Theory and Decision (2006) 61: 277–303
DOI 10.1007/s11238-006-9002-6

PETER SLEZAK

DEMONS, DECEIVERS AND LIARS: NEWCOMB'S *MALIN GÉNIE*

ABSTRACT. A fully adequate solution to Newcomb's Problem (Nozick 1969) should reveal the source of its extraordinary elusiveness and persistent intractability. Recently, a few accounts have independently sought to meet this criterion of adequacy by exposing the underlying source of the problem's profound puzzlement. Thus, Sorensen (1987), Slezak (1998), Priest (2002) and Maitzen and Wilson (2003) share the 'no box' view according to which the very idea that there is a right choice is misconceived since the problem is ill-formed or incoherent in some way. Among proponents of this view, Richard Jeffrey (2004) recently declared that he renounces his earlier position that accepted Newcomb problems as genuine decision problems. Significantly, Jeffrey suggests that "Newcomb problems are like Escher's famous staircase on which an unbroken ascent takes you back where you started" (Jeffrey (2004; 113)). Jeffrey's analogy is apt for a puzzle whose specific logical features can be precisely articulated. Along the lines of these related approaches, I propose to improve and clarify them by providing such a deeper analysis that elucidates their essential, related insights.

KEY WORDS: Descartes's demon, decision theory, Newcomb's problem, paradoxes of self-reference

1. MAN AS INTELLECTUAL CRIPPLE?

A truly insightful solution will reveal why Newcomb's Problem has been so perplexing as to have eluded the best philosophical minds who have toiled over it. Recently, Maitzen and Wilson (2003) suggest that the problem arises from a kind of underdetermination that is "more radical than any previously identified" (2003, p. 160). This claim is not strictly correct in view of the other independent accounts of the same kind. Nevertheless, joining proponents of such views, Richard Jeffrey (2004), too, recently declared that he renounces

his earlier position that accepted Newcomb problems as genuine decision problems. Significantly, Jeffrey suggests, "Newcomb problems are like Escher's famous staircase on which an unbroken ascent takes you back where you started" (Jeffrey (2004, p. 113)). He adds that we know there can be no such things, though we see no local flaw in the puzzle. Jeffrey's merely suggestive analogy is apt for a puzzle whose specific logical features should permit precise articulation. I propose just such a deeper analysis that elucidates the conceptual structure of Newcomb's Problem and, thereby, the source of its notorious recalcitrance.

During 30 years of inconclusive discussion, Newcomb's Problem has been most widely seen as exposing inadequacies of the current standard theory of decision-making. The most plausible normative principles give conflicting recommendations. Subjective expected utility considerations prescribe a choice diametrically opposed to that prescribed by the principle of dominance. As in the case of other paradoxes, seemingly impeccable reasoning gives rise to contradictions. Thus Jeffrey said that Newcomb's Problem may be seen "as a rock on which … Bayesianism … must founder" (Jeffrey (1983, p. 23)). In a similar vein, Resnik (1987, p. 111) declared "this paradox has shaken decision theory to its foundations" and Campbell (1985, p. 3) says "Quite simply, these paradoxes … cast in doubt our understanding of rationality." Despite a vast literature of great technical subtlety and complexity sharing these diagnoses, no clear solution has emerged. Returning to the problem he presented nearly 30 years before, Nozick (1993, p. 43) observed that the controversy continues unabated and judged that "No resolution has been completely convincing." For his part, Nozick remained within the standard framework and aimed to "formulate a broadened decision theory to handle and encompass this problem adequately" (1993, p. 41).

Despite having been neglected by psychologists (though see Shafir (1995)), Newcomb's Problem appears to suggest ill-understood features of rational choice behaviour and appears to reveal further anomalies in our tacit principles of decision-

making like the tradition of research on 'heuristics and biases' (Tversky and Kaheneman (1974)). In this case Newcomb's Problem becomes an addition to the list of such famous "paradoxes" as those of Allais (1953) and Ellsberg (1961). On standard accounts, Newcomb's Problem appears to support Slovic's jaundiced view that research into such phenomena "has led to the sobering conclusion that, in the face of uncertainty, man may be an intellectual cripple, whose intuitive judgements and decisions violate many of the fundamental principles of optimal behaviour" (quoted in Wright (1984, p. 114)). Although available analyses of Newcomb's problem suggest its relevance to cognitive science in this way, I will argue that such general relevance is illusory since the problem is, after all, merely a pseudo-problem and, therefore, needing not to be solved but to be dissolved. On the other hand, the extraordinary elusiveness of a solution is shown to have arisen from a specific kind of cognitive bias or error to which theorists, in particular, are prone.

Indeed, Nozick's judgement that no resolution has been completely convincing suggests that something essential has been overlooked. As some 'no box' accounts suggest, a radically new approach is needed that can meet the important desideratum of revealing the source of the problem's peculiar obduracy. In particular, I will suggest that, contrary to the nearly universal view, Newcomb's Problem does not raise any questions concerning rationality or decision theory. However, if this solution in the spirit of Jeffrey's apostasy suggests good news about our capacity for rational choice, it suggests bad news about other cognitive abilities of interest. As an instance of a familiar class of paradoxes, Newcomb's Problem is revealed as a manifestation of problems that have plagued theorizing about the mind (Slezak (1983, 1998, 2005)). In the extensive literature on Newcomb's Problem, Sorensen (1987, 1988), Priest (2002) and Maitzen and Wilson (2003) are among the few who appear to have noted the affinity with a family of problems having nothing essentially to do with rationality or decision theory. As I show along these lines, Jeffrey's late insight can be given precise formulation which reveals an unnoticed, but

not entirely coincidental, similarity between Newcomb's demon and that of Descartes: Just as Descartes' demon systematically thwarts our beliefs, so Newcomb's demon systematically thwarts our choices. Descartes's demon defeats our attempt to understand the world, while Newcomb's demon defeats our attempt to change it. The analogy between these cases has not been remarked upon undoubtedly because of the peculiar elusiveness of the separate problems in each case. As Freud (1953, p. 89) remarked of unrelated phenomena whose aetiology appeared strikingly similar, "So far-reaching an agreement can scarcely be a matter of chance." (On the other hand, of course, it might rather be as Nietzsche (1887, p. 228) said: "seeing things as similar and making things the same is the sign of weak eyes.")

