NOÛS 35:2 (2001) 239–259

Evidentialism Reconsidered

DAVID PAPINEAU
King's College London

1 Introduction

In this paper, I want to re-examine the debate between evidential and causal decision theory. My immediate motivation for returning to this familiar issue derives from two recent papers, by Clark Glymour and Christopher Meek (1994), and by Christopher Hitchcock (1996). I shall argue that these papers are seduced by a plausible line of thought into a mistaken sympathy for evidentialism. A second motivation is to extend an account of non-causal probabilistic rationality developed earlier with Helen Beebee (Beebee and Papineau, 1997) to deal with causal contexts as well. By the end of the paper, I hope to have deepened our understanding of exactly why rational decision theory needs to attend to causes.

2 Maximizing Expected Utility

Once upon a time, before Newcomb's problem, it was easy to theorise about rational decision. Agents were simply advised to maximize conditional expected utility. In the simplest case, which will suffice for all the purposes of this paper, with only one end, R, and one choice, A or not A, they were told to do A just in case $P(R/A) > P(R/-A)$. (Cf. Jeffrey, 1983.)

3 Newcomb Problems

The deficiencies of this model were first exposed by the original Newcomb paradox, which made play with super-predictors and the like. (Cf. Nozick, 1969.) But the essential point is better displayed with more everyday examples.

Suppose there is a correlation between eating Mars Bars (A) and sleeping well (R), that is, $P(R/A) > P(R/-A)$. However, suppose also that you know that this correlation is due to the prior presence of some hidden hormone (H),

© 2001 Blackwell Publishers Inc., 350 Main Street, Malden, MA 02148, USA, and 108 Cowley Road, Oxford OX4 1JF, UK.

which independently conduces both to eating Mars Bars and to sound sleep, and which therefore "screens off" any direct association between these occurrences, that is, $P(R/A\&H) = P(R/H)$ and $P(R/A\&{-}H) = P(R/H)$.

In this kind of case, the natural conclusion is that eating Mars Bars (A) does not itself cause you to sleep well (R). Rather, these two events are joint effects of the common cause, the hormone (H). Eating Mars Bars is a symptom that you are going to sleep well (since it is a symptom that you have H), but it is not a cause of sound sleep.

Even so, the simple theory of rational decision outlined in the last section advises you to eat Mars Bars if you want to sleep well. For the "raw" conditional probability $P(R/A)$ is still uncontroversially greater than $P(R/{-}A)$. You are more likely to sleep well when you have eaten a Mars Bar than when you haven't (after all, you will then be more likely to have H).

True, it is agreed on all sides that (a) this is a "spurious correlation" which disappears when we consider cases with and without the hormone (H) separately, and that therefore (b) the Mars Bar doesn't cause the sound sleep. But the trouble is that the simple decision theory didn't say anything about spurious correlations or causes. It simply compares the raw conditional probabilities $P(R/A)$ and $P(R/{-}A)$, and recommends action A whenever the former is greater.

4 Causal Decision Theory

Obviously the simple theory is wrong. In the just-described case, you should not eat a Mars Bar in order to get a good night's sleep.

According to "causal decision theory", the remedy is to introduce causal notions into the theory of rational decision. In its simplest version, causal decision theory requires that we first "partition our reference class" by all combinations of presence and absence of other possible causes of the desired result R, and that we then act on the weighted average of the difference action A makes to R *within* each cell of this partition. If, as in our simple example, there is only one other cause H, this means taking the weighted average of the difference A makes given H and not H respectively, and then choosing A just in case this average is positive:

$$\begin{aligned}(\mathrm{I})\ & [P(R/A\&H) - P(R/{-}A\&H)] \times P(H) \\ & + [P(R/A\&{-}H) - P(R/{-}A\&{-}H)] \times P({-}H) > 0.\end{aligned}$$

The idea here is that rational agents need to consider all the different ways in which A might causally influence R, and then take a weighted average of these different differences A can make to R, with the weights corresponding to the probability of each kind of difference. Partitioning by "other causes" will necessarily achieve this result, since anything which makes a difference to the way A influences R will thereby qualify as another cause of R. (Cf. Skyrms, 1980; Lewis 1981.)

The problem with the simple "raw" A-R correlation, from this causal point of view, is that it can be misleading about whether A makes any real *causal* difference to R. In particular, as in our example, we can find a positive "raw" correlation between A and R even when A is itself causally irrelevant, simply because A is probabilistically symptomatic of the presence of some other cause of R.

5 Objective Probabilities

At this point it will be helpful to say something about the *kind* of probabilities I take to be involved in the theory of rational decision. By no means everything which follows will depend on my particular views about this, but in the interests of clarity it is worth being explicit.

When I mention probabilities in this paper, I shall always mean *objective* probabilities. However, these objective probabilities won't necessarily correspond to *chances*, or single-case probabilities. For the purposes of this paper, I shall take the basic notion of probability to be manifested in statements of probabilistic law, such as "*The probability of a* $\varnothing$ *being an* Ω *is p*", which I shall write as "$\Pr_{\varnothing}(\Omega) = p$". Given such a law, the single-case probabilities of particular $\varnothing$s being Ω might themselves always be *p* (when the reference class $\varnothing$ is "homogeneous"). But it is equally consistent with such a law that there be different kinds of $\varnothing$ which fix varying single-case probabilities for Ω (then $\varnothing$ is "inhomogeneous").[1]

This focus on objective probabilities will contrast with most other work in decision theory. Decision theorists standardly work with subjective notions. When they refer to probabilities, they mean subjective probabilities. They are interested in how you ought to act given that you attach certain *subjective degrees of belief* to given outcomes, whatever the objective probabilities of those outcomes. (Indeed a similar point applies to many of those who introduce causal notions into decision theory. When they demand a partition by other causes, they often mean a partition determined by the agent's *subjective beliefs about other causes*, whatever the accuracy of those causal beliefs.)

I think that this subjective approach leaves out the most interesting issues. For we can also ask about which features of the objective world rational actions ought ideally to be sensitive to. From the point of view of getting what you want, which *objective quantities* ought to be tracked by your subjective degrees of belief and hence your actions? (Moreover, do you need to partition by *genuine causes* to arrive at the right decisions?)

