The Lion, the 'Which?' and the Wardrobe – Reading Lewis as a Closet One-boxer

Huw Price*

September 15, 2009

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 exhibit a tension in Lewis's views on these two matters. I present a class of decision problems – actually, I argue, a species of Newcomb problem – in which Lewis's view of the relevance of inadmissible evidence seems to recommend one-boxing, while his causal decision theory recommends two-boxing. I propose a diagnosis for this dilemma, and suggest 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 dissolves many apparent Newcomb problems, and makes one-boxing non-controversial in those that remain.

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

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

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."

	Opaque box empty	Opaque box filled
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**).

Maximise Evidentially-grounded Utility ('V-utility'):

$$EV(Act_i) = V(O_1)P_{evidential}(O_1|Act_i) + V(_2)P_{evidential}(O_2|Act_i)$$

Maximise Causally-grounded Utility ('U-utility'):

$$EU(Act_i) = V(O_1)P_{causal}(O_1|Act_i) + V(O_2)P_{causal}(O_2|Act_i)$$

Here $P_{evidential}(O_j|Act_i)$ is the "purely epistemic" conditional probability of the outcome O_j given the act Act_i; while $P_{causal}(O_j|Act_i)$ is what we may call the *causal* conditional probability.

It is 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,

$$\begin{aligned} EV(One\text{-}box) &= \$0 \times 0.01 + \$1,000,000 \times 0.99 \\ &= \$990,000 \end{aligned}$$

while

$$\begin{aligned} EV(Two\text{-}box) &= \$1,000 \times 0.99 + \$1,001,000 \times 0.01 \\ &= \$11,010 \end{aligned}$$

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

$$\begin{aligned} EU(Two\text{-}box) &= \$1{,}000 \times \alpha + \$1{,}001{,}000 \times (1-\alpha) \\ &= \$1{,}000 + EU(One\text{-}box) \end{aligned}$$

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

Philosophers disagree about which of these two decision rules provides the rational strategy. Among the famous "two-boxers" is the lion of twentieth century metaphysics (and my title), David Lewis. Lewis describes the issue as follows:

> 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)

My aim in this paper is to call attention to an apparent tension between this aspect of Lewis's views, 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 endorsing one-boxing, in a particular

class of Newcomb problems; and that there seems to be no principled way of distinguishing these Newcomb problems from Newcomb problems in general. If I am right, then Lewis's commitments about the two matters are in tension with one another.

In the final section of the paper I suggest a resolution of this tension which extends a suggestion due to 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, Hall argues, to say that there is no such thing as inadmissible information: properly understood, chance tracks expert credence in such a way that such cases simply don't arise. If Hall is right, the effect is that in my chance-based Newcomb problems, chance-grounded and evidence-grounded reasoning must coincide – both recommend one-boxing.

I want to point out that Hall's move is analogous to an option that a number of people – me, for one – 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. I think that this approach can be seen as arguing that causation is best understood as a codification of an "expert function" for a deliberative agent, in the way that Hall treats chance as a codification of an expert function for a betting agent – an *evidential* agent, in both cases.

In neither case (for chance nor causation) is Hall's proposal compulsory, in my view. In either case, we might have grounds – e.g., perhaps, from physics – to postulate a modal notion which might in principle float free of rational agency, in unusual cases. But in this eventuality, once we recognise it for what it is, it will be immediate that the rational policy is to one-box, in the corresponding Newcomb problems. The apparent force of the Newcomb puzzles 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 grasped in the case of chance: rationality and modal metaphysics part company, and the rational choice is to one-box.

2 A chancy Newcomb problem?

On the face of it, Newcomb problems turn on a conflict between *causal beliefs* and *evidential beliefs.* Hence it is natural to ask whether the same kind of conflict can arise for other kinds of objective modality.[1] In particular, can it arise for chance?

[1]This is the question that led me into this topic.

It is easy to see that it can. Suppose God offers you the following options on the toss of what He assures you is a fair coin. Obviously, it is rational to choose to bet Heads.

	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 when He told you that it is a fair coin, he didn't tell you the whole truth. Of course, you knew that already. 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.

So you are not impressed by Satan's revelation. You say, "Tell me something I didn't know!" "Okay," says Satan, rising to the bait, "I betcha didn't know this. On those actual occasions on which you yourself bet on the coin, it comes up Tails about 99% of the time!"

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.

	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)

What is the rational policy in this case? 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 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.

