function as primitive, and go on as I did to define expected utility by means of it. That is the decision theory of J. Howard Sobel, *op. cit.*

Sobel starts with the images of worlds, which he calls *world-tendencies*. (He considers images on all propositions possible relative to the given world, but for purposes of decision theory we can confine our attention to images on the agent's options.) Just as we defined C_A in terms of the W_A's, so Sobel goes on to define images of the agent's credence function. He uses these in turn to define

expected utility in the manner of (10), and he advocates maximising the utility so defined rather than expected value.

Sobel unites his decision theory with a treatment of counterfactual conditionals in terms of closest antecedent-worlds.[13] If $W_A(W')$ is positive, then we think of W' as one of the A-worlds that is in some sense closest to the world W. What might be the case if it were the case that A, from the standpoint of W, is what holds at some such closest A-world; what would be the case if A, from the standpoint of W, is what holds at all of them. Sobel's apparatus gives us quantitative counterfactuals intermediate between the mights and the woulds. We can say that if it were that A, it would be with probability p that X; meaning that $W_A(X)$ equals p, or in Sobel's terminology that X holds on a subset of the closest A-worlds whose tendencies, at W and on the supposition A, sum to p.

Though Sobel leaves the dependency hypotheses out of his decision theory, we can perhaps bring them back in. Let us say that worlds *image alike* (on the agent's options) iff, for each option, their images on that option are exactly the same. Imaging alike is an equivalence relation, so we have the partition of its equivalence classes. If we start with the dependency hypotheses and define the imaging function as I did, it is immediate that worlds image alike iff they are worlds where the same dependency hypothesis holds; so the equivalence classes turn out to be just the dependency hypotheses.

The question is whether dependency hypotheses could be brought into Sobel's theory by defining them as equivalence classes under the relation of imaging alike. Each equivalence class could be described, in Sobel's terminology, as a maximally specific proposition about the tendencies of the world on all alternative suppositions about which option the agent realises. That sounds like a dependency hypothesis to me. Sobel tells me (personal communication, 1980) that he is inclined to agree, and does regard his decision theory as causal; though it is hard to tell that from his written presentation, in which causal language very seldom appears.

If the proposal is to succeed technically, we need the following thesis: if K_W is the equivalence class of W under the relation of imaging alike (of having the same tendencies on each option) then, for any option A and world W', $W_A(W')$ equals $C(W'/AK_W)$. If so, it follows that if we start as Sobel does with the imaging function, defining the dependency hypotheses as equivalence classes, and thence define an imaging function as I did, we will get back the same imaging function that we started with. It further follows, by our results in Section 6, that expected utility calculated in my way from the defined dependency hypotheses is the same as expected utility calculated in Sobel's way from the imaging function. They must be the same, if the defined dependency hypotheses introduced into Sobel's theory are to play their proper role.

Unfortunately, the required thesis is not a part of Sobel's theory; it would be an extra constraint on the imaging function. It does seem a very plausible constraint, at least in ordinary cases. Sobel suspends judgement about imposing

[13] As in my *Counterfactuals* (Oxford: Blackwell, 1973), without the complications raised by possible infinite sequences of closer and closer antecedent-worlds.

a weaker version of the thesis (Connection Thesis 1, discussed in his Section 6.7). But his reservations, which would carry over to our version, entirely concern the extraordinary case of an agent who thinks he may somehow have foreknowledge of the outcomes of chance processes. Sobel gives no reason, and I know of none, to doubt either version of the thesis except in extraordinary cases of that sort. Then if we assume the thesis, it seems that we are only setting aside some very special cases — cases about which I, at least, have no firm views. (I think them much more problematic for decision theory than the Newcomb problems.) So far as the remaining cases are concerned, it is satisfactory to introduce defined dependency hypotheses into Sobel's theory and thereby render it equivalent to mine.

8. Factors Outside our Influence: Skyrms

Moving on to the version of causal decision theory proposed by Brian Skyrms, *op. cit.*, we find a theory that is formally just like mine. Skyrms' definition of *K-expectation* — his name for the sort of expected utility that should be maximised — is our equation (6). From that, with a trivial partition of Z's, we can immediately recover my first definition of expected utility. Skyrms introduces a partition of hypotheses — the K's which give K-expectation its name — that play just the same role in his calculation of expected utility that the dependency hypotheses play in mine. (Thus I have followed Skyrms in notation.) So the only difference, if it is a difference, is in how the K's are characterised.

Skyrms describes them at the outset as maximally specific specifications of the factors outside the agent's influence (at the time of decision) which are causally relevant to the outcome of the agent's action. He gives another characterisation later, but let us take the first one first.

I ask what Skyrms means to count as a 'factor'. Under a sufficiently broad construal, I have no objection to Skyrms' theory and I think it no different from mine. On a narrower and more literal construal, I do not think Skyrms' theory is adequate as a general theory of rational decision, though I think that in practice it will often serve. Insofar as Skyrms is serving up a general theory rather than practical rules of thumb, I think it is indeed the broad construal that he intends.

