Causation, Chance and the Rational Significance of Supernatural Evidence

Huw Price*

June 24, 2010

Abstract

Newcomb problems turn on a tension between two principles of choice: roughly, a principle sensitive to the causal features of the relevant situation, and a principle sensitive only to evidential factors. Two-boxers give priority to causal beliefs, and one-boxers to evidential beliefs.

A similar issue can arise when the modality in question is chance, rather than causation. In this case, the conflict is between decision rules based on credences guided solely by chances, and rules based on credences guided by other sorts of probabilistic evidence. Far from excluding cases of the latter kind, Lewis's Principal Principle explicitly allows for them, in the form of the caveat that credences should only follow beliefs about chances in the absence of "inadmissible evidence".

In this paper I begin by exhibiting a tension in Lewis's views on these two matters. I present a class of decision problems – some of them themselves Newcomb problems – in which Lewis's view of the relevance of inadmissible evidence seems in tension with his causal decision theory. I offer a diagnosis for this dilemma, and propose a remedy, based on an extension of a proposal due to Ned Hall and others from the case of chance to that of causation. The remedy suggests a new view of the relation between causal decision theory and evidential decision theory, namely, that they stand to each other much a chance stands to credence, as objective and subjective faces of the same practical coin.

1 Two decision rules

The original Newcomb problem goes something like this. God offers you the contents of an opaque box. Next to the opaque box is a transparent box containing $1,000. God says, "Take that money, too, if you wish. But I should tell you that it was Satan who chose what to put in the opaque box. His rule is to put in $1,000,000 if he predicted that you wouldn't take the extra $1,000, and nothing if he predicted that you would take it. He gets it right about 99% of the time."

*Centre for Time, Department of Philosophy, University of Sydney, NSW 2006, Australia.

	Opaque box empty	Opaque box full
Take one box	\$0 (0.01)	\$1,000,000 (0.99)
Take both boxes	\$1,000 (0.99)	\$1,001,000 (0.01)

Table 1: The standard Newcomb problem, with evidential probabilities.

Famously, this problem brings to a head a conflict between two decision rules. In the original presentation of the problem, these rules were *Dominance* and *Maximise Expected Utility,* but for many purposes it has turned out to be more interesting to represent the disagreement as a clash between two different ways of calculating expected utility (and hence two different versions of the rule Maximise Expected Utility).

(i) *Evidentially-grounded* expected utility ('V-utility'):

$$\mathrm{EV}(A_i) = \Sigma_j \mathrm{V}(O_j)\mathrm{P}_{evidential}(O_j|A_i)$$

(ii) *Causally-grounded* expected utility ('U-utility'):

$$\mathrm{EU}(A_i) = \Sigma_j \mathrm{V}(O_j)\mathrm{P}_{causal}(O_j|A_i)$$

Here $\{O_j\}$ and $\{A_i\}$ are the relevant sets of Outcomes and Acts, respectively. $\mathrm{P}_{evidential}(O_j|A_i)$ is the "epistemic" conditional probability of the Outcome O_j given the Act A_i. And $\mathrm{P}_{causal}(O_j|A_i)$ is what we may call the *causal* conditional probability – intuitively, the intent is that $\mathrm{P}_{causal}(O_j|A_i) \neq \mathrm{P}_{causal}(O_j|\neg A_i)$ only if O_j is *causally dependent* on A_i (positively or negatively, as the case may be).[1]

It is a simple matter to show that in the decision problem like that described above, these two rules give different recommendations. On the one hand,

$$\mathrm{EV}(\textit{One-box}) = \$0 \times 0.01 + \$1{,}000{,}000 \times 0.99 = \$990{,}000$$

while

$$\mathrm{EV}(\textit{Two-box}) = \$1{,}000 \times 0.99 + \$1{,}001{,}000 \times 0.01 = \$11{,}000$$

so that the rule Maximise V-utility recommends taking only the opaque box. On the other hand,

$$\mathrm{EU}(\textit{Two-box}) = \$1{,}000 \times \alpha + \$1{,}001{,}000 \times (1-\alpha) = \$1{,}000 + \mathrm{EU}(\textit{One-box})$$

(where $\alpha = \mathrm{P}_{causal}(\textit{the opaque box is empty})$), so that – by Dominance reasoning – the rule Maximise U-utility recommends taking both boxes.

Philosophers disagree about which of these two decision rules provides the rational strategy. Among famous "two-boxers", or "Causalists", is David Lewis, who describes the issue as follows:

[1] In other words, $\mathrm{P}_{causal}(O|A)$ is what Joyce (1999, 161) calls the "causal probability" for O given A, and writes as $\mathrm{P}(O\backslash A)$. Joyce notes that while this this notion "has been interpreted in a variety of ways in the literature, ... the common ground among causal decision theorists is that [it] should reflect a decision maker's judgements about her ability to causally influence events in the world by doing A."

> Some think that in (a suitable version of) Newcomb's problem, it is rational to take only one box. These one-boxers think of the situation as a choice between a million and a thousand. They are convinced by indicative conditionals: if I take one box I will be a millionaire, but if I take both boxes I will not. Their conception of rationality may be called *V-rationality;* they deem it rational to maximize *V,* that being a kind of expected utility defined in entirely non-causal terms. Their decision theory is that of Jeffrey [(1965)].
>
> Others, and I for one, think it rational to take both boxes. We two-boxers think that whether the million already awaits us or not, we have no choice between taking it and leaving it. We are convinced by counterfactual conditionals: If I took only one box, I would be poorer by a thousand than I will be after taking both. . . . Our conception of rationality is *U-rationality;* we favor maximizing *U,* a kind of expected utility defined in terms of causal dependence as well as credence and value. Our decision theory is that of Gibbard and Harper [(1978)] or something similar. (Lewis 1981b, 377)

Elsewhere, Lewis affirms his commitment to two-boxing like this:

> [S]ome—I, for one—who discuss Newcomb's Problem think it is rational to take the thousand no matter how reliable the predictive process may be. Our reason is that one thereby gets a thousand more than he would if he declined, since he would get his million or not regardless of whether he took his thousand. (Lewis 1979, 240)

In this paper I call attention to an apparent tension between this aspect of Lewis's views – his Causal Decision Theory (CDT) – on the one hand; and his professed position concerning chance, evidence and rational credence, on the other. In his discussion of the Principal Principle, Lewis allows that chance does not provide an exceptionless constraint on rational credence: on the contrary, he holds, an agent who has access to "inadmissible information" may be rational to allow her credences to be guided by that information, rather than by her knowledge of the relevant objective chances. I want to argue that this amounts to recommending Evidential Decision Theory (EDT) rather than CDT, in a particular class of decision problems. Some of these problems are themselves Newcomb problems, and in these cases, Lewis's view of the relevance of inadmissible information seems literally to support one-boxing. Lewis's commitments about these two matters thus seem in conflict with one another.

As we shall see, Lewis himself was certainly aware of the class of decision problems in question. He qualifies his own version of CDT by stipulating that it is not intended to apply to them. But if I am right that these cases include a particular class of Newcomb problems – a class in which Lewis's own views on the relevance of inadmissible evidence recommend one-boxing – then excluding them by fiat from CDT is hardly a satisfactory solution, from a two-boxer's point of view. It amounts to withdrawing from the field, in some of the cases in which the conflict with EDT matters most.

In the latter part of this paper, I suggest a resolution of this tension, extending a proposal by Ned Hall concerning the Principal Principle. Hall argues that

Lewis's qualification of the Principal Principle to deal with inadmissible information is unnecessary and undesirable. Better, he argues, to say that there is no such thing as inadmissible information: properly understood, chance relates to expert credence in such a way that such cases simply don't arise.

I want to point out that this move is analogous to a view of causation that some writers have found attractive in standard Newcomb cases, viz., that of arguing that where evidential reasoning really does recommend one-boxing, so too does causal reasoning, properly understood.[2] This view thus interprets causation in such a way that CDT and EDT make the same recommendations, even in Newcomb cases. I shall propose that this approach be seen as arguing that causation is an "expert function" for a deliberating agent, in much the way that Hall treats chance as an expert function for a betting agent – an *evidential* agent, in both cases.

In the case of chance, the expert function's outputs are prescriptions for credences. In the case of causation, according to my proposal, the outputs are prescriptions for the conditional credences of Outcomes given Acts, as required by an agent who acts in accordance with EDT. I shall call these *agentive* conditional credences, or *agentive* conditional probabilities.[3] My proposal is thus that causal dependence stands to agentive conditional dependence just as chance stands to credence according to Hall's proposal.

Hall's view of how chance stands to credence is very close to Lewis's own, of course. They differ, essentially, only in their treatment of cases of inadmissible evidence. Similarly for causation, I think. Someone sympathetic to the analogy I wish to draw might nevertheless prefer an analogue of Lewisean chance to an analogue of Hallean chance, in the case of causation. That is, it is compatible with the view that these modal notions (chance and causation) are both experts, first and foremost, that we might have grounds (from physics, perhaps) to prefer a conception of the modal facts which allows they may in principle float free of rational agency, in unusual cases. Exceptional cases, by their very nature, force us to make a trade-off between accuracy and conceptual tidiness. Lewis's picture of chance is tidier than Hall's, but pays for it by having to admit exceptions to the Principal Principle, in some (very) unusual cases.

This trade-off needs to be negotiated for causation, too, according to my proposal. In this case, the very unusual cases are Newcomb problems.[4] Note that in the case of chance, our ranking of tidiness compared to accuracy does not affect our judgements about rational credence and rational action. Hall and Lewis agree what credences are rational, in the presence of what Lewis calls inadmissible evidence; even though they disagree about whether the chances

[2] See, e.g., Price (1991, 1993) for a view of this kind.

[3] As we shall see, the label "agentive" here does triple duty: it marks the fact that these are probabilities an *agent* needs, according to EDT; the fact that they are probabilities conditional on *acts;* and, crucially, the fact that they are assessed from the *agent's* distinctive epistemic perspective.

[4] More precisely, *some* of the various decision puzzles called Newcomb problems; including the classic Predictor case described above, under at least some of its possible disambiguations. I shall have more to say later about other cases, such as the more realistic "medical" Newcomb problems, and other versions of the Predictor case.

are such that these credences follow from the Principal Principle itself. Similarly for causation, I shall argue. A preference for accuracy will deliver a view of causation such that CDT recommends one-boxing in the standard Newcomb problem; while a preference for tidiness will deliver the verdict that Newcomb problems are strange cases in which causal beliefs and rational decision behaviour do not keep step. But the rational policy is to one-box, in either case.

