Realizing Newcomb's Problem

Peter Slezak
Philosophy Program
School of Humanities, University of New South Wales

p.slezak@unsw.edu.au

1 NEWCOMB'S PROBLEM

Box A is transparent and contains $1,000. Box B is opaque and contains either a million dollars or nothing. If a 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 demon is known to make correct predictions 95 percent of the time. He either places the million dollars in Box B or not and can no longer influence the outcome when you make your choice. The principle of subjective expected utility recommends taking only box B since there is almost certainty of winning a million dollars. However, the principle of dominance recommends taking both boxes since you will be $1,000 better off regardless of what the demon has done. (Nozick 1969)

2 GOOFBALL CASE?

David Lewis (1979) remarked that some have dismissed Newcomb's Problem as a "goofball" case unworthy of serious attention. However, similarities have been noted with realistic problems such as common cause or 'medical Newcomb' cases and Prisoner's Dilemma and the intellectual effort has been mainly devoted to reconciling the scenario of the predicting demon with some plausible causal structure. Nevertheless, I will suggest that these real-life analogs have been crucially misleading by diverting attention from the essential function of the science-fiction. I challenge the most widely held assumption that Newcomb's Problem may be given a coherent and consistent description and, therefore, realized in some way as a meaningful decision. That is, I suggest that Newcomb's Problem is worthy of attention precisely *because* it is a "goofball" case of a certain special kind. Neglecting the character and implications of the Predictor's mysterious power has led to missing its precise role in generating the perplexity and its recalcitrance. By analogy, Zeno's paradox of Achilles and the Tortoise is not resolved by trying to reconcile the anomalous conclusion with a calculation of their relative positions over time. The conclusion that Achilles cannot overtake the tortoise is taken as a *reductio* of the argument, and the intellectual task is to expose its fatal flaw. Newcomb's problem has not generally been approached in an analogous way, since the philosophical problem has been conceived as that of reconciling the scenario with some plausible causal structure (Eells 1982, Schmidt 1998, Burgess 2004, 2012). However, I suggest that, ironically, taking the puzzle seriously entails embracing the paradox rather than trying to avoid it. Of course, embracing the paradox requires clearly articulating the incoherence involved. Toward this end, I will consider the shortcomings of several recent accounts and thereby support an alternative analysis that reveals the incoherence of the puzzle. I suggest an experiment in which the choice problem can be actually realized in a straightforward but surprising way that confirms the proposed analysis.

The efforts to seek a causally plausible account of the Newcomb story have led some to conclude that the problem is under-determined or too obscure to permit univocal solution. Thus, McKay (2004) concludes "the right choice depends on extra information about the actions of the predictor not given in standard descriptions of the case." Levi (1975, 1982) also blames under-specification of the choice for the perplexity of Newcomb's problem. Levi says that the conditions of choice are "too indeterminate to render a verdict between the two options considered" (Levi 1975, 161) and "the details given in standard formulations of Newcomb's problem are too sparse to yield a definite solution according to Bayesian standards" (Levi 1982, 337). Levi concludes that, in the light of such under-specification and "obscurities" in the presentation of the problem, it is understandable that there should be a radical division of opinion on what to do and, therefore, he declines to be classified as either a one-boxer or a two-boxer.

On the contrary, however, I will suggest that the problem is neither obscure nor ill-defined but rather clear, though formally paradoxical. That is, the circumstances of the choice are incoherent in a precisely specifiable logical sense. The predictor is not merely an inadequately explained fiction that might be reconciled with a meaningful choice given further information, as both McKay and Levi suggest. Rather, the notorious perplexity may be shown to arise from a familiar paradox. It is in this broad sense that Sorensen's (1987) "instability," Slezak's (2005, 2006) "disguised self-reference," Priest's (2002) "rational dilemma" and Maitzen and Wilson's (2003) "hidden regress" share a "no box" view according to which the problem is ill-formed or incoherent in some way.[1] Belatedly joining the few no-boxers, along these lines Richard Jeffrey (2004) renounced his earlier position that accepted Newcomb problems as genuine decision problems. Jeffrey suggests cryptically "Newcomb problems are like Escher's famous staircase on which an unbroken ascent takes you back where you started" (Jeffrey 2004, 113). He adds that we know there can be no such things, though we see no local flaw in the puzzle. Jeffrey did not explain his suggestive remark further but his analogy is apt for a puzzle whose logical features can be precisely articulated.