(I also ask what Skyrms means by 'relevant to the outcome'. I can't see how any factor, broadly or narrowly construed, could fail to be relevant to some aspect of the outcome. If the outcome is that I win a million dollars tomorrow, one aspect of this outcome may be that it takes place just one thousand years after some peasant felled an oak with ninety strokes of his axe. So I suppose Skyrms' intent was to include only factors relevant to those features of the outcome that the agent cares about, as opposed to those that are matters of indifference to him. That would parallel a like exclusion of matters of indifference in my definition of dependency hypotheses. In neither case is the exclusion important. Richer hypotheses, cluttered with matters of indifference, ought to give the same answers.)

On the broad construal, a 'factor' need not be the sort of localised particular

occurrence that we commonly think of as causing or being caused. It might be any matter of contingent fact whatever. It might indeed be some particular occurrence. It might be a vast dispersed pattern of occurrences throughout the universe. It might be a law of nature. It might be a dependency hypothesis. On the broad construal, Skyrms is saying only that the K's are maximally specific propositions about matters outside the agent's influence and relevant to features of the outcome that the agent cares about.

A dependency hypothesis is outside the agent's influence. It is relevant to features of the outcome that he cares about. (*Causally* relevant? — Not clear, but if we're construing 'factor' broadly, we can let that by as well.) Any specification of something outside the agent's influence is included in a dependency hypothesis — recall that they cover what doesn't depend on the agent's actions as well as what does — unless it concerns something the agent doesn't care about. I conclude that on the broad construal, Skyrms' K's are nothing else than the dependency hypotheses. In that case his theory is the same as mine.

On the narrow construal, a 'factor' must be the sort of localised occurrence — event, state, omission, etc. — that we normally think of as a cause. In the medical Newcomb problems, for instance, the lesion or the nascent cancer or the weak heart is a causal factor narrowly and literally. In motivating his theory, it is factors like these that Skyrms considers.

Our topic is rational decision according to the agent's beliefs, be they right or wrong. So it seems that we should take not the factors which really are outside his influence, but rather those he thinks are outside his influence. But what if he divides his credence between several hypotheses as to which factors are outside his influence, as well he might? Skyrms responds to this challenge by redescribing his partition of hypotheses. On his new description, each hypothesis consists of two parts: (i) a preliminary hypothesis specifying which of the relevant causal factors are outside the agent's influence, and (ii) a full specification of those factors that are outside his influence according to part (i).

That is a welcome amendment, but I think it does not go far enough. Influence is a matter of degree, so shouldn't the hypotheses say not just that the agent has some influence over a factor or none, but also how much? And if the hypothesis says that the agent has influence over a factor, shouldn't it also say which way the influence goes? Given that I can influence the temperature, do I make it cooler by turning the knob clockwise or counterclockwise? Make Skyrms' amendment and the other needed amendments, and you will have the dependency hypotheses back again.

To illustrate my point, consider an agent with eccentric beliefs. He thinks the influence of his actions ramifies but also fades, so that everything in the far future is within his influence but only a little bit. Perhaps he thinks that his actions raise and lower the chances of future occurrences, but only very slightly. Also he thinks that time is circular, so that the far future includes the present and the immediate past and indeed all of history. Then he gives all his credence to a single one of Skyrms' two-part hypotheses: the one saying that no occurrence whatever — no factor, on the narrow construal — is entirely outside

his influence. That means that on Skyrms' calculation his U(A)'s reduce to the corresponding V($A5$)'s, so V-maximising is right for him. That's wrong. Since he thinks he has very little influence over whether he has the dread lesion, his decision problem about eating eggs is very little different from that of someone who thinks the lesion is entirely outside his influence. V-maximising should come out wrong for very much the same reason in both cases.

No such difficulty threatens Skyrms' proposal broadly construed. The agent may well wonder which of the causal factors narrowly construed are within his influence, but he cannot rationally doubt that the dependency hypotheses are entirely outside it. On the broad construal, Skyrms' second description of the partition of hypotheses is a gloss on the first, not an amendment. The hypotheses already specify which of the (narrow) factors are outside the agent's influence, for that is itself a (broad) factor outside his influence. Skyrms notes this, and that is why I think it must be the broad construal that he intends. Likewise the degrees and directions of influence over (narrow) factors are themselves (broad) factors outside the agent's influence, hence already specified according to the broad construal of Skyrms' first description.

Often, to be sure, the difference between the broad and narrow construals will not matter. There may well be a correlation, holding throughout the worlds which enjoy significant credence, between dependency hypotheses and combinations of (narrow) factors outside the agent's influence. The difference between good and bad dependency hypotheses may in practice amount to the difference between absence and presence of a lesion. However, I find it rash to assume that there must always be some handy correlation to erase the difference between the broad and narrow construals. Dependency hypotheses do indeed hold in virtue of lesions and the like, but they hold also in virtue of the laws of nature. It would seem that uncertainty about dependency hypotheses might come at least partly from undertainty about the laws.

