HUW PRICE

AGAINST CAUSAL DECISION THEORY

ABSTRACT. Proponents of causal decision theories argue that classical Bayesian decision theory (BDT) gives the wrong advice in certain types of cases, of which the clearest and commonest are the medical Newcomb problems. I defend BDT, invoking a familiar principle of statistical inference to show that in such cases a free agent cannot take the contemplated action to be probabilistically relevant to its causes (so that BDT gives the right answer). I argue that my defence does better than those of Ellery Eells and Richard Jeffrey; and that it applies, where necessary, to other types of Newcomb problem.

Many people now believe that classical Bayesian decision theory (BDT), based on the principle of maximising conditional expected utility, gives the wrong advice[1] in a class of cases known as Newcomb problems. To deal with such cases, it is argued, BDT needs to be supplemented with a principle ensuring sensitivity to the causal properties of actions.[2] The need for such a *causal decision theory* (CDT) has recently been denied by Ellery Eells (1981, 1982) and by Richard Jeffrey (1981, 1983), whose (1965) formulation of BDT is now the standard one. However, it seems to me that neither Eells nor Jeffrey has pinpointed the weakness in the CDT case. Here I attempt to do better.

Newcomb problems come in three main varieties: the original two-box case (see Nozick, 1969), the prisoner's dilemma (see, e.g., Lewis, 1979), and the medical Newcomb problems (e.g., Lewis, 1981). I shall concentrate on the last. These have the advantage over the two-box cases of actually existing, and over prisoners' dilemmas of possessing a clear (and, so far as I know, universally agreed) correct answer. To defend BDT here is hence to defend it in the most difficult case. In the final section I shall look briefly at the other cases, to consider whether my defence of BDT does or should work for them.

Until then, the paper goes like this: in section 1 I describe a typical medical Newcomb problem, sketching the usual case against BDT and illustrating the considerations on which my reply depends; in section 2 I present the defence; and in section 3 I compare my argument to those of Eells and Jeffrey.

Synthese **67** (1986) 195–212.

1. DOCTORING THE EVIDENCE: THE MEDICAL NEWCOMB PROBLEMS

Many physiological states produce quite specific symptomatic effects on choice behaviour. Pregnancy, for example, often affects what a woman chooses to eat. Any such physiological effect is a potential source of a medical Newcomb problem.[3] All else that is needed is that the attitudes of the person concerned about possession of the underlying physiological state and the performance of the symptomatic act should conflict, in the following sense: either the person would like to perform the act, but wouldn't like to be in the state believed to cause (or make more likely) that act, or vice versa. If the agent is someone who knows that pregnancy tends to make her decline garlic, then the problem will arise if either she likes garlic (*ceteris paribus*) but wants to be (now) pregnant; or doesn't like garlic (*ceteris paribus*) and doesn't want to be pregnant. Clearly she shouldn't refuse the delectable garlic in order to be already pregnant, in the former case; or eat the distasteful garlic in order not to be, in the latter. Either she's already pregnant or she isn't. Whichever it is, and whether that's good or bad, she'll be better off than not following the dictates of taste.

Or (to turn to the less realistic industry-standard case), consider Fred, who believes that the correlation between smoking and lung cancer is caused by the existence of a gene which, in those who have it, both inclines them to smoke and gives them lung cancer. He thus believes some statistical generalisation such as

(1) The cancer gene occurs in 20% of smokers, and in 2% of nonsmokers;

(Fred might not put it in terms of frequencies. He might say 'Smokers have a 20% chance of having the gene.' Either way, the belief is a generalisation, and the frequency version prevents our ignoring this crucial fact).

Fred would prefer to smoke than not, other things being equal, but would hate to get cancer; he values the four possibilities as follows:

	Smoking (S)	Not smoking ($-S$)
Cancer (C)	-99	-100
No cancer ($-C$)	1	0

Fred should smoke. Either he has the gene or he hasn't, and whether he smokes won't affect that; and either way, he'll be happier if he smokes.

However, the critics of BDT claim that it advises Fred not to smoke. Their argument requires

(2) In virtue of (1), Fred should take the relevant probabilities of outcomes conditional on actions to be $P(C/S) = 0.2$, $P(C/-S) = 0.02$, $P(-C/S) = 0.8$ and $P(-C/-S) = 0.98$.

Granting (2), the standard formulae of Jeffrey's version of BDT assign (for Fred) the following conditional expected values to smoking and to not smoking:

$$\mathrm{CEU}(S) = (0.2)(-99) + (0.8)(1) = -19$$
$$\mathrm{CEU}(-S) = (0.02)(-100) + (0.98)(0) = -2.$$

In other words: 'Don't smoke!'