3 SCHMIDT: PHYSICALLY PLAUSIBLE REALIZATIONS?

If Jeffrey's analogy is apt, it suggests that a realization of Newcomb's Problem will not have a plausible causal structure as commonly assumed. Nevertheless, J.H. Schmidt (1998) has been among those concerned to rebut the suggestion that Newcomb's Problem is 'incredible' or cannot occur and seeks to "prevent this beautiful paradox from being classified as physical nonsense" by providing a "physically plausible way in which it can be realized in a classical universe" (Schmidt 1998, 68). Schmidt claims to show that without causal paradox the player's choice "*influences* whether or not, in the past, the predictor put a million pounds into the second box" (1998, 67). Schmidt takes his physically plausible realization of Newcomb's Problem to show that Newcomb's Problem "actually involves backward causation" (1998, 82). However, despite such strong claims, Schmidt relies on an equivocation on the notion of causation to establish his central claim that backward causation may be involved.

Even charitably conceding that Schmidt's science-fiction story of futuristic miniature physicists might be realizable in a way that is consistent with physical laws, his account delivers rather less than it appears to suggest, and has the distinct air of question-begging when his caveats are fully taken into account. In a telling qualification at the outset, Schmidt says "I will not deny that there are other senses of 'causation', according to which there is no backward causation in this scenario" (1998, 69). At the very least, this is an ironic concession since these other senses of 'causation' are, in fact, the standard physical ones and, indeed, perhaps the only ones that have legitimacy at all. Obliquely, Schmidt is acknowledging that his own sense of

[1] I have used "no boxer" as shorthand for someone who recognizes the problem as incoherent and,

causation has dubious provenance and is open to question. Despite his talk of philosophical problems concerning singular event causation, it is clear that these are irrelevant to his concerns which are misleadingly couched in the language of causation but are only about an agent's subjective impressions. Thus, Schmidt explains:

> I therefore will not embark on the enterprise of constructing a general account of causation of my own. Instead, I will restrict myself to the discussion of a particular case, and argue that under the – admittedly rather extraordinary – circumstances described, we would have the *intuition* that we can, by an action in the present, influence an event in the past: i.e. that there is backward causation in that particular case. (Schmidt 1998, 69; *original emphasis*.)

However, the psychological predicament of the decision maker and his subjective impressions concerning backward causation *constitute* the problem and not the solution to the puzzle. It becomes evident that the sense of backward causation that Schmidt claims is merely "whether or not there is considered to be backward causation " as a "matter of personal judgement." His *façon de parler* in talking of "anthropically oriented causal description" (1998, 82) and asserting "that for our ordinary human purposes, there *is* backward causation" (1998, 77) is simply a way of saying, on the contrary, that there only *appears* to be backward causation as a matter of the agent's subjective impressions and avowed intuitions. Nevertheless, as we will see, Schmidt's discussion is of interest through reflecting the predicament of other theorists who also fail to distinguish the subjective problem facing the decision-maker from the problem facing the theorist or philosopher.

4 Burgess: Appointment in Samarra

Like Schmidt, and for analogous reasons, Simon Burgess (2004, 2012) also suggests that as decision-maker you are presently in a position to influence the contents of the boxes. Burgess, too, supposes that the predictor may be imagined as "an extremely technologically advanced fellow" who also relies on a brainscan to make his prediction. Whereas Schmidt thinks that he can influence the Predictor and recommends choosing one box, Burgess thinks he can outsmart the Predictor and recommends two boxes.

Burgess seeks to defend 'causal decision theory' over its rival 'evidential decision theory,' proposing that Newcomb's problem must be understood as a 'common cause' problem following Eells (1982). On this basis he argues "the evidence unequivocally supports two-boxing as the rational option" (2004, 261). It is important to note that Burgess does not consider the possibility that there might be no right choice at all in principle – the 'no box' alternative that has been independently raised by several authors for various reasons (Levi 1975, Sorensen 1987, Slezak 1998, Priest 2002, Maitzen and Wilson 2003).

We see a characteristic difficulty in Burgess' analysis when, like causalists Gibbard and Harper, (1988) he is forced to count the lesser expectation of two-boxing as what is, nonetheless, "most desirable." Thus, Burgess says "It must not be imagined that the option with the conventional conditional expected outcome of greatest monetary value will necessarily be the most desirable and hence rational for the agent" (2004, 269). At the very least, it is to strain ordinary usage to claim that greater monetary expectation is not necessarily the most desirable. Of course, this strained usage is symptomatic of a deeper difficulty. Burgess cites Lewis' similar response to the taunt 'if you're so smart, why ain't ya rich?' – namely, that riches are reserved for the irrational, and that the irrationality of one-boxers is richly "pre-rewarded" (Burgess 2004, 279). Gibbard and Harper (1978) too, 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. (Gibbard & Harper 1978, 369)

One is inclined to reply that, if "irrationality", so-called, is richly and consistently rewarded, it must be *rational* to act in such ways. What principle of rationality would recommend a course of action for every decision in life even if it were known to reliably result in a worse outcome? Inevitably such accounts may be seen as rehearsing Nozick's (1969) original scenario, repeating one position loudly and slowly to opponents.