Skyrms is sympathetic, as am I,[14] to the neo-Humean thesis that every contingent truth about a world — law, dependency hypothesis, or what you will — holds somehow in virtue of that world's total history of manifest matters of particular fact. Same history, same everything. But that falls short of implying that dependency hypotheses hold just in virtue of causal factors, narrowly construed; they might hold partly in virtue of dispersed patterns of particular fact throughout history, including the future and the distant present. Further, even if we are inclined to accept the neo-Humean thesis, it still seems safer not to make it a presupposition of our decision theory. Whatever we think of the neo-Humean thesis, I conclude that Skyrms' decision theory is best taken under the broad construal of 'factor' under which his K's are the dependency hypotheses and his calculation of utility is the same as mine.[15]

[14] Although sympathetic, I have some doubts; see my 'A Subjectivist's Guide to Objective Chance', in R. C. Jeffrey, ed., *Studies in Inductive Logic and Probability,* Volume 2 (Berkeley and Los Angeles: University of California Press, 1980): pp. 290-292.

[15] The decision theory of Nancy Cartwright, 'Causal Laws and Effective Strategies', *Noûs,* 13 (1979): pp. 419-437, is, as she remarks, 'structurally identical' to Skyrms' theory for the case where value is a matter of reaching some all-or-nothing goal. However, hers is not a theory of

9. Counterfactual Dependence: Gibbard and Harper

If we want to express a dependency hypothesis in ordinary language, it is hard to avoid the use of counterfactual conditionals saying what would happen if the agent were to realise his various alternative options. Suppose that on a certain occasion I'm interested in getting Bruce to purr. I could try brushing, stroking, or leaving alone; pretend that these are my narrowest options. Bruce might purr loudly, softly, or not at all; pretend that these alternatives are a rich partition. (Those simplifying pretences are of course very far from the truth.) Much of my credence goes to the dependency hypothesis given by these three counterfactuals:

I brush Bruce $\Box\!\!\rightarrow$ he purrs loudly;
I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly;
I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr.

($\Box\!\!\rightarrow$ is used here as a sentential connective, read 'if it were that . . . it would be that . . .'. I use it also as an operator which applies to two propositions to make a proposition; context will distinguish the uses.) This hypothesis says that loud and soft purring are within my influence — they depend on what I do. It specifies the extent of my influence, namely full control. And it specifies the direction of influence, what I must do to get what. That is one dependency hypothesis. I give some of my credence to others, for instance this (rather less satisfactory) one:

I brush Bruce $\Box\!\!\rightarrow$ he doesn't purr;
I stroke Bruce $\Box\!\!\rightarrow$ he doesn't purr;
I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr.

That dependency hypothesis says that the lack of purring is outside my influence, it is causally independent of what I do. Altogether there are nine dependency hypotheses expressible in this way, though some of the nine get very little credence.

Note that it is the pattern of counterfactuals, not any single one of them, that expresses causal dependence or independence. As we have seen, the same counterfactual

I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr

figures in the first hypothesis as part of a pattern of dependence and in the second as part of a pattern of independence.

It is clear that not just any counterfactual could be part of a pattern expressing causal dependence or independence. The antecedent and consequent must

---

subjectively rational decision in the single case, like Skyrms' theory and the others considered in this paper, but instead is a theory of objectively effective generic strategies. Since the subject matters are different, the structural identity is misleading. Cartwright's theory might somehow imply a single-case theory having more than structure in common with Skyrms' theory, but that would take principles she does not provide; *inter alia,* principles relating generic causal conduciveness to influence in the single case. So it is not clear that Cartwright's decision theory, causal though it is, falls under my claim that 'we causal decision theorists share one common idea'.

specify occurrences capable of causing and being caused, and the occurrences must be entirely distinct. Further, we must exclude 'back-tracking counterfactuals' based on reasoning from different supposed effects back to different causes and forward again to differences in other effects. Suppose I am convinced that stroking has no influence over purring, but that I wouldn't stroke Bruce unless I were in a mood that gets him to purr softly by emotional telepathy. Then I give credence to

I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly

taken in a back-tracking sense, but not taken in the sense that it must have if it is to be part of a pattern of causal dependence or independence.

Let us define *causal counterfactuals* as those that can belong to patterns of causal dependence or independence. Some will doubt that causal counterfactuals can be distinguished from others except in causal terms; I disagree, and think it possible to delimit the causal counterfactuals in other terms and thus provide noncircular counterfactual analyses of causal dependence and causation itself. But that is a question for other papers.[16] For present purposes, it is enough that dependency hypotheses can be expressed (sometimes, at least) by patterns of causal counterfactuals. I hope that much is adequately confirmed by examples like the one just considered. And that much can be true regardless of whether the pattern of counterfactuals provides a noncircular analysis.