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. Ned Hall's exposition of Lewis's view makes this clear:

> Example: we have, on thousands of occasions before this one, consulted the Oracle about what the chancy future would bring—and every time, her predictions have been vindicated, in minute detail. This time, she tells us that the coin will land heads. All of our information is purely historical, concerning only the record of her past successes, plus her most recent prediction. What should we believe about the outcome of the toss, on the supposition that it has a 50% chance of landing heads? Answer: we should be certain, or nearly so, that it will land heads, favouring the reliable words of the Oracle over the guidance of objective chance. We should treat the Oracle as a crystal ball—even though she might merely be lucky, and even though the evidence guiding our opinions contains no information directly about the future. (Hall 1994, 508–509)[2]

Or from Lewis himself:

> The fatal move that led from Humeanism to contradiction is no better than the *obvious blunder*:
>
> > C(the coin will fall heads/it is fair and will fall heads in 99 of the next 100 tosses) = 1/2
>
> or even
>
> > C(the coin will fall heads/it is fair and it will fall heads) = 1/2.
> > (Lewis 1994, 485, my emphasis)

Thus Lewis takes it for granted that someone who takes themselves to have 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.

[2] We'll return to Hall's own view below.

However, it turns out that the Chewcomb problem is just a Newcomb problem, in (light) disguise; and this policy – the second policy, above – amounts to one-boxing, when the disguise is removed. So Lewis's qualification to the Principal Principle thus seems to be in tension with his allegiance to two-boxing, at least in certain kinds of Newcomb problem.

In §3 below I'll introduce a new Chewcomb problem, which makes more explicit the fact that Chewcomb problems are a species of Newcomb problem. But we can already highlight the tension in Lewis's position in the present case, by considering this decision problem in the light of Lewis's own (1981a) formulation of causal decision theory. Lewis's version of CDT depends 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:

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

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 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 both boxes and taking one box as follows:

$$U(Two\text{-}box) = P(K_0)V(Two\text{-}box\ \&\ K_0) + P(K_1)V(Two\text{-}box\ \&\ K_1)$$
$$U(One\text{-}box) = P(K_0)V(One\text{-}box\ \&\ K_0) + P(K_1)V(One\text{-}box\ \&\ K_1).$$

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

If we wish to apply this framework to the Chewcomb problem, the first issue is what we should take the the dependency hypotheses to be. If we take the causal structure to follow the structure of the objective chances – i.e., take it that because the outcome (H or T) is the result of a toss of fair coin[3] – then the relevant dependency hypotheses are:

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

[3]Whose behaviour isn't influenced by the way we choose to bet, presumably!

The next issue concerns the probabilities $P(K_H)$ and $P(K_T)$. Lewis himself 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 $P(K_H) = .01$, for example, but only that $P(K_H|Bet\ T) = .01$.

If this isn't already clear, we can easily modify the example 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" (as Lewis puts it) of the act of choosing to bet rather than not to bet. Accordingly, Lewis's CDT then seems to require that we use $P(K_H) = P(K_T) = 0.5$ for calculating $U(Bet\ H)$, $U(Bet\ T)$ and $U(No\ bet)$, 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 H in both cases, the unconditional case appears to be in violation of his own policy on the relevance of inadmissable evidence. If he recommends T 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.

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

Table 4: Two Chewcomb games – policies and problems.

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 evidential policies seem to give different recommendations from causal policies, and CDT is the decision theory that cleves 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 cleve 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 cleving are liable to yield different recommendations in these cases. At least, they are liable to do so as long as our causal judgements cleve 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).

Lewis himself seems to have been aware 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. Nor do I know whether he saw the fact that serves to highlight this difficulty, viz., that these cases are themselves a species of Newcomb problem.[4]

3 Making the analogy closer

I referred to the Chewcomb problem above as a Newcomb problem in light disguise. Let's remove the disguise. Suppose 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, "Psssst! It is definitely a fair coin, but my crystal ball tells me that in 99% of cases in which people choose to one-box in this game, the coin actually lands Tails."

	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.[5] How could it do so, when that is determined by the result of a toss of a fair coin? To say that the chances are $50/50$ is surely to say that *nothing* prior to the toss can have a causal effect on the outcome.

[4]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)

(He goes on to give a nice example of such a case.) 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.

[5]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.

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.[6]

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.

4 One-boxing the Hall way?

Ned Hall (1994, 2004) recommends that we replace Lewis's Principal Principle with a new principle, requiring that rational credences track *conditional* chances: chances *given our evidence.* On the face of it, 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.[7] 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.

Like many irenic proposals, however, this one is a little too good to be true. To see this, we only have to imagine a proponent of a view of chance according to which it makes no difference what Satan whispers in one's ear: the real meta-

[6] We could make the analogy with the original Newcomb problem even tighter, by specifying that the coin toss (which determines whether the opaque box contains $\$1{,}000{,}000$) has *already taken place,* when the agent makes her choice. (She does not yet have access to the result, of course.)