While these objective questions introduce epistemological issues, the epistemology can be put to one side in the context of decision theory. The prior question is which objective facts we need to know about to make the right decisions. How we might find out these facts is a further issue.

In a previous paper, "Probability as a Guide to Life" (1997), Helen Beebee and I addressed this prior question in connection with decisions where Newcomb-

like complications do not intrude, such as betting on games of chance. Since (unrigged) games of chance preclude any spurious correlations between actions (bets) and desired outcomes (pay-offs), we could ignore causal complexities, and concentrate instead on the question of which objective probabilities rational agents ought ideally to match their degrees of belief to.

A plausible first thought here might be that rational agents ought ideally to be guided by the *single-case probabilities* or *chances* which A and not-A would give R in the circumstances obtaining at the time of their decision. But Beebee and I argued against adopting this as a basic principle. It does not deal happily with gambles where the outcome is already determined but unknown to the agent, such as a bet on a symmetrical coin which has already been tossed but whose face is concealed (for here you should have a 0.5 degree of belief in heads, even though the chance is already either 0 or 1). Instead Beebee and I argued that the theoretically fundamental principle, somewhat surprisingly, is that rational agents should be guided by *relative probabilities*, by which we meant the probabilities of the desired results (R) given A and not-A *in the reference classes defined by agents' possibly limited knowledge (K) of their situation*. (For example, the probability of winning if you bet heads on-a-symmetrical-coin-which-has-been-tossed-but-not-yet-exposed.) Thus, in our view, agents should do A in pursuit of R just in case $P_K(R/A) > P_K(R/-A)$. We called this the "Relative Principle".

This reference to agents' knowledge might seem to imply that the relevant probabilities are subjective after all. But this does not follow. Agents' limited subjective knowledge K may determine *which* relative probabilities matter to their choice, but there is nothing subjective about those probabilities themselves. It is not a matter of opinion that 0.5 is the probability of winning if you bet heads on-a-symmetrical-coin-which-has-been-tossed-but-not-yet-exposed. This is an objective fact about the world, indeed a paradigm of an objective probabilistic law.

As advertised above, the probabilities I mention in this paper will all be of this kind—the probability of a given type of outcome in a given reference class, as specified by the relevant objective probabilistic law. If there is no explicit identification of a reference class, then it should be taken to be the K given by the agent's possibly limited knowledge of his or her situation.

Beebee and I defended our Relative Principle for cases where there is no question of any spurious correlation between actions and desired results. This by no means decides how agents ought to choose when the question of spurious correlations does arise. The contention of causal decision theory, as I understand it, is that in such cases Beebee's and my Relative Principle does not suffice. To do A in pursuit of R, according to causal decision theory, it is not enough that $P_K(R/A) > P_K(R/-A)$, where K includes everything you know about yourself. Action A should remain positively relevant to R when we take a weighted average of the difference A makes to R in each cell of the partition defined by

the presence and absence of other causes. Thus, in the simple case discussed above, you should do A only if inequality (I) holds.

It is worth noting that causal decision theory, as I am construing it, will share with the Relative Principle a commitment to probabilities which are objective but not necessarily single-case. This can be seen in connection with the P(H) and P(−H) which enter into inequality (I). In our simple example, let us specify that the hormone H is in fact either present or absent by the time agents decide whether to eat the Mars Bar. The single-case probability of H will thus either be 0 or 1. But these aren't the numbers that causal theorists recommend for calculating (I). When they say that agents should weigh the difference A makes in H and not-H by the probability of H and not-H respectively, they mean (or anyway should mean, if you ask me) the *probability of H and not-H among people of the kind the agents know themselves to be*, the kind of objective probability that would be evidenced by statistical surveys of such people.

6 Evidential Decision Theory

Not all theorists are persuaded that causal notions are needed to deal with Newcomb-like cases. An alternative line is advocated by "evidential decisions theorists". Evidential theorists, as I shall understand them, argue that we don't need anything more than the Relative Principle outlined above to account for rational action.[2]

These theorists stick to the simple recommendation we started with—namely, act on the "raw" correlation given by the simple comparison P(R/A) vs P(R/−A)—and aim to deal with the counter-examples by devoting appropriate attention to the *requirement of total knowledge*.

This requirement, as embodied in the Relative Principle, specifies that you should always act on the probabilities you assign to various results *given your total knowledge of your situation*. The strategy of evidential decision theory is thus to argue that rational agents in Newcomb-like situations will always know something extra (K) about themselves, such that within K the spurious correlation between A and R will disappear. Within K, A makes no probabilistic difference to R—$P_K(R/A) = P_K(R/-A)$—and so the original simple recommendation, plus attention to the requirement of total knowledge, suffices to stop agents acting stupidly. Evidential decision theory thus maintains that, if we attend carefully enough to the kind of self-knowledge K possessed by rational agents, it turns out that Beebee's and my Relative Principle is all we need after all, without any supplementation by causal partitioning.

By way of illustration, consider the Mars Bar example again. Suppose, for the sake of the argument, that you *knew* you had the hormone H (or that you didn't). Then, by the principle of total evidence, you should act on the probabilistic difference A makes to R within H (or within not-H). And since it is agreed on all sides that this is zero—once you know H (or not-H), knowing A doesn't

make R any more probable—evidential decision theory thus avoids recommending that you eat the Mars Bar.

7 Evidential Motivations

This strategy means that evidential decision theorists face an extra commitment. They need to show that rational people are always sufficiently self-aware to be able to place themselves in a category where any spurious A-R correlation disappears.

It is natural to ask why evidential theorists should wish to enter into this commitment. It is not obvious, to say the least, that rational agents will always know themselves to be in some category within which any spurious correlation between A and R will disappear. Moreover, causal decision theory seems to offer a straightforward alternative account of how, even so, sensible people can avoid acting stupidly—even if they don't know whether or not they are H, they can see that in either case A won't make any difference to R. What have evidentialists got against this eminently sensible "either-or" reasoning?

The official answer is that this reasoning is implicitly causal (since it only works if we partition specifically into the different ways in which A may *causally* influence R), and that it is preferable not to build such a metaphysically doubtful notion as causation into the theory of rational decision.

