STEPHEN MAITZEN and GARNETT WILSON

NEWCOMB'S HIDDEN REGRESS

ABSTRACT. Newcomb's problem supposedly involves your choosing one or else two boxes in circumstances in which a predictor has made a prediction of how many boxes you will choose. We argue that the circumstances which allegedly define Newcomb's problem generate a previously unnoticed regress which shows that Newcomb's problem is insoluble because it is ill-formed. Those who favor, as we do, a "no-box" reply to Newcomb's problem typically claim either that the problem's solution is underdetermined or else that it is overdetermined. We are no-boxers of the first kind, but the underdetermination we identify is more radical than any previously identified: it blocks the very set-up of the problem and not just potential solutions to the problem once it has been set up. The defect is subtle, but it cripples every genuine version of the problem, regardless of variations in such things as the predictor's degree of reliability, the basis on which the prediction is made, or the amount of money in each box. The regress shows that, surprisingly enough, no one can understand Newcomb's problem, and so no one can possibly solve it.

"Suppose that you have two options: ... to take the contents of an opaque box in front of you or ... to take the contents of the opaque box *plus* the contents of another box which is transparent and obviously contains $1,000 in cash. Since you cannot see the contents of the opaque box, choosing it alone may result in getting nothing. ... In the opaque box is either one million [dollars] or nothing, depending on whether a certain being, called the Predictor, has or has not placed $1,000,000 there *prior* to the time at which you are to make your decision. You know that the Predictor will have placed $1,000,000 there if, and only if, the Predictor has predicted that you will choose the opaque box alone. ... You know, moreover, that the Predictor is almost 100 per cent reliable. Imagine that the evidence to confirm this reliability is enormous, and the Predictor has done a detailed study of your personality and past behavior. You are practically certain that the Predictor will be right this time. ... Which option is it rational for you to choose?"[1]

Theory and Decision **54:** 151–162, 2003.

1. THE REGRESS

Newcomb's problem supposedly involves your choosing one or else two boxes in circumstances in which the Predictor has made a prediction[2] of how many boxes you will choose. We argue that the circumstances which allegedly define Newcomb's problem generate a previously unnoticed regress which shows that Newcomb's problem is insoluble because it is ill-formed.

Which circumstances do you face in Newcomb's problem? Imagine, first, circumstances (C1) in which you confront the two boxes – the transparent box obviously containing $1K, and the opaque box whose contents you cannot see – but you lack any belief about whether the contents of the opaque box depend on the Predictor's prediction of how many boxes you will take. Since Newcomb's problem stipulates that you believe the opaque box's contents *do* depend on such a prediction, it is clear that you do not face a Newcomb's choice in C1. In C1, moreover, it is equally clear that your *only* rational action is to take *both* boxes: you have absolutely no reason to leave behind the box obviously containing $1K! C1, then, cannot be the right circumstances: you lack a belief necessary for facing a Newcomb's choice, and your rational action is so obvious as to be completely uninteresting. This is not yet Newcomb's problem.

Granted, we are assuming, as Lawrence Davis puts it, "that what is rationally prescribed for an agent is *relative to the information he has*."[3] The information you *have* depends, in turn, on the beliefs you have. Indeed, starting with Robert Nozick's seminal article,[4] discussions of Newcomb's problem, including the one quoted in our epigraph, typically assume that you *know* that the opaque box's contents depend on the Predictor's prediction about your choice. Although only the weaker propositional attitude of *belief* is required for you to face a Newcomb's choice, our criticism is unaffected if your belief also counts as knowledge. Now, one can distinguish the rationality of your action *given* your beliefs from the rationality, or advisability, of your action *irrespective of* your beliefs. In the latter sense, it can be rational for you to decline the transparent box (if, say, it will explode when touched) while remaining rational in the former sense for you to take the transparent box (if you have no reason to think it will explode when touched). Traditionally, Newcomb's problem has concerned the rationality of your action in light of your beliefs – it would be an entirely different problem if it did not – and we continue that tradition here.

The claim that there is no Newcomb's problem in C1, besides being entailed by the conditions of the problem itself, gains independent support when one considers, again, what it is always rational for you to do in C1. Facing the two boxes without believing that the contents of the opaque box depend on the Predictor's prediction about your choice, you have only one rational option: take both boxes. Having no reason to leave behind the clear $1K, and having some reason to take it, you ought to take it. Thus, again, if C1 described Newcomb's problem, the two-box solution would be obviously and trivially correct. But even if the two-box solution to Newcomb's problem is ultimately correct, it is not obviously or trivially so.