[7] Or the information *that* he has done so, perhaps.

physical chance of a fair coin's landing Tails is independent of such supernatural vocalisations (our objector insists), and so the shift to conditional chances makes no difference.[8] 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 (as he notes) 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

> 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 the available evidence (including, in particular, what Lewis treats as "inadmissible" evidence). And it is this stipulation, rather than the conditionalisation move itself, that finally ensures that there cannot be a genuine Chewcomb problem – a genuine case in which chance and evidential reasoning come into conflict. Failure to conditionalise certainly gives rise to one class of (apparent) Chewcomb problems, and Hall's diagnosis correctly eliminates those. As we observed a moment ago, however, the conditionalisation move does not deal with a second class of potential Chewcomb problems, viz., those in which a metaphysical view of chance simply "disconnects" from expert credence (conditional on all the evidence, in each case).

I've stressed this point because it is the latter aspect of Hall's view that seems to me analogous to (what I find) an attractive resolution of the original Newcomb

[8] I think that the possibility of this objection is obscured in Hall's (1994) discussion of "crystal balls" by his failure to treat the case in which the ball's prediction is itself probabilistic in nature, as in my Satanic example.

case. In Hall's terminology, it is the view that causal judgments should be regarded as the judgments of experts about effective strategies, where these are a matter of maximising conditional V-utility (properly asssessed from the agent's point of view). Once again, the effect is to support one-boxing, but to see this as what maximising U-utility recommends, too, when causal dependence is seen for what it really is.

We can develop the analogy very directly using our Chewcomb example. The original 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 States on Acts – in the sense of those terms that now seems appropriate, given that chance is to be understood as an expert function.

In neither case is Hall's proposal or its causal analogue compulsory, in my view. In either case (for chance or for causation) we might have grounds – e.g., perhaps, from physics – to postulate a modal notion which could drift apart from expert credence and strategy, in unusual cases. But in this eventuality, once we recognise it for what it is, it will be immediate that the rational policy is to one-box, in the corresponding Newcomb problems.

The apparent force of the Newcomb puzzles seems to derive 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 there is no real dilemma. The right option, trivially, is the one that Lewis grasped in the case of chance: rationality and modal metaphysics part company (and rationality follows rationality – what else? – rather than metaphysics). There may be other puzzles and surprises in the vicinity: a puzzle that rationality and some kind of metaphysics do not keep step, in some sense, in the way that we have come to expect; or a puzzle about how there can be an expert function at all, of the chance or the causation variety, in a world of a certain kind.[9] But these are not the original puzzle of the Newcomb

[9]Another very interesting kind of puzzle that certainly survives these conclusions concerns the

problem, which seems to evaporate from this perspective, along with the problem of inadmissible evidence.[10]

Bibliography

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

Gibbard, Allan and 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.

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

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

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

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

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

Lewis, David 1982: Letter to Rabinowicz, 11 March 1982. (Attached below.)

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

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

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

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

nature of the rational expert's recommendation, in difficult cases. Some people – I, for one (Price 1986, 1991) – have argued that while we do make a mistake if we one-box in the so-called medical Newcomb problems, the mistake is a failure of evidential rationality – a failure to assess the evidential probabilities correctly, from the agent's own point of view – rather than a failure to use the correct decision rule. This diagnosis, and the intriguing issue about rationality it tries to address, both seem untouched by the present conclusions,

[10]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. I am also grateful to 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 and the University of Michigan, for much discussion and many helpful comments. My research is supported by the Australian Research Council and the University of Sydney.

Appendix

With the kind permission of Wlodek Rabinowicz, Stephanie Lewis and the Estate of David K. Lewis, the following pages reproduce correspondence between Rabinowicz and Lewis in 1982; including the letter from Lewis dated 11 March 1982, to which I referred in fn. 4 above.

23.01.82

Filosofiska Institutionen
vid
Uppsala Universitet
Villavägen 5
752 36 Uppsala

Dear Professor Lewis,

In the paper that follows (to be published in the Festschrift for Lennart Åqvist), I discuss the relationship between Sobel's version of causal decision theory and the one proposed by you. In particular, I argue that the extra constraint you impose on Sobel's theory

— in order to establish the equivalence between two versions — does not seem to be plausible (at least insofar as we treat it as a constraint on the agent's credences, and not as a simultaneous constraint on both the agent's credences and the tendency structure). I would be very glad to hear your comments. Perhaps it doesn't show in my paper, but I admired your article in the Australasian Journal very much.

With kindest regards,

Wlodek Rabinowicz

Princeton University DEPARTMENT OF PHILOSOPHY
1879 HALL
PRINCETON, NEW JERSEY 08544

11 March 1982

Re: "Causal Decision Theories"

Wlodzimierz Rabinowicz
Filosofiska Institutionen
Uppsala Universitet
Villavägen 5
S-752 36 Uppsala
Sweden

Dear Rabinowicz,

Thank you very much for your paper "Two Causal Decision Theories". I've read it with great interest, and it has increased my understanding of the relation of my treatment and Sobel's. Let me comment on several of the points you raise.