Though the assumption (2) has usually been taken for granted by the proponents of CDT, it is clearly critical. (1) is a statistical generalisation, and so doesn't *always* license an inference to corresponding probabilities in particular cases. Consider some examples.

Fred's friend Benson thinks that he might not inherit his father's tobacco fortune if he hasn't been a smoker. So Benson hedges, smoking to ensure his inheritance. Knowing this, Fred doesn't infer from (1) that *Benson* has a 20% chance of having the gene, and hence getting cancer.

Benson Senior smokes too, but Fred knows that he's rich enough to have had his genes checked, and that he's in the clear. So Fred doesn't think that he has a 20% chance of cancer.

Closer to home, consider Freda (a statistician). She actually believes that smoking *causes* cancer. She values smoking and cancer just as Fred does, but has a reason not to smoke (as she puts it, 'To keep my chance of developing cancer as low as possible') such that Fred cannot take her decision as evidence (via (1)) that she doesn't have the cancer gene. Rather Fred thinks that Freda's 'misguided' reasons for not smoking are strong enough to override the effect of her genes on her decision whether to smoke.

Finally, what about Fred's own case. If Fred accepts BDT, he can argue as follows: 'Suppose that (1) does license my adoption of the conditional credences described in (2). Then by BDT I have a strong reason not to smoke, the expected value of not smoking being much higher than that of the alternative. But then just as with Benson Jr., and Freda, the existence of this overriding reason will mean that my resulting decision not to smoke will not comprise evidence (via (1)) as to my genes; i.e., the supposition is false.'

This reductio seems to show that if Fred is an adherent of BDT, he cannot make the inference from (1) to his own case on which the proponents of CDT rely. But how strong does a reason for smoking or not smoking have to be, to override the effect of the gene? The argument doesn't tell us. Nor does it tell us what conditional credences (of cancer, given smoking or not smoking) Fred should adopt. To block an argument that BDT gives the *wrong* advice, is not to show that it gives the *right* advice.

We shall see that in so far as it needs to be, Fred's reductio can be cured of both these defects. The cure will indeed appeal to Fred's causal beliefs, but with no concession to CDT. By the time Fred applies his decision theory, the causal beliefs will have done their work. Causation will turn out to be associated not with decision theory, but with the principles of statistical inference in contexts of free decision.

2. THE EVIDENTIAL SIGNIFICANCE OF CONTEMPLATED ACTION

Probabilistic judgements about causes are often based on knowledge of *effects*. Seeing a fire in an empty building, I think that it was probably started by an arsonist for insurance purposes. Why? Because I believe that most such fires around here are started that way. Such judgements are sensitive to the fact that many types of events can have various causes. Building fires can be caused by lightning strikes and many sorts of human accidents, as well as by arson. If I had reason to think that some such cause was present in this case, then I wouldn't apply my *general* belief that most such fires are the result of arson to make the *particular* judgement that *this* fire is probably the result of arson. I might make some other particular judgement. ('Given that there was a fierce thunderstorm earlier in the evening, the fire was probably started by lightning.') Or I might not feel able to make any judgement on the matter, if different parts of my evidence conflicted.

The relevance of new evidence depends on its relation to the old. Suppose for example that I discover that the fire in question began in a rubbish bin. I have no reason to think that the correlation between fires in empty buildings started in rubbish bins and fires started by arsonists, is any weaker than that between fires in empty buildings in general and fires started by arsonists. So in this case (unlike the thunderstorm one) the new evidence doesn't conflict with the old, and doesn't require me to revise my judgement that this fire is probably the result of arson.

This sort of constraint on probabilistic inference is quite unexceptional, and far from confined to inference from effect to cause. The reverse inference from cause to effect, for example, is similarly sensitive to the fact that the same kind of event can have different effects in different circumstances. Both sorts of case are examples of the use of a very general, fundamental and familiar principle of probabilistic reasoning, the *principle of total evidence*. Roughly,

(3) In making a probabilistic judgement, take into account all the relevant available evidence.

The precise formulation of this principle is difficult, and goes to the heart of problems concerning the interpretation of probability.[4] But its effect is dramatic: it can lead us to abandon a previous probabilistic judgement, on acquiring new evidence relevant to the matter in question; even if the new evidence conflicts with the old, in such a way that we now cannot make *any* assessment of the probability concerned. Should such evidence come to hand, the original judgement becomes, so to speak, 'inoperable' (a political euphemism which here has just the weight the press secretaries would like it to bear in real life – for these judgements, though *retracted*, are not *falsified*[5]).

