'A new take from Nozick on Newcomb's Problem and Prisoners' Dilemma'
by S. L. Hurley

Analysis, 1994, vol. 54, no. 2, pp. 65-72.

In a salient part of his recent book *The Nature of Rationality* (Princeton: Princeton University Press, 1993), Robert Nozick offers a new argument about Newcomb's Problem and Prisoners' Dilemma. This new argument is salient in its own right, but also in part because Nozick himself was responsible for the first article on Newcomb's Problem, in 1969 ([3]), and this is his first return to the problem since then. It's illuminating to compare some of his earlier remarks with his present view.

In Newcomb's Problem, I'm offered the choice of taking either just the opaque box in front of me, or both it and another transparent box, in which I can see $1000. I am told, and believe, that a predictor of human behaviour has already put either nothing or $1,000,000 into the opaque box. He has done so on the basis of whether he predicts that I will take both boxes, or only one: he has put $1,000,000 into the opaque box if and only if he has predicted I will take just it, and not the other transparent box as well. Moreover, I know that in 99% (or some other very high percentage) of the many other cases of choices people have made in the same situation, the predictor has predicted correctly.

Newcomb's Problem has been of great interest to decision theorists and more generally to those interested in rationality because it seems to pit two principles against one another and forces us to choose between them. On the *causal* view, you may as well take both boxes, whatever the predictor has done. If he's put a million in, you may as well get a thousand as well, and if he hasn't, you may as well get at least a thousand. Either way, you are better off taking both. By refraining from taking the transparent box, you do nothing to bring it about that you get the million, since the predictor has already acted. But the *evidential* rebuttal is, in effect: 'Well, if you're so clever, why aren't you rich?' People who take just the one box in this game tend to be rich, while people who take both tend not to be. It would be good news to learn that you are the kind of person who takes just one box, because that would mean the predictor has probably put the million in, and you are probably about to get rich. The question is whether that 'new value' gives you any reason to take just one box. The jargon is: should we try to maximize causal expected utility or evidential expected utility?

While each side has some intuitive appeal, theorists tend to divide sharply on the issue of normative rationality. But Nozick attends closely to what our intuitions are before instructing us about what they should be. His new move is intriguingly ecumenical. He points out that the intuitive appeal of the 2-box, causal solution and the 1-box, evidential solution to Newcomb's Problem varies as the amount of cash on view in the transparent box changes. Roughly, the causal view loses its intuitive appeal for very small amounts, and the evidential view loses its intuitive appeal for very large amounts. This variation of intuition with size of payoff calls for some explanation, since the change in payoff does not affect the structure that generates the

reasoning for the causal and evidential solutions. Nozick explains it in terms of an overall decision value, within which each of the causal and evidential principles get some weight.

Nozick makes a similar move for the well-known problem in game theory, the Prisoners' Dilemma (PD), which can also be interpreted to illustrate the conflict between causal and evidential reasoning. Causal reasoning tells us that, whatever the other person does, I am better off defecting than cooperating. Evidential reasoning tells us that, since we are in exactly the same position and equally rational, it is probable that if I cooperate, so will the other person, and that if I do not, neither will the other. Since we're both better off if we both cooperate than if neither do, my cooperating would again be good news, even though it has no tendency to bring about the other's cooperation. Now Nozick points out that the intuitive appeal of cooperating as opposed to defecting in the Prisoners' Dilemma also varies with the amounts of the payoffs, even though the essential structure that generates evidential reasons to cooperate and causal reasons to defect does not vary with payoffs. Again, roughly, the causal argument for defecting loses intuitive appeal if the benefits of defecting are very small relative to benefits of cooperation, while the evidential argument for cooperating loses intuitive appeal if the benefits of defecting are very large relative to benefits of cooperation. And again, Nozick explains this variation in terms of weighted component principles within overall decision value.

These parallel moves illustrate a general strategy, of explaining the variation of intuition with payoffs in terms of weighted component principles, with which I have no quarrel. I do have doubts, however, about the particular component values Nozick proposes in these two cases: more specifically, I doubt the evidential principle will do the explanatory work he gives it. This is the point at which it is instructive to go back to his original 1969 article. In that article he also attended closely to the way intuitions vary while the structure that generates evidential and causal reasoning does not.

In that article he considered some cases other than Newcomb's Problem, in which evidential reasoning also applies but in which, unlike Newcomb's Problem, it has no intuitive appeal. That is, he considered another variation of intuition between cases, though this time not a variation with payoffs.

Let's take as our example of a hypothetical case in which evidential reasoning has no intuitive appeal the *smoking gene case*. Suppose, counterfactually, that smoking is caused by a gene that also causes a deadly disease. Then the fact that my smoking would be bad news about my tendency to get the disease gives me no reason at all not to smoke. My refraining from smoking, whether on one occasion or over the long run, does not influence whether I get the disease or not; nor, if there were many people in my position, would refraining from smoking by all of us influence our chances of getting the disease or not. Rather, both the smoking and the disease are products of a common cause, the gene.