Concerning centering. Certainly I am committed to some form of centering: if A and B hold at world w, then $A \Box\!\!\rightarrow B$ holds at w, and that is because no other A-world is as close to w as w itself is. But if A and B hold at w, I do not want to conclude that, at w, there is a tendency of 100% to w and hence to B, and zero tendency to any other world, given A. At least, this is so if I follow Sobel's lead and understand that "tendency" is to be related to our ordinary talk about chance. For instance, suppose a coin will shortly be tossed. Suppose that in fact it will fall heads. If it were tossed, it might with 50% chance fall heads; if it were tossed, it might with 50% chance fall tails. (For suppose it is a fair coin in an indeterministic world.) But if it were tossed, it would fall heads--for it will be tossed and will fall heads--and it is not so that if it were tossed, it might fall tails. In Sobel's terminology, I think I should say that at this world, there is 50% tendency for it to fall heads if tossed, 50% tendency for it to fall tails if tossed.

So, if "tendency" connects to ordinary talk of chance, and if I am to hold on to centering for the similarity relation that governs would and might counterfactuals, then I cannot also hold on to the connection between that similarity relation and tendency that you state as (ii) on page 7. So let that connection be cut. Then I have a disagreement with Sobel 1979 about would-counterfactuals; but not a disagreement that affects our decision theories. I try to assimilate Sobel's tendency function to my imaging function given by

$$W_A(W') = C(W'/AK_W),$$

and that function will, on my view, not be (more than weakly) centered (in the sense of your page 7) in all cases. That is, it may happen that $C(W/AK_W) \neq 1$. In the sense of centering relevant to use of imaging or tendency functions in calculating utility, I think I join Sobel in centering only weakly.

I think this may answer the difficulties posed by Theorems 1 and 2 of page 15, since I think these use the centering of the imaging function that I would not accept.

Concerning dependency hypotheses. Let a practical dependency hypothesis be "a maximally specific proposition about how the things [the agent] cares about do and do not depend causally on his present actions"--that is, let it be what I called simply a "dependency hypothesis". Let a full dependendency hypothesis be a maximally specific proposition about how all things do and do not depend causally on the agent's present actions. Then I agree completely with your observation on page 13 that a tendency proposition (an equivalence class under the relation of having the same tendencies) is, if anything, not a practical but a full dependency hypothesis. However this will not lead to further differences between the treatments if, as I think, I could just as well have formulated mine throughout in terms of full rather than practical dependency hypotheses. The difference in formulations won't matter if, whenever K is a practical d.h. and H is a full d.h. compatible with--and hence implying--K, then V(AH)=V(AK). And that will be so if, for every value-level proposition V=v , C([V=v]/AH) = C([V=v]/AK). And how can that not be so if, indeed, K covers all respects of dependence that the agent cares about?

Concerning agents who think they may have foreknowledge about the outcomes of chance processes. Sobel has nothing to say against Connection Thesis 1 except in cases of this sort; I take it--though of course I should not take Sobel to be committed to this--that he finds it plausible as applied to other cases. (Let me call these cases normal.) So do I. Sobel finds it doubtful in abnormal cases. So do I. Sobel doesn't want to reject the possibility that a fairly reasonable agent might find himself in an abnormal cases. Neither do I. So far, no disagreement. But we handle the problem differently. I impose the principle (restricted to the agent's options) and accordingly agree to set aside "some very special cases--cases about which I, at least, have no firm views". Sobel doesn't impose the principle and doesn't set aside the abnormal cases. But--unless Sobel has some other reason than he gave to avoid Connection Thesis 1--we seem to be in agreement both (1) in our treatment of the normal cases, and (2) in our hesitancy to commit ourselves to much concerning the abnormal cases.

Now, I don't dismiss the abnormal cases for the reason you consider on page 18: that an agent would have to be irrational to be in such a case. As you say, I want the theory--so far as possible--to apply to imperfectly rational agents. Besides, I believe in the logical possibility of time travel, precognition, etc., and I see no reason why suitable evidence might not convince a perfectly rational agent that these possibilities are realized, and in such a way as to bring him news from the future. My worry is a different one. 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. You know that spending all you have on armour would greatly increase your chances of surviving the coming battle, but leave you a pauper if you do survive; but you also know, by news from the future that you have excellent reason to trust, that you will survive. (The news doesn't say whether you will have bought the armour.) Now: is it rational to buy the armour? I have no idea--there are excellend reasons both ways. And I think that even those who have the correct two-box'st intuitions about Newcomb's problem may still find this new problem puzzling. That is, I don't think the appeal of not buying armour is just a misguided revival of V-maximizing intuitions that we've elsewhere overcome.

