Rational Decision Theory: The Relevance of Newcomb's Paradox.

Peter Slezak (p.slezak@unsw.edu.au)

Program in Cognitive Science
University of New South Wales
Sydney NSW 2052 AUSTRALIA

Abstract

Among data implying pessimistic conclusions about human rationality, one might include evidence from the notorious Newcomb's Problem (Nozick 1969), which has hitherto, however, been largely confined to the philosophical literature. After nearly thirty years of inconclusive discussion, Newcomb's Problem is still widely seen as exposing inadequacies of the current standard theory of rational decision since the most plausible principles of choice give conflicting recommendations. Thus, Jeffrey (1983) says that Newcomb's Problem may be seen "as a rock on which ... Bayesianism ... must founder". Despite a staggeringly vast literature of great technical subtlety and complexity, no solution has emerged. I offer a novel analysis which goes beyond merely giving the right answer to the choice problem by also revealing the source of its persistent intractability. If my solution suggests good news about human capacity for rational choice, it entails bad news about other important problem-solving abilities central to cognitive science.

Introduction

The issue of rationality in reasoning and decision-making has become a hot topic in cognitive science largely through the stimulus of research by Kahnemann, Tversky, Wason, Johnson-Laird, St John Evans, and others. In particular, certain decision problems have become famous for being able to elicit choices that conflict with those recommended by utility theory. The two most famous of such "paradoxes" are those of Allais (1953) and Ellsberg (1961) which, like the cases of Kahnemann and Tversky, have led some to raise serious doubt about the extent of a much-vaunted human capacity for rational thought. Among such data implying pessimistic conclusions about human rationality, one might include evidence from the notorious Newcomb's Problem (Nozick 1969) which has, however, been largely confined to the philosophical literature. Nevertheless, like the research on various paradoxes, 'biases' and 'heuristics', Newcomb's Problem suggests deep and ill-understood puzzles in rational choice behavior. Despite almost complete neglect among cognitive psychologists (though see Shafir 1995), Newcomb's Problem should be of crucial relevance to cognitive science through revealing anomalies in our tacit principles of decision making.

After nearly thirty years of inconclusive discussion, Newcomb's Problem is still widely seen as exposing inadequacies of the current standard theory of decision making since 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, logician Richard Jeffrey (1983) says that Newcomb's Problem may be seen "as a rock on which ... Bayesianism ... must founder". In a similar vein, M. Resnik (1987, 111) declares "... this paradox has shaken decision theory to its foundations" and R. Campbell (1985, 3) says "Quite simply, these paradoxes ... cast in doubt our understanding of rationality...". Despite a staggeringly vast literature of great technical subtlety and complexity, no solution has emerged.

I propose a novel analysis going beyond the usual solutions which merely give a preferred or "right" answer to the choice problem. My account provides an essential further element of any satisfactory solution by revealing the source of the persistent intractability of the problem itself.

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 agent who places the money there. If the agent 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 percent 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 agent'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 agent 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 agent has done. There is no point leaving a certain gain of $1,000 when it can not influence the outcome of the choice.

Prima facie, at least, it is surprising that Newcomb's Problem has persisted in the literature for so long without any decisive solution. Recently returning to the problem he presented nearly thirty years ago, Nozick (1993) aims to "formulate a broadened decision theory to handle and

encompass this problem adequately" (1993, p. 41) and a wide variety of other solutions have been proposed. Among these, Richard Jeffrey shares David Lewis' view that Newcomb's Problem is "a prisoners' dilemma for space cadets: a secular, sci-fi successor to the problems of predestination" (1983, p. 25). However, none of these succeed in demystifying the problem and explaining why a solution has been so elusive. A truly insightful solution will reveal why the problem should have evaded the best philosophical minds who have toiled over it. That is, none of the many solutions proposed appear to decisively expose the source of the puzzlement.