Imagine, then, circumstances just like C1 except that you *do* believe that the contents of the opaque box depend on the Predictor's prediction of how many boxes you will take. Only in the latter circumstances can you face a Newcomb's choice. It is just these circumstances, however, that generate a vicious regress. For consider the following partial specification of them:

(C2) Circumstances in which you confront the two boxes, believing that the Predictor has made a prediction of how many boxes you will take.

C2 ends with the phrase "how many boxes you will take," and the question immediately arises, "take in which circumstances?" "*Those* circumstances!" is no answer at all, for what *are* those circumstances? Are they circumstances in which you fail to believe that the opaque box's contents depend on the Predictor's prediction of how many boxes you will take? Surely not: those are not the circumstances of a Newcomb's choice, and, again, in those circumstances taking two boxes is trivially correct. The answer, of course, is "circumstances in which you believe that the opaque box's contents depend on the Predictor's prediction of how many boxes you will take." To repeat, though: "how many boxes you will take" in which circumstances? The answer "*Those* circumstances!" is no more adequate here than before, and so the regress continues.

To put it another way, Newcomb's problem asks you to decide *how many boxes you will take* if you intend to choose rationally. If the answer to that question were obviously and trivially "two," then, without injustice to any side in the debate over Newcomb's problem, we could partially specify the circumstances of a Newcomb's choice as follows, where "a prediction of two boxes" is short for "a prediction that you will take both boxes":

(C3) Circumstances in which you confront the two boxes, believing that the Predictor has made a prediction of two boxes.

In Newcomb's problem you believe, or even know, that the Predictor has determined the contents of the opaque box just as described in our epigraph. Therefore, if C3 even partially specified the circumstances of a Newcomb's choice, you would be plainly irrational to take only the opaque box. If you believed that the Predictor had made a prediction of two boxes, and accordingly had left the opaque box empty, you would choose the opaque box alone only if (a) you failed to draw the obvious inference that the transparent box offered you your only chance at money, or else (b) you didn't wish to maximize your winnings. Option (a) is incompatible with your making a choice which is rational in any interesting sense; if we cannot assume that you will draw an inference as obvious as the one just described, there is no interesting way to theorize about your rational choices. Option (b) violates the crucial assumption of Newcomb's problem that you *do* wish to maximize your winnings. In sum, if C3 even partially specified the circumstances of a Newcomb's choice, then Newcomb's problem would have a trivial two-box solution and would not deserve the attention it has received; in short, it wouldn't be Newcomb's problem.

But C3 *does* partially specify the circumstances of a Newcomb's choice – that is, C2 just *is* C3 – unless C2's phrase "how many boxes you will take" is understood as elliptical for the longer phrase "how many boxes you will take when you believe that the Predictor has made a prediction of how many boxes you will take." Like C2, however, the longer phrase also ends with "how many boxes you will take," and either *that* occurrence of "how many boxes you will take" trivially means "two boxes," or it does not. If it trivially means "two boxes," then Newcomb's problem has a trivial two-box solution which deprives the problem of any interest; if it does not trivially mean "two boxes," then it must be elliptical for the longer phrase "how many boxes you will take when you believe that the Predictor has made a prediction of how many boxes you will take." And so on *ad infinitum*. There is no way to stop the regress without making the two-box solution to Newcomb's problem trivially correct. On the assumption that the two-box solution (even if correct) is not trivially correct, the circumstances of a Newcomb's choice turn out to be impossible to describe in finitely many words. Since none of us can understand an infinitely long description, none of us can understand the circumstances which allegedly define a Newcomb's choice. In that case, none of us can understand, let alone solve, Newcomb's problem.

Notice that a similar regress arises in the two-player, one-shot Prisoner's Dilemma,[5] which stipulates that

(P) Player A's payoff depends on whether player B cooperates.

Given the standard conditions of the problem, including the stipulation that both players are rational maximizers, no player cooperates unless she believes that her opponent will also cooperate; without such a belief, any rational maximizer defects.[6] Thus,

(Q) Any player cooperates only if she believes that her opponent will cooperate.

Prisoner's Dilemma also standardly assumes that each player's rationality includes believing the logical consequences of everything she believes. Given this assumption, propositions P and Q entail the following:

(R) Player A's payoff depends on whether player B believes that player A believes that player B believes that . . .