Concerning constraints on the tendency model. I found your Theorem 4 extremely interesting. The moral, I think, is that Sobel should have imposed a constraint on tendency models that would rule out such models as you produce! This is not because the agent's credences somehow limit the tendencies of the world--I'll have none of such idealism--but because chance has more structure to it than Sobel's treatment captures. So I put to him a dilemma. If we are to understand his primitive tendency function by way of the connections he establishes between it and our ordinary talk of chance, then we must have an additional constraint. (And if we do, this difficulty for L2.2 goes away.) If not, on the other hand, then how are we to understand what a tendency is? (In this case, indeed, Sobel's treatment would diverge from my own.)

The constraint I have in mind resembles the "chances aren't chancy" principle of my "Subjectivist's Guide to Objective Chance". In terms of tendency functions, it comes to this. It is never the case that world w has some nonzero tendency, given A, to w' such that w' differs from w itself in its tendencies given A.

I find the following unintelligible, if "chance" really means chance: this coin has 50% chance (or tendency) of falling heads if it is tossed now; but also, it has 50% chance, if it is tossed now, of having 100% chance of heads if it is tossed now; and likewise it has 50% chance, if it is tossed now, of having no chance of falling heads, if it is tossed now.

If Sobel's theory makes sense of that, so much the worse for Sobel's theory, unless it gets further constrained in a way that accords with his talk of chance. And, as far as I can see, it very well could be further constrained.

Your four-world model illustrates the problem. Suppose that the coin is tossed at worlds w_1 and w_4; it falls heads at w_1, tails at w_4. Consider the situation from the standpoint of w_2. If the coin were tossed, there would be equal chance of the two maximally similar toss-worlds, w_1 with heads and w_4 with tails. But w_1 is the only maximally similar toss-world to w_1, likewise w_4 is the only maximally similar toss-world to w_4. So (again from the standpoint of w_2), if the coin had been tossed, it would have had equal chance of being such that, if tossed, it would have had 100% chance of heads (as at w_1) and of being such that, if tossed, it would have had 0% chance of heads (as at w_4). Now, I submit that so long as Sobel's theory allows that, it admits of models that do not conform to his intention to speak of chance. Contrariwise, if we take him at his word we should regard such models as unintended.

Again, thank you very much for a most instructive paper. I shall be interested to see Sobel's comments, if any--doubtless you sent him your paper, and I shall send him a copy of this.

Sincerely,

David Lewis

Uppsala University
Department of Philosophy
Villavägen 5
S-752 36 UPPSALA
Sweden

Professor David Lewis
Princeton University
Dept of Philosophy
1879 Hall
Princeton, New Jersey 08544, USA

5 April 1982

Dear Professor Lewis,

Thank you very much for your most interesting letter. I agree with many of your remarks. In particular, I now realize that:

(1) I should have made it clear to the reader that you cannot accept Sobel's definition of (the truth conditions for) subjunctive conditionals in terms of tendencies. This is so, since you ascribe centering to the former but not to the latter.

(2) I should have noticed that the difference between the 'practical' dependency hypotheses and the 'full' ones is irrelevant as far as the calculation of expected utility is concerned. Thus, even though your assumption

L 1. For any world w, $K_w = S_w$,

is incorrect, you could have replaced it by the obviously correct claim that,

for any w, $S_w \subseteq K_w$ and $V(AK_w) = V(AS_w)$.

The effect, in terms of expected utility, would be the same. You would still be able to derive from

your definition of expected utility,

(i) $U(A) =_{df} \Sigma_K C(K) V(AK)$,

the equality:

(ii) $U(A) = \Sigma_W C(W) \Sigma_{W'} C(W'/AS_W) V(W')$.

($\Sigma_K C(K)V(AK) = \Sigma_W C(W) V(AK_W) = \Sigma_W C(W) V(AS_W) = \Sigma_W C(W) \Sigma_{W'} C(W'/AS_W) V(W')$.)

And (ii), together with your assumption C 2, implies the Sobelian formula:

$$U(A) = \Sigma_W C(W) \Sigma_{W'} T_{W,A}(W') V(W').$$

(3) I should have noticed that you don't intend to apply your decision theory to 'abnormal' cases (cases in which the agent takes himself to have foreknowledge about the outcome of a chance process). Thus, it is wrong to suggest (as I have done on p. 18) that your theory and Sobel's might generate conflicting prescriptions in abnormal cases. Instead, what I should have said is that your theory, contrary to Sobel's, is not meant to give any guidance at all in cases like that.

(4) In connection with my Theorem 4, I should have discussed the plausibility of my four-world tendency model. As you point out, this model fails to satisfy the following very natural constraint on tendencies:

(α) For all options A, and all W and W', if $T_{W,A}(W') > 0$, then $T_{W',A}(W') = T_{W,A}(W')$.