Pseudo-Problem

Despite having been neglected, available analyses of Newcomb's Problem suggest its crucial relevance to cognitive science as a symptom of further anomalies in our heuristics for rational decision making. However, in venturing to offer a new approach here I must, at the same time, admit to sharing something like the sentiments of Wittgenstein's Preface to his *Tractatus*. At the same time as claiming to have solved all the problems of philosophy, Wittgenstein confessed that little had been accomplished since the problems were, after all, not real, but only pseudo-problems. It is in this spirit that I propose my "solution" to Newcomb's Problem. In this respect, the relevance of Newcomb's Problem to psychology turns out, after all, to be somewhat indirect and less serious than it appears on current accounts. However, demonstrating this is not without interest for cognitive science. Somewhat like Wittgenstein's ladder, it must be thrown away only after one has climbed up with it. Thus, I am concerned to make two inter-related points. First, I wish to draw attention to the unnoticed relevance of Newcomb's Problem to cognitive science given the currently available analyses. This is to place it among the paradoxes such as those of Allais and Ellsberg which are taken to raise questions about human reasoning and its normative foundations. Second, however, I will suggest that the standard approaches have failed to capture the precise source of the paradox in Newcomb's dilemma. If my own analysis is correct, then the decision problem and its consequences for cognitive science are, in fact, quite different from what is entailed by existing theories. If my solution suggest good news about our capacity for rational choice, it suggests bad news about other cognitive abilities of importance. As an instance of a certain familiar class of paradox, it is revealed as a further manifestation of deep cognitive illusions which have plagued theorising about the mind (see Slezak 1982, 1983, 1984).

The lack of empirical or theoretical research may be due, in part, to the science-fictional nature of the choice situation which appears impossible to realise in an experimental set-up, omniscient predictors being hard to find in practice. Close real-life analogs of Newcomb's Problem have been discussed in the literature, but these have been mentioned only anecdotally and not studied systematically for what they might reveal about human choice behaviour.

Due to the limitations of space, I will proceed here by ignoring most of what has been written in the vast literature on the subject because my resolution of the problem appears to be radically distinct from available analyses and is, therefore, neither derived from, nor illuminated by, them. In particular, Newcomb's Problem is universally construed, following Nozick's (1969) original presentation as raising problems of decision theory, however I suggest that these problems are merely symptomatic of difficulties which lie elsewhere. In the thirty years of extensive analysis it appears not to have been noticed that the circumstances of the choice, though seemingly plausible if somewhat extravagant, are in fact incoherent and logically impossible. That is, the choice involving the predictor is not merely fantastic in a science-fiction sense, but paradoxical in a strict logical sense.

What appears as a conflict between two principles of rational decision is, in fact, a clue to the peculiar nature of the problem. The vacillation between two impeccable but contradictory recommendations is induced by certain specific characteristics of the conditions of choice. The scenario involving a super-predicting demon has served to disguise certain crucial logical features of the problem. The extensive and subtle discussions of expected utility and dominance principles of rational choice, have merely distracted attention from the real locus of the problem. In this sense the vast literature spawned by Nozick's original paper has served only as a misdirection from the real sleight-of-hand.

Self-Referential Paradoxes

In particular, a vast amount of speculation concerning conditional probabilities and causal decision theory has been precipitated by the mysterious link between the choice and its consequences via the earlier action of the Predictor, even though this cannot be a causal connection. However, the link need not be as mysterious as widely believed. The deliberations leading to a choice of boxes must make inescapable reference to the Predictor and his criteria for placing the $M, since the choice necessarily involves calculating the consequences of the Predictor's actions. However, the Predictor's actions themselves are predicated upon these very deliberations and their outcome. Thus, we see an inescapable circularity in the conditions of the choice as the subject attempts to incorporate features of his own reasoning into themselves. Calculating his preferred choice involves taking into account the demon's prediction which, in turn, must take into account the subject's own calculations. In short, the subject is caught in a self-referential loop of trying to predict his own actions. His choice attempts to anticipate the outcome of the choice itself.

When the self-referential nature of the subject's deliberations are noticed, it becomes clear that the situation is closely analogous to the notorious Liar Paradox and the family of related conundrums. Noticing this fact is illuminating through assimilating the seemingly independent puzzle to a familiar class of problems. Furthermore, of course, noticing the relatedness of Newcomb's Problem to the class of self-referential paradoxes explains not only why it should have remained so recalcitrant, but also the peculiar vacillation it induces.