We contend that the infinitely complex proposition R is incomprehensible,[7] in which case no one genuinely understands the conditions that allegedly define the one-shot Prisoner's Dilemma. To put it another way, if defection is trivially required of any rational player, then the one-shot Prisoner's Dilemma hardly deserves to be called a "dilemma." But defection *is* trivially required of any rational player unless she believes that her opponent will cooperate, and the belief that her opponent will cooperate cannot be spelled out without launching an infinite regress.[8]

This pessimistic conclusion need not, however, apply to the *iterated* Prisoner's Dilemma, because in the iterated version proposition Q is false: a rational maximizer can cooperate even without believing that her opponent will cooperate. The remarkably successful strategy of Tit-for-Tat illustrates the latter point. Tit-for-Tat begins by cooperating and then simply mimics the opponent's moves; it need not assume that the opponent will ever cooperate, and it is not an irrational strategy even if the opponent never cooperates. By contrast, iteration does not rescue Newcomb's problem from the infinite regress we have identified, since it is the very set-up of Newcomb's problem which launches the regress.

In essence, our main argument is as straightforward as it seems, but it can be further clarified by our answering three objections, to which we now turn.

2. OBJECTIONS AND REPLIES

Objection A. Newcomb's problem does not require the concept of *prediction* at all. All it requires is that you, the player, accept these statements of conditional probability:

(S) Prob(Opaque box contains $1M|You take only opaque box) is high.

(T) Prob(Opaque box contains $0|You take only opaque box) is low.

(U) Prob(Opaque box contains $1M|You take both boxes) is low.

(V) Prob(Opaque box contains $0|You take both boxes) is high.

Provided that S–V do not depend on your causally influencing the contents of the opaque box, we have all the ingredients that give Newcomb's problem its importance and interest. Likewise, classic cases of common cause – such as the fanciful "gene" example in which a gene causes both lung cancer and the desire to continue smoking – produce the same conflict of decision-theoretic principles, all without invoking prediction.

Reply: It is by no means clear that Newcomb's problem can dispense with the concept of prediction. After all, you must believe S–V while *also* believing that your choice does not causally influence the contents of the opaque box. This puzzling set of beliefs needs some motivation if it is to be rationally held, and the story of the Predictor allegedly provides that motivation. Notoriously, some who profess sympathy for the one-box solution, given the story of the Predictor, nevertheless profess no sympathy for the one-box analogue in the gene example (*viz*., quitting the symptomatic behavior), even given the story of a common genetic cause.[9] This fact suggests to us that some notion of prediction is crucial to motivating the one-box solution.

Nevertheless, even if the objector is correct about the dispensability of prediction, the regress still arises. A conditional probability "Prob($\phi \mid \iota$)" is defined as "Prob($\phi \& \iota$) $\div$ Prob(ι)," provided, of course, that Prob(ι)>0. But there's the rub. As before, if you choose rationally, the probability of your taking only the opaque box must be *zero* – and thus the conditional probabilities in S and T must be *undefined*, making S and T both false – unless you accept S and T. But you will rationally accept S and T only if you regard the conditional probabilities in S and T as well-defined, as capable of being high or low. Your doing so requires, in turn, that you assign non-zero probability to your taking only the opaque box. Again, though, any interesting theorizing about your rational choice presupposes that you draw all obvious logical inferences, and so you assign *zero* probability to your taking only the opaque box *unless* you accept S and T. But you accept S and T only if you regard the conditional probabilities in S and T as well-defined. Your regarding them as well-defined requires that you assign non-zero probability to your taking only the opaque box. . . and so on. S–V thus generate their own infinite regress.

Interestingly, even as staunch a one-boxer as Terence Horgan does not consider the one-box analogue in the gene example – quitting smoking – to be rational unless the example is modified as follows:

> Let the agent believe that the genetic factor in question induces in smokers a tendency to choose to continue smoking *when confronted with the present decision problem*; and let him believe (implausible though this may be) that smokers who lack the genetic factor have a tendency to choose to stop smoking *when confronted with this problem*.[10]

The phrases we have italicized represent Horgan's crucial modifications of the original gene example, and they too generate an infinite regress. In the phrase "when confronted with the present decision problem," which decision problem is being referred to? Obviously, it is a problem in which the agent believes that a gene causes both (a) lung cancer and (b) the desire to continue smoking when confronted with *the present decision problem*. Again, though: which decision problem is that? In short, the regress crops up even in cases which do not involve prediction.