I have always found this answer puzzling. After all, evidential decision theorists typically take causal notions for granted when they analyze Newcomb-type counterexamples, even if they don't attribute causal thinking to their agents. Evidential *agents* may be able to avoid "either-or" reasoning, but evidential *theorists* are standardly found arguing that sufficiently self-aware agents will *either* know themselves to be in some category where the hidden *cause* is present, *or* in a category where it is absent, and either way these agents will have grounds for judging that A is no longer correlated with R.

The notion of causation may not be fully understood by metaphysicians. But if it is good enough for evidential theorists, it is hard to see why it should be denied to rational agents.

Still, perhaps there is a better way of motivating evidential decision theory, albeit not one that is prominent in the literature.[3] Evidentialists shouldn't say that causal thinking is bad. Rather they should say that evidential thinking is better, since it can justify causal thinking, while not standing in need of any justification itself.

Note first how the evidential recommendation seems self-evident. Evidential theory simply recommends that agents perform those actions that make desired results *most probable*. This recommendation doesn't seem to need any further justification. Doesn't everybody want it to be probable that they will get what they want? (True, the desired results will be probable specifically in reference classes defined by rational agents' knowledge of themselves. But then there is arguably an independent warrant for wanting actions which makes de-

sired results probable in this specific sense: according to the Relative Principle, it is just these probabilities which pick out rational decisions in contexts where Newcomb-like problems do not arise.)

By contrast, it is not at all self-evident why rational agents should act on causes. Suppose it is indeed granted, in line with both causal decision theory and pre-theoretical intuition, that you ought to do A in pursuit of R only to the extent that A *causes* R. Even so, it still seems reasonable to ask *why* it is a good idea to act on causes in this way. It doesn't look as if this ought to be a basic truth about rational decision. What's so good about causes, that they should be acted on? It would be nice if we could give some explanation of the wisdom of acting on causes, some account of what makes causes good to act on.

Now, evidential decision theory promises an answer to just this question. For, if evidentialism can be made to work, it will in effect show that agents who act on causes are *more likely* to get desired results than those who don't. After all, standard evidential decision theory aims to legitimate just the same class of actions as its causal opposition. Yet it hopes to do so by showing those actions are the ones which make desired results *most probable*. This then offers an immediate explanation of why it is wise to act on causes. Those who act on causes are more likely to get desired results.

In short, if evidential decision theory is viable, then it will *justify* acting on causes, by showing that such actions make you into the kind of person who is *most likely* to enjoy desired results. This would make evidentialism preferable to causal decision theory, which can offer no such justification.

8 "Other-Cause Independent" Agents

Two recent papers by Hitchcock, and by Glymour and Meek, both embrace versions of this line of thought. (Actually, Glymour and Meek don't yield an evidential account of why it *always* right to act on causes, since they don't think it is, but only of why it *usually* is. Let us leave this complication until later.)

These papers also both appeal to the thesis that agents whose choices are *probabilistically independent of the other causes* of desired results will never be led astray by acting in line with evidential recommendations. This thesis is indeed true, as the following calculations will show. (Intuitively, the point is obvious enough. There is only a danger of action A being *spuriously* correlated with R if it is itself correlated with the other causes of R; if it's not so correlated with any other causes of R, then any probabilistic association between A and R must be genuinely causal.)

Recall inequality (I), which specified this causal requirement for action A:

$$\text{(I)}\quad [P(R/A\&H) - P(R/-A\&H)] \times P(H) \;+\; [P(R/A\&-H) - P(R/-A\&-H)] \times P(-H) \;>\; 0$$

Compare this with the simple evidential requirement that the raw correlation $P(R/A) - P(R/-A)$ should be positive. Since elementary probability theory tells us that

$$P(R/A) = P(R/A\&H)P(H/A) \quad + \quad P(R/A\&-H)P(-H/A)$$

and

$$P(R/-A) = P(R/-A\&H)P(H/-A) \quad + \quad P(R/-A\&-H)P(-H/-A),$$

we can rewrite the evidential requirement as

$$\text{(II)} \quad \begin{array}{l}[P(R/A\&H)P(H/A) - \\ \quad P(R/-A\&H)P(H/-A)]\end{array} \quad + \quad \begin{array}{l}[P(R/A\&-H)P(-H/A) - \\ P(R/-A\&-H)P(-H/-A)]\end{array} \quad > \quad 0.$$

Comparing (I) with (II), it is obvious that the two recommendations will coincide as long as A and H are probabilistically independent, that is, if $P(H/A) = P(H/-A) = P(H)$, and $P(-H/A) = P(-H/-A) = P(-H)$.

It may be helpful briefly to consider (II) from the point of view of causal decision theory. Causal theorists will say that the "raw correlation" in (II) weighs the difference A makes to R, in H and not-H respectively, by the "wrong" factors—instead of using the unconditional $P(H)$ and $P(-H)$, it is in danger of "confounding" any real difference A makes to R with the tendency for A to occur when H is present.

Still, this danger will be absent if A is *not* so probabilistically associated with any other causes of R. In such cases the "raw correlation" will give us a measure which can be agreed on all sides to be appropriate for rational action.

Given this, it may seem attractive to reason as follows. Surely the choices of *genuinely rational* agents *are* independent of the other causes of desired results. Presumably genuinely rational agents can choose *freely*, can arrive at decisions in ways that are unconstrained by extraneous causal pressures. Maybe the choices of unthinking, non-deliberative agents are indeed contaminated by inappropriate influences. But surely we can expect genuine deliberators to choose in ways that are probabilistically independent of the other causes of their desired results.

If this is right, then we will be able to run an evidential justification of causal decision-making, of the kind outlined in the last section. Agents who reason in causal terms, using the quantities involved in inequality (I), won't of course be led astray. But this is simply because the causal (I) gives the same answers as the evidential (II) gives for genuinely rational and therefore other-cause-independent agents. Agents who reason in causal terms can thus be confident of doing the right thing, since they are thus guaranteed to reach the same decisions as rational agents who reason evidentially.

It is crucial, however, to realize that this is not the only possible reaction to the coincidence of (I) and (II) for "other-cause-independent" agents. Let it indeed be agreed that other-cause-independent agents don't need anything beyond evidential decision theory to reach the right decisions. This needn't mean that all rational agents *are* "other-cause-independent". It might simply reflect the fact that, *when* agents are "other-cause-independent", then the "raw correlations" in (II) are guaranteed to measure the *causal* influence of A on R as in (I). And so, in such cases, but not when agents aren't "other-cause-independent", evidential theory will succeed in shadowing the correct recommendations of causal theory.