In my view, much of the force of the Newcomb puzzle derives from the fact that we have allowed our modal and evidential notions to drift apart in this way, without being aware of the diagnosis. Once we understand these facts, we can either eliminate these cases altogether, via Hall's prescription and its causal analogue, or we can choose to live with them. But in the latter case the right option is the one that Lewis himself grasped for chance: rationality and modal metaphysics part company, and the rational choice is to one-box.

Finally, I note that although my proposal is motivated by an apparent tension in Lewis's views, and disagrees with Lewis about the rational policy in the classic Newcomb problem, it is in other respects Lewisean in spirit. In particular, it aims to extend to causation the well-judged balance between subjectivism and objectivism, or pragmatism and metaphysics, that Lewis himself offers us in the case of chance.

2 A chancy Newcomb problem?

On the face of it, Newcomb problems turn on a conflict between *causal beliefs* and *evidential beliefs.* It is natural to ask whether the same kind of conflict can arise for other kinds of objective modality. In particular, can it arise for chance?

It is easy to see that it can. Suppose God offers you the payoffs shown in Table 2 on a bet on the outcome of a toss of a fair coin. It is a good bet either way, obviously, but a better bet Heads than on Tails.

	Heads	Tails
Bet Heads	\$100	\$0
Bet Tails	\$0	\$50

Table 2: A free lunch?

Now suppose that Satan informs you that although God told you the truth, and nothing but the truth, about the coin, He didn't tell you the *whole* truth. So far, this revelation shouldn't impress you. You were well aware that – as in the case of any event governed by (non-extreme) chances – there is a further truth about the actual outcome of the coin toss, not entailed by knowledge of the chances. "Tell me something I didn't know," you think to yourself.

"Okay," responds Satan, rising to this silent bait, "I bet you didn't know this: on those actual future occasions on which *you yourself* bet on the coin, it comes up Tails about 99% of the time. (On other occasions, it is about 50% Tails.)"

What strategy is rational at this point? Should you assess your expected return in the light of the objective chances? Or should you avail yourself of Satan's further information? Call this the chancy Newcomb problem, or *Chewcomb problem,* for short. (See Table 3.)

Presumably we should use our rational credences to calculate the expected values of the available actions, but there are two views as to what the rational credences are. According to one view, the rational credences are given to us by our knowledge of the objective chances, in accordance with the Principal Principle. In this case, Satan's contribution makes no difference to the rational expected utility, and we should bet Heads, as before. According to the other view, our rational credence should take Satan's additional information into account, in which case (as it is easy to calculate), our rational expected return is \$1 if we choose Heads and \$49.50 if we choose Tails.

	Heads	Tails
Bet Heads	\$100 (0.01)	\$0 (0.99)
Bet Tails	\$0 (0.01)	\$50 (0.99)

Table 3: The Chewcomb problem (with Satanic evidential probabilities)

Which policy should we choose? If we turn for guidance to the masters, we find that Lewis's discussion of the constraint that a theory of chance properly places on rational credence – the discussion in which he formulates the Principal Principle – seems initially to recommend the second policy in such a case. What it explicitly recommends – the point of Lewis's exclusion to the Principal Principle for the case in which one take oneself to have "inadmissible evidence" – is that in such a case one's rational credences follow one's beliefs about the new evidence, rather than remaining constrained by one's theory of chance. As Lewis puts it, it would be an "obvious blunder" to take the Principal Principle to dictate the following credence:

> C(the coin will fall heads/it is fair and will fall heads in 99 of the next 100 tosses) = 1/2. (Lewis 1994, 485)

So Lewis takes it for granted that someone who has inadmissible evidence should base their credences on that evidence, rather than on their beliefs about the relevant chances. In the present case, then, this suggests that we should assess our options in the Chewcomb problem simply by replacing credences based on chances with credences based on the 'Satanic' evidential probabilities.

However, it is easy to configure the Chewcomb problem so that this recommendation seems in tension with that of CDT. Lewis's own (1981a) formulation of CDT is based on a partition $K = \{K_0, K_1, \dots\}$ of "dependency hypotheses", each of which specifies how what an agent cares about depends on what she does. The expected U-utility of an act act A, is then calculated as a sum of the values of each option allowed by this partition, weighted by the corresponding unconditional probabilities:

$$\mathrm{U}(A) = \Sigma_i \mathrm{P}(K_i)\mathrm{V}(A\&K_i).$$

Thus in a standard Newcomb problem, where it is specified that the agent has no causal influence over the contents of the opaque box, the dependency hypotheses may simply be taken to be:

K_0: The opaque box is empty.
K_1: The opaque box contains \$1,000,000.

We then calculate the causal utilities for taking both boxes and taking one box as follows:

$$\mathrm{U}(\textit{Two-box}) = \mathrm{P}(K_0)\mathrm{V}(\textit{Two-box} \,\&\, K_0) + \mathrm{P}(K_1)\mathrm{V}(\textit{Two-box} \,\&\, K_1)$$

$$\mathrm{U}(\textit{One-box}) = \mathrm{P}(K_0)\mathrm{V}(\textit{One-box} \,\&\, K_0) + \mathrm{P}(K_1)\mathrm{V}(\textit{One-box} \,\&\, K_1).$$

By dominance reasoning, the result is that $\mathrm{U}(\textit{Two-box}) > \mathrm{U}(\textit{One-box})$.

To apply this framework to the Chewcomb problem, what should we take the dependency hypotheses to be? If we assume that because the outcome is the result of a toss of fair coin, it is not causally influenced by the way we choose to bet, then again the dependency hypotheses seem to take a simple form:

K_H: The coin lands Heads
K_T: The coin lands Tails.

As we shall see, Lewis's own formulation of dependency hypotheses in such a case is a little more complicated; but it produces the same results, for present purposes, so for the moment we may work with this simpler alternative.

The next issue concerns the probabilities $\mathrm{P}(K_H)$ and $\mathrm{P}(K_T)$. Lewis stresses that if CDT is to remain distinct from EDT, we need to use unconditional probabilities at this point, not probabilities conditional on action:

> It is essential to define utility as we did using the unconditional credences $C(K)$ of dependency hypotheses, not their conditional credence $C(K|A)$. If the two differ, any difference expresses exactly that news-bearing aspect of the options that we meant to suppress. Had we used the conditional credences, we would have arrived at nothing different from V. (1981a, 12)

This means that if we set up the example so that Satan's inadmissible evidence yields *unconditional* probabilities, Lewis can consistently allow that CDT yields the recommendation to bet on Tails. But it doesn't have to be set up like this. We can specify that the information that we learn from Satan doesn't tell us that $\mathrm{P}(K_H) = 0.01$, for example, but only that $\mathrm{P}(K_H|\textit{We bet}) = 0.01$.

If this isn't already clear, we can easily modify the example slightly to make it explicit. As I set things up above, Satan's information concerns the class of cases in which the agent bets at all (either on H or T) – and it might be argued this yields an unconditional probability for an agent who already knows herself to be taking part in the game. However, if we specify that the agent has a third choice – viz., not to bet at all – then the situation is unambiguously

of the new sort (i.e., it involves conditional probabilities). In this case, Satan's information certainly concerns a "news-bearing aspect" of the act of choosing to bet rather than not to bet. Accordingly, Lewis's CDT then seems to require that we use $\mathrm{P}(K_H) = \mathrm{P}(K_T) = 0.5$ for calculating $\mathrm{U}(\textit{Bet } H)$, $\mathrm{U}(\textit{Bet } T)$ and $\mathrm{U}(\textit{No bet})$, for there are no other *unconditional* probabilities available. The upshot is that CDT recommends the first of the two policies we distinguished above: it recommends betting on H, on the grounds (i) that H pays a higher return, and (ii) that K_H and K_T are taken to be equally likely, *in the only sense this decision theory allows to be relevant.*

We thus have two versions of the Chewcomb game, the Conditional and the Unconditional version, where the difference consists in the availability of the No Bet option. The problem for Lewis takes the form of a trilemma (see Table 4). If he recommends betting Heads in both cases, the Unconditional case appears to be in violation of his own policy on the relevance of inadmissible evidence. If he recommends Tails in both cases, the Conditional case appears to be in violation of his own version of CDT. While if he recommends different policies in each case, the difference itself seems implausible. After all, the case has been set up so that it seems obvious that a rational agent will choose to bet – it's a free lunch. And the mixed case seems to yield different recommendations, depending on whether the agent is allowed first to choose to bet and then to choose *which* bet, or has to make both choices at the same time.[5]

Unconditional	**Conditional**	**Problem for Lewis**
Heads	Heads	Conflict with policy on inadmissible evidence
Tails	Heads	Implausible difference in recommendations
Tails	*Tails*	Conflict with CDT

Table 4: Two Chewcomb games – policies and problems.

2.1 Following Lewis more closely

I noted above that in applying Lewis's CDT to the Chewcomb game, we used a different choice of dependency hypotheses. When Lewis considers the formulation of CDT in indeterministic worlds, he takes the relevant dependency hypotheses to be counterfactual conditionals whose antecedents are the actions an agent is considering, and whose consequences are full specifications of chances for relevant outcomes. In the Chewcomb game, these counterfactuals take a simple form. In the Unconditional version of the game they are simply:

$\textit{Bet Tails} \mathrel{\square\!\!\rightarrow} \mathrm{Ch}(H) = \mathrm{Ch}(T) = 0.5$
$\textit{Bet Heads} \mathrel{\square\!\!\rightarrow} \mathrm{Ch}(H) = \mathrm{Ch}(T) = 0.5.$

In the Conditional version of the game, where the agent has the option not to bet, we need to add:

[5] As Arif Ahmed pointed out to me, this amounts to a violation of Independence. In the mixed case, the agent prefers betting on Tails to betting on Heads if she does not have the option not to bet at all, but betting on Heads to betting on Tails if she does have the latter option.

No Bet $\Box\!\!\rightarrow \mathrm{Ch}(H) = \mathrm{Ch}(T) = 0.5$.

Since the consequent is identical in all three cases, we may take the dependency hypothesis to be simply a specification of the chances – i.e., in this case, the proposition (FC) that the coin is fair. It follows that:

$$\begin{aligned} \mathrm{U}(\textit{Bet Tails}) &= \mathrm{V}(\textit{Bet Tails} \,\&\, \textit{FC}) \\ \mathrm{U}(\textit{Bet Heads}) &= \mathrm{V}(\textit{Bet Heads} \,\&\, \textit{FC}) \\ \mathrm{U}(\textit{No Bet}) &= \mathrm{V}(\textit{No Bet} \,\&\, \textit{FC}) \end{aligned}$$

How should we calculate the utilities on the right hand side of these expressions? The first two expressions require a calculation of expected utility, and here, once again, we face the issue of what probabilities we use in the calculation. If we use simply the chances, the option of betting Heads will maximise U-utility. If we use the Satanic evidential probabilities, the option of betting Tails will do so.