Objection B. Every game involves the regress, or circularity, you've identified, because the notion of rational agents that is presupposed in game theory is circular. Consider any game between two players, A and B. Then

(W) A is rational if A chooses a strategy that maximizes A's expected payoff, given the assumption that B is rational.

Precisely the same definition applies to B, interchanging "A" and "B". So, if we like, we can substitute that definition for the phrase "B is rational" in W and produce an expanded definition which is explicitly circular, since it now contains the phrase "A is rational." The very concept of Nash equilibrium is thus a circular concept.[11]

Reply: While it is not entirely clear, apparently the objector intends W as a *definition* of the phrase "is rational," but in that case W is fatally uninformative. No one can grasp this definition of "is rational" without already knowing what the phrase means. If the objector's idea is that our regress is benign because definition W is perfectly adequate, then we reject the presupposition: W is nowhere near adequate. Indeed, it seems that the objector has made our point for us.

Objection C. Why can't we rely on our *intuitive* grasp of the notion of prediction in coming to understand Newcomb's problem? After all, we rely on just such an intuitive grasp in order to understand, say, the concept of an omniscient, foreknowing God despite the fact that fully reflective belief in such a God produces a regress: If you believe that an omniscient God foreknows your actions, then you are committed to God's foreknowing your actions *given* your belief in God's foreknowledge of your actions *given* your belief... and so on. Yet presumably we can understand this concept of God in spite of the regress. The circumstances of a Newcomb's choice are no less comprehensible.

Reply: We do not share the objector's assumption that the concept of omniscient foreknowledge is obviously comprehensible; we ourselves would not claim to comprehend it. A number of philosophers have forcefully challenged the coherence of omniscience and therefore, by extension, omniscient foreknowledge,[12] and although we do not have an independent argument against the comprehensibility of the latter concept, to say that Newcomb's problem is *no less* comprehensible is hardly to rescue Newcomb's problem. While an infinite divine intellect, if such exists, presumably grasps the concept of omniscient foreknowledge, it seems likely that no one among us finite intellects *obviously* grasps it, even if some among us in fact do manage to grasp it. Attacking the coherence or intelligibility of omniscience is nothing new, but we know of no previous arguments claiming that the circumstances of a Newcomb's choice are impossible to grasp. The objector, then, is defending the comprehensibility of Newcomb's problem by equating it to a concept whose comprehensibility is at least as controversial.

3. CONCLUSION

At this point in the debate, we find, objections often take forms that do not merit further response, such as (a) merely repeating a version of C2, the partial specification which we have already argued is elliptical (and viciously regressive when one attempts to spell it out), or (b) merely expressing bewilderment or incredulity. Granted, it does *look* as if the circumstances of a Newcomb's choice are entirely comprehensible, but something can look comprehensible without being so. According to one standard approach to the Liar paradox, for instance, the classical Liar sentence – "This sentence is false" – illustrates this fact. Every constituent of the sentence is comprehensible, but, arguably, the sentence itself is not, since, arguably, it fails to express a proposition at all. On this approach, the classical Liar sentence makes trouble only because people mistakenly take it to mean something, namely, what it seems to mean. In a similar way, students of Newcomb's problem have understandably but mistakenly assumed that they grasp the circumstances of a Newcomb's choice well enough to solve the problem.[13]

Those who favor, as we do, a "no-box" reply to Newcomb's problem typically claim either that the problem's solution is underdetermined or else that it is overdetermined.[14] We are no-boxers of the first kind, but the underdetermination we identify is more radical than any previously identified: it blocks the very set-up of the problem and not just potential solutions to the problem once it has been set up. The defect is subtle, but it cripples every genuine version of the problem, regardless of variations in such things as the Predictor's degree of reliability,[15] the basis on which the prediction is made,[16] or the amount of money in each box. The regress is consistently suppressed in discussions of Newcomb's problem, but once made explicit it shows that, surprisingly enough, no one can understand the circumstances presupposed in the problem, and so no one can possibly solve the problem. It is time to stop pretending otherwise.

ACKNOWLEDGEMENTS

For helpful comments, we thank Ann Levey, William Vanderburgh, Ian Wilks, and Sheldon Wein.