I shall eventually defend this second causalist response to the coincidence of (I) and (II) for "other-cause-independent" agents. And I shall adopt the obvious strategy, of seeking out agents who are not other-cause-independent to serve as test cases. That is, I shall aim to show that such agents do indeed exist, and that for them the right recommendation is the causal (I) rather than the evidential (II). Before coming to this, however, let me comment on the papers by Christopher Hitchcock (1996), and Clark Glymour and Christopher Meek (1994), both of which go the other way, and offer versions of the first, evidential response.

9 Hitchcock and Fictional Evidentialism

In his "Causal Decision Theory and Decision-theoretic Causation" (1996), Christopher Hitchcock begins by noting that standard probabilistic accounts of causation take the causal influence of a putative cause C on some putative effect E to be measured by the probability that C gives E *within* the cells of the partition created by presence and absence of other factors which are causally relevant to E. Hitchcock then asks why is this such an interesting relationship between C and E? What is so special about conditional probabilities within this "elaborately constructed partition" (p. 509)? "[W]hy should we be so interested in the conditional distributions that obtain *relative to the c-partition*..., which is so baroque in its construction?" (p. 512)

In order to answer this question, Hitchcock makes the assumption that causes are recipes for achieving results. C is a cause of E just in case it is advisable to do C (if you can) if you want E. As Hitchcock explains, this is to adopt a version of the "manipulability" theory of causation. We aim to analyse causation in terms of rational choice.

However, as Hitchcock immediately observes, this strategy would be unattractively circular if we need to appeal to a prior notion of causation in explaining rational choice. If the definition of a rational action were simply that it be apt to cause the desired results, then the manipulability theory of causation would simply take us round a small circle.

Hitchcock thinks we can break out of this circle with the help of evidential decision theory. This theory will tell us which actions are rational *without* as-

suming anything illegitimately causal. Rational actions are those which bear such-and-such probabilistic relationships to desired results. We can then use this, plus the manipulability theory of causation, to infer that *this* kind of probabilistic relationship is characteristic of causal relationships. This thus yields a non-circular answer to Hitchcock's original question, of why probabilistic theories of causation should focus on just those probabilistic relationships. In short, they do so because those relationships are good to act on.

Hitchcock is thus committed to a version of the evidential explanation of why causes are good to act on, as outlined in sections 7 and 8 above. True, he doesn't aim to explain the wisdom of causal choices as such, but rather why probabilistic theories of causation pick out certain specific probabilistic relationships as causal. But since the answer he offers, via his manipulability thesis, is that these peculiar probabilistic relationships are evidentially good to act on, his overall story also commits him to the evidential explanation of the wisdom of acting on causes.

So far, Hitchcock's project is entirely cogent. Things go wrong, however, when he explains how he understands evidential decision theory. In his version, evidential theory does not make the recommendation that agents should act on the "raw correlation" $P(R/A) - P(R/-A)$. Hitchcock does not deny that this correlation may be spurious in many cases, even after we have conditioned on the total knowledge of self-aware rational agents. And he accepts that when raw correlations are spurious in this way, they will be a bad guide to rational decision.

So instead Hitchcock reads evidential decision theory as recommending that you should act on the A-R correlation that *would* obtain under the assumption that you *were* other-cause-independent. He accepts that this assumption will be false for many agents. But even so, he suggests, rational agents can reason *as if* they were other-cause-independent, and consider whether A would still be correlated with R in the fictional distribution fixed by this assumption.

In effect, then, Hitchcock is advising agents to act on the comparison given by inequality (I) above, rather than the actual raw correlation (II). He wants them to hold fixed the dependency of R on A within H and not-H, but imagine away any correlation between A and H.

We can all agree that this recommendation will give the intuitively right answers. What is not clear, however, is why Hitchcock thinks it a version of evidential decision theory. I would say that it is just causal decision theory in disguise.

Suppose we ask Hitchcock *why* people should act on the assumption they are other-cause-independent, given that for many of them it will be an inaccurate "fiction". Causal decision theorists of course have an immediate answer. For them correlation (I) simply identifies the overall causal difference that A makes to R. The "fictional correlations" are thus guaranteed to equal the weighted causal difference A makes to R given the presence and absence of other causes.

However, there is no corresponding answer available to evidential theorists. If it is indeed a fiction that some agents are other-cause-independent, then why,

according to evidential theory, should they choose as if they were? For such agents, the fictional correlations won't correspond to the raw correlations evidential theory favours. The fictional correlation will be a bad guide to how often the good result R accrues to agents who do A. If Hitchcock were really appealing to evidential thinking, then surely he ought to urge such agents to act on the spurious correlation (II), not the casual difference (I).

This shows that Hitchcock cannot give any non-causal rationale for acting on his version of "evidential theory". Once we ask *why* anybody should act under the *fiction* they are other-cause-independent, the only available answer is that this comes to the same thing as acting on causes.

This is why I said Hitchcock's decision theory is really causal decision theory in disguise. And, given this, his overall project collapses. Since his explanation of the significance of the relevant probabilistic relationships must in the end be that they are the relationships to which we are directed by the recommendation to act on *causes*, he can't give any non-circular explanation for why probabilistic theories of causation focus on just those probabilistic relationships. (All he can say is that "those probabilistic relationships are important because they mean that C causes E"). This is disappointing, for Hitchcock's original question is certainly worth asking. But disappointment will be inevitable, if there is no non-causal explanation of why we should act on causes.

10 Glymour, Meek and "Intervention"

Clark Glymour's and Christopher Meek's paper, "Conditioning and Intervening" (1994), is motivated by research about the possibility of deriving causal claims from correlations. In a number of recent works (Glymour, Scheines, Spirtes and Kelly, 1987; Spirtes, Glymour and Scheines, 1993), Glymour and others have shown that survey-derived unconditional and conditional correlations between sets of variables will often determine directed causal connections between those variables (given a few plausible assumptions about the relationship between causes and probabilities).