Once again, the problem is that both policies seem defensible, in Lewisean terms. Lewis's views on the relevance of inadmissible information seem to recommend betting Tails. But in the Conditional game, at least, this again has the effect of making U-utility sensitive to "that news-bearing aspect of the options that we meant to suppress" (as we saw that Lewis himself put it, in the case of the weights on the dependency hypotheses).

2.2 Discussion

So, as I say, there seems to be a tension here, from Lewis's point of view. I offer the following diagnosis of the difficulty. Newcomb problems are decision problems in which evidential policies seem to give different recommendations from causal policies, and CDT is the decision theory that cleaves to the causal side of the tracks. Cases of inadmissible evidence are cases in which chance-based credences lead to different recommendations from (total-)evidence-based credences, and Lewis takes it for granted that the rational policy is to cleave to the evidential side of the tracks. Chewcomb problems are decision problems in which both these things happen at once. It follows that the two kinds of cleaving are liable to yield different recommendations in these cases. At least, they are liable to do so as long as our causal judgements cleave to our judgements about objective chance. But to give that up – to allow, instead, that causal judgements might properly follow the "merely evidential" path – would be to abolish the very distinction on which Newcomb problems rely (or at least to move in that direction).

As I noted, Lewis recognised that cases like the Chewcomb problem lead to special difficulties. In the paper in which he presents his own version of CDT, he compares it to several earlier proposals by other writers. One of these proposals had been presented in unpublished work by Sobel, and Lewis's discussion of Sobel's theory closes with the following remarks:

> But [Sobel's] reservations, which would carry over to our version, entirely concern *the extraordinary case of an agent who thinks he may somehow have foreknowledge of the outcomes of chance processes.* Sobel gives no reason, and I know of none, to doubt either version of the thesis except in extraordinary cases of that sort. Then if we assume the thesis, it seems that we are only setting aside some very special cases – cases about which I, at least, have no firm views. (I think them much more problematic for decision theory than the Newcomb problems.) So far as the remaining cases are concerned, it is satisfactory to introduce defined dependency hypotheses into Sobel's theory and thereby render it equivalent to mine. (Lewis, 1981a, 18, my emphasis)

However, I don't know whether Lewis saw the difficulty that these cases pose for his own views – a difficulty that turns on a tension between his attitude to the relation between causal judgements and evidential judgements, on the one hand, and chance judgements and evidential judgements, on the other.[6]

In any case, the move of simply setting aside these cases can hardly be regarded as satisfactory, by Lewis's own lights. His own policy on inadmissible evidence seems to yield a clear recommendation in the Unconditional version of the game; and hence a clear recommendation in the Conditional case, too, given the implausibility of the mixed strategy. We thus have a class of Newcomb-like problems in which Lewis's policy on inadmissible evidence concurs with EDT; and in which CDT escapes defeat only by withdrawing from the field.[7]

3 Making the analogy closer

So far, our Chewcomb problems have been Newcomb-like in two respects. Even the Unconditional version of the Chewcomb game is analogous to a Newcomb problem, in that it provides a case in which modal beliefs and evidential beliefs yield different recommendations. (The difference between Chewcomb and Newcomb is that the modality concerned is chance rather than causality.) But the introduction of the Conditional game produced a decision problem which is Newcomb-like in a more direct sense, namely, that it involves an apparent conflict between CDT and evidential reasoning.[8]

On the face of it, we can go even further. We can produce a Chewcomb game whose decision table looks exactly like that of the classic Newcomb problem.

---

[6] Lewis also notes the difficulty posed by these cases in correspondence with Wlodek Rabinowicz in 1982, saying:

> It seems to me completely unclear what conduct would be rational for an agent in such a case. Maybe the very distinction between rational and irrational conduct presupposes something that fails in the abnormal case. (Lewis, 1982: 2)

(I am grateful to Howard Sobel for alerting me to the existence of this correspondence, and to Wlodek Rabinowicz, Stephanie Lewis and the Estate of David K. Lewis, for giving me access to it.)

[7] True, they are "extraordinary cases." But so, too, is the classic Newcomb problem. Once CDT has become fickle in this way, what reason do we have to trust it in that case?

[8] Provided, at least, that the latter is understood in the light of Lewis's policy on inadmissible evidence.

Suppose that God offers you the contents of an opaque box, to be collected tomorrow. He informs you that the box will then contain $0 if a fair coin to be tossed at midnight lands Heads, and $1,000,000 if it lands Tails. Next to it is a transparent box, containing $1,000. God says,"You can have that money, too, if you like." At this point Satan whispers in your ear, saying, "It is definitely a fair coin, but my crystal ball tells me that in 99% of future cases in which people choose to one-box in this game, the coin actually lands Tails; and ditto for two-boxing and Heads."

	Heads	Tails
Take one box	$0 (0.01)	$1,000,000 (0.99)
Take two boxes	$1,000 (0.99)	$1,001,000 (0.01)

Table 5: A better free lunch?

Assuming you are convinced that both God and Satan are telling the truth, what is the rational decision policy in this case? Here the evidential and causal recommendations seem to be exactly as in the original Newcomb problem, as presented above. Your action will not have any causal influence on whether there is money in the opaque box, apparently. How could it do so, when that is determined by the result of a toss of a fair coin?[9] Yet you have (or, what is relevant here, you *believe* yourself to have) evidence of a strong *evidential* correlation between your action and the result of the coin toss, such that you are much more likely to get rich if you one-box.

In this case there is no unconditional version of the game, to highlight the tension in Lewis's position in the way that we did above. (The parallel with the original Newcomb problem depends on the fact that the high evidential probability of money in the opaque box is conditional on the agent's only choosing that box.) However, a similar effect can be achieved in a different way. Suppose that the agent makes her choice by choosing a ticket – the one-box ticket, or the two-box ticket – and is then free to sell the ticket and associated expected returns on the open market. How much is each ticket worth, to someone who has access to the inadmissible evidence provided by Satan? Lewis's policy concerning inadmissible evidence dictates that the one-box ticket would be more valuable than the two-box ticket; and hence that an agent with access to this option has a clear reason to one-box. But if the market value of the ticket is itself based on rational expectations, how could the addition of this factor make a difference to the rationality of the original choice? Without such a difference, the policy concerning inadmissible evidence leads to a recommendation in tension with CDT.

[9] In the next section I suggest an understanding of causation which challenges this claim, but the moment I simply want to point out that someone who says that the agent has no causal influence on the contents of the opaque box in the standard Newcomb problem, should say exactly the same here.

3.1 Remembering the counterfactuals

When this version of a Chewcomb game is played without the tickets, two-boxers will make their standard response. They will argue that whatever payout the one-boxer receives, it will always be true that *had* she two-boxed, she would have received the same payout plus $1,000. Later (in §6), I want to suggest a reply to this move, which also depends on an analogy with the case of chance and inadmissible evidence. First, however, to the relevance of the tickets: granting the Causalist this response in the game without the tickets, do the tickets make a difference?

We can phrase the issue in terms of regret. Consider a one-boxer, who sells her one-box ticket for its market value, in the light of Satan's information – i.e., for $\$990,000$. What should she believe about what her return *would have been*, had she two-boxed; does she have grounds for regret?

It is clear that in that case she would not have had a one-box ticket to sell, but rather a two-box ticket. At that point, she could have sold the ticket for its market price, or held it and waited for the outcome of the game itself. Its market price in the counterfactual case depends on whether we take Satan's information to be available to the market in that world (in the same form as in the actual world). The best case for the value of the two-box ticket is that we do not, in which case its value is $501,000. So the option of selling a two-box ticket does not make it the case that the agent would have been better off if she had two-boxed.

But what about the option of waiting for the outcome of the game itself? In this case, the agent's return would have been $1,001,000 if the coin had landed Tails, and $1,000 if it had landed Heads. How should the agent weight these possibilities, in considering the counterfactual case? Presumably, the guiding thought is supposed to be that the result of the coin toss *would have been* just what it *actually* is – this is where the lack of causal influence makes itself felt. If the coin toss has not yet taken place in the actual world, the agent's expectation about what it will be will derive from Satan's inadmissible information. In other words, she will think that there is a probability 0.99 that the actual result is Tails, in which case her return in the counterfactual case would have been $1,001,000; and a probability 0.01 that it would have been $1,000. This yields a better expectation than the actual market value of her one-box ticket, by $1,000.

So a Causalist is entitled to object that the introduction of tickets really makes no difference. An agent who one-boxes and then sells her ticket for its market value has still foregone an even more valuable option: viz., that of two-boxing the very same game. The tickets simply make vivid something that the two-boxer long since acknowledged, namely, that one-boxers will in fact do better in Newcomb games. But the sense in which one-boxing is nevertheless irrational remains intact. In any individual case, a one-boxer *could* have done even better.

This argument turns on the fact that the present case is crucially different from the Conditional version of the previous Chewcomb game, where there is no such Dominance argument. A possible strategy for CDT is therefore to try

to hold the line here, while conceding the previous games (both Conditional and Unconditional) to the Evidentialist. It would still need to be explained how CDT can be formulated so as to follow EDT in the Conditional version of previous Chewcomb problem, without also endorsing one-boxing in the present case. But the counterfactuals associated with Dominance seem to mark a line at which a stand might be made.

I want to respond to this suggestion by calling attention to another possible analogy with the issue of the relation between chance and inadmissible evidence. It seems to me that Evidentialists typically concede too much concerning counterfactuals to their Causalist opponents. The analogy with the case of chance, and a proposal made in that context by Ned Hall, together suggest a more forceful response.

4 One-boxing via the Hall way?

Ned Hall (1994, 2004) recommends that we replace Lewis's Principal Principle with a modified principle, requiring that rational credences track *conditional* chances: chances *given our evidence.* At first sight, this may seem to eliminate the problem cases. What matters isn't simply the chance of the coin coming up Tails, but the chance of it doing so given the extra information that Satan has whispered in our ear.[10] On the face of it, then, this seems to be irenic resolution of the dilemma posed by the Chewcomb problems: they are pseudo-problems, artifacts of a mistaken rule for aligning credence with one's beliefs about chance: in one sense a victory for Evidentialism, but a face-saving victory for the Evidentialists' opponents, too, in that it maintains that they never had any good reason to disagree.

But things aren't so simple. To see this, we only have to imagine a proponent of a view of conditional chance according to which it makes no difference what Satan whispers in one's ear: the real metaphysical chance of a fair coin's landing Tails is insensitive to such supernatural vocalisations (our objector insists), and so the shift to conditional chances makes no difference.In such a case, it remains an issue whether rational (conditional) credence should be guided by chance alone, or by other kinds of information.