Turning from language to propositions, what we want are causal counterfactuals $A \Box\!\!\rightarrow S$, where A is one of the agent's options and S belongs to some rich partition. The rich partition must be one whose members specify combinations of occurrences wholly distinct from the actions specified by the agent's options. It seems a safe assumption that some such rich partition exists. Suppose some definite one to be chosen (it should make no difference which one). Define a *full pattern* as a set consisting of exactly one such counterfactual proposition for each option. I claim that the conjunction of the counterfactuals in any full pattern is a dependency hypothesis.

Conjunctions of different full patterns are contraries, as any two dependency hypotheses should be. For if S and S' are contraries, and A is possible (which any option is), then also $A \Box\!\!\rightarrow S$ and $A \Box\!\!\rightarrow S'$ are contraries;[17] and any two full patterns must differ by at least one such contrary pair.

What is not so clear is that some full pattern or other holds at any world, leaving no room for any other dependency hypotheses besides the conjunctions of full patterns. We shall consider this question soon. But for now, let us answer it by fiat. Assume that there is a full pattern for every world, so that the dependency hypotheses are all and only the conjunctions of full patterns.

That assumption yields the causal decision theory proposed by Allan Gibbard and William Harper, *op. cit.*, following a suggestion of Robert Stalnaker. My

[16] In particular, my 'Causation', *Journal of Philosophy*, **70** (1973): pp. 556-567; and 'Counterfactual Dependence and Time's Arrow', *Noûs*, **13** (1979): pp. 455-476.

[17] Here and henceforth, I make free use of some fairly uncontroversial logical principles for counterfactuals: namely, those given by the system CK+ID+MP of Brian F. Chellas, 'Basic Conditional Logic', *Journal of Philosophical Logic*, **4** (1975): pp. 133-153.

statement of it amounts to their Savage-style formulation with conjunctions of full patterns of counterfactuals as act-independent states; and their discussion of consequences in their Section 6 shows that they join me in regarding these conjunctions as expressing causal dependence or independence. Although they do not explicitly distinguish causal counterfactuals from others, their Section 2 sketches a theory of counterfactuals which plainly is built to exclude back-trackers in any ordinary situation. This is essential to their purpose. A theory which used counterfactuals in formally the same way, but which freely admitted back-trackers, would not be a causal decision theory. Its conjunctions of full patterns including back-trackers would not be causal dependency hypotheses, and it would give just those wrong answers about Newcomb problems that we causal decision theorists are trying to avoid.[18]

Consider some particular A and S. If a dependency hypothesis K is the conjunction of a full pattern that includes $A \mathrel{\square\!\!\rightarrow} S$, then AK implies S and $C(S/AK)$ equals one. If K is the conjunction of a full pattern that includes not $A \mathrel{\square\!\!\rightarrow} S$ but some contrary $A \mathrel{\square\!\!\rightarrow} S'$, then AK contradicts S and $C(S/AK)$ equals zero. *Ex hypothesi,* every dependency hypothesis K is of one kind or the other. Then the K's for which $C(S/AK)$ equals one comprise a partition of $A \mathrel{\square\!\!\rightarrow} S$, while $C(S/AK)$ equals zero for all other K's. It follows by the Rule of Additivity for credence that

$$(12) \qquad C(A \mathrel{\square\!\!\rightarrow} S) = \Sigma_K \, C(K)C(S/AK).$$

(Comparing (12) with (8), we find that our present assumptions equate $C(A \mathrel{\square\!\!\rightarrow} S)$ with $C_A(S)$, the credence of S imaged on the option A.)
Substituting (12) into (7) we have

$$(13) \qquad U(A) = \Sigma_S C(A \mathrel{\square\!\!\rightarrow} S)V(AS),$$

which amounts to Gibbard and Harper's defining formula for the 'genuine expected utility' they deem it rational to maximise.[19]

We have come the long way around to (13), which is not only simple but also intuitive in its own right. But (13) by itself does not display the causal character of Gibbard and Harper's theory, and that is what makes it worthwhile to come at it by way of dependency hypotheses. No single $C(A \mathrel{\square\!\!\rightarrow} S)$ reveals the agent's causal views, since it sums the credences of hypotheses which set $A \mathrel{\square\!\!\rightarrow} S$ in a pattern of dependence and others which set $A \mathrel{\square\!\!\rightarrow} S$ in a pattern of independence. Consequently the roundabout approach helps us to appreciate what the theory of Gibbard and Harper has in common with that of someone like Skyrms who is reluctant to use counterfactuals in expressing dependency hypotheses.

[18] Such a theory is defended in Terence Horgan, 'Counterfactuals and Newcomb's Problem', *Journal of Philosophy* (forthcoming).

[19] To get exactly their formula, take their 'outcomes' as conjunctions AS with 'desirability' given by $V(AS)$; and bear in mind (i) that $A \mathrel{\square\!\!\rightarrow} AS$ is the same as $A \mathrel{\square\!\!\rightarrow} S$, and (ii) that if A and A' are contraries, $A \mathrel{\square\!\!\rightarrow} A'S$ is the empty proposition with credence zero.