It is possible to query the practical significance of such findings. Can these causal conclusions be used as a basis for action? Do they tell us what will happen if we "wiggle" one variable in order to influence another? Let us accept, for example, that we can use correlational data derived from surveys to infer existing causal connections between parental income, type of school, pre-school test scores, and school-leaving abilities. Does this tell us what will happen if, say, the government tries to improve school-leaving performance by changing which types of schools children attend?

To answer this question, Glymour and his associates have introduced the notion of an "intervention" (Spirtes, Glymour and Scheines, 1993; Meek and Glymour, 1994). They define an intervention as something which directly controls some "manipulated" variable, in such a way as to render that manipulated variable probabilistically independent of all its other causes, while leaving the rest of the causal structure the same. For example, a government "intervention"

could fix the type of school a child attends, independently of usual causes like parental income and pre-school test scores, while leaving constant the connection between school type itself and leaving abilities. Glymour then shows how initial causal conclusions derived from surveys can allow us to infer the difference such an intervention will make to any further effect of interest.

For example, suppose we want to know how much government manipulation of types of school attended would affect school-leaving abilities, based on our prior knowledge of the causal connections between these and other relevant variables. Glymour's solution is to look at the probabilistic association between school type and leaving abilities that *would* be found if school type *were* controlled by an "intervention" in the above sense. The point here is that we don't want to make policy decisions on the basis of the "raw correlation" between school type and leaving abilities in the original probability distribution. Instead we need to work out what that correlation *would* be in a new distribution that preserves the original conditional probabilities of each variable given its direct causes (and the unconditional probabilities of independently caused exogeneous variables), but decorrelates the manipulated variable from all variables that it doesn't itself affect. This will enable us to eliminate any element of the "raw correlation" which doesn't reflect a genuine causal influence of school type on leaving abilities, but is due rather to prior probabilistic associations between school type and other causes of leaving abilities, such as parental income or pre-school test scores.[4]

There are obvious connections between Glymour's analysis and our earlier discussion. In terms of our simple A-R-H example, his analysis implies that the influence of A, eating the Mars Bar, on R, sleeping well, should be measured by the difference used in inequality (I), that is, the correlation we *would* find between A and R if A *were* independent of the other cause of R (the hidden hormone, H). Correlatively, his analysis implies that it would be a mistake to measure the influence by the "raw correlation" displayed in (II), since that will compound any genuine influence A may have on R with the tendency of A to be more common among people with the hidden hormone H.

However, as I pointed out in section 8, there are two ways to respond to the generally-agreed superiority of (I) over (II) as a guide to action. One response—the first, evidential response—is to say that genuinely rational choices *are* independent of other causes of desired results, and so (I) is simply the special case of (II) that applies to rational agents. The other response—the causal response—is simply to say that (I), but not (II), measures the *causal* influence of A on R, and so is the appropriate guide to action, even for rational agents who are not "other cause independent".

Though it is not immediately obvious, a careful reading of Meek and Glymour reveals that they are simply assuming that the first, evidential reading is the correct one. Thus, in discussing the fact that causal and evidential decision theories can in principle issue in different recommendations, they say:

"The difference in the two recommendations does not turn on any difference in normative principles, but on a substantive difference about the causal

processes at work in the context of decision making—the causal decision theorist thinks that when someone decides when to smoke, an intervention occurs, and the 'evidential theorist' thinks otherwise" (p. 1009).

Glymour and Meek are here suggesting that the rationale for the recommendations of causal decision theory is the causal theorist's commitment to the "other-cause-independence" of agents. This follows from the definition of the notion of an "intervention". Remember, an "intervention" is something that decorrelates a manipulated variable (smoking, in the above quotation) from any other causes it may have. So Glymour and Meek are here taking it that causal recommendations are justified just in case the deliberations of rational agents actually render their actions independent of the other causes of desired results.

11 Actions and "Interventions"

Let us now focus on the two competing responses to the comparison of (I) and (II). The evidential line, recall, was that the causal recommendation (I) yields good advice simply because it is the special case of the evidential recommendation (II) for other-cause-independent agents—and rational agents *are* indeed generally other-cause-independent. The causal line, by contrast, was that the evidential recommendation (II) will indeed agree with the causal recommendation (I) for other-cause-independent agents—but that rational agents are *not* all other-cause-independent, and that when they are not the evidential (II) will lead them astray.

The obvious way to decide this issue, as I signalled earlier, is to consider whether there in fact are any rational agents who are not other-cause-independent (that is, whose actions are *not* "interventions" in the sense specified by Glymour and Meek). If we cannot find any such rational agents, then evidentialists will be able to stick to their story that the worth of acting on causes derives from the fact that causal choices will always coincide with evidentially recommended choices. But if some agents are *not* other-cause-independent, if some rational actions are *not* interventions in Glymour's and Meek's sense, then the recommendations of causal and evidential theory will diverge, and the only option left to evidential theory will be the radical step of insisting that it is sometimes right to act contra-causally, that is, on spurious correlations.

Is there really an issue about whether actions are "interventions"? Surely, one might feel, it is just obvious that humans "intervene" in nature when they act. But this isn't obvious at all, in the sense which matters for present purposes. The terminology of "intervention" is very misleading here. In an everyday sense, it is indeed uncontentious that governments can act, or intervene, to standardise schools, and individuals can act, or intervene, to get Mars Bars into their stomachs. But this by no means shows that they can "intervene" in Glymour and Meek's sense. For Glymour and Meek define "interventions" as requiring *other-cause-independence*, and it remains to be shown that agents who act or intervene in an everyday sense are indeed "interveners" in this technical sense.[5]

In fact, it is quite obvious that many rational choices are *not* other-cause-independent, and so not "interventions" in the relevant sense, at least when the *population at large* is our reference class. This simply reflects that fact that many choices can be positively influenced, in a rational way, by the presence of a factor which also exerts an independent influence on the desired result.

I earlier aired the thought that the choices of free, deliberative agents must *per se* be other-cause-independent. But this thought does not stand up, if it is intended to establish other-cause-independence in the population at large. Maybe an extreme kind of libertarian freedom would decorrelate agents entirely from any other causes of their desired results. But there is no reason whatsoever to suppose that all deliberative, rational agents must be free in this extreme libertarian sense.