Burgess defends causal decision theory by assimilating Newcomb's Problem to realistic 'common cause' problems where the recommendation of causal theory is acknowledged to be the rational one. But this assimilation is to stack the deck in favour of two-boxing in Newcomb's Problem unless the differences are inessential. However, I will suggest that the differences are crucial and Newcomb's problem can only be categorized as a case of common cause on certain untenable, question-begging assumptions.

Burgess follows Eells' (1982) analysis of the common cause structure, typified by the case of smoking and the cancer gene – often referred to as a 'medical Newcomb' case. It is assumed that, although there is a high statistical correlation between smoking and lung cancer, smoking does not actually cause cancer, but, rather, both are caused by a particular gene. Thus, despite providing unwelcome evidence that one has the cancer gene, smoking is the rational choice if it provides pleasure – the analog of choosing two boxes in Newcomb's problem. Since backwards causation is ruled out, these choices cannot affect the earlier facts for which they merely provide evidence. Like the Calvinist who believes in a predetermined soul, you might as well sin, just as you might as well smoke or choose both boxes. In Eells' terminology, the 'symptomatic acts' and 'symptomatic outcomes' in such cases are highly correlated but causally independent.

Burgess (2004, 283) takes Newcomb's problem to be "a distinctive kind of common cause problem in that you are presently in a position to influence the nature of the common cause." Burgess suggests "all you have to do to influence it appropriately is to make a commitment to one-boxing" – a strategy unavailable in the 'medical' cases because in those "the common cause is something genetic and thus effectively immutable."

We may understand Burgess's picture from a revealing remark in which he supposes that we might distinguish the *commitment* to one-boxing from the *actual choice* itself, thereby contriving a means to avoid the predictor's mysterious powers. Burgess divides the deliberation process into two stages – first, "the point at which the predictor gains the information used as the basis of his predictions" (2004, 279) by means of a brain-scan, and second, when Burgess alleges "the evidence unequivocally supports two-boxing as the rational option." However, this attempt to split the commitment from the actual choice is clearly a futile attempt to outwit the Predictor in a way that is ruled out by the specifications of the problem. The Predictor cannot be assumed to base his prediction on the wrong or irrelevant, earlier diagnostic brainstate – Burgess' BATOB in Figure 1. Of course, without such a spurious assumption, the parallel with common cause cases cannot be maintained. The two-stage strategy is futile because it misconceives or reformulates the problem, thereby avoiding it rather than solving it.

Burgess (2004, 284,5) suggests "Practically all those who fail to use the problem to become rich are simply ill-prepared." However, the very idea that one could get rich by choosing two boxes is the clearest symptom of the flaw in Burgess's account. As we have seen, even causalist advocates of two-boxing concede that you must fare worse, while pretending that it is somehow rational, nonetheless. Since the statistical pattern of the Predictor's success is a stipulation of the problem, there can be no question of getting rich by two-boxing. The suggestion that we might switch commitments in order to trick the predictor recalls the famous story told by W. Somerset Maugham: A servant is frightened when encountering Death in the market place of Baghdad and, taking the master's horse, flees to Samarra. Recounting the meeting to the master, Death says: "I was astonished to see him in Baghdad, for I had an appointment with him tonight in Samarra." Despite Burgess's attempted switcheroo, Newcomb's Demon, like Death, must be assumed to know the truth about your final choice and not merely an irrelevant precursor to it.

Following Eells (1982), Burgess' picture may be represented in the schema of Figure 1 in which the common cause is the 'Brainstate At the Time of the Brainscan' or 'BATOB'.

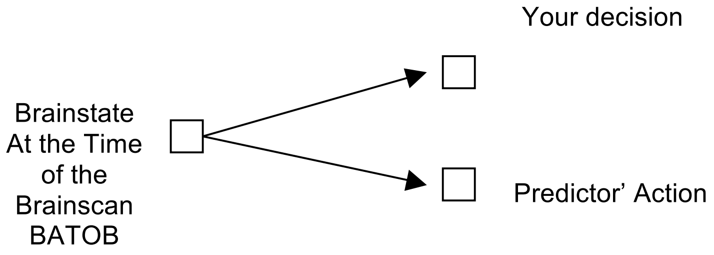

Figure 1.