The application of (3) which concerns us here is to probabilistic judgements about the causes of *actions*. I shall assume (though it isn't critical) that we are talking about actions whose immediate causes are the agent's *reasons* for acting in that way. This doesn't mean that an action cannot have causes which are not reasons; but only that the effect of such causes must be mediated by an effect on the beliefs and desires of the agent concerned. I am going to show that in virtue of the fact that probabilistic inference from effect to cause is constrained by (3), BDT gives Fred the right advice – it tells him to smoke. The application of (3) will depend on the supposition that Fred does reason by BDT, or something like it. There is no circularity here: all we want to show is that someone who *does* use BDT gets the right answers.

If Fred uses BDT, then if the fact that he would hate to get cancer is to influence his decision as to whether to smoke, he has to decide what probabilistic relevance (if any) smoking would have to his possession of the gene (and hence to whether he develops cancer). He believes that the gene is the typical cause of a decision to smoke (at least to the extent described in (1)). But in virtue of the constraint just described, the relevance of this belief depends on what else he believes about (inter alia) his own reasons for acting in a particular way. To make the probabilistic judgement on which his choice will depend, he must, in effect, list the causally relevant factors. These factors include the beliefs and desires which would be the immediate cause of his hypothetical action. And at some point he has to close the list, deciding that such-and-such a body of evidence forms the basis for the judgement.

The judgement gives Fred the conditional credences he is after: his assessment of the probability of cancer, given either that he smokes or that he abstains. Can these conditional credences appear on the list that Fred drew up in making this judgement? Clearly not; the list had to be closed, before the judgement could be made. Yet the conditional credences are potentially among the reasons for his action, and hence the kind of thing that needed to be listed. (Indeed, Fred acquired them specifically in order to decide how to act.) So having made the judgement, Fred is liable to find that it is no longer valid. He has some new beliefs, which, being causally relevant to his choice, may need to be considered in deciding the probabilistic relevance of his eventual action.

Whether they do need to be considered depends on whether the new evidence they constitute conflicts with the old. As in the arson case, the mark of conflict is a reason to think that the newly recognised causal factors will affect the correlation between the relevant effects and the cause whose presence is in question. In other words, does Fred have reason to think that his newly acquired conditional credences make a difference to the correlation between a decision to smoke and possession of the cancer gene? In a class of Fred-alikes, would acquisition of these conditional credences be expected to alter the ratio of smokers to gene-possessors? He does, and it would: in anyone who (like Fred) would rather not get cancer, (2)-guided conditional credences will be negatively correlated with smoking; they will increase the attractiveness of not smoking compared to smoking. So the new evidence which would become available to Fred if he were to decide that $P(C/S) >$

$P(C/{-S})$, would conflict with the evidence (as described in (1)) on which such a judgement would be based. The same would apply if he decided that $P(C/{-S}) > P(C/S)$, except that this decision would make him more likely to smoke. Either judgement would therefore be self-defeating.

The argument shows that Fred cannot (rationally) take his contemplated action to be probabilistically relevant to his possession of the gene: that opinion is inherently unstable. Fred has two alternatives: he can judge that his smoking (or not) would be probabilistically irrelevant to his possession of the gene; or he can simply have no opinion on the matter. However, from the point of view of BDT these two states of opinion come to the same thing. BDT tells us what to do with judgements of probabilistic relevance. It isn't and couldn't be invoked whenever what we have is a judgement of probabilistic irrelevance, or no opinion at all on such a matter; for there is no end to such cases, and the simplest action would depend on calculations ad infinitum. So whether Fred believes that the decision he is contemplating is probabilistically irrelevant to his possession of the gene, or has no opinion on the matter, BDT will simply (and properly) *ignore* his desire not to get cancer (and hence not to have the gene). It will no more be a factor that he should take into account in deciding whether to smoke than, say, his desire to avoid sunstroke. As the case has been described the only *relevant* feature of his desires will be his preference for smoking, *ceteris paribus*; in virtue of which BDT will of course recommend that he smoke.

Fred's powerful desire to avoid cancer thus turns out to play no part in his final BDT-guided choice. (The same would apply to a CDT-guided choice, of course.) However, it is crucial earlier, in de-stabilising any judgement on Fred's part of a probabilistic relevance of his action on his prospects of cancer. For it is in virtue of Fred's desire to avoid cancer that any such judgement would be a contributing cause to his decision; and in virtue of this causal role that the judgement is self-defeating in the way described.