Consider again the connection between types of school and school-leaving abilities, this time not from the perspective of government policy, but from the perspective of individual parents deciding whether or not to send their children to fee-paying schools in order to enhance their leaving abilities. Assume, as seems highly plausible, that while school type makes some difference to leaving abilities, the wealth of parents is *another* distinct cause of this result. Now note that, on average rich parents are more likely to send their children to fee-paying schools than poorer parents, for the obvious reason that they can better afford it.

These facts alone mean that choices to send children to fee-paying schools will *not* be other-cause-independent in the population at large. Such choices will obviously be more likely among rich people who can afford the fees, that is, more likely when another cause (parental wealth) of the desired result (high leaving abilities) is present.

Of course, this in itself is not a problem for evidentialists generally, nor perhaps for Glymour and Meek, since evidentialists do not generally regard the population at large as the appropriate reference class.[6] As explained earlier, evidentialists appeal to the general principle that agents should always act on probabilities in the reference class defined by their total knowledge of their situation. The standard evidentialist contention is that within this narrower class there will no longer be any spurious correlations between actions and desired results. In effect, standard evidentialists thus hold that total knowledge always restores other-cause-independence, and that all rational actions therefore qualify as "interventions" within the appropriate reference class.

The obvious problem with this argument is the assumption that agents will generally know whether or not they have any common causes of prospective choices and desired results. Do rich people always know how much they are worth, or insomniacs whether they have the hidden hormone H?

The standard evidential response at this point is to appeal to "tickles". The above examples all share a structure in which the common cause creates a spurious action-result correlation *by influencing the motives of agents*. Wealth affects choice of school because it means you don't *mind* spending the fees

so much. The hidden hormone gets you to eat the Mars Bar by making you *want* to eat it. So, provided agents know their own minds, argues "the tickle defence", they are guaranteed to know something that will render them other-cause-independent and so stop them acting on spurious correlations. Maybe they won't know directly whether they have the original common cause or not, but they will know whether or not they have the psychological "tickle" which mediates between that cause and the decision, and so will be able to appreciate that, among people who are like them in this respect, the prospective decision will now be other-cause-independent, and so can't be spuriously correlated with the desired result.

At this point the argument gets more intricate. A natural objection is that there is no obvious reason why rational agents need to be so perfectly self-knowing. David Lewis (1981) complains that decision theory ought also to guide agents who are unsure of their own motives. To which evidentialists have responded that agents surely cannot help but be aware of their motives in the kind of case at issue, since we are explicitly concerned with agents who are *choosing* actions which they *believe* to be correlated with *desired* results (Horwich, 1987, p. 183). A further question which has been discussed is whether awareness of belief and desires is guaranteed to yield reference classes which render subsequent choices other-cause-independent, given that you could know your beliefs and desires, yet remain unsure, in complex cases, how they will lead you to decide (cf. Eells, 1982, Horwich, 1987).

12 Compatibilist Unfreedom

I am not going to pursue this familiar line of debate. This is because there is a quite different kind of case which shows clearly, contra evidentialism, that agents are *not* always other-cause-independent, even within the reference classes defined by everything they know about their reasoning, and moreover that even such agents should choose causally, not evidentially.

So far we have been considering agents who are at least free in a compatibilist sense, even if not a libertarian sense. That is, we are supposing that their actions are entirely controlled by their motives and subsequent deliberations, even if those motives are in turn affected by other factors (including factors that may exert a distinct influence on the desired results).

But what about agents who are not free even in this compatibilist sense? In particular, what about agents whose actions are partly influenced by their motives and deliberations, but also partly influenced by some entirely non-psychological route, some route that quite by-passes their self-conscious reasoning?[7]

It is not hard to construct plausible cases of this kind. Suppose you are considering whether to have a cigarette, and are concerned, inter alia, to avoid getting lung cancer. Whether or not you have a cigarette is a chance function of two factors: first, whether you consciously decide to smoke, D; second, the probabilistically independent presence of a certain psychologically undetectable

addictive chemical, H, in your bloodstream. (Thus, for example, you're 99.9% certain to smoke if you decide to, and have H; 95% likely to smoke if you decide to, and lack H; still 40% likely to smoke if you decide not to, yet have H; and 1% likely to smoke if you decide not to, and don't have H.) Now suppose further that H causes lung cancer, quite separately from inducing people to smoke. Smoking itself, however, doesn't cause lung cancer. Among people with H, cancer is equally likely whether or not they smoke, and similarly among people without H. And suppose, finally, that you know all this.

Should you aim to smoke or not (assuming that you'd quite like a cigarette, very much don't want cancer, and don't care about anything else)? I say obviously you should smoke. You know that smoking doesn't cause cancer. What matters is whether you've got H or not.

However, there seems no good way for evidentialists to recommend smoking. Given the above specifications, there will be a raw correlation between smoking and cancer (since smoking provides some positive probabilistic evidence that you have H, and H causes cancer). Moreover, this correlation will remain, however much you narrow the reference class by reference to aspects of your reasoning. For, whatever decision you reach, your actually ending up smoking would still provide some extra evidence that you have H, and thus that you are likely to get cancer.

Can't evidentialists say that agents in this kind of situation are not *fully rational*, since their actions are influenced by non-psychological factors, and that it is therefore unsurprising that "rational decision theory" does not explain what they should do? But this will not serve. Even agents who lack full compatibilist freedom are in need of advice about how to reach decisions. Whether or not we call such agents "rational" is a red herring. Whatever we call them, there is a right and wrong way for them to reason, and a normative theory of correct decision-making ought to account for this.

Imagine you are actually facing the issue of whether or not to aim to smoke, knowing you are the kind of agent specified in the above smoking-H-cancer example. (This shouldn't be too hard for anybody with addictive inclinations, or with a tendency to weakness of will.) Of course, you know that your resolving to smoke, say, will not be decisive on whether you actually smoke, since it is just one factor, along with H, which influences whether you smoke or not. But still, your decision will still have *some* influence in whether or not you smoke. And, given this, you would still like to get this decision right. So you face a live issue, on which normative decision theory ought to advise you, about *which* smoking-cancer dependency ought to provide an input to your practical reasoning. Should you aim not to smoke, because cancer is commoner given smoking among people who share your known characteristics? Or should you aim to carry on smoking regardless, because you know this correlation is spurious, and that smoking won't cause you any harm. I say the latter answer is clearly right, even if it runs counter to evidentialism.