You suggest that Sobel should have added this constraint to his theory. I must admit that I am not convinced. I shall return to this problem below.

It might be interesting to note that there is

another seemingly natural constraint on tendencies that my model does not satisfy:

(β) For all options A, and all w and w', if $AS_w \neq \emptyset$ and $T_{w,A}(w') > 0$, then $w' \in S_w$.

(In my model, $T_{w_1,A}(w_2) = \frac{1}{2}$, even though $w_2 \notin AS_{w_1} = \{w_3\} \neq \emptyset$.)

(β) is of some interest, since it can be shown that the following relationship obtains:

(Unrestricted) Connection Thesis 1 & (β) $\Rightarrow$ L2.2.

(To prove this result, note that, insofar as T is weakly centered, (β) implies

(β⁺) For all options A, and all w, if $AS_w \neq \emptyset$, then
$$T_{w,A} = T_{w,AS_w}.$$

It follows from (β⁺) that, if $T_{w,A}(w') = n$,
$$AS_w \subseteq AS_w \diamond\!\!\rightarrow \{w'\}.$$
Therefore, if $C(AS_w) > 0$, then
$$C(AS_w \wedge (AS_w \diamond\!\!\rightarrow \{w'\}) > 0$$
and
$$C(w'/AS_w) = C(w'/AS_w \wedge (AS_w \diamond\!\!\rightarrow \{w'\})).$$
But then (unrestricted) Connection Thesis 1 implies that
$$C(w'/AS_w) = n = T_{w,A}(w'). \;)$$

This means that, in the presence of (β), your assumption L2.2 and Sobel's Connection Thesis 1 turn out to be closely related to each other. L2.2 entails Connection Thesis 1 (restricted to the agent's options) — cf my Theorem 3 (p. 17); and, given (β), (unrestricted) Connection Thesis 1 entails L2.2.

Of course, (α) does not imply (β). Nor does (α) follow from (β). But the following weakening of (α) directly follows from (β):

(α⁻) For all options A, and all w and w', if $AS_w \neq \emptyset$ and $T_{w,A}(w') > 0$, then $T_{w',A}(w') = T_{w,A}(w')$.

Now I turn to an issue that seems to me to be crucial when one compares your theory with Sobel's. I now think that the main difference between these two theories consist in that they express different attitudes to centering. You assume that tendencies are not (generally) centered. Sobel, in his work from 1979, does not assume centering either. But, and here comes the difference, neither does he want to assume that tendencies are <u>not</u> centered (cf p. 9 in my paper, and Sobel, section 6.42). He wants to stay <u>neutral</u> on this issue, while you are prepared to <u>commit</u> yourself. This policy of neutrality means that Sobel's theory of tendencies must be so weak as to allow of different possible extensions. One such possible extension is a theory in which tendencies are centered without being (in general) fully determinate. If Sobel's neutral theory is to allow of this extension, it cannot contain conditions from which it would follow that,

(C) If T is centered, then T is fully determinate.

Now, your condition L2.1 entails (C) (cf my Theorem 1 on p. 15). Thus, Sobel cannot assume L2.1 as a part of his theory. It is different with you. Since you are already committed to the view that tendencies are not centered, the fact that L2.1 implies (C) will not bother you at all.

This explains also my reluctance to admit that Sobel should incorporate your constraint (α) into his (neutral) theory. It is easy to see that (α), just

like L2.1, implies (C). (Assume that T is centered and that (K) holds. Suppose that $T_{w,A}(w') > 0$. Then w' is an A-world, so that, by centering, $T_{w',A}(w') = 1$. But then (K) implies that $T_{w,A}(w') = T_{w',A}(w') = 1$.)

On the other hand, (B) does not imply (C). Thus, it seems that Sobel might have added (B) to his theory and still remain neutral. Therefore, I now think that my objection against your L2.2 (as expressed in Theorem 4) was too hasty. Given (B), L2.2 becomes as plausible (or as implausible) as Connection Thesis 1; and (B) seems to be a very reasonable constraint on tendencies. But I still think that your theory and Sobel's are significantly different, insofar as Sobel does not share your commitment to non-centering.

Is it, however, reasonable to allow for the possibility that tendencies might be centered (even in indeterministic worlds)? As you point out, in our ordinary talk of chance, we treat chances (tendencies) as non-centered. Thus, for instance, the mere fact that a coin falls heads is not supposed to show that the chance of heads was one. I think, however, that this argument from the ordinary parlance is not conclusive. In particular, when an act-utilitarian defines the utility of an action, A, as the weighted sum of the values of its possible consequences, with weights being the chances of different consequences given the performance of A, then his concept of chance might very well differ from the ordinary one. It is often said that what is characteristic for the act-utilitarian approach is its realism – its tendency to take into consideration 'all the facts'.