2. NEWCOMB'S PROBLEM

The Problem involves a choice between two alternatives: Of two boxes A and B, you may choose either to take Box B only, or you may choose to take both boxes A and B. Box A contains $1,000; Box B contains either a million dollars or nothing depending on the prediction of the demon who places the money there. If the demon predicts you will choose only Box B, then he will place the million dollars in it. If he predicts that you will choose both boxes, he will leave Box B empty. This predictor is known from previous experience to be extremely reliable, making correct predictions 95% of the time. He makes his prediction, and depending on what he predicts about your choice, either places the million dollars in Box B or not. He departs and can no longer influence the outcome, and then you make your choice. Given the high reliability of the demon's predictions, the principle of subjective expected utility recommends taking only box B since there is almost certainty of winning a million dollars. However, since the demon either places the money or not prior to your choice and can no longer influence the situation, the principle of dominance recommends taking both boxes since

you will be $1,000 better off regardless of what the demon has done. There is no point leaving a certain gain of $1,000 when it cannot influence the outcome of the choice.

3. THE SHADOWY PREDICTOR

A central feature of Newcomb's Problem is the peculiarity of the apparent link between one's choice and the previously determined contents of the second box. A measure of the perplexity generated by this feature of the problem is noted by Lewis (1979) observing that some dismiss the puzzle as a "goofball" case unworthy of serious attention. Others like Gibbard and Harper (1978) propose simply ignoring the link and recommend the 'two-box' solution as rational despite being forced to admit that you will fare worse in choosing it. They explain:

> We take the moral of the paradox to be something else: If someone is very good at predicting behavior and rewards predicted irrationality richly, then irrationality will be richly rewarded. (1978, p. 369)

For reasons we will see, this "solution" has a distinct air of question-begging. One is inclined to reply that, if "irrationality", so-called, is richly rewarded, it must be *rational* to act in such ways. Gibbard and Harper may be seen as rehearsing Nozick's original amusing scenario, merely repeating one position loudly and slowly to opponents.

In this regard, Gibbard and Harper are not alone. The crucial issue is illuminated by a recent defence of the causalist position (McKay (2004)) that inadvertently brings into relief the fatal flaws in such a defence. In particular, it is important to notice that McKay uncritically assumes that there must be, in fact, a "right choice" – one box or two boxes. McKay suggests that recent attention to the predictor has been along the right lines, but "has not yet identified the right question" (2004, p. 187) which she says must concern the causal influence of one's choice on the predictor's decision. McKay

has correctly identified the locus of the malaise, but I suggest that she has given the wrong diagnosis of the pathology. The predictor's mysterious abilities and the connection with the agent's choices are, indeed, at the heart of the puzzle, but cannot be as McKay suggests. McKay correctly points out that "Intuitions waver in Newcomb cases because the one-box choice is attractive" (2004, p. 189), an attractiveness due to "implicit *causal* reasoning" (2004, p. 189) about the success rate of the "shadowy figure of the predictor" (2004, p. 187). While McKay is undoubtedly right about the profound *psychological* pull of the one-box choice, nevertheless she is doing little more than restating the problem.

For McKay, the "right question" is "can your action now in choosing one or both of the boxes have a causal influence on the predictor's decision?" (2004, p. 187) The burden of McKay's paper is to emphasize the predictor's unfathomable ability even though "the action of the predictor is in the past, and backwards causation is impossible" (2004, p. 187). Undeniably, this is the central enigma that has led others, too, to speculate about the anomalous correlation between one's choice and the demon's prediction. McKay suggests that the reliability of the predictor is so extraordinary "that it undermines your belief that your choice can have no causal influence on the action of the predictor" and even "challenges the conviction that the action of the predictor is genuinely in the past" (2004, p. 188). McKay's struggle to make sense of the predictor's uncanny ability is not only universal, but also symptomatic of a curious, self-imposed, and gratuitous limitation. McKay says that faced with the predictor's reliability, "it is not impossible that you would come to believe that there is some cleverly arranged cheating going on" (2004, p. 188). Indeed, if it were not *science-fiction* but a *real* case, we would be desperate to find some plausible basis for the phenomenon. Equally, if we encountered an Escher staircase in real life, we would be anxious to resolve the anomaly in a way that is consistent with geometry and physics, as Gregory (1981, p. 409) has actually demonstrated with an apparent real-life impossible Penrose triangle. However, McKay's anal-

ysis leads us astray by insisting that the appeal of the one-box choice "is due to implicit *causal* reasoning" (2004, p. 189) because she thereby fails to take seriously the fiction of an inherently occult "*acausal* synchronicity" between our choice and the predictor's actions. McKay is not at liberty to retell the story in a way that eliminates the puzzle arising from the predictor's supernatural ability. In common with almost all theorists, McKay takes the demon's miraculous powers as if they must be reconciled somehow with physical possibility (Eells (1982); Schmidt (1998)). But this is akin to wondering how the Road Runner can pass through solid rock or does not fall after running off a cliff until he notices. Philosophers have succumbed to the "passion of surprise and wonder" that Hume wrote about, leading the credulous to believe stories of miracles.