The inadequacy of this schema as an analysis of Newcomb's problem is now easily seen. Among the crucial differences between Newcomb's Problem and common cause cases is the temporal sequence according to which the common cause must precede its two effects. It is only by separating some earlier diagnostic brainstate from the one directly responsible for (i.e., identical with) the actual decision that Newcomb's problem may be characterised in terms of Eells' schema for common cause problems. That is, Burgess assumes that the Predictor bases his action on some brainstate earlier than the one actually constituting the decision itself, but this opens a questionable gap in the causal sequence between the brainstate and the decision. As in the story of the precursor presentiment or 'tickle' (Eells 1984) this gap provides the room for Burgess' two-stage strategy and supposing that the actual decision might somehow deviate from the one indicated by the brainstate being scanned, thereby evading the demon's prediction. Clearly, however, this must be ruled out since, *ex hypothesi*, as a reliable predictor, the demon will anticipate such a sneaky strategy.

Burgess (2012) defends his 'two stage' account of the decision problem, but his account misses the force of criticisms along the lines just noted. Burgess has taken the brain scan story too literally and, thereby, introduces new, extraneous assumptions which significantly alter the problem. The brain scan story has been an inessential dramatization and embellishment of the Demon's method of prediction but it cannot be used in Burgess' manner to introduce irrelevant features of the scenario. In his original article introducing the problem, Nozick (1969) made no mention of brain scans but only "One might tell a science-fiction story about a being from another planet, with an advanced technology and science" (1969, 114). He adds "One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation … will be correct" (1969, 114). That is, the story of the brain scan is a purely illustrative heuristic device to dramatize the Predictor's abilities. His method of predicting choices is not specified and is not an essential part of the Problem or relevant to the source of its special puzzlement. Since the prediction is inescapable as a stipulation of the problem, Burgess has simply changed the story and thereby made his solution irrelevant. Burgess cannot give his analysis without the special features of the brain scan and its timing. However, the only essential, non-negotiable feature of the problem is the near-infallibility of the predictions. Therefore, Burgess' two-stage account, it can in principle have no bearing on the problem as he insists. In particular, the time of the brain scan BATOB cannot be separated from the *decision and actual action* since it is this inescapable prediction that justifies Jeffrey's (1983) characterization of the problem as "a secular, sci-fi successor to the problems of predestination." Nozick says only "The being gathers his data on the basis of which he makes his prediction" (1969, 132). Thus, we might change the illustrative embellishment to imagine that the Predictor uses a high-tech crystal ball to foretell future states of the physical world and, therefore, does not rely on any monitoring of the subject's brain at all. The problem is unchanged, but Burgess cannot run his version of the story.

Although futile, Burgess' ruse captures something of the inescapable paradox of trying to avoid one's self – to flee one's fate. As we will see, the assumption of the demon introduces an additional step in what is, in fact, the anticipation of one's own decision. We will see that, if not through backward causation, the 'state of nature' is, nonetheless, not independent of my choice. Thus, it may be acknowledged that a prior brainstate could be a highly reliable, even if not perfect, basis for predicting an agent's decision, but not for the agent himself.

5 Not Taking Science Fiction Seriously

In view of such efforts in the same vein, Eells' (1982) original assimilation of Newcomb's Problem to common cause cases is instructive. Significantly, Eells notes, "At first sight, it seems that there may be important differences between the decision situations of Newcomb's paradox and those of other, less fantastic, Newcomb situations" but he concludes "I do not think that these are really important differences" (1982, 210). We may examine Eells' revealing grounds for this assimilation. He writes:

> It seems that if the *agent* is rationally to have enormous confidence in the accuracy of the predictor ... then *the agent must believe* that there is a causal explanation for his success, though he may not know what that explanation is, and neither may the predictor. Indeed, it seems presupposed by much of our inductive reasoning that a high statistical correlation has a causal explanation. (Eells 1982, 210-11; emphasis added.)

Up to this point, Eells presents the predicament facing *the agent* in making sense of the decision problem according to the canons of inductive, causal reasoning. However, it is significant and typical that Eells slips from the perspective of the agent to the perspective of the philosopher or theorist in the immediately following remarks. He continues:

> The only *kind* of causal explanation of the predictor's success *that I can think of* that is consistent with the set-up of Newcomb's paradox is one that invokes a common cause ... Indeed, ... that is the only possibility, since the predictions do not cause the acts and the acts do not cause the predictions. ... Also, it seems that *on any plausible account* of any kind of successful prediction, the causal structure must be of this form. (Eells 1982, 210-11; emphasis added after first.)