It might seem as if evidentialists could argue that in this example we ought to think of agents deciding whether to *aim to smoke*, rather than whether to *smoke*. After all, this "basic action" is within the agents' control, rather than smoking per se. Moreover, and precisely because *aiming to smoke* is within the agent's control, this shift of focus will restore the standard evidential ability to mimic causal decision theory and deliver the intuitively right answers. In the above example, there won't be any spurious correlation between *aiming*-to-smoke and getting cancer to start with, since H only affects whether you smoke, not whether you aim-to-smoke. And if we complicate the example so that H does affect whether you aim-to-smoke, by somehow also affecting your motives, then the spurious correlation which results will disappear when you narrow the reference class with the help of your knowledge of your own practical reasoning, as per the standard "tickle" argument.

This is a reasonable response. But now let me change the example slightly, so that you become compatibilist-unfree "all the way down", with no "basic action" fully under the control of your motives. Thus make the undetectable addictive chemical H more insidious. It doesn't just affect what you do, but what you *aim* to do. It surreptitiously and undetectably biases you towards the decision to aim to smoke. So now aiming-to-smoke is itself the outcome of a probabilistic process, influenced partly by H, and partly by your desire to avoid cancer and your belief about the dependency of cancer on smoking. Readers can fill in some numbers themselves if they wish. The point is that even in this example there will be a question about *which* cancer-smoking dependency ought to influence your decision about whether to aim to smoke. True, even this decision will now be affected probabilistically by H as well. But, taking this as given, should the mere belief that smoking is *correlated* with cancer also weigh probabilistically against aiming-to-smoke? Or should you only be less likely to aim-to-smoke when you believe that smoking actually *causes* cancer?

Once more, this seems a clear normative question, on which agents of this kind could use some good advice. Maybe their decisions (even about whether to aim-to-smoke) are less under the control of their beliefs and desires than those of fully rational agents. But, still, these agents would like to know, just as much as fully rational agents, *which* beliefs would provide the better inputs to their decision-processes. And here it seems clear that the causal beliefs would be better, and that beliefs about correlations among people who share their known characteristics would direct them to choose badly. These agents may be in a sad state. But they aren't so sad as to want a smoking-cancer correlation to influence them to stop smoking even when they know full well this correlation is spurious.

13 Biting the One-Boxing Bullet

The only option that seems left to evidentialists at this stage is to bite the bullet and deny the causal intuitions. They can admit that there are examples where

the tickle defence doesn't work, and non-causal correlations cannot be made to disappear. And they allow that in such cases it may initially *seem* as if agents ought not to be swayed by these evidential connections. But they can argue that we should not necessarily trust these initial intuitions, and should instead stand by the principle that agents do best by acting so as to render desired results likely.

This was of course the line adopted by the "one-boxers" in the original discussion of Newcomb's paradox.[8] An even clearer case is Paul Horwich, in his *Asymmetries in Time* (1987), where he insists that there are indeed possible cases where no amount of conditioning on self-knowledge will make non-causal correlations disappear, and who maintains that in such cases agents ought still to choose those actions which will render desired results non-causally most probable.

It is also the line adopted by Glymour and Meek in "Conditioning and Intervening" (1994). I explained above how, in their view, causal recommendations are justified just in case actions are "interventions", that is, other-cause independent. The obvious corollary is that agents who are *not* other-cause-independent ought to act on "raw correlations", even when this would violate causal recommendations.

Glymour and Meek explicitly embrace this corollary. Speaking about Teddy Seidenfeld's view that an agent in the original Newcomb paradox ought to act evidentially, and take one box rather than two, they say that

> Seidenfeld's judgement is fully in accord with [Glymour's and Meek's analysis]; were it stipulated with Seidenfeld that there is no intervention, his judgement is also that which causal decision theory ought to give (p. 1014).

And a bit later they say that, in cases where evidentialists like Seidenfeld differ from the causalists, this

> ...is because they differ about whether an action is an intervention...If so, then a different event must be conditioned on than if not, and a different calculation results (p. 1015).

The implication is clear. If an action is *not* an "intervention", as in cases of compatibilist unfreedom, then we "must condition on" an event which *is* associated with the other causes of the desired result, and so act on spurious correlations.

It is difficult to accept the contra-causal line here being advocated by Glymour and Meek, along with Horwich and one-boxers generally. Surely it is wrong for agents to act on correlations that they know to be causally spurious. In my original example, I took it to be simply obvious that you shouldn't eat a Mars Bars just because this is spuriously correlated with sound sleep. Standard evidentialists responded by bringing in total knowledge, tickles, and so on. But now we are told that, when this story runs out, as in cases of compatibilist unfreedom, then agents should be influenced by spurious correlations after all. This

still seems absurd. Surely there is no virtue in an action that can make no causal difference to the result you want.

Radical one-boxing evidentialism does have one last arrow in its quiver. Recall the motivation for evidentialism discussed in section 7. Evidentialism appealed to the simple thought that you ought to render yourself the kind of person who is most likely to get good results. By contrast, causal decision seemed to offer no independent justification for acting on causes.

Radical evidentialists can hold onto this thought even in the hard cases where evidential and causal recommendations diverge. They will point out that in these cases causal theory actually advocates actions that make it *less* likely you will prosper. (In Newcomb's paradox, those who take two boxes will find there is nothing in the opaque box, where one-boxers get the million pounds.) Given this, they will argue, surely we ought to question the causal intuitions. If causal theorists are so smart, they can ask, how come they are so likely to stay poor?

Of course, evidentialists can allow, causal theory *normally* makes you rich, even if not in the hard cases of compatibilist unfreedom. Causal answers shadow evidential answers in the vast majority of cases. Given this, it is unsurprising that everyday intuition should favour causal choices. It is a good rule of thumb to trust causal theory. But it would be foolish, and unfaithful to the rationale for those intuitions, to stand by them even in those cases where they recommend actions that make you less likely to prosper, such as in cases of compatibilist unfreedom. Or so bullet-biting evidentialists can argue (cf. Horwich, 1987).