10. Counterfactual Dependence with Chancy Outcomes

The assumption that there is a full pattern for each world is a consequence of Stalnaker's principle of Conditional Excluded Middle,[20] which says that either $X \,\square\!\!\rightarrow\, Y$ or $X \,\square\!\!\rightarrow\, -Y$ holds at any world (where $-Y$ is the negation of Y). It follows that if $Y, Y', \ldots$ are a partition and X is possible, then $X \,\square\!\!\rightarrow\, Y$, $X \,\square\!\!\rightarrow\, Y', \ldots$ also are a partition. The conjunctions of full patterns are then a partition because, for any option A, the counterfactuals $A \,\square\!\!\rightarrow\, S$, $A \,\square\!\!\rightarrow\, S', \ldots$ are a partition.

Conditional Excluded Middle is open to objection on two counts, one more serious than the other. Hence so is the decision theory of Gibbard and Harper, insofar as it relies on Conditional Excluded Middle to support the assumption that there is a full pattern for each world. Gibbard and Harper themselves are not to be faulted, for they tell us that their 'reason for casting the rough theory in a form which gives these principles is that circumstances where these can fail involve complications which it would be best to ignore in preliminary work.' (*Op. cit.:* 128.) Fair enough; still, we have unfinished business on the agenda.

The first objection to Conditional Excluded Middle is that is makes arbitrary choices. It says that the way things would be on a false but possible supposition X is no less specific than the way things actually are. Some single, fully specific possible world is the one that would be actualised if it were that X. Since the worlds $W, W', \ldots$ are a partition, so are the counterfactuals $X \,\square\!\!\rightarrow\, W$, $X \,\square\!\!\rightarrow\, W', \ldots$ saying exactly how things would be if X. But surely some questions about how things would be if X have no nonarbitary answers: if you had a sister, would she like blintzes?

The less specific the supposition, the less it settles; the more far-fetched it is, the less can be settled by what carries over from actuality; and the less is settled otherwise, the more must be settled arbitrarily or not at all. But the supposition that an agent realises one of his narrowest options is neither unspecific nor far-fetched. So the Arbitrariness Objection may be formidable against the general principle of Conditional Excluded Middle, yet not formidable against the special case of it that gives us a full pattern for each world.

Further, Bas van Fraassen has taught us a general method for tolerating arbitrariness.[21] When forced to concede that certain choice would be arbitrary, we leave those choices unmade and we ask what happens on all the alternative ways of making them. What is constant over all the ways of making them is determinate, what varies is indeterminate. If the provision of full patterns for certain worlds is partly arbitrary, so be it. Then indeed some arbitrary variation may infect the $C(K)$'s, $C(S/AK)$'s, $C(A \,\square\!\!\rightarrow\, S)$'s, and even the $U(A)$'s. It might even infect the set of U-maximal options. Then indeed it would be

[20] Robert C. Stalnaker, 'A Theory of Conditionals', in N. Rescher, ed., *Studies in Logical Theory* (Oxford: Blackwell, 1968) gives a semantical analysis in which Conditional Excluded Middle follows from ordinary Excluded Middle applied to the selected antecedent-world.

[21] See Bas van Fraassen, 'Singular Terms, Truth-Value Gaps and Free Logic', *Journal of Philosophy,* **63** (1966): pp. 481-495. Use of van Fraassen's method to concede and tolerate arbitrariness in counterfactuals was suggested to me by Stalnaker in 1971 (personal communication) and is discussed in my *Counterfactuals:* pp. 81-83.

(wholly or partly) indeterminate which options the Gibbard-Harper theory commends as rational. All of that might happen, but it needn't. The arbitrary variation might vanish part way through the calculation, leaving the rest determinate. The less arbitrary variation there is at the start, of course, the less risk that there will be any at the end.

I conclude that the Arbitrariness Objection by itself is no great threat to Gibbard and Harper's version of causal decision theory. We can well afford to admit that the theory might fail occasionally to give a determinate answer. Indeed, I admit that already, for any version, on other grounds: I think there is sometimes an arbitrary element in the assignment of C and V functions to partly rational agents. No worries, so long as we can reasonably hope that the answers are mostly determinate.

Unfortunately there is a second, and worse, objection against Conditional Excluded Middle and the Gibbard-Harper theory. In part it is an independent objection; in part an argument that van Fraassen's method of tolerating arbitrariness would be severely overloaded if we insisted on providing full patterns all around (and *a fortiori* if we insisted on saving Conditional Excluded Middle generally), and we could not reasonably hope that the answers are mostly determinate. Suppose the agent thinks — as he should if he is well-educated — that the actual world may very well be an indeterministic one, where many things he cares about are settled by chance processes. Then he may give little of his credence to worlds where full patterns hold. In fact he may well give little credence to any of the $A \mathrel{\square\!\!\rightarrow} S$ counterfactuals that make up these patterns.