Now Eells offers the common cause analysis as the only possibility *he can think of* as plausible according to the usual standards of scientific reasoning, thereby conflating the agent's perspective with that of the theorist seeking to explain the source of the apparent conflict among decision principles. This surreptitious collapse of the two perspectives has the effect of imposing irrelevant constraints on the theorist who need not, after all, be bound by those of the science-fiction story. Above all, there can be no requirement that an analysis must conform with what is a plausible causal structure such as Eells' illustration of weather prediction. The science-fictional nature of the problem frees us, indeed *precludes* us, from wondering how such a predictor could possibly accomplish his success. Furthermore, it is clear that imposing such requirements of plausibility on our account of the predicting demon must be entirely gratuitous. Eells continues:

> A successful predictor must have – consciously or unconsciously – a method, in the sense that the predictions are based on observations, conscious or unconscious. And if we look far enough back in the causal chain culminating in the relevant observations, we must be able to find factors that are causally relevant to the event predicted. It is easy to see that this is the causal structure involved in weather prediction, for example. (210-11; latter two emphases added.)

While Eells is surely correct in his remark that, on any *plausible* account, the structure must be of the form of common causes, the point is precisely that we, *as theorists*, are not required to seek a plausible account in this sense for a fantastic fictional concoction. The effort to do so has not merely involved inventions that go beyond the story's specifications, but has also, consequently, diverted attention from the source of the puzzlement generated by the science-fiction.

It is important to notice that Eells' concern with finding a 'causal structure' has been a consistent *leitmotif* in subsequent discussions. For example, Schmidt's "empathic" focus on the subjective intuitions of the decision-maker is not *merely* an error, but a seductive conflation of agent and theorist. *Qua* decision-maker constrained by the stipulations of the story, we inevitably seek to impose some plausible causal structure onto the subjectively puzzling situation we are confronted with, but *qua* theorist we obviously need not be confined in the same way. Similarly, McKay (2004, 188) writes "The right way to approach the Newcomb problem is to attempt to work out the underlying causal structure, just as the causalists prescribe." Evidently McKay, too, seeks to makes sense of Newcomb's problem along the line of the foregoing remarks by Eells, but my suggestion is that this has been a crucial mis-step.

Eells' final analogy with our reasoning in weather prediction is telling, since it indicates the effort to make the puzzle comprehensible in keeping with his view that there are no "really important differences" between Newcomb's problem and less fantastic situations. Significantly, Eells says that the common cause structure is the only kind of causal explanation of the predictor's success that *he* can think of. However, the question that remains unasked is why we should seek a plausible explanation at all in this way, rather than assume that the problem may be inherently incapable of being reconciled with *any* causal structure.

Similarly revealing in this regard is Burgess' suggestion that we consider the problem on the basis "that *you* are the subject of predictions" (2004, 262) – an invitation to consider the philosophical problem from the point of view of the deciding agent. For Burgess, this stance is adopted "simply for ease of exposition," but resort to the first person is not merely an innocent expository device. Notoriously, intractable puzzles arise when deliberations attempt to accommodate prior determination or fore-knowledge of the choice itself (Popper 1950).[2] As we will see, the significance of first-person reflection on

[2] Burgess (2012) dismisses my reference to Popper but he evidently fails to recognize the direct relevance of the logical, explanatory incoherence to which Popper is drawing attention.

one's own decisions has been discussed (Levi 1997, Schick 1979, Rabinowicz 2002), though its specific implications for Newcomb's problem appear not to have been fully recognized.

Thus, McKay (2004) suggests that the reliability of the "shadowy predictor" is so extraordinary, even though backwards causation is impossible, "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, 188). 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, 188). Indeed, if it were not *fiction* but a real case, we would be desperate to find some plausible basis for the phenomenon. McKay, like Eells, does not take the science-fiction seriously because she insists on reconciling it with science-fact. To be sure, taking Jeffrey's (2004) suggestive metaphor, if we encountered what appeared to be 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 Richard Gregory (1981) has actually demonstrated with an apparent real-life Penrose triangle. However, finding a respectable scientific analysis of Newcomb's Problem is surely not required for the solution of a science-fiction puzzle.

By seeking implicit causal structure in the problem, Eells. McKay and most other philosophers have failed to accept the inherently occult nature of the correlation between our choice and the predictor's actions. In particular, we are not at liberty to retell the story in a way that eliminates the puzzle arising from the predictor's mysterious ability, as many accounts do, for example, by gratuitous appeal to unspecified additional information. The demon's ability and the peculiar link between one's choice and the previously determined contents of the opaque box is the central, defining feature of Newcomb's problem. It is this mysterious link that prompted Jeffrey's (1983) original characterization of the problem as "a secular, sci-fi successor to the problems of predestination." The science-fictional nature of the problem frees us, indeed *precludes* us, from wondering, as both Eells and McKay do, how such a predictor could possibly accomplish his success. That is, solving Newcomb's problem may be achieved not only by showing how it may be reconciled with some plausible causal structure, but also by revealing exactly *why it can't* and, thereby, the inherent source of its paradox. Thus, McKay's (2004) insistence on the relevance of a causal connection is to miss the characteristic point of the puzzle. However, if we are not misled by analogous problems that *are* realizable, we may accept the inexplicable, occult correlation between our choice and the predictor's earlier action, and thereby focus on the *logical* structure of the puzzle, rather than its supposed *causal* structure.