Nozick asked, in effect: given that the evidential solution has no appeal in such cases, what could give it any appeal in Newcomb's case? What is the difference? This of course is one classic strategy of argument used against the evidential view by causalists. But Nozick gave it an essential twist. He went on to point out that it is not enough merely to challenge the evidentialist to produce a difference between the cases that justifies evidentialism in Newcomb's case but not the others. Such a challenger should also grant that at least it is not *so immediately and intuitively*

obvious that the evidential view is wrong in Newcomb's case as in the other cases, and should himself be able to produce a difference that explains why. What is the difference that makes for conflicting intuitions in some cases and not in others, given that evidential and causal reasoning both apply in all cases? It may seem that in Newcomb's case an *illusion of influence* arises in support of the one-box solution, while no parallel illusion arises in the other cases, even though evidential reasoning would equally apply. But we still want to know why: what gives rise to this illusion?

Nozick doesn't offer an answer to this last question in that earlier paper. But even without an answer to that question, his earlier discussion has a bearing on his present view. Moreover, bringing his old and new arguments together may help to find an answer to that question. In both the old and the new arguments he notices a puzzling variation of intuition, and suggests an explanation. But the explanation of one variation will not work for the other variation. We need to consider what might explain both variations.

Nozick now wants to explain the variation of intuition with payoffs he observes in terms of the weights given to each of the evidential and causal principles within decision value. No mere illusion of influence will explain the variation of intuition with payoffs that he now points to, since there is no reason for the illusion of influence to decrease with the amount in the transparent box. Therefore, some weight within decision value does seem to be given to the evidential principle.

On the other hand, such a weighting seems inadequate to explain the variation of intuition he noted in his earlier article. In Newcomb's Problem, Nozick suggests the evidential principle gets some weight in order to explain the intuitive appeal of the 1-box solution, at least given small enough amounts in the transparent box. But in the smoking gene case the evidential principle seems to get no weight: varying the analogue of the amount in the transparent box, namely, the value forgone by not smoking, has no tendency to shift our intuitions in favour of not smoking in this case. That is, even if we reduce the value of smoking to a very small amount, there is no parallel tendency here to shift over to evidentialist principles. If evidentialism gets some weight in Newcomb's problem, why doesn't it get some comparable weight in the smoking problem, as payoffs vary? Here we do still need, as he pointed out in the early article, some account of this difference, whether in terms of an illusion of influence that arises in some cases but not others, or something else. Evidentialism doesn't get to the heart of the matter.

So, the position we're left in by bringing his old and new arguments together is this. An illusion of influence won't explain the variations of intuition with payoff Nozick is now concerned with. But Nozick's explanation of variation of intuition with payoff in terms of evidential principles *per se* getting some weight within decision value won't explain the variation between cases he pointed out in the earlier article. What we'd like is a version of weighted decision value that will explain both variations.

Let me now go back to the question: what explains the illusion of influence in Newcomb's problem and the lack of it in other cases, such as the smoking gene case? That way of putting the question, of course, is tendentious: maybe the difference isn't one with respect to an illusion of influence at all, but something else. However, I think it is, in a certain sense that I'll now explain.

My explanation requires us to look again at the relationship between Newcomb's Problem and Prisoners' Dilemma. It is well known that evidential reasoning can be applied to the Prisoners' Dilemma, in the way sketched above, and that in this sense the Prisoners' Dilemma is a Newcomb Problem (see Lewis [2]). What seems not to be so well recognized is that Newcomb's Problem can be interpreted as a Prisoners' Dilemma (see Hurley [1]). That is, Newcomb's Problem, unlike other problems that also support evidentialist reasoning, may seem to support cooperative reasoning. Cooperative reasoning is not the same thing as evidentialist reasoning. Cooperative reasoning applies in Prisoners' Dilemma and to a certain natural (though not necessarily correct) interpretation of Newcomb's Problem. But cooperative reasoning does not apply in the smoking gene case, nor is there any illusion or natural interpretation to the contrary, even though evidentialist reasoning does apply. I therefore suggest that cooperative reasoning, not evidentialist reasoning, is what needs to be given some weight within decision value to explain the variation of intuition between Newcomb's Problem and the smoking gene case as well as the variation of intuition with payoffs in Newcomb's Problem and in Prisoners' Dilemma.

When does cooperative reasoning apply? I suggest: when there is a collective causal power to bring about a (mutually) preferred result on the part of a group of acts, despite the absence of power on the part of any one of those acts either to bring about the preferred result or to bring about the other acts. This at any rate is a necessary, if not a sufficient condition, for the application of cooperative reasoning. This is clearly the case in Prisoners' Dilemma: the prisoners *together* have the causal power to bring it about that they each get their mutual second best outcome rather than their mutual third best outcome, even if neither alone has the power to bring this about or to influence what the other does. After all, the outcome depends on nothing else but what they each do. The standard Prisoners' Dilemma Matrix makes this clear. Let one prisoner be called 'P' and the other 'C'. If both cooperate, they get their mutual second-best outcome; if both defect the get their mutual third-best outcome:

Now notice that there is a natural illusion of a parallel collective causal power--an illusion of *collective* influence-- in Newcomb's Problem, despite the absence of any power on the part of an individual act of taking one box. That is, it is natural, if unwarranted, to interpret Newcomb's problem by supplying the predictor with a preference ordering such that the predictor and the predictee together are indeed in a Prisoners' Dilemma. Think of the predictee as an intelligent Child and the predictor as a Parent who wants the Child not to be greedy on a particular occasion. Let us help ourselves explicitly to the preference orderings of each. (Of course, we're not given any such preference ordering in Newcomb's problem, which is why it is unwarranted if natural to apply cooperative reasoning.) The Child simply prefers getting more money to less. The Parent doesn't mind about whether his prediction is right or not; what he most prefers is that the Child not be greedy on this occasion, that is, that he take one box rather than two; this concern has priority over concern with saving money. But as between two situations in which the Child takes the same number of boxes, the Parent prefers the one that costs him less money. Apart from the amounts of money involved, this basic pattern of concerns is very familiar: a Parent who doesn't mind spending money in relation to a Child, though he doesn't want to throw it away. Thus, **both** Parent and Child prefer the child's taking one box and getting a million to the Child's taking two and getting a thousand; this shared preference is the basis for cooperation. But the **Child** would **most** like to take two boxes and get a million plus a

thousand and would **least** like to take one and get nothing. The **Parent**, on the other hand, would **most** like for the Child to take only one box and get nothing (the Child has not been greedy and this result has cost the parent nothing) and would **least** like the child to take both boxes and get a million plus a thousand. Their assumed preference rankings are then as follows:

Parent	Child
Child takes one, gets $0	Child takes two, gets $M+T
Child takes one, gets $M	Child takes one, gets $M
Child takes two, gets $T	Child takes two, gets $T
Child takes two, gets $M+T	Child takes one, gets $0

On these assumptions, the pair are in a standard Prisoners' Dilemma:

As before, P and C do have a collective causal power to bring it about that their shared preference with respect to the two middle outcomes is fulfilled, even though neither has the power to bring that result about through his separate acts, nor to bring about the act of the other. It's irrelevant to the PD structure that one party, the predictor, acts before the other; this could be the case in the original PD and it would make no difference to the availability of cooperation. I suggest that there is a temptation to project something like these familiar parental motivations onto the predictor in the original Newcomb's Problem in order to make sense of the game he is playing. The intuitive appeal of the one-box solution could then be explained in terms of the urge to co-operation in a Prisoners' Dilemma of this Parent-Child type.

Here, then, is an explanation of how the illusion of influence arises in Newcomb's Problem, but it is an illusion of *collective* influence: the influence in question is a collective causal power, as in Prisoners' Dilemma, not an influence on the part an individual act by itself. Moreover, it is an *illusion* of a collective influence because we are not actually given the essential preference orderings that put the pair into a PD, even though it is natural to think along such lines. This would explain why the illusion of influence arises in Newcomb's problem, but not in the smoking gene case: no collection of acts, whether by one person or by different persons *could* have the relevant causal power there, no matter what preference orderings are assumed. Moreover, this explanation would also fit into Nozick's suggested weighted decision value structure. However, it would substitute for the component evidential principle a component cooperative principle, as a rival of causal reasoning. The cooperative principle would recommend doing one's part in exercising a collective causal power under certain conditions (which I haven't tried to specify fully here). The cooperative principle might apply to Newcomb's Problem, given further information, such as the above preference orderings; but it does not apply in the smoking gene case. As far as I can see everything else Nozick says about

the way in which the weighting of the rival principles explains the variation of intuition as payoffs vary would still apply. We have complete confidence in neither the causal principle nor the cooperative principle, which explains why we shift between them as the payoffs vary. But it is the cooperative principle that underlies the intuitive appeal of evidentialism in some cases, not the other way round. And the inapplicability of the cooperative principle accounts for evidentialism's lack of any intuitive appeal in other cases. These points about the explanation of intuitions hold whether or not one ultimately endorses cooperative reasoning as normatively rational, but they also help to see what issues of normative rationality are at stake in these cases. This revised version of weighted decision value explains both intuitive variations: the variations of intuition with payoffs Nozick is now concerned with, as well as the intuitive difference between cases he pointed out in the earlier article.

PAIS
University of Warwick, Coventry, CV4 7AL

REFERENCES

[1] S. L. Hurley, 'Newcomb's Problem, Prisoners' Dilemma, and Collective Action', *Synthese* 86 (1991) 173-96.

[2] David Lewis, 'Prisoners' Dilemma is a Newcomb Problem', in *Paradoxes of Rationality and Cooperation*, edited by Richmond Campbell and Lanning Sowden, (Vancouver: University of British Columbia Press, 1985).

[3] Robert Nozick, 'Newcomb's Problem and Two Principles of Choice', in *Essays in Honor of C. G. Hempel*, edited by N. Rescher et al (Dordrecht: Reidl, 1969); reprinted in *Paradoxes of Rationality and Cooperation*, edited by Richmond Campbell and Lanning Sowden, (Vancouver: University of British Columbia Press, 1985).