Consider again my problem of getting Bruce to purr. I think that Bruce works by firing of neurons, I think neurons work by chemical reactions, and I think the making or breaking of a chemical bond is a chance event in the same way that the radioactive decay of a nucleus is. Maybe I still give some small credence to the nine full patterns considered in Section 9 — after all, I might be wrong to think that Bruce is chancy. But mostly I give my credence to the denials of all the counterfactuals that appear in those patterns, and to such counterfactuals as

> I brush Bruce $\mathrel{\square\!\!\rightarrow}$ a chance process goes on in him which has certain probabilities of eventuating in his purring loudly, softly, or not at all;

and likewise for the options of stroking and leaving alone. A diehard supporter of the Gibbard-Harper theory (not Gibbard or Harper, I should think) might claim that I give my credence mostly to worlds where it is arbitrary which one of the nine full patterns holds, but determinate that some one of them holds. If he is right, even this easy little decision problem comes out totally indeterminate, for the arbitrary variation he posits is surely enough to swing the answer any way at all. Nor would it help if I believe that whichever I did, all the probabilities of Bruce's purring loudly, softly, or not at all would be close to zero or one. Nor would a more realistic decision problem fare any better: unless the agent is a fairly convinced determinist, the answers we want vanish into indeterminacy. The diehard destroys the theory in order to save it.

Anyway, the diehard is just wrong. If the world is the chancy way I mostly

think it is, there's nothing at all arbitrary or indeterminate about the counterfactuals in the full patterns. They are flatly, determinately false. So is their disjunction; the diehard agrees that it is determinate in truth value, but the trouble is that he thinks it is determinately true.

Unlike the Arbitrariness Objection, the Chance Objection seems to me decisive both against Conditional Excluded Middle generally and against the assumption that there is a full pattern for each world. Our conception of dependency hypotheses as conjunctions of full patterns is too narrow. Fortunately, the needed correction is not far to seek.

I shall have to assume that anyone who gives credence to indeterministic worlds without full patterns is someone who — implicitly and in practice, if not according to his official philosophy — distributes his credence over contingent propositions about single-case, objective chances. Chance is a kind of probability that is neither frequency nor credence, though related to both. I have no analysis to offer, but I am convinced that we do have this concept and we don't have any substitute for it.[22]

Suppose some rich partition to be chosen which meets the requirement of distinct occurrences laid down in Section 9. Let the variable p range over candidate probability distributions for this rich partition: functions assigning to each S in the partition a number $p(S)$ in the interval from zero to one, such that the $p(S)$'s sum to one. Let [P=p] be the proposition that holds at just those worlds where the chances of the S's, as of the time when the agent realises his chosen option, are correctly given by the function p. Call [P=p] a *chance proposition,* and note that the chance propositions are a partition. Now consider the causal counterfactuals $A \mathbin{\square\!\!\rightarrow} [P=p]$ from the agent's options to the chance propositions. Define a *probabilistic full pattern* as a set containing exactly one such counterfactual for each option. I claim that the conjunction of the counterfactuals in any probabilistic full pattern is a causal dependency hypothesis. It specifies plain causal dependence or independence of the chances of the S's on the A's, and thereby it specifies a probabilistic kind of causal dependence of the S's themselves on the A's.

Here for example, are verbal expressions of three chance propositions.

$[P=p_1]$ The chance that Bruce purrs loudly is 50%; the chance that he purrs softly is 40%; and the chance that he purrs not at all is 10%.
$[P=p_2]$ (similar, but with 30%, 50%, 20%).
$[P=p_3]$ (similar, but with 10%, 10%, 80%).

(The chance is to be at the time of my realising an option; the purring or not is to be at a certain time shortly after.) And here is a dependency hypothesis that might get as much of my credence as any:

I brush Bruce $\mathbin{\square\!\!\rightarrow} [P=p_1]$ holds;
I stroke Bruce $\mathbin{\square\!\!\rightarrow} [P=p_2]$ holds;
I leave Bruce alone $\mathbin{\square\!\!\rightarrow} [P=p_3]$ holds.

[22] For a fuller discussion of chance and its relations to frequency and credence, see 'A Subjectivist's Guide to Objective Chance'.

Observe that this hypothesis addresses itself not only to the question of whether loud and soft purring are within my influence, but also to the question of the extent and the direction of my influence.

If a chance proposition says that one of the S's has a chance of one, it must say that the others all have chances of zero. Call such a chance proposition *extreme*. I shall not distinguish between an extreme chance proposition and the S that it favours. If they differ, it is only on worlds where something with zero chance nevertheless happens. I am inclined to think that they do not differ at all, since there are no worlds where anything with zero chance happens; the contrary opinion comes of mistaking infinitesimals for zero. But even if there is a difference between extreme chance propositions and their favoured S's, it will not matter to calculations of utility so let us neglect it. Then our previous dependency hypotheses, the conjunctions of full patterns, are subsumed under the conjunctions of probabilistic full patterns. So are the conjunctions of mixed full patterns that consist partly of $A \mathrel{\Box\!\!\rightarrow} S$'s and partly of $A \mathrel{\Box\!\!\rightarrow} [P{=}p]$'s.