The peculiarity of the apparent link between one's choice and the previously determined contents of the second box is the central, defining feature of Newcomb's Problem. It is this mysterious acausal link that prompted Jeffrey's (1983, p. 25) earlier characterization of the problem as "a secular, sci-fi successor to the problems of predestination." Thus, the science-fictional nature of the problem frees, indeed *precludes*, us from the need to wonder about *how* such a predictor could possibly accomplish his success. McKay's approach is question-begging because the peculiarity of the link has been the motivation for distinguishing between a causal decision theory and evidential decision theory (Gibbard and Harper (1978), Lewis (1981)), the former discounting such spurious acausal connections. Indeed, the very difference between these two approaches was motivated by Nozick's original discussion (Levi (2000, p. 391)). As Gärdenfors (1988, p. 337) notes, "Newcomb's problem shows that causal independence may occur without probabilistic independence." Accordingly, McKay's insistence on the relevance of the causal connection is to miss the characteristic point of the puzzle. Unless it is construed purely as the *psychological* basis for our intuitions, McKay's claim that the one-box choice depends on "implicit causal reason-

ing" defeating evidentialism (2004, p. 189) is merely avoiding the problem by re-stating it.

4. UNDERDETERMINATION

In effect, McKay postulates an underdetermination in the weaker sense identified by Maitzen and Wilson due to lack of sufficient information. That is, her account boils down to the assertion that "there is no set answer to the Newcomb problem" because it depends entirely upon what we believe about the "underlying causal structure" (2004, p. 188). She says "You must decide whether or not there is a concealed causal connection" (2004, p. 188) and there are two possible answers depending the underlying causal structure, both answers being prescribed by the causalist. McKay concludes that the indeterminacy of Newcomb's problem is one of equipoise between the belief that there is, and that there is not, a causal connection between your choice and the demon's prediction. She says that recent concerns have been misplaced to the extent that they have not been "directed at wondering whether this counterfactual dependence implies a causal relation." Accordingly, she concludes "There are two possible answers and always will be, because the right choice depends on extra information about the actions of the predictor not given in standard descriptions of the case" (2004, p. 188, 189)

On McKay's side, we may note that Levi (1975, 1982) also blames under-specification of the choice for the perplexity of Newcomb's Problem. Levi regards the conditions of choice as "too indeterminate to render a verdict between the two options considered" (1975: 161). Levi suggests that "the details given in standard formulations of the Newcomb problem are too sparse to yield a definite solution according to Bayesian standards" (1982: 337) and he concludes that it is understandable that there should be a radical division of opinion on what to do in view of "the obscurities in Nozick's engaging presentation of Newcomb's problem" (1975, p. 164).

Accordingly, Levi declines to be classified as either a "one-boxer" or a "two-boxer" on the grounds of agnosticism.

I will suggest that such analyses are symptomatic of difficulties that lie elsewhere. Whereas, Levi and McKay see the choice problem as obscure, ill-defined or under-specified, on the contrary, I suggest that it is perfectly clear, fully specified but formally paradoxical. As Sorensen, Priest, and Maitzen and Wilson have noticed, the circumstances of the choice are not merely fantastic, but incoherent in a logical sense. That is, the predictor is not merely a fiction providing insufficient "extra information" as both McKay and Levi suggest. Rather, the choice is paradoxical in a strict and familiar logical sense, and thereby permits a precise specification revealing the source of the notorious perplexity. It is along these lines that Sorensen's (1987) "instability," Slezak's (1998) "disguised self-reference," Priest's (2002) "rational dilemma," and Maitzen and Wilson's (2003) "hidden regress" have important affinities. For their part, Matizen and Wilson go too far in their suggestion that "none of us can understand the circumstances which allegedly define a Newcomb's choice" (2003, p. 155). When properly identified, the incoherence of the decision problem is perfectly comprehensible and, indeed, familiar from other, related, cases. Nevertheless, these approaches promise to break the long-standing stalemate by rejecting standard assumptions about the nature of Newcomb's problem. The scenario involving a super-predicting demon has served to disguise these crucial logical features of the problem. The subtle discussions of conditional probabilities, expected utility and dominance principles of rational choice, have merely distracted attention from the real locus of the difficulty. In this sense the vast literature spawned by Nozick's original paper has served as a misdirection from the sleight-of-hand. What appears as a conflict between two principles of rational decision is, in fact, a clue to the peculiar nature of the problem. In particular, the vacillation between two impeccable but contradictory recommendations is induced by certain logical features of the conditions of choice. Intuitively, when contemplating the decision problem one feels a temptation to reverse one's initial deci-

sion in a futile attempt to outwit and thwart the demon's prediction (see Burgess (2004)). In Sorensen's formulation, we are faced with an 'instability' "If it is judged that p then ∼p; and if it is judged that ∼p then p" (1987, p. 307). This vacillation is a kind of indeterminacy that is familiar from analogous logical problems and a symptom of highly specific, pathological, conditions, as we will see.

5. SCHRÖDINGER'S CASH?

Thus, McKay is right to emphasize the mystery of the demon's abilities arising from *apparent* causality, the reason that some have dismissed the problem as a 'goofball' case, as Lewis (1979) noted. This anomalous "causation" is illuminated by an admittedly eccentric account that resorts to a familiar case of correlation without causation. Undoubtedly the silliest, and at the same time perhaps most insightful, analysis of Newcomb's problem is the one proposed by Wolf (1981) which appeals to Heisenberg uncertainty, quantum theoretic superposition of states and observer effects. Although the resort to quantum mysteries cannot be taken seriously in this context, it is a revealing symptom of desperation in the face of the extraordinary recalcitrance of the problem. However, Wolf's analysis is not *merely* absurd or desperate, and its appeal to quantum effects is analogous to McKay's resort to extra information about unknown classical "causal structure" (see also Eells (1982)). Both seek to deal with the mysterious reliability of the predictor by gratuitous appeal beyond the actual specifications of the puzzle. However, despite the manifest absurdity of invoking quantum effects, by taking the occult link between choice and box contents seriously, Wolf's analysis is actually, in a certain sense, the most illuminating of all:

> The answer is choose box R [second box only] if you want the million bucks. Your reward isn't caused by the Being's omnipotence or clairvoyance, however. It only appears that way to our Western preconditioned minds. ... the million dollars is in paradox-land where it is in the box and not in the box at the same time. Your act of observation creates the

choices – money there or money not there, according to whichever you choose. It is your act of observation that resolves the paradox. Choosing both boxes creates box R empty. Choosing box R creates it one million dollars fuller. (Wolf 1981, p. 150)

Like Schrödinger's cat, which is both dead and alive, the money is in a superposition of states, being both in the box and not in the box until you make your choice, whereupon the wave function collapses and *voilà*! However, when stripped of 'New Age' excrescences and the gratuitous invocation of quantum effects, Wolf's resort to an observer-induced effect captures the crucial peculiarity of the problem of uncaused correlation that McKay and others have grappled with. Though not through any quantum effects, the act of choice itself does, in a certain sense, create the state of box B.

6. NEWCOMB'S PROBLEM AS TWIN PRISONER'S DILEMMA

This feature of the puzzle may be appreciated by recasting Newcomb's problem in a way that reveals its formally paradoxical character. Jeffrey (1983) endorsed Lewis' (1979) insight that there is a formal isomorphism between Prisoner's Dilemma and Newcomb's Problem (see also Sobel (1985), Priest (2002)). Of particular interest is the case of duplicates or 'twin' prisoners who are assumed to be identical. The strategic or normal form representation of the game (Figure 1) shows your payoffs (your choices in rows, twin's choices in columns).

In this game you are confronted with a decision that is formally identical with that in Newcomb's paradox, assuming that the other player is an (almost) identical replica of yourself. To preserve the parallel, we assume that the *doppelgänger* is only very likely to make a choice identical with your own. Accordingly, maximizing expected utility requires that you choose one box, since your twin will almost certainly make the same choice and you get one million dollars. Obviously this is identical with the case in which Newcomb's

Your choices \ Twin's choices	2 Boxes	1 Box
2 Boxes	\$1,000 + 0	\$1,000 + \$1 Million
1 Box	0	\$1 Million

Figure 1. Twin Prisoner's Dilemma.

demon predicts your choice and places the million dollars in the box. If you choose both boxes, your twin will almost certainly make the same choice and you will both receive only \$1,000, just as if the demon accurately predicts your choice. Your hope of gaining \$1,000 plus one million is the hope of choosing both boxes while your counterpart chooses only one, just as in the case of the predictor wrongly predicting your choice.

Despite recognizing the formal analogy with twin prisoner's dilemma, commentators appear not to have drawn an obvious consequence for Newcomb's problem. The isomorphism of the two problems reveals that Newcomb's problem is a way of contriving a prisoner's dilemma against one's self. The other player in Newcomb's science-fiction version of prisoner's dilemma is actually one's self mediated by the predicting demon. From the isomorphism we see that Levi's and McKay's appeal to extra information, like Wolf's appeal to quantum mysteries, is a misdirection from the mischief – namely, the hidden self-referentiality that underlies Priest's (2002) diagnosis of "rational dilemma," Sorensen's (1987) instability and Maitzen and Wilson's (2003) hidden vicious regress.

7. HIDDEN REGRESS

Maitzen and Wilson (2003, p. 155, 160) suggest that the puzzlement arises from an infinitely long, infinitely complex

proposition that is incomprehensible and, therefore, "no one can understand the circumstances presupposed in the problem." They suggest that "something can look comprehensible without being so" as in their significant, but misleading, illustration of the Liar paradox. As we will see, their analogy with the Liar is closer than they appear to think, and the moral to be drawn from it is quite different. Maitzen and Wilson suggest, "the classical Liar sentence makes trouble only because people mistakenly take it to mean something. " They explain, "Every constituent of the sentence is comprehensible, but, arguably, the sentence itself is not" (2003, p. 159). On the contrary, however, the classical problem of the Liar arises precisely because the sentence is perfectly meaningful and appears to be both true and false. The paradox with its contradictory truth values would not arise if the Liar sentence were meaningless. Maitzen and Wilson miss the precise way in which the Liar paradox does, indeed, illuminate Newcomb's problem when the analogy and its analysis is properly understood. The regress they note is a symptom of a familiar circularity and paradox which is, nonetheless, perfectly intelligible.

Before returning directly to this perceptive analogy, we may first pursue the idea that Newcomb's problem is a Prisoner's dilemma against one's self arising from the paradoxical situation in which one's choice is based on deliberations that attempt to incorporate the outcome of this very choice. Newcomb's demon is simply a device for externalizing and reflecting one's own decisions. This hidden circularity facing the decision-maker in Newcomb's problem may be illuminated more directly by considering its standard formulation in extensive form (Figure 2), assuming that the demon makes his decision following the agent's choice, though without knowing what the agent's move was. This makes no material difference to the problem, but permits representing its logic more clearly. Thus, we are to choose either one box or two, whereupon the demon makes his move, not knowing what we have decided, basing his own action on his reliable knowledge of our behaviour as in the usual formulation.

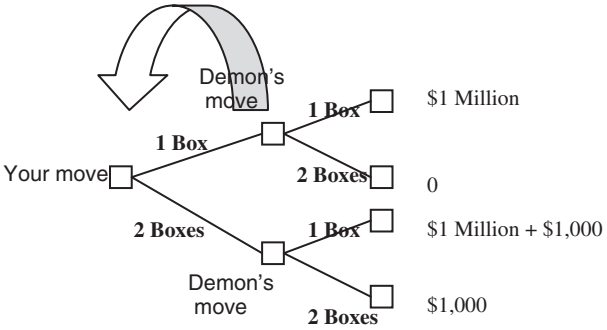

Figure 2. Newcomb decision tree.