I think that the real relevance of Hall's treatment of the Principal Principle to our present concerns lies in a different feature. Drawing on earlier proposals by Gaifman (1988) and van Fraassen (1989, 197–201), Hall suggests that "chance plays the role of an expert":

> Why should chance guide credence? Because—as far as its *epistemic* role is concerned—chance is like an expert in whose opinions about the world we have complete confidence. (1994, 511)

In his (2004) paper Hall elaborates on this idea by distinguishing two kinds of expert—roughly, the kind of expert (a "database-expert", as Hall puts it) who simply knows a lot, and

[10] Or the information *that* Satan has provided this information, perhaps.

> the kind of expert who earns that status not because she is so well-informed, but rather because she is extremely good at *evaluating the relevance* (to claims drawn from the given subject matter) *of different possible bits of evidence.* (2004, 100)

"Let us call the second kind an analyst-expert," Hall continues. "[S]he earns her epistemic status because she is particularly good at evaluating the relevance of one proposition to another." (2004, 100) Hall takes chance to be the second kind of expert: "I claim that *chance is an analyst-expert*." (2004, 101)

Thus for Hall it simply becomes a matter of definition that chance and reasonable credence cannot come apart, once we have conditionalised on all our evidence (including, in particular, what Lewis treats as "inadmissible" evidence). And it is this stipulation, rather than the conditionalisation move itself, that ensures that there cannot be a genuine Chewcomb problem – a genuine case in which chance and evidential reasoning come into conflict.

I've stressed this point because it is the latter aspect of Hall's view – the view that chance is an analyst-expert – that seems to me analogous to an attractive resolution of the original Newcomb case. In Hall's terminology, the resolution turns on the idea that causal dependence should be regarded as an analyst expert about conditional *evidential* dependence, *assessed from the agent's point of view.*[11] In other words, B is causally dependent on A just to the extent that an expert agent would take $\mathrm{P}(B|A) \neq \mathrm{P}(B)$, in a calculation of the V-utility of bringing it about that A, in circumstances in which the agent is not indifferent to B.

Evidently, the effect of this proposal is going to be to support one-boxing – at least in certain cases – but to regard this as what maximising U-utility properly recommends, too, when causal dependence is seen for what it really is. Consider our last Chewcomb game (Table 5), for example. The argument for the causal independence of the outcome (Heads or Tails) on our choice of one or two boxes was that in either case, the chance of Heads and Tails remains the same. (How could we exert a causal influence, we reasoned, if we couldn't influence the chances of the outcomes concerned?) According to Hall's prescription, however, the conditional chance of Tails given one-boxing *is* higher than conditional chance of Tails given two-boxing (and higher than the conditional chance of Heads given one-boxing). And since we can choose which antecedent to "actualise" in these various conditional chances, we can also influence the resulting unconditional chance, in the obvious sense. Thus the intuitive connection between chance and causation now works in the opposite direction. It suggests that we do have influence and causation – in particular, causal dependence of Outcomes on Acts – in the sense of those terms that now seems appropriate, given that chance is to be understood as an expert function.

Let's call this proposal *Causation-linked Evidentialism* – or *Clevidentialism,* for short (to remind us of the role of expert functions). As we have just seen, the proposal holds that the classic Newcomb problem is not a case in which CDT and EDT come apart, but simply a case in which the causes are not

[11] More on the importance of this qualification in a moment.

what they seem. We might take this to imply that it is not really a Newcomb problem at all, on the grounds that as Joyce (1999, 152) puts it: "It is part of the definition of a Newcomb problem that the decision maker must believe that what she does will *not* affect what the psychologist has predicted." But this is a terminological matter, on a par with that as to whether, in the light of Hall's proposal, we want to continue to speak of "inadmissible evidence." The substantial point is that in the classic (so-called) Newcomb problem, the view proposes an understanding of the causal structure of the case such that CDT and EDT agree in recommending one-boxing.

Of course, it is not news that CDT recommends one-boxing if the agent's choice affects what the predictor puts in the boxes. Retrocausal variants of the original Newcomb problem are familiar. (Indeed, they date back to Nozick's original (1969) paper.) What Causation-linked Evidentialism adds to this background is a proposal about the nature of causal dependence itself, such that the Newcomb problem cannot *but* be retrocausal, if there is genuine evidential dependence of the predictor's behaviour on the agent's choice, from the agent's point of view.[12]

This proposal may seem an obvious non-starter, blocked by familiar and ordinary cases "medical" cases, in which it is uncontroversial that causal dependence and evidential dependence do not align with one another, in the way that "Clevidentialism" suggests. I turn to this objection in a moment. But before that, I want to stress one more lesson to be drawn from the analogy with Hall's view of chance: in neither case, for chance or for causation, is Hall's view or its causal analogue the only game in town. In either case, we might have grounds to prefer a modal notion which could drift apart from expert credence and strategy, in unusual cases. I merely want to claim that in this eventuality, once we recognise it for what it is, it should seem clear that the rational goes with the evidential notion, not with the modal notion. This already seems unremarkable to us in the case of chance, where Lewis's offers one such modal notion and regards it as obvious that it diverges from rational credence, in exceptional cases involving inadmissible evidence. I am proposing (and will be arguing) that it should seem just as unremarkable in the case of causation.

5 A cigarette at bay?

Whatever the appeal of Causation-linked Evidentialism in Chewcomb cases, it may seem that there are familiar Newcomb problems in which causal dependence and evidential dependence are clearly distinct. Consider the famous case of the Smoking Gene, for example, in which an agent believes that there is a gene which predisposes both to smoking and cancer, ensuring that these two outcomes are positively correlated. In general, the fact that someone is a smoker this indicates that she is more likely than otherwise to have the gene, and hence more likely than otherwise to develop cancer. EDT is therefore held

[12] And if the agent really has a choice in the matter, of course.

to recommend that even if such an agent prefers smoking to not smoking, other things being equal, she should decide not to smoke, in order to minimise the evidential probability that she will develop cancer (and thereby maximise her expected V-utility). But it would add idiocy to irrationality, surely, to try to justify this recommendation by claiming that causation should be understood in such a way that this agent can cause herself to lack the gene.

Indeed it would, and I make no such claim. Instead, I propose that in these familiar cases, the agent is making a mistake – a mistaken probabilistic inference, not a mistaken decision – if she concludes that her choice as to whether to smoke is probabilistically relevant to whether she carries the gene in question, *from her own point of view.*

In support of the claim that this proposal is at least not obviously absurd, I appeal first to the authority of some of my (traditional) Causalist opponents, who recognised long ago that Evidentialists could get at least close to this claim. Here is Brian Skyrms, for example, describing what was then becoming known as the Tickle Defence – an argument that in cases such as the Smoking Gene, an agent should indeed regard her action as probabilistically independent of whether she carries the gene: "There is a defense for [the Evidentialist] which can be pushed very far, but not, I think, far enough."[13]

Lewis himself goes even further:

> I reply that the Tickle Defence does establish that a Newcomb problem cannot arise for a fully rational agent, but that decision theory should not be limited to apply only to the fully rational agents. Not so, at least, if rationality is taken to include self-knowledge. May we not ask what choice would be rational for the partly rational agent, and whether or not his partly rational methods of decision will steer him correctly? (1981a, 10)

It seems to me that at least in hindsight, this assessment positively *invites* a response framed in terms of expert functions. More about this in a moment, but before that, a couple of preliminary points.

Obvious no longer

First, a remark on the relevance of the dialectic of these old discussions to the present case. As noted, the acknowledged successes of the Tickle Defence (and its successors) do much to meet the objection that there are cases in which it is *obvious* that my proposed analogue of Hall's suggestion will attribute causal dependency, where actually there is none. These successes force the Evidentialist's opponents to retreat in one of two directions: either to less familiar and less realistic examples, in which it is correspondingly less plausible to say that the causal structure is not a matter for debate; or, as noted, to less rational agents, about whom there is inevitably an issue about the nature of their irrationality. So long as we Evidentialists can find an alternative interpretation to *decision-theoretic* irrationality, these agents need not trouble us.

[13] See Skyrms (1980, 130). Skyrms adds, "I have heard this defense independently from Frank Jackson, Richard Jeffrey, David Lewis, and Isaac Levi." More recent versions of this argument include those of Horgan (1981), Eells (1981, 1982, 1984), Horwich (1985) and Price (1986, 1991).

Both these points are well made by Paul Horwich. Writing about analogues of medical Newcomb problems that might evade the Tickle Defence, Horwich notes that

> such scenarios do not constitute clear counterexamples to the evidential principle because they are extremely unrealistic—in exactly the same way as New comb's problem itself—and cannot, therefore, provide the material for authoritative intuitions. (1985, 435)

Later, taking up the Lewis's objection that "decision theory should not be limited to apply only to the fully rational agents," Horwich points out that

> this criticism neglects a certain systematic equivocation in the evaluation of actions. They are always judged in relation to desires and beliefs which are themselves susceptible to evaluation. Therefore, an act may be criticized as irrational because it was based on irrational beliefs, even though it was correct relative to those beliefs. (1985, 438)

We'll return to this observation below, and amplify it with reference to the analogy with chance and credence.

A causal shortcut to evidential virtue

Next, an advantage of the Clevidentialist proposal, with respect to medical Newcomb problems, not shared by more orthodox versions of Evidentialism, such as that of Horwich. Suppose, as the Clevidentialist claims, that causal information *just is* information about evidential dependencies, from the agent's point of view. Then at least in familiar and uncontroversial cases, such as that of the Smoking Gene, an agent with a proper grasp of the causal concept and firm beliefs about the causal structure of a particular case, can no more be confused about the evidential dependencies, than, according to Lewis, an agent with firm beliefs about chance, and a good understanding of the concept, can be confused about the associated credences. For causation as for chance, the Clevidentialist insists, confusion in typical cases is simply an indication that the agent in question does not have a proper grasp of the concept in question.

Indeed, the Clevidentialist can go even further. Having interpreted causal information in this evidential manner, she can allow that it is a considerable advantage of CDT, in many cases, that it operates directly with this encoded form of evidential information. Like computers programmers more comfortable in C++ than in machine code, ordinary fallible agents find it much easier to operate at the causal level of description – much easier, thereby, to avoid the perils of probabilistic inference, a task which most of us are prone to get wrong.

But this convenience comes with a cost. In unfamiliar circumstances, it may *seem to us* that the causal facts and evidential facts pull in opposite directions. In familiar cases, we rely on various associations between causal facts and other features of situations – in other words, we take various criteria to be grounds for ascribing or withholding causal claims (i.e., really, on this view, evidential claims). But in unusual circumstances, these criteria can be a poor guide to the

evidential structure of the case in question. We are habituated to regarding them as good guides to causal structure, and so it seems that causal and evidential dependency are coming apart. But it is an illusion, generated by the mistaken assumption that we were dealing with two distinct kinds of information in the first place – by the fact that we have allowed the causal realm to take on a life of its own, distinct from our evidential point of view.