Dare we assume that there is a probabilistic full pattern for every world, so that on this second try we have succeeded in capturing all the dependency hypotheses by means of counterfactuals? I shall assume it, not without misgivings. That means accepting a special case of Conditional Excluded Middle, but (i) the Chance Objection will not arise again,[23] (ii) there should not be too much need for arbitrary choice on other grounds, since the options are quite specific suppositions and not far-fetched, and (iii) limited arbitrary choice results in nothing worse than a limited risk of the answers going indeterminate.

So my own causal decision theory consists of two theses. My main thesis is that we should maximise expected utility calculated by means of dependency hypotheses. It is this main thesis that I claim is implicitly accepted also by Gibbard and Harper, Skyrms, and Sobel. My subsidiary thesis, which I put forward much more tentatively and which I won't try to foist on my allies, is that the dependency hypotheses are exactly the conjunctions of probabilistic full patterns.

(The change I have made in the Gibbard-Harper version has been simply to replace the rich partition of S's by the partition of chance propositions $[P{=}p]$ pertaining to these S's. One might think that perhaps that was no change at all: perhaps the S's already were the chance propositions for some other rich partition. However, I think it at least doubtful that the chance propositions can be said to 'specify combinations of occurrences' as the S's were required to do. This question would lead us back to the neo-Humean thesis discussed in Section 8.)

Consider some particular A and S. If a dependency hypothesis K is the conjunction of a probabilistic full pattern, then for some p, K implies $A \mathrel{\Box\!\!\rightarrow} [P{=}p]$. Then AK implies $[P{=}p]$; and $C(S/AK)$ equals $p(S)$, at least in any ordinary case.[24] For any p, the K's that are conjunctions of probabilistic full

[23] Chances aren't chancy; if $[P{=}p]$ pertains to a certain time, its own chance at that time of holding must be zero or one, by the argument of 'A Subjectivist's Guide to Objective Chance': pp. 276-277.

[24] That follows by what I call the Principal Principle connecting chance and credence, on the

pattern including $A \mathrel{\square\!\!\rightarrow} [P{=}p]$ are a partition of $A \mathrel{\square\!\!\rightarrow} [P{=}p]$. So we have

$$(14) \qquad \Sigma_p C(A \mathrel{\square\!\!\rightarrow} [P{=}p])p(S) = \Sigma_K C(K)C(S/AK).$$

Substituting (14) into (7) gives us a formula defining expected utility in terms of counterfactuals with chance propositions as consequents:

$$(15) \qquad U(A) = \Sigma_S \Sigma_p C(A \mathrel{\square\!\!\rightarrow} [P{=}p])p(S)V(AS).$$

For any S and any number q from zero to one, let $[P(S){=}q]$ be the proposition that holds at just those worlds where the chance of S, at the time when the agent realises his option, is q. It is the disjunction of those $[P{=}p]$'s for which $p(S)$ equals q. We can lump together counterfactuals in (14) and (15) to obtain reformulations in which the consequents concern chances of single S's:

$$(16) \qquad \Sigma_q C(A \mathrel{\square\!\!\rightarrow} [P(S){=}q])q = \Sigma_K C(K)C(S/AK),$$

$$(17) \qquad U(A) = \Sigma_S \Sigma_q C(A \mathrel{\square\!\!\rightarrow} [P(S){=}q])qV(AS).$$

There are various ways to mix probabilities and counterfactuals. I have argued that when things are chancy, it isn't good enough to take credences of plain $A \mathrel{\square\!\!\rightarrow} S$ counterfactuals. The counterfactuals themselves must be made probabilistic. I have made them so by giving them chance propositions as consequents. Sobel makes them so in a different way: as we noted in Section 7, he puts the probability in the connective. Under our present assumptions (and setting aside extraordinary worlds where the common asymmetries of time break down), the two approaches are equivalent. Sobel's quantitative counterfactual with a plain consequent.

If it were that A, it would be with probability q that S

holds at W iff $W_A(S)$ equals q. Given my derivation of the imaging function from the dependency hypotheses, that is so iff $C(S/AK_W)$ equals q. That is so (setting aside the extraordinary worlds) iff K_W implies $A \mathrel{\square\!\!\rightarrow} [P(S){=}q]$. Given that there is a probabilistic full pattern for each world, that is so iff $A \mathrel{\square\!\!\rightarrow} [P(S){=}q]$ holds at W. Hence the Sobel quantitative counterfactual with a plain consequent is the same proposition as the corresponding plain counterfactual with a chance consequent. If ever we must retract the assumption that there is a probabilistic full pattern for each world (or if we want to take the extraordinary worlds into account), the two approaches will separate and we may need to choose; but let us cross that bridge if we come to it.