As we contemplate our best move, we consider the next level in this game tree representing the demon's decision, which is actually a representation of the very same game tree and, in particular, the first-node that we currently occupy. That is, the branches from the second-level nodes are copies of the first node branches, since they represent the demon's *supposed* reflection on our decision at the first-node. As we deliberate, taking this situation into account, we are, in effect, representing the demon's deliberations as incorporating our own. The very hypothesis of such a demon requires conceiving that he is representing our current representations since the nodes and branches at the second level are actually duplications of the left-most node. From the diagram, we readily see the vicious circularity implicit in Newcomb's problem. The incoherence of this self-referentiality is hidden by the usual formulations.

8. "DELIBERATION CROWDS OUT PREDICTION"

Levi (1997: 80) and Schick (1979) have drawn attention to the problem arising when a deliberating agent adopts the posture of a spectator concerning his own performances. In such cases, the agent cannot adopt a predictive or explanatory attitude towards his own choices at the very time that deliberation is taking place. Levi says the agent cannot coherently

assign unconditional probabilities to hypotheses as to what he will do for "Deliberation crowds out prediction" (1997, p. 81) or as Schick puts it, "logic alone rules out our knowing the whole truth about ourselves" (1979, p. 243). Levi (1997, p. 32) remarks that "Prediction is precluded only for X predicting his own rational choice in the current context of deliberation." However, neither Levi nor Schick appear to offer this analysis of self-referential paradox specifically as a diagnosis of Newcomb's problem and the source of its puzzlement as I have been suggesting.

This alternative to the standard approach meets the important desideratum of revealing the source of the problem's peculiar obduracy. The present analysis confirms Priest's (2002) account of 'rational dilemmas' in which one is required to do the impossible, and gives a diagnosis of such pathologies for which, as Priest says "*Ex hypothesi*, rationality gives no guidance on the matter" (2002, p. 15). Furthermore, with the present account we see the specific mechanism giving rise to the impossible choice, thereby dissolving the pseudo-problem it presents.

As McKay notices, the temptation to take both boxes according to the dominance principle is the futile effort to win the extra $1,000 by outsmarting the infallible predictor. Although it is absurd, inevitably one thinks that one might outwit the demon by intending to take just box B all along, and switching choice at the last moment to take both boxes. Something like this has, in fact, been seriously discussed regarding "Tickles" and "Metatickles" (Eells (1984)) in which a presentiment is taken as evidence of the choice and the basis for the demon's prediction. The *actual* choice is then taken contrary to the presentiment or 'tickle'. Clearly, this is absurd since, *ex hypothesi*, as a reliable predictor, the demon would anticipate this sneaky strategy, but it captures something of the inescapable paradox of trying to avoid one's self – to flee one's own shadow. The futile strategy of attempting to trick the demon is explicitly recommended by Burgess (2004), but is ruled out by the stipulated conditions of the problem (see Slezak (2005)). The demon introduces an inessential step

in what is, in fact, the anticipation of one's own decisions. Thus, if not through backward causation, nor in the manner supposed by Wolf, the 'state of nature' is, after all, not independent of my choice. The demon serves as a sort of mirror of one's own deliberations.

9. SKYRMS' 'MEAN DEMON'

The essential logical features of the problem can be seen in a reformulation by Skyrms (1982), which eliminates the usual complexities of the conflict between expected utility and dominance principles. Instead, in this reformulation, a decision is required between two simple alternatives, one of which results in a reward. As before, however, the alternative which secures the reward depends on the prediction of a 'mean demon' concerning your choice. Consider, then, two boxes X and Y, one of which will contain a million dollars depending on the mean demon's anticipation of your decision. If the mean demon expects you to choose box X, he will put the money in box Y and vice versa.

Your reasoning must be as follows: On first deliberation, your choice is to take box X, but if you do, then the mean demon will almost certainly have anticipated this and put the money in box Y. Therefore, you should choose box Y. But, of course, the mean demon will have anticipated that you will make this second-level calculation involving his reasoning and put the money in box X after all. Therefore, you should choose box X as you originally intended, and so on, *ad infinitum*. The vacillation between choices is precisely parallel with the familiar vacillation of truth values in the Liar paradox where the sentence is alternately true and false, each one leading directly or indirectly to its opposite. Here, you should choose the opposite of whatever the mean demon thinks you will choose. If the mean demon is reliable, this means that you should choose the opposite of whatever you would choose! The best choice is whatever you decide not to do. This reformulated problem eliminates the complexities of the original

Newcomb Problem, which serve to distract attention from the essential logical features of the choice. In both cases, via the intermediary role of the predictor demon, the subject is placed in the position of trying to choose whatever he does not wish to choose. The demon merely serves to extend the loop, making the self-reference indirect. In attempting to anticipate his choice the subject is, in fact, attempting to anticipate his own.

The problem is beguiling because the assumption of the predictor seems merely an extravagant fiction, but appears to have camouflaged the circular reasoning it abets. Despite their differences, Eells, McKay, Burgess and Levi, *inter alia*, have sought some plausible causal structure, but this quest is like expecting cartoon characters to obey the laws of physics. Reflection on the demon in the course of deliberating requires considering a prediction about one's choice as part of the very act of making the choice itself. As Levi and Schick have noted, this is the illegitimate attempt to anticipate one's choice in the very act of making it.