The classic Newcomb problem plays on this danger, by presenting us with just such as case. It gives us causal information, or simply allows us to arrive with our ordinary causal picture of the kind of case it describes; and then presents us with conflicting evidential information – "conflicting", in the sense that were we aware of the true evidential significance of the causal information, we would see that we have simply been presented with an incoherent example. No wonder it is so hard to decide what to do!

Is CDT to EDT what chance is to credence?

Once again, the analogy with chance is helpful at this point. According to my Clevidentialist, *causal dependence* stands to the conditional subjective probabilities needed by EDT, much as *chance* stands to the subjective probabilities, or credences, required by decision makers whose rational behaviour is modelled by an unconditional decision theory of Savage's sort. Savage's (1954) theory is a *subjective* rational decision theory: it prescribes rational behaviour for agents with a given set of credences and preferences, but remains silent about the rationality of those credences and preferences themselves. The Principal Principle steps into the latter gap (in the case of credence), imposing a rationality constraint on credences themselves, in the light of the agent's beliefs about chances (or in the light of the *facts* about chances, if we wish to interpret the Principal Principle as an objective constraint on rational credence).

Note that in principle we could combine these two levels, formulating an analogue of Savage's theory directly in terms of beliefs or even facts about chances. Why wouldn't that be preferable? Well, because it would formalise Horwich's "systematic equivocation in the evaluation of actions", for one thing; and thereby, arguably (more on this in §6.1), obscure something very important about the "subjective", "pragmatic" or "practical" foundations of the concept of chance itself – the sense in which the concept has its roots in subjective decision.

Let SDT_{ch} be such an "objective" version of Savage's decision theory, formalised in terms of chances, and SDT_{ev} the familiar subjective version. The Clevidentialist regards the relation between CDT and EDT as closely analogous to that between SDT_{ch} and SDT_{ev}. CDT is simply the "objectified" version of EDT, which runs together two issues: the *subjective* issue of the rationality of a decision policy, given certain preferences and conditional credences, and the *objective* (or at least *less subjective*) issue of the rationality of certain conditional credences – credences of outcomes given actions – given the facts, or the agent's beliefs, about causation. (CDT then has the analogous disadvantage to SDT_{ch}, in that it obscures the practical, subjective roots of the concept of causation itself, and invites Horwich's "systematic equivocation.")

Two dimensions of expertise

With all this in hand, let us now return to Lewis's remark that "decision theory should not be limited to apply only to the fully rational agents." Lewis is right, of course, that (subjective) decision theory should not simply fall silent, in the case of an agent whose beliefs and preferences are not fully rational. But this is compatible with the insight that decision theory itself is supposed to be an expert, and therefore *intolerant* of irrationality in decisions made *on the basis* of those beliefs and preferences. Agents themselves may be irrational at this step, of course, but decision theory aims to codify the standard that rational agents are *trying* to meet. So there is a sense in which decision theory does "apply only to the fully rational agents." Taken *descriptively,* it does not apply – at least not strictly – to agents who are not fully rational in making decisions on the basis of their credences and preferences. Taken *prescriptively,* it does tolerate irrationality; but only in the *acquisition* of beliefs and preferences, not within its own domain.

Subjective decision theory is the expert we consult as we try to do the best with the credences and preferences we actually possess. But to what experts do we turn to avoid the kind of irrationality that decision theory itself tolerates? That is, for help with the credences themselves? Hall has already given us a large part of the answer, perhaps all of it. We need two experts: first, the database-expert, who knows all the evidence that we ourselves would have, under idealisation; and second, the analyst-expert, who knows what credences to assign on the basis of that evidence.

However, in the cases for which we need Jeffrey's subjective decision theory rather than Savage's – cases with conditional dependence of States and hence of Outcomes on Acts – these two Hallean experts have a special job to do. They need to collaborate to consider the special epistemic situation of the deliberating agent – to determine the rational conditional credences, from her own point of view, of States given contemplated Acts. Because the task involves this collaboration, it will be helpful both for the two experts and for their clientele to create a single "shopfront", through which requests for guidance may conveniently be channelled. The Clevidentialist proposes that this expert shopfront – the Agency Guidance Agency, perhaps – is causal dependence.

5.1 Resuscitating the medical objections?

This program for aligning CDT and EDT would be undermined by a genuine medical Newcomb problem – i.e., a realistic case in which it was clear that the relevant causal dependencies really differed from the evidential dependencies, from the agent's point of view. With such a case in hand, critics could fairly object that Clevidentialism amounts, at best, simply to changing the meaning of "causal dependence", in a way that obscures the genuine difference between CDT and EDT.

Realistic cases seem to be hard to find, however, and this is certainly good news, from the Clevidentialist's point of view. But shouldn't the Clevidentialist expect even better news? After all, if the Clevidentialist wants to claim that

there is some sort of conceptual tie between causal dependence and agentive evidential dependence, shouldn't it be more than a contingent matter that there are no cases in which these notions clearly diverge, in the way that the proposal seeks to disallow?

In response to this challenge, I want first to emphasise, once again, that Clevidentialism need not claim that it offers the *only* acceptable understanding of causal dependence. On the contrary, it should acknowledge that the notion of causation has other conceptual ties, and – in a good Quinean spirit – allow that the preservation of these ties might seem preferable, in some quarters, when the concept comes under pressure for revision in strange cases.[14] By the resulting lights, it will indeed seem that causal dependence can "come apart" from conditional evidential dependence, even when the latter is assessed from the agent's point of view.

The significance of this loophole should not be exaggerated, however. For one thing, it would not help in the face of a genuine medical Newcomb problem, where it would be implausible to maintain that the notion of causal dependence was under any sort of conceptual pressure. For another thing, Clevidentialism is committed to an "in principle" claim of a weaker sort, even in strange cases. If causal dependence and conditional evidential dependence are allowed to part company in this way – if causation's other conceptual ties are judged to be more worth preserving, in strange cases – the Clevidentialist wants to maintain that it should be nevertheless clear, at least when all the cards are on the table, that rationality goes with conditional evidential dependence, rather than with causal dependence. In other words, it should be clear that if causation is taken this way, such cases provide counterexamples to CDT.

So Clevidentialism has some work to do, and I want to suggest a line of attack. It relies on a feature of the landscape where CDT and EDT already find common ground, in the thought that in certain cases, Evidentialists will actually do better than Causalists. This is the basis of the famous *Why Ain't You Rich?* argument against two-boxing. What will be important for my argument will be that the Clevidentialist and her opponent will agree about when EDT leads to greater riches (i.e., higher expected V-utility) than CDT, under certain specified circumstances (namely, that it is a random matter which decision policy an agent follows).

Agreement on this matter means that we have a criterion acceptable to both sides for dividing Newcomb-like decision problems into two kinds of cases. In one kind of case, where randomly-assigned Evidentialism does lead to riches, Clevidentialism will be able to appeal to a novel response to the Causalist's usual objection to the *Why Ain't You Rich?* argument, to argue that

[14]Thus Michael Dummett (1954, 32ff.), though defending the conceptual possibility of circumstances that would support deliberation for past ends, argues that these would be cases of "quasi-causation", rather than genuine causation. (The principle he thereby preserves is that remote causes should *begin* a process that leads to their effects.) Dummett's terminological choice illustrates the present point very nicely. The fact that he takes quasi-causation to support means–end reasoning makes it clear that in his view, rational decision does not cleave strictly to *causal* dependence, in unusual cases of this kind.

in these cases it is irrational not to follow the Evidential policy, even if one prefers for other reasons not to label the case in question as one of genuine "causal" dependence. In the other kind of case, where randomly-assigned Evidentialism does not lead to riches, the Clevidentialist will be able to argue that there is no agentive evidential dependence, in the relevant sense; and hence that her own version of EDT does not differ from CDT, with causation standardly understood.

If this argument works, it provides both the insurance the Clevidentialist seeks about the non-existence of "realistic" medical Newcomb problems, and support for her claim that rationality goes with EDT, even if causation is understood in such a way that EDT and CDT diverge, in "unrealistic" cases. So first, then, to the Clevidentialist's response to the Causalist's objection to *Why Ain't You Rich?*, in the classic Newcomb case.

6 We're all Causalists now

The standard Causalist response to the *Why Ain't You Rich?* argument goes something like this: "Sure, one-boxer, you're rich. But if you *had* two-boxed in those same games, you would have been even richer." Here is Joyce's version of the point, for example:

> [H]aving gotten the $1,000,000, [the one-boxer] must believe that she would have gotten it whatever she did, and thus that she would have done better had she taken the $1,000. So, while she may feel superior to [the two-boxer] for having won the million, [she] must admit that her choice was not a wise one compared to *her own* alternatives. The "If you're so smart why ain't you rich?" defense does nothing to let [the one-boxer] off the hook; she made an irrational choice that cost her $1,000. (Joyce 1999, 153)

But unlike a traditional Evidentialist, who accepts the Causalist's conception of the modal landscape, my Clevidentialist will simply *deny* that "she would have gotten [the million] whatever she did." On the contrary, as she understands the counterfactuals – regular *causal* counterfactuals, as she sees them, not backtrackers[15] – she would have received only $1,000, had she two-boxed. It is the two-boxer who is irrational in this counterfactual sense, by the Clevidentialist's lights: *had* the two-boxer one-boxed instead, she would have had the million.

At this point, a lively discussion is likely to ensue about who has the "proper" notions of causation and counterfactual dependence. But as we have already seen, the Clevidentialist is prepared for this. "Keep your notions of causation and counterfactual dependence, if you wish," she says to her traditional Causalist opponents:

[15]This is how this line differs from that of Horgan (1981). As Horgan puts it, "I do recommend acting *as if* one's present choice could causally influence the being's prior prediction, but my argument does not presuppose backward causation." (1981, 340–341)

> "But recognise, with me, how we came to the present juncture, where we need to make a choice. In the case of chance, a 'super-natural' source of information about the future (i.e., a source not envisaged by our usual physical theories) would confront us with a choice about how to continue to use the notion of chance: we could hold fixed our notion of chance, and deal with the unusual cases by allowing an exception to the Principal Principle; or we could modify our notion of chance, and preserve the universality of the Principal Principle. But whichever we choose, it is clear that it makes no difference to rational betting behaviour. Either way, we should not ignore the new information – that's why the first choice requires an exception to the Principal Principle, after all. The only wrong option is the choice that muddles and mixes the two right options, by holding fixed the standard notion of chance, *and* insisting on the universality of the Principal Principle.
>
> Similarly for causation. We can imagine cases – the traditional Newcomb problem is one – which confront us with a choice about how to continue to use the notions of causation and counterfactual dependence. Again, we have two choices. We can hold fixed the traditional notions of causation and counterfactual dependence, and allow exceptions to CDT (which is the analogue, here, of the Principal Principle); or we can preserve the universality of CDT, by allowing that the causal structure of these strange cases is not what initially we took it to be. Again, this choice makes no difference to the rational behaviour in such a case: either way, it is to one-box. The only wrong option is the choice that muddles and mixes the two right options, by holding fixed the standard notion of causation, *and* insisting on the universality of CDT."