---

assumption that (i) AK holds or fails to hold at any world entirely in virtue of the history of that world up to action time together with the complete theory of chance for that world, and (ii) the agent gives no credence to worlds where the usual asymmetries of time break down. Part (ii) fails in the case which we have already noted in Section 7 as troublesome, in which the agent thinks he may have foreknowledge of the outcomes of chance processes. See 'A Subjectivist's Guide to Objective Chance': pp. 266-276.

11. The Hunter-Richter Problem

That concludes an exposition and survey of causal decision theory. In this final section, I wish to defend it against an objection raised by Daniel Hunter and Reed Richter.[25] Their target is the Gibbard-Harper version; but it depends on nothing that is special to that version, so I shall restate it as an objection against causal decision theory generally.

Suppose you are one player in a two-person game. Each player can play red, play white, play blue, or not play. If both play the same colour, each gets a thousand dollars; if they play different colours, each loses a thousand dollars; if one or both don't play, the game is off and no money changes hands. Value goes by money; the game is played only once; there is no communication or prearrangement between the players; and there is nothing to give a hint in favour of one colour or another — no 'Whites rule OK!' sign placed where both can see that both can see it, or the like. So far, this game seems not worthwhile. But you have been persuaded that you and the other player are very much alike psychologically and hence very likely to choose alike, so that you are much more likely to play and win than to play and lose. Is it rational for you to play?

Yes. So say I, so say Hunter and Richter, and so (for what it is worth) says noncausal decision theory. But causal decision theory seems to say that it is not rational to play. If it says that, it is wrong and stands refuted. It seems that you have four dependency hypotheses to consider, corresponding to the four ways your partner might play:

K_1 Whatever you do, he would play red;
K_2 Whatever you do, he would play white;
K_3 Whatever you do, he would play blue;
K_4 Whatever you do, he would not play.

By the symmetry of the situation, K_1 and K_2 and K_3 should get equal credence. Then the expected utility of not playing is zero, whereas the expected utilities of playing the three colours are equal and negative. So we seem to reach the unwelcome conclusion that not playing is your U-maximal option.

I reply that Hunter and Richter have gone wrong by misrepresenting your partition of options. Imagine that you have a servant. You can play red, white, or blue; you can not play; or you can tell your servant to play for you. The fifth option, delegating the choice, might be the one that beats not playing and makes it rational to play. Given the servant, each of our previous dependency hypotheses splits in three. For instance K_1 splits into:

$K_{1,1}$ Whatever you do, your partner would play red, and your servant would play red if you delegated the choice;
$K_{1,2}$ Whatever you do, your partner would play red, and your servant would play white if you delegated the choice;
$K_{1,3}$ Whatever you do, your partner would play red, and your servant would play blue if you delegated the choice.

[25] 'Counterfactuals and Newcomb's Paradox', *Synthese*, **39** (1978): pp. 249-261, especially pp. 257-259.

(If you and your partner are much alike, he too has a servant, so we can split further by dividing the case in which he plays red, for instance, into the case in which he plays red for himself and the case in which he delegates his choice and his servant plays red for him. However, that difference doesn't matter to you and is outside your influence, so let us disregard it.) The information that you and your partner (and your respective servants) are much alike might persuade you to give little credence to the dependency hypotheses $K_{1,2}$ and $K_{1,3}$ but to give more to $K_{1,1}$; and likewise for the subdivisions of K_2 and K_3. Then you give your credence mostly to dependency hypotheses according to which you would either win or break even by delegating your choice. Then causal decision theory does not tell you, wrongly, that it is rational not to play. Playing by delegating your choice is your U-maximal option.

But you don't have a servant. What of it? You must have a tie-breaking procedure. There must be something or other that you do after deliberation that ends in a tie. Delegating your choice to your tie-breaking procedure is a fifth option for you, just as delegating it to your servant would be if you had one. If you are persuaded that you will probably win if you play because you and your partner are alike psychologically, it must be because you are persuaded that your tie-breaking procedures are alike. You could scarcely think that the two of you are likely to coordinate *without* resorting to your tie-breaking procedures, since *ex hypothesi* the situation plainly *is* a tie! So you have a fifth option, and as the story is told, it has greater expected utility than not playing. This is not the option of playing red, or white, or blue, straightway at the end of deliberation, although if you choose it you will indeed end up playing red or white or blue. What makes it a different option is that it interposes something extra — something other than deliberation — after you are done deliberating and before you play.[26]

Princeton University

Received April 1980

[26] This paper is based on a talk given at a conference on Conditional Expected Utility at the University of Pittsburgh in November 1978. It has benefited from discussions and correspondence with Nancy Cartwright, Allan Gibbard, William Harper, Daniel Hunter, Frank Jackson, Richard Jeffrey, Gregory Kavka, Reed Richter, Brian Skyrms, J. Howard Sobel, and Robert Stalnaker.