The failure to have noticed that Newcomb's Problem is a variant of familar self-referential paradoxes appears to be due to the fact that its formulation serves to disguise this connection. In order to expose these links we may briefly

attend to the central features of self-referential paradoxes such as that of the Liar. In its most familiar and direct form, the Liar Paradox arises with a sentence (1) which asserts its own falsehood:

(1) This sentence is false

Notoriously, sentence (1) is true if it is false, and it is false if it is true. Without rehearsing two thousand years of inconclusive debate, we may note only that the contradiction arises in this and related cases due to two features of the sentence.

Self-Reference. First, there is a loop or self-reference in all cases of paradox. Thus, for example, Russell's set theoretical paradox turns on essential reference to classes whose members are classes, giving rise to the possibility of classes which are members of themselves. In its popular formulation, Russell's Paradox concerns the village barber who shaves all and only those who do not shave themselves. The question is who shaves the barber? Does he shave himself? Similarly, contradiction arises from a seemingly innocuous definition of the adjectives 'autological' and 'heterological' as those words which do or do not apply to themselves respectively. Thus, most adjectives do not apply to themselves: the words 'red', 'frightening', 'flexible', 'delicious', 'carnivorous' etc., do not possess the properties to which they refer. We may define these as 'heterological'. However, a few adjectives do apply to themselves and we may define these as 'autological': 'English', 'short', 'polysyllabic', 'unambiguous', 'pronounceable' etc. The question now is: To which class does the adjective 'heterological' belong? If it applies to itself, it must be autological, but in that case it must have the property to which it refers and be heterological. And if it is does not apply to itself, being heterological, then it is autological! Aha! Gotcha.

Not all self-reference leads to contradiction. Thus, sentence (2) is not paradoxical:

(2) This sentence is true.

Negation. In addition to self-reference, a further element is required to generate paradox, namely, negation. Thus, the Liar sentence asserts of itself that it is *not* true, Russell's paradoxical class is that which is *not* a member of itself, 'heterological' applies to itself only if it does *not* etc.

Indirect Paradox. Finally, it is crucial here to notice that paradox may be generated in less direct ways. For example, contradiction can arise not only from a sentence which asserts its own falsehood, but indirectly as in the following pair of sentences (3) and (4):

(3) Sentence (4) is true.
(4) Sentence (3) is false.

Neither of these sentences is paradoxical on its own, but together they generate a contradiction and the chain of such sentences could be extended indefinitely.

Newcomb as the Liar

We are now in a position to see more clearly how Newcomb's Problem may be diagnosed as a version of self-referential paradox. It can be seen to have precisely the key features we have just noted which lead to contradiction, namely, self-reference and negation. Moreover, Newcomb's Problem has the structure of indirect paradoxes in which the contradiction is mediated by intervening steps - perhaps accounting partly for the universal failure to have noticed its character.

In Newcomb's problem, the Predictor acts as an intermediary serving to externalise what is, in fact, a loop in one's attempt to second-guess one's self. This is closely 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 (3) and (4) above. Here too, the self-referential nature of the puzzle is obscured by the role of the Predictor, though it only extends the loop and does not essentially alter the self-contradictory nature of the problem.

A Reformulated Variant of Newcomb

The essential logical features of the problem can be seen clearly in a reformulation 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 an agent concerning your choice. Consider, then, two boxes X and Y, one of which will contain a million dollars depending on the Predictor's anticipation of your decision. If the Predictor 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 Predictor will almost certainly have anticipated this and put the money in box Y. Therefore, you should choose box Y. But, of course, the predictor 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 Predictor thinks you will choose. If the predictor 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 concerning utilities etc. which serve to distract attention from the essential logical features of the choice. In this distilled form the contradiction and analogy with the Liar are evident and permit us to see how the original Newcomb Problem is also essentially of the same form. In both cases, via the intermediary role of the Predictor, the subject is placed in the position of trying choose whatever he doesn't wish to choose. The Predictor 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.