The traditional Causalist will want to disagree rather vigorously at this point, of course. She will want to defend this "mixed" option – and to deny that it involves any sort of "muddle"! To do so, she needs to *explain* the relevance of causality, as *she* understands it, to rational strategic deliberation. To see what is at issue here, it is will be helpful, once again, to compare the analogous problem in the case of chance.

6.1 The limits of objectivism

Some philosophers feel that there is a problem about explaining the link between beliefs about objective probabilities and rational credence, of the kind encapsulated in the Principal Principle. David Papineau, for example, calls this connection the "Decision-Theoretical Link":

> We base rational choices on our knowledge of objective probabilities. In any chancy situation, a rational agent will consider the difference that alternative actions would make to the objective probabilities of desired results,

> and then opt for that action which maximizes objective expected utility. (1996, 238)

"Perhaps surprisingly," Papineau continues, "conventional thought provides no agreed further justification [for this principle]":

> Note in this connection that what agents want from their choices are desired *results,* rather than results which are objectively *probable* (a choice that makes the results objectively probable, but unluckily doesn't produce them, doesn't give you what you *want*). This means that there is room to ask: *why* are rational agents well advised to choose actions that make their desired results objectively probable? However, there is no good answer to this question .... Indeed many philosophers in this area now simply take it to be a primitive fact that you ought to weight future possibilities according to known objective probabilities in making rational decisions. ... It is not just that philosophers can't agree on the right justification; many have concluded that there simply isn't one. (1996, 238)

Not all views of probability will agree with Papineau that there is any such a problem, however. One tradition, variously known as *subjectivism, pragmatism,* or *Bayesianism,* regards it as a pseudo-problem, generated, in effect, by starting one's account of probability in the wrong place. Provided we *start* with the insight that probabilistic models are guides to decision-making under uncertainty in particular domains, there's no further mystery as to why they may be used for that purpose. There is no primitive assumption needed, and no decision-theoretical Missing Link. There may be other interesting questions in the vicinity: e.g., about how, and why, such probabilistic models are linked to other kinds of models, such as those provided by physics in various domains; and about whether these links uniquely constrain the associated probabilistic models. But these are not the practical puzzle about why probability properly guides action. That isn't a puzzle at all, from the subjectivist point of view.

Another interesting issue, famously explored in Lewis's own account of chance, is the extent to which this subjectivist insight can be combined with an objectivist, or metaphysically realist, theory of chance. Lewis thought that it could be. In his account the tie to subjectivism consists in the fact that it is *definitional* of objective chances that they support the Principle Principal. If something doesn't do that, it isn't properly called chance. "A feature of Reality deserves the name of chance to the extent that it occupies the definitive role of chance," as Lewis (1994, 489) puts it.

But consider a view of chance of the kind Papineau has in mind, which is prepared to take it to be a primitive fact, needing no justification, that chance constrains rational credence in the manner described by the Principal Principle. We can imagine that such a view – emboldened by its own courage in making a stand on this point – might also dig in its heels concerning the rationality of bettings Heads, in the first version of our Chewcomb game, despite the availability of inadmissible evidence. "To hell with Satan," says this objectivist, thumping the table. "By betting Tails, you *irrationally* forgo an *equal chance* of a *greater reward*." Or in the past tense: "No matter that you actually won; you

were nevertheless irrational, because you sacrificed an equal chance of a greater reward." Or in the long run: "No matter that you have won many times, and are now rich; and that I, betting on Heads, am not rich. This is not a mark of irrationality on my part, but merely a sign that the rewards were reserved for the irrational."

I am not sure whether anyone actually thumps the table to this dialectical end, in the case of chance. But many people, including Lewis, staunchly defend what I take to be its analogue in the case of causation: that is, orthodox two-boxing. The full set of analogies is depicted in Table 6. On the right hand side are views that take modal beliefs to constrain practical rationality, even in cases of exceptional evidence. On the left hand side are views that construe practical rationality in evidential terms; typically combining this preference with some element of subjectivism about the associated modal judgements. In the middle are mixed positions, that allow that there may be exceptional cases in which evidence and objective modality part company, and in which practical rationality goes with the former. Lewis himself holds the mixed view in the case of chance, but the full modal priority view in the case of causation – and the combination creates internal difficulties, as we have already seen.[16]

	Evidential priority	**Modal priority with exceptions**	**Modal priority**
Chance	Hall	Lewis	The table-thumper
Causation	Clevidentialists	Horgan, Horwich	Two-boxers

Table 6: Three views of practical rationality.

Is my comparison of the orthodox two-boxer position to table-thumping objectivism about chance a fair one, or can causal objectivism do better than its probabilistic cousin? Can the causal objectivist *justify* (rather than simply assume as primitive) the claimed link between causal judgement and rational decision (and hence explain why the objective modality takes precedence, in case of conflict with exceptional evidence)?

What does the history of these debates tell us about the prospects for such an argument? It reveals a widespread acceptance, even on the part of two-boxers themselves, that there is no such argument to be found. Lewis himself remarks that the debate is "hopelessly deadlocked." (Lewis 1981a, 5) Elsewhere, he says:

> [I]t's a standoff. We [two-boxers] may consistently go on thinking that it proves nothing that the one-boxers are richly pre-rewarded and we are not. But [one-boxers] may consistently go on thinking otherwise. (1981b, 378)

Hopeless deadlock will be bad enough for present purposes, but I note in passing that Horgan (1981) argues persuasively for an even less promising

[16]The most important distinction here is the one marked by the double line. To the left of this line, evidence rules rationality, and modality plays a vice-regal role. (Evidence is the throne behind the powers, so to speak.) To the right of the this line, modality rule rationality, and evidence defers.

conclusion, from the two-boxers' point of view. He notes an apparently ineliminable circularity in their attempt to *justify* two-boxing, turning on the fact that attempts at justification always return to the same kind of counterfactuals. One-boxers do better, Horgan argues, by confining their attention to deliberation about *actuality*.[17]

Why is even deadlock bad news, from a two-boxer's point of view? For a reason which Lewis himself puts his finger on, with respect to "unHumean" theories of chance, more willing than he himself is to postulate metaphysical primitives:

> Be my guest—posit all the primitive unHumean whatnots you like. ... But play fair in naming your whatnots. Don't call any alleged feature of reality "chance" unless you've already shown that you have something, knowledge of which could constrain rational credence. I think I see, dimly but well enough, how knowledge of frequencies and symmetries and best systems could constrain rational credence. I don't begin to see, for instance, how knowledge that two universals stand in a certain special relation N* could constrain rational credence about the future coinstantiation of those universals. (Lewis, 1994, 484)

My Clevidentialist simply makes the same demand of an account of causation:

> Be my guest—posit all the primitive whatnots you like. But play fair in naming your whatnots. Don't call any alleged feature of reality "causation", or "counterfactual dependence", unless you've already shown that you have something, knowledge of which could constrain rational deliberation.[18]

Had Lewis himself been in a position to meet this challenge to his own account of causation and counterfactuals, he would have had the key required to break the deadlock between one-boxers and two-boxers. The fact that he thought the deadlock hopeless therefore supports my contention that two-boxers occupy a position analogous to that of hardline primitivist objectivists about chance – a further manifestation, in my view, of a deep tension in Lewis's own view.

Clevidentialism, by contrast, has precisely the advantages of Lewis's own subjectivism in the case of probability. By building its account of causality *on* deliberation, evidentially construed, it ensures that causality does not lose conceptual or practical touch with deliberation.

---

[17] Cf. Price & Weslake (2009), who note that Lewis's "deadlock" reflects the difficulty that accounts of counterfactuals such as Lewis's have in explaining the connection between counterfactuals and deliberation. Like Horgan, Price & Weslake argue that we do better if we begin with non-counterfactual modes of deliberation – with "material deliberation", as they term it.

[18] Is this demand is compatible with the concessive policy I have recommended earlier, viz., that of allowing alternate choices about the use of these concepts in exceptional cases, provided it is conceded (as Lewis himself concedes for chance in cases of inadmissible evidence) that the usual ties with rational action are broken in these cases? Yes, provided we read the demand as calling only for an explanation of the concepts' connection with rational action in normal cases. (Lewis must take it this way in the case of chance, of course, if his own view is not to fail the test.)

The Principal Principle can be regarded as a codification of the relation that something must bear to credence, to count as chance, or objective probability. As we might put it:

$\mathrm{SDT}_{ch} = \mathrm{SDT}_{ev}$ + PP.

In other words, PP is the rationality condition one needs to add to SDT_{ev}, to produce the "bundled", two-experts-in-one theory represented by SDT_{ch}.

In the same spirit, the Clevidentialist proposes that we should expect a codification of the relation that something must bear to conditional agentive credence, to count as causality – in other words, a rationality condition one needs to add to EDT to produce CDT (which is the two-experts-in-one version of conditional decision theory). What is this principle CP, such that

CDT = EDT + CP?

Essentially, it is the principle that in assessing one's agentive conditional credences for Outcomes given Actions, one should be guided by one's causal beliefs.

7 Random riches

Now to the task deferred above: using *Why Ain't you Rich?* as a point of agreement between Clevidentialism and traditional Causalism, in order to argue that Clevidentialism is as general as it needs to be – there are no nasty surprises, lurking around the corner.

In the standard Newcomb problem, two-boxers accept that one-boxers will get rich, and that they themselves will not. To make this thought a little more formal, suppose that it is proposed to allocate agents randomly to a one-boxer stream or a two-boxer stream (perhaps with an additional inducement, so that all parties agree that it is rational to play the game). Causalists and Evidentialists will agree that the expected return – the expected V-utility, in fact – for the one-box stream is much greater than for the two-box stream.[19]

In the medical cases, presumably, Causalists take a different view. They will not expect that agents randomly assigned to the No Smoking stream will have a lower incidence of the cancer gene than those assigned to the Smoking stream. If they themselves are the players in this random game, then their own conditional credences of having the gene, conditional on being assigned to the No Smoking stream or to the Smoking stream, are identical. By the Causalist's lights, in other words, there is no *Why Ain't you Rich?* challenge to be answered in this case: those who decline a cigarette will be no "richer" (i.e. healthier) on average than those who do not. On average, they will be worse off, once the denied pleasure of smoking is taken into account.