What then if Fred doesn't care whether he has the gene and hence develops cancer? In this case his choice as to whether to smoke won't be influenced by a belief that smoking would make him more (or less) likely to have the gene. So such a belief won't be self-defeating. Nor, however, will it play any part in his BDT-guided choice. BDT requires not only judgements of probabilistic relevance, but also differential desires. Clearly we don't and shouldn't spend our time calculating the effects of equally pleasant possible outcomes on conditional expected values. In the absence of a preference (on Fred's part) between getting cancer and not, BDT's advice to him ignores any beliefs he may have about the bearing of the contemplated action on the chances that he will get cancer. Indeed, this is why in this case such beliefs are not among the causes of his choice, and hence why such beliefs are not self-defeating, as they would be if he did care whether he got cancer.

It is important to realise that the above argument does not mean that *once Fred has acted*, he cannot take his act to have evidential significance with respect to whether he has the gene, and hence will develop cancer. On the contrary, providing he does act on the basis that $P(C/S) = P(C/-S)$, and has no other grounds for thinking that his reasons for acting are exceptional in such a way as to prevent him applying (1) to his own case, he can then generate single-case probabilities for his own case from (1). The beliefs thus generated, arriving *after* he has acted, cannot be among the reasons for his action. Hence the judgement which yields them is not self-invalidating, as before.

Incidentally, this means that the conditional credence that Fred adopts of, say, developing cancer given that he smokes, need not correspond to the absolute credence of developing cancer that he would adopt, if he came to believe that he had started to smoke. That absolute credence may be based on evidence which, as we have just seen, cannot consistently be available to him before he decides whether to smoke. This explanation means that the case is no serious obstacle to the view that ascriptions of conditional probability express dispositions to adopt corresponding absolute credences.[6]

I have thus argued that as Fred decides whether to smoke, he cannot be justified in taking his contemplated action to be probabilistically relevant to his prospects of cancer. The argument depends on the fact that Fred would use any such judgement of relevance in making his decision; for it is in virtue of this that the judgement would itself be a causal factor, contributing to his choice. This needn't mean that Fred uses BDT. Any decision procedure which requires him to assess these conditional probabilities will have the same consequence. But it does mean that if he does use BDT, he has a reason for not adopting the conditional credences which would lead BDT to give him the wrong advice.

3. EELLS, JEFFREY AND THE EXTENDED TICKLE

As I mentioned earlier, BDT has recently been defended both by Ellery Eells and by Richard Jeffrey. Readers familiar with those arguments may have noted respects in which mine is similar to each. Here, while exhibiting these similarities, I want to show that my argument does a better job.

Eells' defence, particularly, is I think best seen as a development of an earlier reply to the proponents of CDT, the so-called 'Tickle Defence'. The Tickle Defence is based on the observation that if the cancer gene produces its effect on choice behaviour by inducing an identifiable craving (or 'tickle') to smoke, then once Fred knows whether or not he has the tickle, his action becomes probabilistically irrelevant to whether he has the gene. Symbolically,

(4) $P(C/T\&S) = P(C/T\&-S)$ and
$P(C/-T\&S) = P(C/-T\&-S)$
(where 'T' means 'Having the tickle').

From (4), and the fact that once Fred knows whether he has the tickle either $P(T) = 1$ or $P(-T) = 1$, it follows that $P(C/S) = P(C/-S)$; given which, as we know, BDT gives the right answer.

The Tickle Defence exploits a general feature of causal chains. Suppose an event of type X is the typical cause of an event of type Z, and that the causal relation is mediated by an event of type Y. In other words, X typically causes Z by causing Y which then causes Z. (The finger on the button launches the missiles by sending a signal down the wires.) Then knowledge of whether or not Y has occurred, on a particular occasion, will 'screen off' the probabilistic bearing of Z on X. If we don't know whether Y, then $P(X/Z) > P(X/-Z)$; but if we learn that Y (or $-Y$), then $P(X/Z) = P(X/-Z)$. ('If the missiles go up, then probably the button has been pressed'; but 'If there's no signal in the wire, then even if the missiles go up, the button probably wasn't pressed'.) The evidential significance of an effect to a cause can be taken over by an earlier effect of the same cause, if such is available.