In the original Newcomb's Problem, the temptation to take both boxes according to the dominance principle is the futile effort to win the guaranteed extra $1,000 by

outsmarting the predictor and thereby outsmarting one's self. Although it is absurd and even amusing, inevitably one thinks that one might outwit the Predictor by intending to take just box B all along and only switching choice at the last moment to take both boxes. Something like this has, in fact, been seriously discussed in the literature on "Tickles" and "Metatickles" (Eells, 1984). Clearly, this is absurd since the Predictor would anticipate this strategy, but it captures something of the inescapable paradox of trying to avoid one's self - to flee one's own shadow. The Predictor introduces an inessential step in what is, in fact, the calculation and anticipation of one's own decisions. Deliberating about the Predictor is indirectly to deliberate about one's own current deliberations. Thus, the state of nature is not independent of my choice - not in the sense of backward causation, but because the very conception of such a state intrinsically embodies assumptions about that choice. Conversely, the choice I make is itself determined by considerations which attempt to encompass supposed knowledge or information concerning the choice itself. In making the choice I am trying to weigh up facts including those which determine the outcome of the choice - of course, via the intermediary of the Predictor. The Predictor serves as a sort of mirror of my own deliberations. The problem is beguiling because the assumption of the predictor seems merely extravagant in a science-fictional sense, but otherwise innocuous. However, the assumption appears to have camouflaged the circular reasoning it abets. Reflection on the postulated circumstances in the course of deliberating itself requires considering a prediction about the choice as part of the very act of making the choice itself. Thus, attempting to take into account the prediction involves attempting to anticipate the choice in the very act of making it. The situation is impossible not because it involves science fiction speculations, but because it involves self-contradiction of a notoriously familiar sort.

Conditional Probabilities

The failure to recognize the foregoing considerations as relevant to Newcomb's Problem has meant that elaborate analyses have been undertaken in terms of conditional probabilities. However the talk of conditional probabilities in the context of Newcomb's Problem has been seriously misleading since, the link in question here is conditional but not probabilistic. Of course, the difficulty is that the link is not causal either and hence the incentive to develop 'causal decision theory' as a critique of, and alternative to, conditional expected utility theory. Causal decision theory as articulated by Gibbard and Harper (1978) recommends the maximization of an expected utility different from conditional expected utility, this alternative utility being formulated in terms of "nonbacktracking" counterfactual conditionals instead of conditional probabilities. However, this might be seen as an elaborate question-begging. The problem with conditional probabilities in the context of Newcomb's Problem is that it does not explain the nature of the link and the underlying basis for it. In this sense the conditional probability is merely a symptom of a deeper problem rather than a full explanation as such. It follows that existing analyses at this level have failed to diagnose the source of the difficulty in Newcomb's Problem or to deal with it decisively.

The Causal Decision Theory was introduced in order to deal with the fact that causal independence may occur without probabilistic independence. The article of Gibbard and Harper (1978) argues that Bayesian decision theory cannot account for such probabilistic dependence with causal independence and "solves" Newcomb's problem by claiming that the absence of causal connection removes the conflict of principles by making the probabilistic dependence irrelevant. That is, we should act as if we had probabilistic independence and take both boxes as recommended by the Dominance Principle. This has an air of question-begging about it and the paradoxicality of their conclusion seems to make little advance on Nozick's original statement of the paradox. In response to the seeming irrationality of recommending a policy (taking both boxes) which is guaranteed to be worse than the alternative they write:

> 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)

Good News and Bad Choices

A realistic analog to Newcomb's Problem is the situation in which two independent events have a common cause. For example, we might assume that smoking is not a cause of lung cancer but merely a manifestation of a desire or personality trait which is caused by a gene which also predisposes one to cancer. Smoking is then merely a symptom or indication that one has the deadly cancer gene but not itself causally relevant to contracting the disease. The choice whether to smoke or not presents a dilemma similar to Newcomb's Problem since one knows that smoking cannot cause cancer but provides 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.