But then it is natural for such a realistic approach to take into consideration the actual future outcomes of indeterministic processes. And this seems to imply that the practical concept of chance -- the one used in the calculation of utility -- should be a centered one (even though our ordinary concept of chance is not centered at all).

An example: If an agent who is offered a bet on heads with odds five to one accepts the bet and looses, then my act-utilitarian would say that the agent acted wrongly in accepting the bet. He acted wrongly because he lost. And accepting the bet would still be wrong even if the odds offered were even better. Now, if

(a) an action is wrong iff it has lower utility than some of its alternatives,

and

(b) the utility of an action, A, is the weighted sum of the values of its possible consequences, with weights being the 'practical' chances of different consequences given the performance of A,

then the claim that accepting the bet was wrong, independently of the odds offered, must imply that the practical chance of ~~loosing~~ loss was equal to one. This is so even if we assume that the coin was fair and indeterministic, so that, in ordinary parlance, the chance of ~~loosing~~ loss was only 50%. Thus, the act-utilitarian realism leads to the conclusion that the chances used in the calculation of utility should be centered.

How does this relate to decision theory? Well, insofar as decision theory prescribes maximization of expected utility and the expected utility of an action is taken to be the weighted sum of its possible utilities, with weights being the probabilities (credences) of these different utilities, the conclusion is obvious: the concept of chance used in the calculation of expected utility should also be a centered one!

Note that, if practical chances are centered, then it is no longer implausible to define the truth conditions for (centered) subjunctive conditionals in terms of chances, as Sobel does. We must only remember that the chances used in the definition should be 'practical', and not the ordinary ones.

If we ignore 'abnormal' cases, we may define ordinary chances in terms of conditional credences and (full) dependency hypotheses -- just as you suggest. But then we still have the problem of defining practical chances.

Another thing: even though practical chances differ from the ordinary ones, we should still expect that, in all 'normal' cases, probable ordinary chances (i.e., the credence weighted sums of possible ordinary chances) will coincide with probable practical chances.

The above distinction between practical and ordinary chances and the suggestion that practical chances should be centered is due to Sobel. In a letter written in February this year, he writes:

"Weak centering is right for the theory of chance: we want room for agents whose chance-views are weakly centered. But, roughly speaking, strong centering is right for the theory of practical relevance for which we want not 'probable chances' but 'probable centered-chances!'"

It is such 'probable centered-chances' that Sobel nowadays wants to use as weights in calculation of expected utility. He developes this suggestion in more detail in a manuscript called 'Notes on some recent changes in the theory of chance and the theory of probable chance' (21 Febr. 1982). Thus, one might say that Sobel no longer wants to remain neutral. He is prepared to commit himself to centering (in practical contexts). As for myself, I feel that this is the right thing to do.

Sobel thinks that 'centered-chances' are easily definable in terms of the ordinary non-centered ones. He simply lets the centered-chance of q given p be equal to the truth-value of q (one or zero) *if* p is true, and to the ordinary chance of q given p *if* p is false. It seems to me that this definition is wrong. I think that the centered-chance of q given p might deviate from its ordinary chance even in some cases when p is false. In particular, this happens whenever q describes a chance event whose occurrence does not in any way depend on whether p is true or false.

Thus, to take the simplest case, if the agent declines an offered bet on heads, and the coin is subsequently tossed and falls heads, then -- insofar as the outcome of the toss is assumed to be wholly independent of the agent's betting behaviour -- we should let the centered-chance of heads conditional on the agent's accepting the bet be one (even though the agent has, in fact, declined the bet and the ordinary chance of heads was less than one).

As a matter of fact (this is something that I was not aware of when I wrote to Sobel), there may be cases

in which the centered-chance of q given p is neither one nor zero but it still differs from the ordinary chance of q given p.

An example: Suppose that we have two coins, C_1 and C_2, with different ordinary chances for heads. Let these chances be $\frac{1}{2}$ and $\frac{1}{3}$, respectively. The agent is offered the following bet: he will win iff both coins will fall heads or both will fall tails. We assume that C_1 is going to be tossed whatever the agent does -- whether he accepts the bet or not. C_2, on the other hand, will be tossed only if the agent accepts the bet. The ordinary chance of winning conditional on accepting the bet is in this case equal to $\frac{1}{2}$ (this chance equals $\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{2}{3}$). Suppose, however, that the agent declines the bet and C_1 is tossed and falls heads. (C_2 is not tossed since the agent abstained from betting.) Then it seems that the centered-chance of winning conditional on accepting the bet should be taken as equal to $\frac{1}{3}$. (If the agent had accepted the bet, he would have won iff C_2 had fallen heads. Therefore, his centered-chance of winning would equal the ordinary chance of C_2 falling heads.) Thus, the centered-chance of winning is in this case neither one nor zero, but it still differs from the ordinary chance of winning.