The trouble with the Tickle Defence is that there is no guarantee that the effect of a physiological state on choice behaviour which gives rise to a medical Newcomb problem will be mediated by any identifiable craving or tickle. The desire to smoke, for example, might be the net result of many separate positive and negative desires: on the plus side, perhaps an oral fetish and a yen for the cowboy life; on the other, a distaste for yellow fingers and dirty ashtrays. Who is to say which of such desires the gene affects, so as to tip the balance in favour of smoking?

All the same, the tickle defender might reply, we are supposing that the gene acts by influencing rational choice: it affects Fred's actions by affecting his beliefs and desires. So it must come into the open, as it were, before he actually acts. To whatever extent his action provides evidence for the gene, that evidence must already be provided by the beliefs and desires which gave rise to that action. Hence if he knew what these beliefs and desires were, the action itself would no longer be probabilistically relevant to possession of the gene (and hence to cancer) – in which case BDT would give him the right advice.

This, I think, is Eells' argument. As it stands, it is dangerously self-referential. Fred has to know what his relevant beliefs and desires are, including his conditional credences in cancer given smoking and not smoking, in order to conclude that these credences should be equal. However, a reductio formulation meets this objection. Fred can reason that unless he takes $P(C/S) = P(C/-S)$, he will be lead to an assessment of these probabilities which conflicts with the credences he already has. (In this respect the argument will then be similar to mine.)

However, Eells argument is vulnerable to other objections. If Fred's knowledge of his beliefs and desires is to screen off the relevance of his action to whether he has the gene, he must know more than simply what his beliefs and desires *are*. He must know whether they are the sort of beliefs and desires which will lead him to smoke. (It is not enough to know *what* signal is in the wire, in order to infer that the button has been pressed; we need to know that it is the kind of signal that launches missiles.) And if he knows this, his choice is effectively already made. In particular, either $P(S)$ or $P(-S)$ will now be zero, and hence either $P(C/S)$ or $P(C/-S)$ will be undefined. So the argument won't yield the required conditional credences.

To avoid this problem, Eells must insist that $P(S)$ and $P(-S)$ both remain non-zero, even when Fred has made up his mind whether or not to smoke. That is, he requires that agents be less than certain that whatever decision they make will be put into effect. This constraint might seem harmless. In the real world are we ever justified in being *certain* that we could act as we chose? But justification isn't what's at issue. We can imagine that Fred, perhaps mistakenly, does fully believe that at least in this respect, he can do what he chooses. Eells' defence of BDT would then be unavailable to him.

In any case, Eells' argument faces a more serious problem. For it is natural to suppose that what is correlated with the cancer gene is not *actually* smoking but rather *choosing* to smoke. (We can imagine a statistical survey testing this supposition by determining the rate of possession of the gene among people who chose to smoke, but were prevented from doing so, and vice versa.) Here 'choosing to smoke' means having beliefs and desires such that one's decision theory leads one to *try* to smoke. Believing that this is how the correlation goes, Fred in any case believes that once he has made his choice, thrown his hat in the ring, it is actually irrelevant to whether he has the gene whether he *succeeds* in smoking (or not smoking, as the case may be). He then takes $P(C/S) = P(C/-S)$, but this is irrelevant to his decision, which has already been made. Eells' argument is no use to him, because it is unable to yield this equality at any earlier stage. We saw that Fred can make use of his knowledge of his beliefs and desires only when he knows what choice they'll lead him to make – by which time, in this case, it is too late to re-assess their bearing on whether he has the gene.

Jeffrey's defence of BDT does better here. Jeffrey adds to BDT the principle that an agent's decision should be 'ratifiable': i.e., roughly, that the decision should still seem correct, after the agent has chosen but before he or she actually acts. As Jeffrey notes, the principle comes into play in these odd cases in which a person's conditional credences may change in the interval between choice and action. Jeffrey concentrates on the prisoner's dilemma, but in terms of Fred's decision, the argument will go like this. Fred should ask himself, 'If I choose to smoke (or abstain), then before I carry out that choice, will it seem the right one? Or will I then hope to be prevented from carrying out my choice?' Believing, as above, that the gene is correlated with the *choice* to smoke, and only thereby to the *action*, Fred knows that once he has chosen, he will take $P(C/S) = P(C/-S)$. He'll already have the news about the gene, and whatever that is, would then prefer to smoke. So his decision will be ratifiable only if he has chosen to smoke.

Like Eells' argument, Jeffrey's relies on an agent perceiving a gap between choice and action. If on the contrary Fred regards his choices as certain indicators of his actions then he can make no sense of wondering whether it would be better for him to smoke, given that he's decided not to – except as wondering whether he should have *chosen* to smoke, which is not what Jeffrey wants.

