ISSN 0039–7857

SYNTHESE / *Volume 76 No. 1 July 1988*

PHILIP PETTIT

THE PRISONER'S DILEMMA IS AN UNEXPLOITABLE NEWCOMB PROBLEM

1. INTRODUCTION

David Lewis has shown that if a certain similarity between participants can be assumed, as it would seem it can, then the two-party prisoner's dilemma is a Newcomb problem.[1] This result suggests that if it is rational to maximise conditional expected utility then it will sometimes be rational to cooperate in a prisoner's dilemma.[2] But one of the firmest intuitions around is that cooperating in a one-shot prisoner's dilemma is irrational. And so the suggestion must count against the evidential decision theory which supports such maximisation. By the same token it will count in favour of the causal decision theory which Lewis himself supports.[3]

I wish to argue, however, that except under one marginal sort of circumstance, the Lewis result will not support a policy of cooperation for decision-makers who seek to maximise conditional expected utility. From an outside point of view there may be grounds for thinking of a prisoner's dilemma as a Newcomb problem, but those grounds fail to provide a basis on which participants might think of cooperating. The prisoner's dilemma, at best, is an unexploitable Newcomb problem.

In the next section I provide essential background to this argument, giving an account of the Newcomb problem, the prisoner's dilemma and the Lewis result. The prisoner's dilemma is an exploitable Newcomb problem only if there are particular sorts of similarity from which participants can argue that cooperation may maximise conditional expected utility. In the third section I consider how they might try to reason to this conclusion from their common rationality; in the fourth how they might try to do so, more generally, from any 'option-based' similarity; and in the fifth how they might try to run an argument from an 'agent-based' similarity.

I propound three theses, defended respectively in these last three sections. Thesis 1 is that the argument from rationality fails. Thesis 2 is that any argument from such an option-based similarity also fails. And Thesis 3 is that while an argument from an agent-based similarity

Synthese **76** (1988) 123–134.

might succeed, the case is so marginal as to be of vanishing significance.

2. THE NEWCOMB PROBLEM AND THE PRISONER'S DILEMMA

In the classic Newcomb problem I am offered a choice between two options, A and B.[4] The value of each option depends on a causally independent contingency: say, whether *r* or not *r*, since '*r*' usefully suggests 'reward'. A gives me $1,000,000 if *r*, zero dollars if not *r*. B gives me $1,001,000 if *r*, $1,000 if not *r*. The matrix looks like this.

	r	Not *r*
A	$1,000,000	$0
B	$1,001,000	$1,000

Given the causal independence of *r* or not *r*, it seems that I must choose B: this is the dominant option, in the sense that it is better for me whether *r* or not *r*. But here is where the characteristic Newcomb complication enters. It turns out – for our purposes, it does not matter why – that the probability that *r* given A is higher than the probability that not *r* given A, whereas the probability of *r* given B is less than the probability of not *r* given B. This means that the probability of *r* given A is higher than the probability of *r* given B and so, equivalently, that the probability of not *r* given A is less than the probability of not *r* given B.[5] Thus it may be that the way to maximise conditional expected utility is to choose the dominated option A. The problem, then – the Newcomb problem proper – is how to choose: whether by reference to dominance or conditional expected utility.[6]

Recall now the prisoner's dilemma. Considered from my point of view as a participant, the dilemma also offers me a choice between two options, one of which dominates the other. The payoff indeed may be as in the matrix given above. All that is necessary to ensure that the matrix represents my half of a prisoner's dilemma is that '*r*' and 'not *r*' stand respectively for my partner's choosing between counterpart options A′ and B′, where the rewards on offer to him are symmetrical. This condition fulfilled, we have the two crucial features of a prisoner's dilemma: B and B′ are dominant strategies and yet the result B, B′ is Pareto-inferior to A, A.[7]

Is the prisoner's dilemma which corresponds to such a matrix a Newcomb problem? It will be if, despite the causal independence of my partner's behaviour, the probability that he cooperates (i.e., that *r*) when I cooperate is higher than the probability that he defects (that not *r*), in such a case, and if the probability that he defects (that not *r*) when I defect is similarly higher than the probability that he cooperates (that *r*) in such an event.