This approach having nothing to do with rationality or decision theory, meets the important desideratum of revealing the source of the problem's peculiar obduracy. The present analysis is along the lines of Priest's (2002) account of 'rational dilemmas' for which "rationality gives no guidance on the matter" (2002, 15). When we see the specific mechanism giving rise to the impossible choice, we dissolve the pseudo-problem it presents.

6 Hidden Regress: Demon's and Liars

In particular, Newcomb's problem may be understood as a game against one's self in which one's choice is based on deliberations that attempt to incorporate the outcome of this very choice (Slezak 2006). Newcomb's demon is simply a device for externalizing and reflecting one's own decisions. This hidden circularity facing the decision-maker arises because, as we contemplate our best move, we consider the demon's decision, which is actually based on this very choice we are trying to make. As we deliberate, 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. We see the vicious circularity, the self-reference implicit in Newcomb's problem that is hidden by the usual formulations. You are, in effect, attempting to predict your own choice.

In Figure 2 we can see a portrayal of the manner in which the puzzle arises – namely, through the effort to represent the demon's reasoning, which in turn represents our own. Of course, the depiction of the demon on the left is that of a fiction, but the right hand picture captures our reasoning and the source of the puzzle in attempting to internally represent the hypothesized demon on the left. Of course, the puzzle arises in this way only for the decision-making agent or subject faced with the choice which is not necessarily the situation of the philosopher qua theorist.

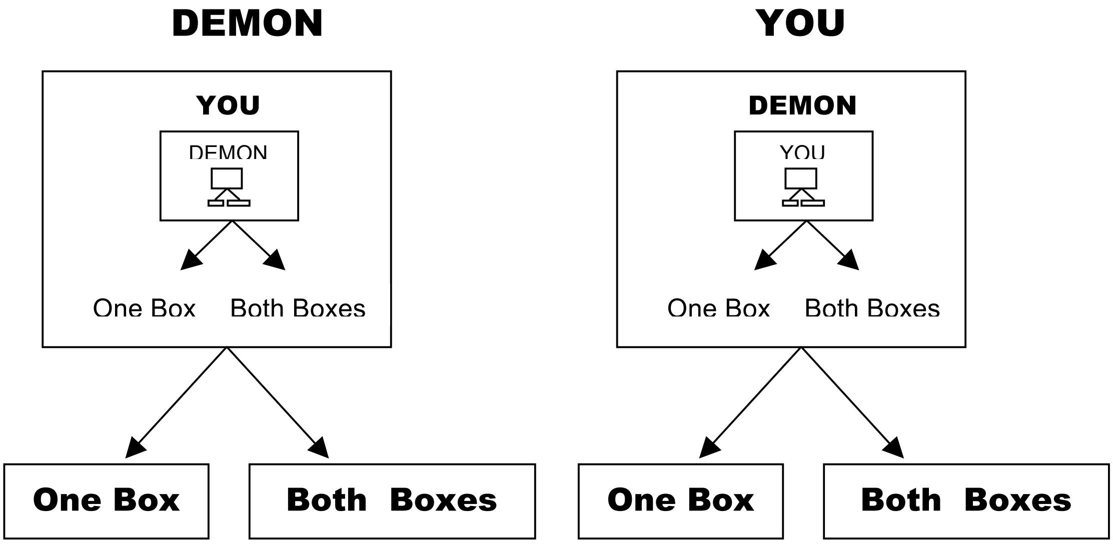

Figure 2.

The foregoing analyses suggest that the puzzle has important affinities with notorious self-referential paradoxes such as the Liar. However, contrary to Maitzen and Wilson (2002, 155), Newcomb's puzzle does not arise by analogy with an infinitely long, infinitely complex proposition that is incomprehensible. They claim that the analogy with the Liar consists in the fact that "Every constituent of the sentence is comprehensible, but, arguably, the sentence itself is not" (2002, 155). On the contrary, however, the 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 way in which the Liar paradox does indeed illuminate Newcomb's problem as we see when the analogy is properly understood.

Thus, the agent's deliberations and predicament may be represented alternatively with the relevant propositions as follows, where we see a familiar schema and paradox arising from self-reference:

(1) The demon predicts whatever I choose.