Ratifiability thus trades on the gap between the choice which is indicative of a physiological state, and the action whose consequences are what the agent values. It is not difficult to imagine other cases in which this gap is closed. Suppose, for example, that Fred's attitude to smoking stems entirely from pleasurable consequences of the *choice*. Fred thinks that choosing to smoke will make him feel like a cowboy (and that this will be enjoyable) even if something – his doctor, perhaps – prevents him carrying out the choice. Once he's chosen to smoke, the cards are down. He couldn't care less whether he's able to put the choice into effect; ratifiability finds no foothold.

Worse still, the principle of ratifiability gives the wrong answer in another quite familiar kind of decision problem. Imagine you are tempted by a deadly sin – by gluttony, say, to take one of the more attractive ones. You must weigh the pleasures of the flesh (or comestibles of your choice) against the likelihood of eternal damnation. Suppose also that you believe that the Judgement will be just, in that your fate will depend on whether you freely chose gluttony, rather than on whether you actually stuffed your face. If you chose gluttony but couldn't get your hands on the food, you'll be damned; if you chose abstinence but somehow were forced to eat, you'll be saved. Hell being what it is, you should clearly abstain. But suppose you choose accordingly. Knowing that you won't be damned (on that score, at any rate), you will sensibly hope to be unable to carry out your choice – i.e., to have the food forced on you. This decision is thus not ratifiable. If on the other hand you choose gluttony then, your soul condemned, you'll ratify your decision to indulge your body. Hence in this sort of case (we might call it a 'Faust problem'), Jeffrey's principle gives the wrong answer.

The distinguishing characteristic of a Faust problem is that an agent has conflicting desires with respect to the effects of an action, on the one hand; and a different effect of the *choice* of that action, on the other. In other words, these cases are identical to the medical Newcomb problems, except in that the evidential significance of the choice here stems from its effects rather than its causes. Evidently, this difference makes all the difference to how we should handle such decisions: as we decide, we can take gluttony to be probabilistically relevant to damnation, and hence should abstain; but we can't take smoking to be probabilistically relevant to our possession of the gene, and hence should smoke. Jeffrey's principle of ratifiability ignores the causal basis of evidential significance, and is thus unable to draw this distinction.

Jeffrey's defence improves on Eells' in one respect: its stress on the need for hypothetical reasoning. However it reasons hypothetically about the wrong sort of thing. What matters is not the ratifiability of a hypothetical *choice* (for the Faust problem shows that correct choices need not be ratifiable), but rather the consistency of a hypothetical set of conditional credences. In other words, hypothetical reasoning is needed, in a reductio argument, to show that Fred can't apply his general statistical belief (1) to his own case.

Eells' defence, on the other hand, seems to me to be correct in both aim and motivation. It aims to show that the statistical generalisations underlying the medical Newcomb problems cannot be applied by agents to their own cases (as the case against BDT requires). And it is motivated by Eells' observation that

> any approach . . . that does not deal directly with the nature of the causal relation between the [physiological state] and the symptomatic act in a Newcomb situation cannot really go to the heart of the matter.[7]

It fails only in appealing to the wrong connection between the causal situation and the agent's statistical inference. What matters is not the fact that (in virtue of the nature of the causal relation) Fred's reasons will 'screen off' the probabilistic relevance of his action to his genes. It is rather the fact that if he were to take his action to be probabilistically relevant to his possession of the gene, this belief would be a causal factor in deciding his action. We saw that the belief would hence invalidate the judgement on which it was itself based. For in reasoning from effect to cause, as in any probabilistic inference, we must take into account all the relevant evidence: in other words, here, all the causal factors of which we are aware.

4. TWO-BOX PROBLEMS AND THE PRISONER'S DILEMMA

I want to finish with a brief discussion of the bearing of the defence of BDT offered here on the other two main types of Newcomb problems. The original Newcomb problem, the two-box case, goes like this: I have a choice of the contents of one or both of two boxes, one transparent and the other opaque. The transparent one contains $1000, and I believe that a very reliable predictor of my actions has placed $10,000 in the opaque box if and only if she has predicted that I will not take the contents of the transparent box. What should I do? 'Two-boxers' argue that I should take both boxes, since I'll then get $1000 more than if I just take the opaque box. 'One-boxers' say that that's foolish, since I'll then get only $1000, instead of $10,000.

If I'm an adherent of BDT, then the answer depends on whether, *as I make my choice*, I can take my action to be probabilistically relevant to the actions of the predictor (and hence to the contents of the opaque box). If I can, so that there's a high probability that if I take only the opaque box it will contain $10,000 (and a low probability that if I take both, the opaque one will contain the money), then BDT tells me to take just the opaque one.