If I am right, then, defining centered-chances in terms of ordinary chances might prove to be quite complicated.

Concerning abnormal cases. I found your armour-example fascinating. Does the information that the agent will survive the coming battle give him any reason to abstain from buying the armour? My first intuitive reaction was negative. It seemed to me that this additional information would not

change anything in his decision-problem: buying the armour would still be the reasonable thing to do. All this providing

(a) that survival is much more important to the agent than not becoming a pauper,

and

(b) that buying the armour significantly raises his chance to survive.

For example, suppose that the values of survival and of becoming a pauper are, respectively, 100 and -10, while the probable (ordinary) chances of survival are, respectively, $\frac{2}{3}$ and $\frac{1}{3}$, depending on whether the agent will buy the armour or not.

But then I started to wonder. After learning that he is going to survive the agent becomes certain that the action he is actually going to perform will have good consequences. And he knows that the chances are that he might not survive <u>if</u> he were to act differently. Thus it seems that everything depends on the agent's credences for his two options. In particular, if he believes that he won't buy the armour, then, perhaps, not buying the armour is the rational thing for him to do. On the other hand, if the agent, before learning that he was going to survive, thought it quite credible that he would buy the armour, then the conditionalization-model for belief-change implies that buying the armour should become even more credible in view of the new information. (I assume here that the agent's

initial credences for survival conditional on his different options go by chances. That is, I assume that

$$C_0(S/A) = \frac{2}{3} \quad \text{and} \quad C_0(S/\neg A) = \frac{1}{3}.$$

C_0 represents the agent's initial credence function, 'S' stands for 'survival' and 'A' denotes 'buying the armour'.

Now, let C_S stand for the agent's credence function after receiving the information that he is going to survive. The conditionalization model implies that

$$C_S(A) = C(A/S) = \frac{C_0(S/A) \cdot C_0(A)}{C_0(A)C_0(S/A) + C_0(\neg A)C_0(S/\neg A)}$$

$$= \frac{\frac{2}{3}C_0(A)}{\frac{2}{3}C_0(A) + \frac{1}{3} - \frac{1}{3}C_0(A)}$$

$$= \frac{2\,C_0(A)}{C_0(A)+1}.$$

Thus, if $0 < C_0(A) < 1$, the ratio $\frac{C_S(A)}{C_0(A)}$ will be higher than one.) But then the new information should not give the agent any reason to abstain from buying the armour. (In fact, the opposite is true.)

The centered extension of Sobel's theory bears out the informal argument presented above. Assuming that the values of survival and of becoming a pauper are 100 and −10, and that the ordinary chances of survival given A and $\neg A$ are, respectively, $\frac{2}{3}$ and $\frac{1}{3}$, the centered approach to expected utility implies that

$$U(A) = C_S(A)(100-10) + C_S(\neg A)(\tfrac{2}{3}\cdot 100 - 10),$$

while

$$U(\neg A) = C_S(A)(\tfrac{1}{3}\cdot 100) + C_S(\neg A)\,100.$$

But then it follows that

$$U(A) > U(\neg A) \text{ iff } C_s(A) > \frac{13}{30}.$$

Thus, assuming that the agent's initial credences go by chances, and supposing that he changes his beliefs by conditionalization,

$$U(A) > U(\neg A) \text{ iff } C_0(A) > \frac{13}{47}.$$

(Other variants of decision theory would lead to different results. Thus, according to the standard, Jeffrey-like approach, the agent should abstain from buying the armour -- quite independently of his credence for that option. Your own theory would have given the same answer, if it were applicable to abnormal cases. But, of course, it is not. And the non-centered extension of Sobel's theory would imply that buying the armour is always the best policy -- quite independently of the agent's credence for that option.)

Of course, it seems very strange to say that, in some cases, the rationality of an action might be dependent on its credence. But then the abnormal cases, in which we have this kind of dependence, are strange. (However, I do not mean to suggest that this kind of dependence might obtain only in the abnormal cases. The story of the man who met death in Damascus, described by Gibbard and Harper in their paper 'Counterfactuals and Two Kinds of Expected Utility', provides an example of a 'normal' case in which we seem to have the same kind of dependence.)

Anyway, the following seems to be true: If decision theory should always treat the agent's credences for his options as irrelevant for choice, then you are right in suggesting that decision theory is not applicable to abnormal cases.

Well, this was an awfully long letter. I hope you don't mind. Of course, I will be grateful for any comments.

Sincerely,

Wlodek Rabinowicz

P.S. I shall send a copy of this letter to Sobel. I suppose that he will send you his comments on my paper, if and when he will have any.