10. GOOD NEWS AND BAD CHOICES

The foregoing analysis enables us to see how arguments based on so-called 'medical' or 'common cause' Newcomb cases have been misleading in providing a justification for causal decision theory (Eells (1982), Burgess (2004)). These cases are taken to be analogous to the original Newcomb Problem, but they differ in a crucial respect. Medical Newcomb problems are ones in which two independent events such as smoking and lung cancer have a common cause. In such a case, for example, smoking is not itself a cause of cancer but merely a manifestation of a personality trait or desire which is caused by a gene which also predisposes one to cancer. Smoking is then merely a symptom or indication of bad news that one has the deadly cancer gene but not itself causally relevant to contracting the disease. The choice whether to smoke or not appears to present a dilemma similar to Newcomb's Problem since one knows that smoking cannot cause cancer but pro-

vides unwelcome evidence that one is likely to be a victim. The temptation is to avoid smoking in order to avoid the disease, but this is plainly irrational since foregoing the pleasure of smoking in no way affects the prior genetic facts.

Relying on such cases, Lewis (1981) thinks that evidential or noncausal decision theory gives the wrong answer because "It commends an irrational policy of managing the news so as to get good news about matters which you have no control over" (1981, p. 377). That is, expected utility considerations appear to dictate the impotent manipulation of the cause by trying to suppress its symptoms. Lewis argues that "To decline the good lest it bring bad news is to play the ostrich" and "The trouble with noncausal decision theory is that it commends the ostrich as rational" (1981, p. 381).

However, Lewis's case rests crucially on the claim that the news in such medical Newcomb cases concerns "matters which you have no control over" – that is, the fact that a policy of managing the news "does not at all tend to prevent the evil" since, as he says, "there's nothing you can do about it now" (1981, p. 381). However, this independence of one's decision from the earlier causes does not hold in the original Newcomb Problem of the predicting demon for logical reasons already indicated. Accordingly, the medical cases are not strictly analogous to the original problem and are crucially misleading for the purpose of seeking a general solution. Specifically, the anomalous inter-dependence of the agent's choice and the demon's prediction in the original Newcomb problem is not preserved in the medical case, though it is crucial to the original problem and its special perplexity. Thus, medical Newcomb cases provide rational grounds for a policy in accordance with causal decision theory, but not the original form of the problem. The divergence of the two kinds of cases arises precisely because of the troublesome link between the agent's decision and the demon's prediction, which has no parallel in the medical case. As I have already noted, it is for this reason that Gibbard and Harper are forced to concede that their kind of "rationality" produces the worse outcome. The question-begging appearance of Gibbard and Harper's

causal decision theory arises because, like Lewis, they decide to ignore the intrinsically occult link between one's choice and the contents of the box – the defining feature of the problem. Ironically, or irrelevantly, this is to ignore a fictional constraint on the grounds that it could not be real. Realistic common cause problems would be a helpful model for dealing with Newcomb's problem if it were not for the fact that the assimilation has involved neglecting the predictor's supernatural power as not constituting an essential difference (Eells (1982)). However, the effort to find a "plausible" framework requires avoiding the very constitutive feature of Newcomb's problem and the source of its paradox. There can be no grounds for insisting on a plausible causal structure for a science-fiction story, and the persistent effort to do so has been largely to blame for the neglect of the underlying conceptual source of the puzzlement created by Newcomb's problem. In particular, contrary to Maitzen and Wilson, it is not that we can not understand the circumstances presupposed in the problem; rather, when we do understand them properly we recognize the logical incoherence of the problem and the pointlessness of the choice. Ironically, when Newcomb's Problem is fully understood in this way, it becomes susceptible to realization in easily contrived circumstances that simulate the choice situation and confirm its self-referential, paradoxical nature (Slezak (2006)).

11. SELF-REFERENTIAL PARADOXES

The source of the problem may be seen readily from a schematization in the same vein as Sorensen and Priest and clarifies the insight of Maitzen and Wilson. We may represent the relevant propositional attitudes as follows:

(x) I choose (a)

(y) The demon predicts (b)

We may then have the following substitutions:

(x*) I choose ~ (y*)
[I choose the opposite of whatever the demon predicts]

(y*) The demon predicts (x*)
[The demon predicts whatever I choose]

Substituting appropriately, we get:

(x*) I choose ~ (The demon predicts (x*))

Assuming that the demon predicts reliably and we may take whatever the demon predicts to be true,

(x*) I choose ~ (x*)

This means, "I choose the opposite of whatever the demon predicts" or "I choose the opposite of whatever I choose".

Of course, when the self-referential nature of the agent's deliberations are noticed in this way, it becomes clear that the situation is precisely analogous to the notorious Liar Paradox and the family of related conundrums arising from diagonalization. The Liar sentence may be given as:

(p) It is not the case that (p)

More generally, the problem is a version of the 'paradoxes of grounding' (Herzberger (1970)). Noticing this convergence is illuminating through assimilating the seemingly independent puzzle to a familiar class of problems, and explains why it should have remained so elusive.

Moreover, versions of the Liar paradox may be generated indirectly showing, contrary to Maitzen and Wilson (2003), that the problem does not arise from meaninglessness or incomprehensibility. Thus, for example, contradiction can arise not only from a sentence that asserts its own falsehood, but also indirectly as in the following pair of sentences:

(q) Sentence (r) is true.
(r) Sentence (q) is false.

Neither of these sentences is meaningless or paradoxical, but together they generate a contradiction. Newcomb's Problem has the structure of such indirect paradoxes in which the contradiction is mediated by intervening steps – perhaps accounting partly for the widespread failure to have noticed its character. In Newcomb's problem, the predicting demon acts as an intermediary serving to externalize what is, in fact, a loop in one's attempt to second-guess one's self. This is analogous to the way in which the Liar Paradox can be extended via intermediary agents whose beliefs extend the loop and thereby avoid a direct contradiction in the manner of sentences (q) and (r) above. In Newcomb's case too, the self-referential nature of the puzzle is obscured by the role of the predictor-demon, though it only extends the loop and does not essentially alter the self-contradictory nature of the decision problem.