NOTES

1. Richmond Campbell, "Background for the Uninitiated," *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, ed. Richmond Campbell and Lanning Sowden (Vancouver: University of British Columbia Press, 1985), pp. 3–41; p. 22.
2. Here we prefer the phrase "has made a prediction" to the simpler "has predicted," since there is a reading of "has predicted" on which it is a success verb, on which whatever the Predictor has predicted must occur. The phrase "has made a prediction" carries no such implication, and we do not wish to *assume* inerrancy or infallibility on the part of the Predictor in Newcomb's problem. As we indicate below, however, our argument goes through whether the predictor is inerrant, infallible, or neither.
3. Lawrence H. Davis, "Prisoners, Paradox, and Rationality," in Campbell and Sowden (*op. cit.*), p. 52, emphasis in original.

4. Robert Nozick, "Newcomb's Problem and Two Principles of Choice," *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher, et al., (Dordrecht: D. Reidel, 1969), pp. 114–146.
5. This result is not terribly surprising given our criticism of Newcomb's problem and given David Lewis's argument in "Prisoner's Dilemma is a Newcomb Problem," *Philosophy and Public Affairs* 8 (1979): 235–240.
6. According to Philip Pettit, "one of the firmest intuitions around is that cooperating in a one-shot prisoner's dilemma is irrational" ("The Prisoner's Dilemma is an Unexploitable Newcomb Problem," *Synthese* 76 (1988): 123–134; 123). If this firm intuition is correct, if no rational player cooperates, then it follows that any rational player cooperates only if she believes that her opponent will also cooperate. This conclusion follows because, in truth-functional logic, if a proposition "ϕ" is false, then the conditional "ϕ only if ι" is true. Even in standard *non*-truth-functional logic, if a proposition "ϕ" is *impossible*, then the conditional "ϕ only if ι" is true – and Pettit's firm intuition implies that it is not only false but impossible that a rational player cooperates. Therefore, proposition Q – because it contains the phrase "only if" rather than "if and only if" – is consistent with the firm intuition which Pettit identifies.
7. Notice that R is not equivalent to a finite proposition such as (R$*$) "Each player's payoff depends on whether she believes that the other player will cooperate," although, together with the other conditions stipulated in Prisoner's Dilemma, R$*$ implies R.
8. Prisoner's Dilemma also often assumes that both players possess strict "common knowledge," an assumption which itself generates a regress: Both players know all of the game conditions, including the condition that both players know all of the game conditions, including the condition that. . . and so on. Common knowledge is the subject of a literature too vast for us to address here, and in any case it is unclear to us that the resulting regress makes all versions of Prisoner's Dilemma incomprehensible. We have argued that one-shot Prisoner's Dilemma is incomprehensible because of a *different* regress, one not afflicting iterated Prisoner's Dilemma, and one which arises not from assuming common knowledge but from assuming that a rational maximizer might cooperate in a one-shot game.
9. For example, Nozick (*op. cit.*, p. 135) concedes that the one-box solution to Newcomb's problem is not obviously wrong, but he describes the one-box analogue in the gene example – *viz.*, intentionally avoiding behavior which is merely symptomatic of the presence of the gene – as "perfectly wild" (p. 126).
10. Terence Horgan, "Counterfactuals and Newcomb's Problem," *Journal of Philosophy* 78 (1981): 331–356; 354, emphases added.
11. We owe this objection, *verbatim*, to an anonymous referee.
12. See, in particular, Patrick Grim, "Logic and the Limits of Knowledge and Truth," *Noûs* 22 (1988): 341–367; *The Incomplete Universe* (Cambridge, MA: MIT Press, 1991); and "The Being that Knew Too Much," *International Journal for Philosophy of Religion* 47 (2000): 141–154.

13. Of course, challenging people's pretensions to understand things they in fact do not understand is a philosophical tradition going back to Socrates; arguably, it is the essence of philosophy.
14. For references to no-boxers of each kind, see Campbell, "Background for the Uninitiated," p. 24.
15. Among those who think that infallibility in the Predictor causes a special problem of overdetermination are Don Hubin and Glenn Ross, "Newcomb's Perfect Predictor," *Noûs* 19 (1985): 439–446.
16. J. L. Mackie, "Newcomb's Paradox and the Direction of Causation," *Canadian Journal of Philosophy* 7 (1977): 213–225, argues that any conceivable explanation of the Predictor's reliability makes the structure of Newcomb's problem incoherent.

Address for correspondence: Stephen Maitzen, Department of Philosophy, Acadia University, Wolfville, Nova Scotia B4P 2R6, Canada
E-mail: smaitzen@acadiau.ca

Garnett Wilson, Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia B3H 1W5, Canada
E-mail: gwilson@cs.dal.ca