14 Conclusion

At first sight it may look as if this leaves the argumentative advantage with the evidentialists. After all, they can offer a principled basis for their position, against which the causal side can offer only intuitions. But this appearance is deceptive. The analysis of this paper shows that the underlying principle to which the evidentialists are appealing simply begs the question against the causal theory.

The fundamental issue here is the status of the thought that rational agents should do what will make desired results most likely in the reference class defined by their total knowledge of themselves. (Equivalent formulations of this thought employed in this paper have been the "Relative Principle", and the recommendation to act on inequality (II).) So far I have not explicitly questioned my earlier suggestion that this evidential thought is self-justifying. "What could be more obvious than that agents should make it the case that good results will probably ensue?" I asked earlier, contrasting this with the apparently unwarranted idea that agents should pick actions apt to "cause" good results.

I hope it is now clear that causal theorists should have resisted this suggestion of self-justification from the start. The basic truth about rational decision, causal theorists should insist, is that you should always perform the action best suited to *causing* good results. There is no independent virtue in the principle

that you should make good results *probable*. When this is true, it is true is only because it is a special case of the causal recommendation.

It might initially have seemed self-evident that you should make desired results most likely in the reference class defined by your total knowledge of yourself (conform to the "Relative Principle", act on inequality (II)). But far from being self-evident, this is often false, and only true when it happens to shadow the causal choice of that action which makes the most causal difference to the desired result, on weighted average over the different causal possibilities. In some special cases, the action so causally chosen will happen also to be one which will renders desired results most probable in the total knowledge reference class. But this will only be because, in these special cases, the action best correlated with the result will also be the one best suited to causing the result.

Earlier in this paper, I aired the possibility of justifying causal decision recommendations evidentially, by showing how causally recommended actions would in general render desired results probable. We can now see that, from the causal point of view, this project had things exactly back to front. Rather, it is evidential decisions that need to be justified causally. Evidential recommended actions are acceptable, when they are, only in virtue of their being apt to cause desired results.

Notes

[1]If you like, you can assume that all objective probabilities which aren't chances any longer were previously chances. (For example, the 0.25 probability of drawing a spade may derive from earlier chance mechanisms, in the brain perhaps, which influenced how the pack was shuffled.) However, I myself am equally happy to take the notion of probability displayed in probabilistic laws as primitive. (This doesn't mean that I favour the ill-conceived frequency interpretation of probability. My idea, to repeat, is that the law-displayed notion of probability is primitive, just as chances are primitive on the chance interpretation of probability.)

[2]In one way this is unfaithful to standard evidentialism, since most actual evidential theorists work with subjective degrees of belief, rather than my objective relative probabilities, as I pointed out in the last section. But this is orthogonal to the issue which interests me in this paper: can we manage with the Relative Principle alone, without appealing to causal partitions? This question arises both for those who set things up subjectively, like standard evidentialists, and for those who assume that all degrees of belief correspond to objective relative probabilities, as I shall throughout this paper.

[3]But see the discussions of Hitchcock, and Meek and Glymour, in sections 9 and 10 below, and of "one-boxing" in section 13.

[4]Of course, this kind of calculation assumes that the "intervention" doesn't change any causal connections in addition to those directed into the manipulated variable. A government decision to eliminate private schools, for example, might make rich parents devote more resources to home tuition, and thereby *enhance* the direct influence of parental income on school-leaving abilities. In this kind of case Glymour's calculation will give the wrong answer.

[5]This conflation can be discerned in ch. 5 of Dan Hausman's *Causal Asymmetries* (1998). Hausman is not concerned with the evidential-causal debate, but in discussing agency theories of causation he assumes without argument that all human actions are "interventions" in the technical sense.

[6]Glymour and Meek don't say much about reference classes. Their one relevant remark is: "We agree with 'evidential' decision theories that nothing but an ordinary calculation of maximum ex-

pected utility is required; we agree with causal decision theorists that sometimes the relevant probabilities in the calculation are *not the obvious conditional probabilities*" (my italics, p. 1015).

[7]There are brief hints at this kind of case in Lewis (1981, p. 312) ("a partly rational agent may well [have] choices influenced by something besides his beliefs and desires") and Horwich (1987, p. 181) ("we could simply have stipulated that cancer be correlated with smoking...*regardless of the agent's inclinations*"). But neither develops these quoted suggestions. And Horwich claims that there are no actual such cases.

[8]Of course "one-boxing" is interesting only when it is specified that the choice does not (backwardly) cause what was placed in the opaque box. All can agree that one-boxing is rational given backwards causation.

References

Beebee, Helen and Papineau, David. (1997) "Probability as a Guide to Life," *Journal of Philosophy 94*, pp. 217–43.

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

Eells, Ellery. (1991) *Probabilistic Causality* (Cambridge: Cambridge University Press).

Glymour, Clark, Richard Scheines, Peter Spirtes, and Kevin Kelly. (1987)
Discovering Causal Structure (New York: Academic Press).

Hausman, Dan. (1998) *Causal Asymmetries* (Cambridge: Cambridge University Press).

Hitchcock, Christopher Read. (1996) "Causal Decision Theory and Decision-theoretic Causation," *Noûs 30*, pp. 508–26.

Jeffrey, Richard. (1983) *The Logic of Decision, 2nd ed.* (Chicago: University of Chicago Press).

Lewis, David. (1981). "Causal Decision Theory," *Australasian Theory of Philosophy 59*, pp. 263–94, reprinted in his (1986) *Philosophical Papers* (Oxford: Oxford University Press) (page references to this reprinting).

Meek, Christopher and Glymour, Clark. (1994) "Conditioning and Intervening," *British Journal for the Philosophy of Science 45*, pp. 1001–21.

Nozick, Robert. (1969) "Newcomb's Problem and Two Principles of Choice," in *Essays in Honor of Carl G. Hempel*, ed. N. Rescher (Dordrecht: Reidel), pp. 114–46.

Price, Huw. (1991) "Agency and Probabilistic Causality," *British Journal for the Philosophy of Science 42*, pp. 157–76.

Skyrms, Brian. (1980) *Causal Necessity* (New Haven: Yale University Press).

Spirtes, Peter, Clark Glymour and Richard Scheines. (1993) *Causation, Prediction and Search* (New York: Springer-Verlag).