Can I take my action to be relevant? It depends on how I see the causal relation between the prediction and my action; on what I take to explain the reliability of the predictor. If the prediction is a cause of my action, or itself the effect of some cause which it and my action have in common, then the case is like a medical Newcomb problem. Hence I am unable to take my action to be probabilistically relevant to the actions of the predictor, for the reason described above. In this case BDT advises me to take both boxes.

If the prediction isn't a cause of my action (or the effect of a common cause), then a judgement concerning the relevance of my action to the contents of the opaque box won't depend on an analysis of the causes of my action. Hence the judgement is not liable to be self-defeating, as in the medical case; and I can rationally act on the basis that by choosing only the opaque box, I *ensure* (or at least make it probable) that it contains $10,000. In this case BDT will advise me to so act. And so I should, it seems to me.

Thus in the two-box Newcomb problem, the correct decision depends on details left unspecified in the usual description of the case. (This point has been made by Mackie (1977), who gives an illuminating survey of the many different readings of the original problem.) Where necessary, the defence I have described is applicable. It shows that BDT gives what is intuitively the correct answer, in those interpretations of the case in which this might be doubted.

The prisoner's dilemma goes like this. Arrested on some charge, I am told that another prisoner and I are to be questioned, separately, about the alleged offence. If neither of us confesses, we'll both get five years in prison. If we both confess, we'll both get ten years. If one of us confesses, he or she will go free, while the other gets fifteen years. I believe that we are likely to act in the same way. What should I do? 'Confessors' say that whatever the other prisoner does, I'm better off if I confess; I can't affect what she does, so confess I should. 'Nonconfessors' say that since the probability that she and I will act differently is low, what matters is the relative value of the two cases in which we act alike; and here the case in which I don't confess is clearly preferable.

Once again, what matters is whether from my *agent's* perspective, my choice of action is probabilistically relevant to hers. This depends on what I take to be the basis of the correlation between us. There are elaborations of the story which do preserve such relevance. An obvious one is that in which I believe that my action has a causal influence on hers, by some physical or psychic means. However, this case is always intended to be ruled out. A more interesting case is that in which I believe that she and I are constructed similarly, in such a way that (being soul-mates, if not cell-mates) we tend to *think* similarly. This means that external factors being equal, I have reason to take what I am thinking as an indication of what she is thinking.

In these circumstances, can I consistently take my choice to be probabilistically relevant to hers? Suppose I do: i.e., that I judge that $P(-S/-I) > P(-S/I)$ (where 'I' is 'I confess' and 'S' is 'She confesses'). As in the earlier cases, this judgement will be a causal factor, influencing my decision. However, in this case the judgement relies not on an assessment of the *particular* causes of my action, but on the *general* principle that the other prisoner's reasons for acting are likely to be the same as mine. It is as if I believe that when there are two fires in the same area on the same night, they probably have the same cause. Inferences resting simply on this belief are insensitive to changes in my beliefs about the causes of *one* of the two fires. If for example I decide that the insurers will probably have to pay either on both buildings or on neither, this judgement is not invalidated when I learn that one of the fires was started by a cowboy smoking in bed (unless, of course, I have independent grounds for thinking that the other fire didn't start this way).

Hence my judgement that $P(-S/-I) > P(-S/I)$ is not self-invalidating. On the contrary, when I decide that my action will probably be the same as that of my fellow prisoner I simply infer that she has probably come to the same conclusion. This conclusion tends to confirm, rather than conflict with, the judgement on which it was based.

The present defence of BDT thus admits cases in which BDT tells the prisoner not to confess. Some people will think that it thus admits too much, for the rational choice is always to confess. I suspect that the dispute here is not about what it is rational to do in any fully specified decision problem. Rather it is about the *possibility* of cases of the kind just described. These cases depend on the prisoner's possessing a belief that the other prisoner will probably reason in the same way as he does himself. The belief must be strong enough to survive the reasoning process, where everything turns on its being correct; and where the penalty, if it is false, may be an extra ten years in prison. Whatever the evidence on which such a belief is based, there will be a strong temptation to ignore it, when it comes to the point, and to confess to minimise one's losses. The temptation is enhanced by the fact – inferred from the belief itself – that the other prisoner is subject to the same temptation.