Gibbard and Harper talk of the tendency or temptation to bring about an *indication* of a desired state of the world, even if it is known that the act that brings the news has no effect in bringing about the desired state itself. However, such talk about the impotent manipulating of news simply misses the point. Specifically it misses the nature of the connection between the "news" and the event it seems merely to announce. Despite the absence of any causal connection, the universal conclusion that there is no relevant connection has been too hasty and has overlooked what is, after all, a crucial link. To appreciate the specific nature and the force of this link it is essential to distinguish the science-fictional or metaphysical aspects of the Newcomb scenario and its logical features. That is, the science-fiction features may be accepted as unproblematic but should not obscure quite different problems namely, the stricly logical issues confronted by the an agent in the supposed circumstances in attempting to reason in order to make a choice. In focussing attention on the reasoning process, I have been concerned to draw attention to purely logical or conceptual issues confronted by one who tries to calculate the best course of action under the supposed circumstances. Such a reasoner is forced to make a sequence of inferences which

are not merely problematic regarding principles of choice as generally assumed.

Demons, Deceivers and Liars

There is an uncanny but not entirely coincidental similarity between Newcomb's demon and that of Descartes. Just as Descartes's demon systematically thwarts his beliefs, so Newcomb's demon systematically thwarts his choices. Descartes' 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 despite the prima facie similarity, but this oversight is undoubtedly due to a curious failure to notice the kind of self-referential paradoxes to which I have been referring. It is perhaps no accident, therefore, that both Newcomb's Problem and Descartes' doubting argument have been unusually recalcitrant. Whether or not the controversy over Newcomb's Problem might persist for another three hundred years like Descartes's, there are grounds for seeing their peculiarity as deriving from common sources. Thus, Slezak (1983) has set out the logical features of Descartes' notorious argument which leads to the conclusion 'Cogito ergo sum' and reveals this to have the structure of 'diagonal' arguments like Gödel's Theorem and the Liar Paradox. The fact that the Cogito argument is a variant of the Liar Paradox helps to explain many of the textual and philosophical features of this notorious argument which remain obscure on other accounts. In particular, the inexplicable certainty of Descartes's insight follows as a strict logical consequence of his attempt to doubt everything: In short, everything may be open to doubt except doubt itself, which thereby comes to be self-certifyingly certain. The interest and relevance of Descartes in the present context is seen from the suggestion (Slezak 1983) that the peculiarities of the Cogito argument reflect certain deep, inherent cognitive mechanisms arising from attempting to understand one's self as part of the world. Such self-reference gives rise to well-known paradoxes in logic which may be 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. An analysis along these lines has been given by K. Gunderson (1970) as an account of the aetiology of certain puzzles about the mind. Specifically, 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 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 be seen as influenced by 'heuristics' and 'biases' of a particular intellectual variety. Happily the domain affected seems to be narrowly confined: The cognitive illusions in question appear to violate the norms of rational thought only in philosophical speculation.

References

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

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

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.

Gibbard, A. & Harper W.L. (1978). Counterfactuals and Two Kinds of Expected Utility. Reprinted in Gärdenfors, P and Sahlin, N. (1988). *Decision, Probability and Utility*, Cambridge: Cambridge University Press.

Gunderson, K. (1970). Asymmetries and Mind-Body Perplexities' in M. Radner and S. Winokur eds., *Minnesota Studies in the Philosophy of Science, Vol. 4*, Minneapolis: University of Minnesota Press.

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

Nozick, R. (1970). Newcomb's Problem and Two Principles of Choice. N. Rescher et al. eds. *Essays in Honor of Carl G. Hempel.* Dordrecht: D. Reidel.

Nozick, R. (1993). *The Nature of Rationality.* Princeton: Princeton University Press.

Resnik, M. (1987). *Choices: An Introduction to Decision Theory*, Minneapolis: University of Minnesota Press.

Shafir, E. (1995). Uncertainty and the difficulty of thinking through disjunctions, in J. Mehler and s. Franck eds., *Cognition on Cognition*, Bradford: MIT Press, 253-280.

Slezak, P. (1982). Gödel's Theorem and the Mind, *British Journal for the Philosophy of Science*, 33, 41-52

Slezak, P. (1983). Descartes's Diagonal Deduction, *British Journal for the Philosophy of Science*, 34, 13-36.

Slezak, P. (1984). Mind, Machines and Self-Reference, *Dialectica*, 38, 1, 17-34.