(2) I choose the opposite of whatever the demon predicts.

Therefore,

(3) I choose the opposite of whatever I choose.

The decision problem involves self-reference in a way that is reminiscent of familiar paradoxes in other domains. These features of the problem may be seen in a formulation by Skyrms (1982) that eliminates the complexities of the conflict between expected utility and dominance principles. As before, however, the choice that secures the reward depends on the prediction of a 'mean demon.' Given a choice between two boxes, if the mean demon expects you to choose box X, he will put the money in box Y and vice-versa. 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.

Your vacillation between choices is precisely parallel with the familiar alternation of truth values in the Liar paradox where the sentence is successively both true and false, each one leading directly or indirectly to its opposite. Thus, the familiar Liar sentence may be given as:

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

More generally, the problem arising from self-reference is a version of the 'paradoxes of grounding' (Herzberger 1970). Thus, contradiction can arise not only from a sentence that asserts its own falsehood, but 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 – the 'deferred' Liar. Newcomb's Problem has the structure of such indirect or deferred paradox in which the contradiction is mediated by intervening steps. The predicting demon acts as an intermediary serving to externalise what is, in fact, a loop in one's attempt to second-guess one's self.[3]

7 EXPERIMENTALLY REALIZING THE PREDICTOR

The foregoing analysis of an incoherent decision problem may be readily confirmed in an empirically realizable arrangement that precisely simulates the choice situation of the agent in Newcomb's Problem. The subjective predicament of confronting the predictive powers of Newcomb's demon may be easily simulated without resorting to the fanciful accounts we have seen.

We noted earlier that Schmidt's "strange but possible story" may be questioned regarding its realizability. His tiny sub-particle, super-predicting "dwarf" creatures whose scientific knowledge is millennia ahead of our own are surely questionable on the grounds of plausibility. However, the essential features of Schmidt's (1998) story, like that of Burgess, involving a brain-scan can be re-cast in a form that is readily realized in practice and tested in an actual experimental set-up with available techniques employing well-known facts of neuroscience.

An experiment permits confirming the foregoing suggestions about the structure of the problem. W. Grey Walter misled surgical patients into thinking that their voluntary action of pressing a button caused an effect such as advancing the carousel of a slide projector. In fact, the button was not connected to the slide projector at all and the slides were advanced directly by amplified signals from implanted cortical electrodes. Dennett (1991, 167) explains that the patients were startled by the effect because "it seemed to them as if the slide projector was anticipating their decisions." Clearly, the same arrangement could provide a startling impression that one's choice of boxes is being predicted in Newcomb's Problem. Instead of advancing a slide carousel, a computer could simply register whether a million dollars is placed in the opaque box or not. There can be little doubt about the subjective effect of such an arrangement. While unremarkable in itself, such an experiment suggests the source of Newcomb's paradox and elusiveness.

However, essentially the same experiment may be performed without such intrusive surgical procedures. Libet's (1985) work on the subjective delay of

[3] See also Sorensen, (1986).

consciousness of intention and the so-called “readiness potential” or “preparatory response” provides a means for a laboratory simulation of Newcomb’s Problem. We may obtain precise, reliable predictions of a subject’s actions from the prior state of their brain in a way that does not depend on any utopian neuroscience. EEG recordings from scalp electrodes show that the instant of a subject’s conscious decisions are between 350 and 400 milliseconds later than the onset of cerebral electrical activity that is the substrate of the voluntary action. These “readiness potentials” show that voluntary actions are preceded by unconscious neural activity and, as Dennett (1991, 163) puts it, “your consciousness lags behind the brain processes that actually control your body.” These data are not particularly surprising despite the seeming paradox for our naïve notions of free-will. However, my concern here is with the opportunity these phenomena provide for generating a laboratory experiment in which a subject is confronted with a precise and revealing simulation of Newcomb’s Problem. It is a trivial matter to connect the scalp electrodes to a computer screen in such a way that detection of the readiness potential for a choice would cause a million pounds or nothing to be placed in the second opaque box – before the subject consciously “makes the decision.” Since this would happen in the milliseconds prior to the subject’s conscious decision, it would amount to a prediction by the computer of the subject’s choice. It is clear that this experimental arrangement is a precise parallel to Schmidt’s elaborate fiction about sub-microscopic dwarfs with a futuristic physics or any standard account of Newcomb’s Predictor. Just like Schmidt’s creatures or Burgess’ brain-scan and other such suppositions, the electrodes rely on neurological activity to predict the subject’s decision, and the money is either placed in the opaque box or not according to the usual rule.