12. DEMONS, DECEIVERS AND LIARS

In concluding, then, we are in a position to indicate the further, remarkable and revealing, similarity between Newcomb's demon and that of Descartes. In short, Newcomb's Problem and Descartes' *Meditations* are both variants of the Liar Paradox, and their notorious elusiveness is explained once this is recognized. Reflecting the universally held view of Descartes' *cogito* argument, Markie (1992) has noted the crucial importance of Descartes' claim to certainty about his thought and existence, but laments that "his account of how he gains this certainty turns out to be one of the most confusing aspects of his philosophy" (1992, p.141). In the same vein, Cottingham (1992, p. 1) states that Descartes' *Cogito ergo sum* "remains the most celebrated philosophical dictum of all time". However, the 'diagonal' analysis of Slezak (1983, 1988) supports a somewhat less reverent attitude in the spirit of a comment by the logician Bar-Hillel (1970). Bar-Hillel remarked in an aside that the confusion created by the phenomenon of indexicality when not fully understood "is partly responsible

for the otherwise almost incomprehensible veneration in which the Cartesian Cogito is held" (1970, p. 199). Bar-Hillel did not explain this cryptic comment, but the 'diagonal' analysis confirms his judgement and thereby reveals yet another instance of the perils of self-reflection. Descartes' systematic doubt may be represented as an enumeration of propositions that are successively subject to denial – that is, the contemplation of their falsity. Thus, his corpus of original beliefs may be a set of propositions:

(a) Roses are red.

(b) Violets are blue.

(c) Sugar is sweet.

etc.

Descartes' systematic doubt may be represented as entertaining successive substitution instances of the following formula:

(p) I doubt (q)

where (q) is (a), (b), (c) and so on, in turn.

Clearly, a possible substitution for (q) is (p) itself, which gives:

(z*) I doubt (z*)

This sentence is striking for its obvious analogy with the earlier Newcomb sentence

(x*) I choose $\sim$(x*)

In this case, Descartes' sentence (z*) says of itself that it is doubtful and has obvious similarities with the Liar sentence that says of itself that it is false. The sentence captures Descartes' insight, for attempting to doubt (z*) means considering it to be false. In turn, since (z*) asserts 'I doubt (z*)', its falsity means that I do not doubt (z*) or, in other words, that (z*) is certain. This diagonal sentence makes perfect sense of remarks which otherwise remain obscure or irrelevant. Just as Descartes says in his *Search for Truth*, "my doubt and

my certainty did not relate to the same objects: my doubt applied only to things which existed outside me, whereas my certainty related to myself and my doubting" (Descartes 1984, p. 418). Of interest here, we see here that Descartes' insight corresponds with Schick's (1979) independent articulation of the problems arising for self-knowledge in relation to choice: Schick's remark that "logic alone rules out our knowing the whole truth about ourselves" explains why Descartes' demon and Newcomb's are one and the same – the former systematically thwarts our beliefs, the latter systematically thwarts our choices. In both cases, the problem arises from the notoriously paradoxical features of self-reference, and in both cases the question of the very coherence of supposing such a demon may be raised.

13. CONCLUSION

Self-reference gives rise to well-known paradoxes in logic, which may be regarded as abstract schemata capturing deep psychological processes. These cognitive mechanisms may be inherent features of our mental representations of the world insofar as they attempt to encompass the self as part of the world. Long forgotten in the philosophical literature, an analysis of self-knowledge along these lines was given by Royce (1900) in his Gifford Lectures *The World and the Individual*, and more recently by Gunderson (1970) as an account of the aetiology of certain puzzles about the mind. Specifically, Gunderson suggests that it is the asymmetry between our perceptual, cognitive relation to our selves and the world which gives rise to the characteristic mind-body perplexities. Newcomb's Problem and its paradoxical features, like Descartes's *cogito* argument, may be due to the operation of the same self-referential schemata and may be yet another manifestation of the peculiarities of such tacit logical reasoning. Such reasoning may, after all, be seen as akin to the 'cognitive illusions' uncovered in the 'heuristics and biases' research programme, though in this case, being of a particular, limited intellectual

variety. Happily the domain of thought affected seems to be narrowly confined: The cognitive illusions in question appear to violate the norms of rational thought only in philosophical speculation.

ACKNOWLEDGEMENTS

I am very grateful for the comments by Paul Thogand and participants at a talk at the University of Waterloo, and to Shira Alqayam, Richard Schiffrin, Steve Sloman and other participants at the 27th Annual Conference of the Cognitive Science Society in Stresa, Italy 2005 where a version of this material was presented. I am also grateful to a referee of this journal for helpful advice.

REFERENCES

Allais, M. (1953), Le comportement de l'homme rationnel devant le risque, *Econometrica* 21, 503–546.

Bar-Hillel, Y. (1970), Indexical expressions, in *Aspects of Language*. The Magnes Press, Jerusalem.

Burgess, S. (2004), The Newcomb problem: An unqualified resolution, *Synthese* 138, 261–287.

Campbell, R. (1985), Background for the uninitiated, in Campbell, R. and Sowden, L. (eds.), *Paradoxes of Rationality and Cooperation*. The University of British Columbia Press, Vancouver.

Cottingham, J. (ed.) (1992), *The Cambridge Companion To Descartes*, Cambridge University Press, Cambridge.

Descartes, R. (1984), The Search for truth in *The Philosophical Writings of Descartes*, Vol. II, Translated by Cottingham, J., Stoothoff, R. and Murdoch, D. Cambridge University Press, Cambridge.

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

Eells, E. (1984), Metatickles and the dynamics of deliberation, *Theory and Decision* 17, 71–95.

Ellsberg, D. (1961), Risk, ambiguity and the savage axioms, *Quarterly Journal of Economics* 75, 643–669.