I don't need to insist on it (to defend BDT), but it seems to me that the evidence could be such as to make this temptation irrational. One thing that makes me think so is that if we alter the values (but not the ordering) of the various punishments, the temptation becomes much less attractive. Suppose, say, that if we both confess we get fourteen years, that if neither of us confesses we get a short suspended sentence, and that otherwise it's nothing and fifteen years, as before. Here I think we would each be much more inclined to trust a belief that we would both reason in the same way and hence act alike. There is little incentive to 'play safe', by confessing; and we are more inclined to appeal to the mentioned belief, to justify not confessing.

To me this suggests that our reluctance to rely on the belief in the usual case does stem from the high cost of being wrong. I don't mean that the reluctance is irrational – on the contrary, it is doubtful whether any belief is immune from reasonable doubt, if enough hangs on it. I mean rather that for any given set of punishments (ordered as usual), it can be rational not to confess, if the evidence that the other prisoner will act as you do is sufficiently strong.

In such a case you are justified in taking your action to be positively relevant to that of the other prisoner. The figures being right, BDT would advise you to confess; and so would I. Otherwise BDT tells you to confess, and its opponents agree.

In summary, it seems to me that whenever one is justified in taking a contemplated action to be probabilistically relevant to some state of affairs, one ought to take this relevance into account, as BDT prescribes, in calculating the expected value of the action. The cases in which this prescription has seemed *clearly* irrational are those in which the apparent relevance stemmed from some common typical cause of the type of action and state of affairs in question. My argument shows that in such cases the relevance is only apparent, resting on a failure to appreciate the peculiarity of the free agent's point of view.

This leaves cases in which the prescription is not *clearly* irrational. Here the lack of clarity seems largely a result of a failure to adequately describe the decision problems concerned. I think that when they are adequately described, the prescription is seen to be worth taking. I cannot show that the prescription will deal with all future Newcomb problems; but I think it handles his present ones.

NOTES

* I am grateful to Peter Menzies for long discussions on this topic; and to Hugh Mellor and to a referee for comments on an earlier version.

[1] 'Advice' suggests that BDT is *normative*, and I follow convention in treating it as such. But nothing here hangs on this. If BDT is really a *descriptive* theory of the origins of action (as Ramsey [1978, pp. 75–76] thought) the Newcomb cases are still a problem; to which I think the argument of this paper provides a solution.

[2] See Lewis (1981) for a recent version of the argument, and for references to others. The supplemented theory may itself be Bayesian in character.

[3] The same role might be played by a nonphysiological state, so that 'medical' is really an inappropriately restrictive description of the kind of decision problem. However, the medical examples are common, and vivid.

[4] See tor example the remarks of Mellor (1971, p. 53), and the references mentioned there.

[5] For more on this feature of probabilistic judgments, and its implications, see my (1983a) and (1983b).

[6] I advocate such a view in 'Conditional Credence', forthcoming in *Mind*, 1986.

[7] Eells (1982), p. 149.

REFERENCES

Eells, E.: 1981, 'Causality, Utility, and Decision', *Synthese* **48**, 295–329.

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

Jeffrey, R. C.: 1965, *The Logic of Decision*, McGraw-Hill, New York.

Jeffrey, R. C.: 1981, 'The Logic of Decision Defended', *Synthese* **48**, 473–492.

Jeffrey, R. C.: 1983, *The Logic of Decision*, 2nd. ed., University of Chicago Press.

Lewis, D.: 1979, 'Prisoner's Dilemma is a Newcomb Problem', *Philosophy and Public Affairs* **8**, 235–240.

Lewis, D.: 1981, 'Causal Decision Theory', *Australasian Journal of Philosophy* **59**, 5–30.

Mackie, J. L.: 1977, 'Newcomb's Paradox and the Direction of Causation', *Canadian Journal of Philosophy* **7**, 213–225.

Mellor, D. H.: 1971, *The Matter of Chance*, Cambridge University Press.

Nozick, R.: 1969, 'Newcomb's Problem and Two Principles of Choice', in N. Rescher (ed.), *Essays in Honour of Carl G. Hempel*, Reidel, Dordrecht, pp. 114–146.

Price, H.: 1983a, '"Could a Question be True": Assent and the Basis of Meaning', *The Philosophical Quarterly* **33**, 354–364.

Price, H.: 1983b, 'Does "Probably" Modify Sense', *Australasian Journal of Philosophy* **61**, 396–408.

Ramsey, F. P.: 1978, *Foundations*, D. H. Mellor (ed.), Routledge, London.

School of Philosophy
University of New South Wales
P.O. Box 1 Kensington,
NSW 2033
Australia