8 CHEATING THE SUBJECT & ELIMINATING THE DEMON

Finally, although realizable in practice, the technical difficulties of exploiting the readiness potential may be circumvented altogether without altering the logic of the scenario. The point may be more conveniently and more convincingly demonstrated by a modification of the experiment that dispenses with the foregoing methods altogether while preserving the essential features of the brain-scan and the decision problem. This time, the computer simulation of the boxes and their contents may be arranged so that a touch-screen or push-button registers the subject’s choice. However, an illusion of predicting the choice may be created by having the 'demon' place the money in the opaque box (or not) at the instant *after* the subject indicates his choice, but before the contents of the box are revealed to him. In other words, the subject's *actual* choice is used to give the *appearance* of having been predicted. Although cheating in an obvious sense, the subjective impression on the chooser (Schmidt's central concern, as we saw) would be identical with the case of genuine prediction by means of a brain scan. It should be clear that there is no essential difference from Newcomb's original problem, but in this case its logical structure and incoherence is now completely transparent: Although he does not know it, the subject is plainly making a futile attempt to incorporate the outcome of his own current decision into the very deliberations about it, unwittingly violating the precept of Schick and Levi. Newcomb's demon is the decision-maker himself.

REFERENCES

Burgess, S. 2004. The Newcomb Problem: An Unqualified Resolution, *Synthese*, 138, 261-287.

Burgess, S. 2012. Newcomb's Problem and its Conditional Evidence: A common cause of confusion, *Synthese*, 184, 319-339.

Chomsky, N. 1962, Explanatory Models in Linguistics. In E. Nagel, P. Suppes, A. Tarski, eds., *Logic, Methodology and Philosophy of Science*. Stanford: Stanford University Press, 528-550.

Dennett, D.C. 1991. *Consciousness Explained*. Harmondsworth: Penguin.
Eells, E. 1982. *Rational Decision and Causality*, Cambridge: Cambridge University Press.

Eells, E. 1984. Metatickles and the Dynamics of Deliberation, *Theory and Decision*, 17, 71-95.

Gibbard, A. & W.L. Harper, W.L. 1988. Counterfactuals and Two Kinds of Expected Utility, in P. Gardenfors and N. Sahlin eds., *Decision, Probability and Utility*, Cambridge: Cambridge University Press.

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

Herzberger, H. 1970. Paradoxes of Grounding in Semantics, *Journal of Philosophy*, 67, 145-67.

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

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

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: Cambridge University Press.

Lewis, D. 1979. Prisoner's Dilemma Is a Newcomb Problem, *Philosophy & Public Affairs*, 8, 3, 235-240; reprinted in R. Campbell and L. Sowden eds., *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, Vancouver: The University of British Columbia Press, 1985, 251-255.

Libet, B. 1985. Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action, *Behavioral and Brain Sciences*, 8, 529-566.

Maitzen, S. & Wilson, G. 2003. Newcomb's Hidden Regress, *Theory and Decision*, 54, 151-162.

McKay, 2004. Newcomb's problem: the causalists get rich, *Analysis* 64.2, 187-89.

Nozick, R. 1969. Newcomb's Problem and Two Principles of Choice, in N. Rescher, ed., *Essays in Honor of Carl G. Hempel*, Dordrecht: Reidel.

Popper, K.R. 1950. Indeterminism in Quantum Physics and in Classical Physics, *The British Journal for the Philosophy of Science*, 1, 117-33.

Priest, G. 2002. Rational Dilemmas, *Analysis* 62, 11-16.

Pylyshyn, Z. 2003. *Seeing and Visualizing: It's Not What You Think*. Cambridge, Mass.: Bradford/MIT Press.

Rabinowicz, W. 2002. Does Practical Deliberation Crowd Out Self-Prediction? *Erkenntnis*, 57, 91-122.

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.

Skyrms, B. 1982. Causal Decision Theory, *Journal of Philosophy*, 79, 11, 686-711.

Slezak, P. 2002. The Imagery Debate: Déja vu all over again? Commentary on Zenon Pylyshyn, *Behavioral and Brain Sciences*, Vol. 25, No. 2, April, 209-210.

Slezak, P. 2005. Newcomb's Problem as Cognitive Illusion, *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, B. G. Bara, L. Barsalou & M. Bucciarelli eds., Mahway, N.J.: Lawrence Erlbaum, 2027-2033.

Slezak, P. 2006. Demons, Deceivers and Liars: Newcomb's Malin Génie, *Theory and Decision*, 61, 277-303.

Sorensen, R.A. 1986. Was Descartes's Cogito a Diagonal Deduction? *British Journal for the Philosophy of Science*, 37, 346-51.

Sorensen, R.A. 1987. Anti-expertise, Instability, and Rational Choice, *Australasian Journal of Philosophy*, 65, 3, 301-315.