Somewhere between these two cases thus lies a boundary, by a regular Causalist's lights. On one side of the line, a randomly prescribed "Evidentialist"

[19] At any rate, I shall assume for the time being that this is the case. More later on variations of the classic Newcomb problem for which it is not the case.

choice leads to higher expected V-utilility. On the other side, it does not. So long as Clevidentialism can cleave to this same line, recommending *only* the choices that have higher expected V-utility in this random game, then when it differs from traditional Causalism, it will always be able to appeal to *Why Ain't you Rich?* (with the response outlined above in hand, to deal with the Causalist's objections).

To guarantee this result, the Clevidentialist needs to show that her agentive conditional probabilities make Outcomes dependent on Acts *only* in the kind of cases that pass this test – in other words, that she is entitled to ignore any evidential dependency that doesn't hold when her action is randomly chosen.

How might this result be established? The most direct option would be to maintain that it is simply a primitive, *constitutive* fact about the free agent's point of view that she regards her actions as "uncaused", in such a way that she is automatically committed to the claim that any evidential dependency between her Actions and Outcomes would survive if her actions were randomly chosen (this being simply one way in which her choices may be uncaused). Such a view of free action (without the claim that it is primitive) is proposed by Price (1993), who attributes it to Ramsey:

> Ramsey [identifies] what he takes to be the crux of the agent's perspective, namely the fact that from the agent's point of view contemplated actions are always considered to be *sui generis,* uncaused by external factors. As he puts it, "my present action is an ultimate and the only ultimate contingency." [Ramsey 1978, 146] I think this amounts to the view that free actions are treated as probabilistically independent of everything except their effects.[20] (Price 1993, 261)

Similar views are defended by Hitchcock (1996) and Joyce (2007). Hitchcock's version perhaps comes closest to regarding this as simply a primitive feature of free action: he suggests that we might regard it as a kind of "fiction", central to our practice of regarding ourselves as free agents.

Can we do better than regarding this as a primitive feature of agency, fictional or not? The Tickle Defence and its descendants comprise a sustained attempt to do better; to show that an agent's evidential perspective is *guaranteed* to have this distinctive character, in virtue of differences between her own epistemic situation and that of external observers. As we have already seen, Lewis himself offers an optimistic assessment of the prospect for this endeavour. In this context, where the task is provide Clevidentialism with a guarantee that there are no cases in which her own policy need differ from that of the random game, we need not be concerned about Lewis's remarks about the unsuitability of the Tickle Defence for imperfectly rational agents.[21]

[20] Price goes on to suggest that this point be read "in reverse", so that we regard the effects of an action, as the Clevidentialist here proposes, as those outcomes properly regard as (positively) conditionally probabilistically dependent on the action, in the context of deliberation.

[21] Though perhaps Lewis's caution about the coherence of free choice for a perfectly rational agent might lend support to fictionalism after all.

7.1 The epistemics of deliberation

Joyce (2007) offers an alternative to the Tickle Defence, which turns on the special epistemic authority of an agent's deliberations concerning her own actions. Joyce takes the crucial point to be that "an agent's beliefs about her own free decisions and actions provide evidence for their own truth." (2007, 558) Such beliefs are "self-supporting", as Joyce puts it.[22]

An alternative way to put this thought, preferable in my view, is to say that there is an important sense in which, as she deliberates, an agent simply *does not have* knowledge, beliefs or credences about the action in question. In this form, Joyce's thought corresponds to familiar view, nicely characterised by Wlodek Rabinowicz in the following passage:[23]

> On this view, the relevant distinction is between the *first-person* perspective of a practical deliberator and the *third-person* perspective of an observer. While the observer can predict what I will do, I can't, insofar as I deliberate upon what is to be done. Deliberating in this way is incompatible with predicting the outcome of deliberation. To put it shortly, *deliberation crowds out prediction.* (Rabinowicz 2002, 91)

One route to this thesis, as to Joyce's version, turns on the special authority of the deliberating agent, concerning her own actions. This authority trumps any merely predictive knowledge claim about the same matters, rendering it necessarily unjustified.

A familiar application of this point, in a superficially different guise, is Dummett's (1964) observation that an agent can coherently believe that she can *affect* some past state of affairs only if she takes herself to be unable to *know* whether the state of affairs in question obtains, before she decides whether to perform the action she takes to be required to bring it about. Dummett's point applies equally to effects in any temporal relation to actions, of course. It is especially striking in the retrocausal case only because we typically assume that we do have epistemic access, at least in principle, to states of affairs in the past. In the usual future-directed case, the presumption goes the other way. We assume that we do not have epistemic access, in advance, to the future effects of contemplated actions. Joyce's point – already implicit, I think, in Dummett's discussion – is that the epistemic character of deliberation mandates this assumption.

Concerning the source of this special epistemic authority of the deliberating agent about her own actions, Jenann Ismael proposes that it is simply a special case of a familiar form of "epistemic degeneracy", as she calls it, characteristic of self-representing representations in general:

[22] Joyce himself makes these points in order to block an objection from Richard Jeffrey, to the effect that Newcomb problems are not really cases of free choice at all, because the agents involved know too much about their own actions. Joyce is thus defending Causalism against an Evidentialist objection. In my view, however, the point ultimately counts in favour of Evidentialism, by showing how the Evidentialist can justifiably ignore spurious evidential correlations: they fall into the category of evidence properly ignored by the deliberating agent.

[23] Rabinowicz himself opposes this view; which he attributes particularly to Spohn and Levi.

> Alethic constraints . . . on representational activity are empty when applied reflexively, i.e., when what is being represented is the representational act itself. The most familiar examples of this degeneracy are self-representing linguistic performances: "I promise to X", "I declare that Y". Such performances are perfectly good representational acts. They have truth conditions that can fail to obtain; someone else can certainly falsely ascribe a promise to me, and I can misrepresent my own past promises and declarations. But because they provide their *own* truthmakers, they are unconstrained at the time that they are made. They are self-fulfilling. . . .
>
> [This] degeneracy is unavoidable for any system that includes its own activity in the field of representation. The desire to tell the truth in general will not guide my answer the question "Will I A?" and the ordinary epistemic procedures for getting information about whether [someone] A'd will not apply. Guidance has to come from elsewhere. . . .
>
> The emptiness, or degeneracy of alethic constraints . . . when applied to one's own actions opens up the space for deliberation. I believe that it captures the sense in which, from the point of view of the participant in a dynamical process, her own actions have the status of what Ramsey called "an ultimate contingency". (Ismael 2007, §3)

As Ismael remarks, the epistemic authority of deliberation seems to explain the striking feature of action noted by Ramsey, viz., its apparent "contingency," from the agent's point of view. Indeed, the point is easily made by adapting Dummett's condition for the coherence of a belief in retrocausality. An agent who took her own future actions to be *caused by* an earlier state of affairs of which she had knowledge, before she made up her mind what to do, would be in exactly the same incoherent epistemic position as Dummett's agent, who took herself to have knowledge of the past *effects* of a future action. (The difference between the two cases is simply the direction of the causal link, which makes no difference to what matters here, namely, the evidential significance of the link in question.)

A little more generally, this argument establishes that as she deliberates, a free agent must take her action to be uncorrelated with anything of which she might in principle have knowledge, before she makes up her mind what to do. (For any such correlation could be "bilked", to put the point in the terminology familiar from discussion by Dummett and others of the retrocausality case.)

7.2 Discussion

This appeal to the special epistemic situation of a deliberating agent provides much of what the Clevidentialist needs. It offers a powerful argument, grounded on what is arguably an essential feature of deliberation, for the conclusion many evidential correlations are quite properly ignored from a deliberating agent's point of view.

But does it go far enough? Couldn't there be some *non-causal* correlation between an agent's actions and some state of affairs of which she could not

have knowledge, even in principle, as she deliberates? The epistemically inaccessibility of the state of affairs in question would then enable to correlation to survive under deliberation, from the agent's point of view, precisely as in Dummett's examples of coherent conceptions of retrocausality. But if it is not *really* a causal correlation, isn't the Clevidentialist still in trouble?

The Clevidentialist will agree that such cases are possible by the traditional Causalist's lights, but deny that they are possible by her own lights. On the contrary, she insists, such a correlation would automatically count as causal, by her standards. That's what causal dependency *is*, by her lights, after all: conditional evidential dependency, from the agent's point of view.

Moreover, since the correlation in question survives (by assumption) under deliberation, it survives in particular in the random choice case (which is simply a special case of deliberation, a choice to be guided by the outcome of a random event); which means that the traditional Causalist will have to admit that it leads to riches, when linked to suitable Outcomes. So it falls on the right side of the line, from the Clevidentialist point of view.

If this line of argument succeeds, it will give the Clevidentialist the guarantee she sought, that the conditional credences to which her version of EDT appeals will yield the same assessments of expected V-utility as in the random game. From this point, the argument is straightforward. On one side of the line, where the classic Newcomb problem lies, Clevidentialism recommends one-boxing, responding to the two-boxer's objection to *Why Ain't you Rich?* in the way described above, and challenging the traditional Causalists to offer a non-question-begging defence of the rationality of their own policy. On the other side of the line, where the familiar medical problems lie, Clevidentialism recommends "two-boxing" – i.e., the same choice as traditional CDT.

In both kinds of cases, the Clevidentialist stresses that by her lights – i.e., according to her understanding of causality, and her view of the probabilities required by EDT – CDT and EDT actually coincide. She recognises that traditional Causalists understand the term "causation" somewhat differently, and hence that by their lights, the cases in which Clevidentialism recommends one-boxing are cases in which CDT differs from EDT. Her challenge to these opponents is to defend the claim that CDT remains rational, in these exceptional cases, if "causation" is understood as they prefer. To meet this challenge, traditional Causalists need a response to the *Why Ain't you Rich?* objection – but now in the hands of an opponent who, unlike meeker traditional Evidentialists, is not prepared simply to concede the Causalist her counterfactuals.[24]

8 Conclusion

This paper has covered a lot of ground. I close with a summary of the main points, and some remarks about the limits of the present conclusions.

[24]Not, at least, until the Causalist concedes that counterfactuals need not guide rationality, in exceptional cases.

8.1 Summary

(i) The Chewcomb problems reveal a significant tension in a popular combination of views (a combination exemplified by Lewis himself, amongst others) concerning the rational practical significance of exceptional evidence, in the case of chance on the one hand, and causation on the other.

(iii) The tension can be resolved by adopting the same degree of subjectivism with respect to causation that Lewis adopts with respect to chance – accepting that causation, too, has its roots in evidential decision making, at least in the sense that nothing deserves the name causation, unless we can explain its relevance to decision. For causation as for chance, the required degree of subjectivism is nicely captured by the proposal that the modal notion in a question is an expert, intended to represent ideal evidential practice.