Freud, S. (1953), *Standard Edition of Complete Works of Sigmund Freud*, Vol. VIII. Strachey J. (ed.), Hogarth Press, London.

Gärdenfors, P. (1988), Causal decision Theory, in Gärdenfors, P. and Sahlin, N. (eds.), *Decision, Probability and Utility*, Cambridge University Press, Cambridge.

Gibbard, A. & Harper W.L. (1978), Counterfactuals and two kinds of expected utility, Reprinted in Gärdenfors, P. and Sahlin, N. (eds.), (1988) *Decision, Probability and Utility*, Cambridge University Press, Cambridge.

Gregory, R. (1981), *Mind in Science: A History of Explanations in Psychology and Physics*, Cambridge University Press, Cambridge.

Gunderson, K. (1970), Asymmetries and mind-body perplexities, in Radner, M. and Winokur, S. (eds.), *Minnesota Studies in the Philosophy of Science*, Vol. 4, University of Minnesota Press, Minneapolis.

Herzberger, H. (1970), Paradoxes of grounding in semantics, *Journal of Philosophy* 67, 145–167.

Jeffrey, R.C. (1983), *The Logic of Decision*. 2nd revised edition, University of Chicago Press, Chicago.

Jeffrey, R.C. (2004), *Subjective Probability: The Real Thing*, Cambridge University Press, Cambridge.

Levi, I. (1975), Newcomb's many problems, *Theory and Decision* 6, 161–175.

Levi, I. (1982), A note on Newcombmania, *The Journal of Philosophy* 79(6), 337–342.

Levi, I. (1997), *The Covenant of Reason: Rationality and the Commitments of Thought*, Cambridge University Press, Cambridge.

Levi, I. (2000), Review of James M. Joyce 'The foundations of causal decision theory', *Journal of Philosophy* 97, 387–402.

Lewis, D. (1979), Prisoner's dilemma is a Newcomb problem, *Philosophy & Public Affairs* 8(3), 235–240.

Lewis, D. (1981), Causal decision theory, *Australasian Journal of Philosophy*, 59, 5–30; reprinted in Gärdenfors, P. and Sahlin, N. (eds.), (1988). *Decision, Probability and Utility*, Cambridge University Press, Cambridge.

Maitzen, S. and Wilson, G. (2003), Newcomb's hidden regress, *Theory and Decision* 54, 151–162.

Markie, P. (1992), The cogito and its importance, in Cottingham, J. (ed.), *The Cambridge Companion to Descartes*, Cambridge University Press, Cambridge, pp. 140–173.

McKay, P. (2004), Newcomb's problem: the causalists get rich. *Analysis* 64(2), 187–89.

Nietzsche, F. (1887/1974), *The Gay Science*, W. Kaufman Trans. Random House, New York.

Nozick, R. (1969), Newcomb's problem and two principles of choice. in Rescher, N. (ed.), *Essays in Honor of Carl G. Hempel*. D. Reidel, Dordrecht.

Nozick, R. (1993), *The Nature of Rationality*. Princeton University Press, Princeton.
Priest, G. (2002), Rational dilemmas, *Analysis* 62, 11–16.
Resnik, M. (1987), *Choices: An Introduction to Decision Theory*, University of Minnesota Press, Minneapolis.
Royce, J. (1900), Supplementary Essay, in *The World and the Individual: Gifford Lectures First Series*; reprinted 1959, Dover, New York.
Schick, F. (1979), Self-knowledge, uncertainty, and choice, *British Journal for the Philosophy of Science* 30, 235–252.
Schmidt, J.H. (1998), Newcomb's paradox realized with backward causation, *British Journal for the Philosophy of Science* 49, 67–87.
Shafir, E. (1995), Uncertainty and the difficulty of thinking through disjunctions, in Mehler, J. and Franck, S. (eds.), *Cognition on Cognition*, MIT Press, Bradford, Cambridge, Mass. pp. 253–280.
Skyrms, B. (1982), Causal decision theory, *Journal of Philosophy* 79(11), 695–711.
Slezak, P. (1983), Descartes's Diagonal Deduction, *British Journal for the Philosophy of Science* 34, 13–36.
Slezak, P. (1988), Was Descartes a liar? diagonal doubt defended, *British Journal for the Philosophy of Science* 39, 379–388.
Slezak, P. (1998), Rational decision theory: The relevance of Newcomb's paradox, in Gernsbacher, M. A. and Derry, S. J. (eds.), *Proceedings of 20th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, NJ, pp. 992–996.
Slezak, P. (2005), Newcomb's problem as cognitive illusion, in Bara, B. G., Barsalou, L. and Bucciarelli, M. (eds.), in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, NJ, pp. 2027–2032.
Slezak, P. (2006), Realizing Newcomb's problem, *submitted*.
Sobel, J.H. (1985), Not every prisoner's dilemma is a Newcomb problem, in Campbell, R. and Sowden, L. (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, University of British Columbia Press, Vancouver pp. 263–274.
Sorensen, R. A. (1987), Anti-expertise, instability, and rational choice, *Australasian Journal of Philosophy* 65(3), 301–315.
Sorensen, R.A. (1988), *Blindspots*, Clarendon Press. Oxford.
Tversky, A. and Kahneman, D. (1974), Judgement under uncertainty: Heuristics and biases, *Science* 185, 1124–1131.
Wolf, F.A. (1981), *Taking the Quantum Leap*, Harper & Row. New York.
Wright, G. (1984), *Behavioural Decision Theory: An Introduction*, Penguin Books, Harmondsworth.

Address for correspondence: Peter Slezak, School of History & Philosophy of Science, University of New South Wales, Anzac Parade, Kensington, Sydney, NSW 2052, Australia E-mail: p.slezak@unsw.edu.au; Phone: +61-2-9385-2422; Fax: +61-2-9385-8003.