(iv) As in the case of chance, there are two ways to develop this thought, which differ in their treatment of certain cases of exceptional evidence. For Lewis, chance is not an infallible expert, and is rationally set aside by someone who believes herself to have inadmissible evidence. For Hall, there can be no such cases: no evidence is inadmissible, from chance's point of view. The difference is largely a matter of taste, and the two views agree about the rational credences, in the exceptional cases. Similarly in the case of causation, in circumstances such as the classic Newcomb problem. The Hall-like view treats these as cases in which the causal structure is abnormal (e.g., in involving retrocausality). The Lewis-like view treats them as cases in which rational choice does not follow CDT. But the two views agree on the rational policy: it is to one-box.

(v) If we go Hall's way in the case of causation – the Clevidentialist proposal, as I have called it – then EDT and CDT now coincide everywhere. But this does not imply that CDT need endorse the equivalent of one-boxing (i.e., not smoking, in the Smoking Gene problem) in medical cases. There, someone who believes that EDT recommends one-boxing is simply someone confused about the proper evidential bearing of her actions, as she deliberates. In normal cases, in which the causal structure is uncontroversial, our knowledge about it provides our best protection against such confusion; causation being precisely the expert we need to consult, in order to get these evidential judgements right.

(vi) This alignment between CDT and EDT entails that the usual Causalist response to *Why ain't you rich?* is powerless. Unlike conventional Evidentialists, the Clevidentialist rejects the Causalist's claims about the relevant counterfactuals (insisting that had she two-boxed, she would have been $1,000,000 poorer, not $1,000 richer).

(vii) At this point, the traditional Causalist will wish to defend her own reading of the counterfactuals, in the form of some objectivist rival to the Clevidentialist's account of causal dependence. But such views are vulnerable to the charge that Lewis himself makes against analogous views of chance: what they offer us does not deserve the name causation, unless the Causalist can explain

its relevance to rational deliberation. (And if the Causalist could do that, she would already have a response to the Clevidentialist.)

(viii) Viewed by these lights, the original Newcomb problem is pathology of rational deliberation, induced, in the main, by excessive objectivism about causality. By obscuring the practical foundations of causal thought, this objectivism makes it hard to see that the Newcomb puzzle presents us with what amounts to conflicting information about the causal structure of a decision problem. The puzzle's intractability then rests on its incoherence. The trick we played on ourselves was to treat causation and agentive evidential dependence as independent degrees of freedom, in order to imagine a case in which they seem to conflict. The right response is simply to call the illusion's bluff. If we wish to treat these notions as independent variables (in such extreme cases, at least), then we have no right to insist that causality always constrains rational action; if not, then the case as described is impossible, for the evidence implies a different causal structure.

(ix) A secondary contributing factor to the power and longevity of Newcomb's puzzle, if this diagnosis is correct, is the subtlety of the special epistemic status of the deliberating agent, as needed to ground an adequate subjectivist view of causation, immune from these illusions. The rough shape of the terrain is relatively familiar, the key feature being that fact that deliberation "crowds out" prediction, as we put it earlier. But a detailed elucidation of this idea, in a well worked-out model of the epistemic dynamics of deliberation, remains an open challenge.

8.2 Limitations

I want to emphasise, first, that I am not proposing that Clevidentialism offers a one-stop solution to decision puzzles in the Newcomb tradition. The alignment Clevidentialism allows between causal dependency and agentive evidential dependency does not imply, by any means, that it will always easy to determine what these dependencies are, in a particular decision problem. Clevidentialism tells us that they lie in the same place, but not in *which* place – and that, of course, can still be hard to determine.

Among the hard cases are various variants of the classic Newcomb problem. I have been taking for granted that the classic version of the problem is one in which Evidentialism really does recommend one-boxing. But it is easy to construct variants – including more realistic variants – in which this is far from obvious. Random experimentation remains the best guide, but it is not foolproof. We can never be certain that we have genuine randomness, for there might always be a lurking common cause, of which we are unaware. Nor can we be sure that the random mechanism itself does not have perturbing effects.[25] No matter – we inch our way forward, as in science in general, prepared always

[25] As it does in Nozick's own presentation of the original case, in which the Predictor penalizes agents who choose by random means.

to retreat if necessary. The difference that Clevidentialism makes is simply that as we do so, causal dependence and agentive evidential dependence keep step. Hypotheses about one are hypotheses about the other.

A related difficulty which survives Clevidentialism is that it may be unclear whether a proposed decision problem is really a *decision* problem at all – that is, do its constraints really allow us to regard ourselves as making a choice? Here Clevidentialism presumably has some bearing, for it reduces the potential parameters of a decision problem in a new way. But it seems unlikely to eliminate such puzzles altogether. As an extreme example, think of the variant of the classic Newcomb problem in which both boxes are transparent. Can an agent believe the usual story about the evidential significance of one-boxing, and yet believe that she has a choice in the matter? For a Clevidentialist this is the same puzzle as a claimed case of causation in which the effect of a contemplated action is known in advance. But reducing two puzzles to one is not the same as eliminating them altogether.

This question connects with one which is both deeper and much broader, and raises potential challenges to Clevidentialism, that of the status of agency itself. As we have seen, the Clevidentialist relies heavily on the idea that the epistemic viewpoint of an agent is distinctive in certain ways. Roughly, it requires that agents see their own actions as "uncaused", at least in the midst of deliberation about those same actions. This not only binds the fate of the Clevidentialist, at least in some sense, to that of free will. It also means, potentially even more uncomfortably, that Clevidentialism becomes a rope that binds *causation* to the fate of free will – no problem, perhaps, if these notions turn out to share the same fate, but a potential disaster if they do not. Clevidentialism thus has a stake in some large issues about such matters as the metaphysics of causation, and here there are general questions, of interest on all sides, that connect with the central strategy of this paper, that of comparing chance and causation. Why has subjectivism seemed less attractive in the case of causation than in case of chance, for example? And is the difference well-grounded?

So many puzzles certainly remain, in the general vicinity of the Newcomb case, even if we accept the Clevidentialist proposal. Nevertheless, it offers a solution to the central puzzle: it mends, and ends, the strange divergence of causation and evidence which lies at the conflicted heart of Newcomb's famous problem. And it brings an attractive unity to chance and causation, extending the pragmatism of Lewis's treatment of the former to a treatment of the latter. It would be nice, on balance, if it were true.

Acknowledgements

The beginnings of this paper were much indebted to Jossi Berkovitz, whose work on Newcomb problems prompted me to ask the question at the beginning of §2; and to Rachael Briggs, who suggested the link with Hall's response to Lewis on chance and inadmissible information (and gave me many useful comments at later stages). I am also very grateful to Arif Ahmed, Helen Beebee,

Steve Campbell, Mark Colyvan, Andy Egan, Adam Elga, Jenann Ismael, Jim Joyce, Peter Menzies, Wlodek Rabinowicz, Brian Skyrms, Nick Smith, Howard Sobel and Hong Zhou, and to audiences at the University of Sydney, ANU, MIT, Michigan and Oxford, for much discussion and many helpful comments. My research is supported by the Australian Research Council and the University of Sydney.

Bibliography

Dummett, Michael 1954: 'Can an effect precede its cause?', *Proceedings of the Aristotelian Society Supplementary Volume* 38, 27–44.

———1964: 'Bringing about the past', *Philosophical Review* 73, 338–359.

Eells, Ellery 1981: 'Causality, utility, and decision', *Synthese* 48, 295–329.

———1982: *Rational Decision and Causality* (Cambridge: Cambridge University Press).

———1984: 'Newcomb's many solutions', *Theory and Decision* 16, 59–105.

Gaifman, H. 1988: 'A theory of higher order probabilities', in B. Skyrms, et. al. (eds.), *Causality, Chance, and Choice* (Dordrecht: Reidel).

Gibbard, Allan & Harper, William 1978: 'Counterfactuals and two kinds of expected utility', in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory,* Vol. 1 (Dordrecht: Reidel), 125–162.

Hall, Ned 1994: 'Correcting the guide to objective chance', *Mind* 103, 505–517.

———2004: 'Two mistakes about credence and chance', *Australasian Journal of Philosophy* 82, 93–111.

Hitchcock, Christopher 1996: 'Causal decision theory and decision-theoretic causation', *Noûs* 30, 508–526.

Horgan, Terry 1981: 'Counterfactuals and Newcomb's problem', *Journal of Philosophy* 78, 331–356.

Horwich, Paul 1985: 'Decision theory in light of Newcomb's problem', *Philosophy of Science* 52, 431–450.

Ismael, Jenann 2007: 'Freedom, compulsion, and causation', *Psyche* 13.

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

Joyce, James 2007: 'Are Newcomb problems really decisions?', *Synthese* 156, 537–562.

Lewis, David 1979: 'Prisoners' dilemma is a Newcomb problem', *Philosophy and Public Affairs* 8, 235–240.

———1981a: 'Causal decision theory', *Australian Journal of Philosophy* 59, 5–30.

———1981b: '"Why ain'cha rich?"', *Noûs* 15, 377–380.

———1982: Letter to Wlodek Rabinowicz, 11 March 1982.

———1994: 'Humean supervenience debugged', *Mind* 103, 473–490.

Nozick, Robert 1969: 'Newcomb's problem and two principles of choice', in Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel* (Dordrecht: Reidel), 107–133.

Papineau, David 1996: 'Many minds are no worse than one', *British Journal for the Philosophy of Science* 47, 233–241.

Price, Huw 1986: 'Against causal decision theory', *Synthese* 67, 195–212.

———1991: 'Agency and probabilistic causality', *British Journal for the Philosophy of Science* 42, 15–76.

———1993: 'The direction of causation: Ramsey's ultimate contingency', in David Hull, Micky Forbes and Kathleen Okruhlik (eds.), *PSA 1992, Volume 2* (East Lansing, Michigan: Philosophy of Science Association), 253–267.

Price, Huw & Weslake, Brad 2009: 'The time-asymmetry of causation', in Helen Beebee, Christopher Hitchcock and Peter Menzies (eds), *The Oxford Handbook of Causation* (Oxford: Oxford University Press), 414–443.

Rabinowicz, Wlodek 2002: 'Does practical deliberation crowd out self-prediction?', *Erkenntnis* 57, 91–122.

Ramsey, Frank 1978: 'General propositions and causality', in D. H. Mellor (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics and Economics* (London: Routledge & Kegan Paul), 133–151.

Savage, Leonard 1954: *The Foundations of Statistics* (New York: Wiley).

van Fraassen, Bas 1989: *Laws and Symmetry* (Oxford: Oxford University Press).