David Lewis argues that if the participants in the dilemma can count as replicas, however imperfect, of one another, then these conditions will be fulfilled. After all, the fact that I cooperate makes it more probable that a replica will cooperate rather than not in a symmetrical situation, and similarly for defection. Lewis takes it as obvious that the participants do satisfy such a similarity relationship. "The most readily available sort of replica of me is simply another person, placed in a replica of my predicament. For instance: you, my fellow prisoner".[8]

It would seem, therefore, that if I seek to maximise conditional expected utility, then the thing for me to do in a prisoner's dilemma may be to cooperate; whether it is or not will depend on the relative values of the different possible outcomes. But this appearance is misleading. It dissipates once we consider the precise arguments for cooperation that I might try to run.

3. THESIS 1

The most obvious ground on which I might try to argue that my partner is a replica and that cooperation may maximise conditional expected utility is that we are each rational, in the sense which connects with the choice of rational options: say, that we are each rational so far as we maximise conditional expected utility.

The argument envisaged can be nicely summarised with the help of some notation. Let 'Ci' and 'Di' mean that I cooperate and I defect; let 'Cy' and 'Dy' predicate the counterpart choices of you; let 'Ri' and 'Ry' mean respectively that I and you are rational; and let 'Sw' mean that we are placed in symmetrical choice situations. The argument can now be formulated with the help of the expression for the probability of a state of affairs *m* conditional on another state of affairs *n*: 'p(*m*/*n*)'.

The argument is this.[9]

1. If Ri & Ry & Sw, then p(Cy/Ci) > p(Dy/Ci) and p(Dy/Di) > p(Cy/Di).

2. Ri & Ry & Sw.
3. Therefore, $p(Cy/Ci) > p(Dy/Ci)$ and $p(Dy/Di) > p(Cy/Di)$.
4. And therefore, depending on the desirabilities involved, it may be that cooperation maximises conditional expected utility: viz., that

$$p(Cy/Ci)d(Ci \& Cy) + p(Dy/Ci)d(Ci \& Dy) > p(Cy/Di)d(Di \& Cy) + p(Dy/Di)d(Di \& Dy).$$

Suppose that I believe that it is always rational to maximise conditional expected utility? Ought I to find this argument convincing in a prisoner's dilemma, allowing it to affect what I do? The argument is valid, so the question is whether I am entitled to endorse the premises. I want to show that I am not: specifically, that in such a decision-making predicament I may not assume that Ri.

Rationality is required to have a number of features if the argument given is to be one that I may use in making my decision.

a. It must be a property of agents.
b. It must be a property which requires particular choices of agents, at least in prisoner's dilemmas.
c. It must be a property such that an agent is entitled simultaneously to assume that he is rational and to let that assumption figure in the derivation of what he should do.

I deny that rationality can meet this third condition and so I deny that I can legitimately endorse the second premise if at the same time I am using the argument to determine what I should do.[10] I will support my case in two stages. First I will describe a scenario under which rationality certainly does not meet condition c. And then I will argue that this scenario actually obtains.

Stage 1

If I may assume that I am rational, then I must be in a position to believe that the option I choose is always what rationality requires; it is *the*, or *a*, rational option. The question raised by condition c is whether I remain in a position to believe this if I sometimes make a choice on the basis of the assumption that I am rational. Does the fact of choosing on the basis of that assumption deprive me of good reason for believing that I am then choosing rationally?

It will do so under at least the following scenario. Suppose that at bottom there is one sort of good reason, and one sort only, for me – or anyone else – to believe that I am rational and that this consists in the fact that I am sensitive to the features of options which make them rational alternatives to choose. The sensitivity might come about through facets of my biological make-up that are not under my control; equally it might result from my scrupulously following certain procedures of deliberation. If this scenario obtains, then it cannot be that I ever choose what to do in the light of the assumption that I am rational.

Here's why. If one of my choices is made in the light of that assumption, there is still a good reason for me to believe that I am being sensitive to the rational features of options only if there is a good reason for me to think that the assumption is true; sensitivity to such features can hardly come of arguing from a premise which there is no good reason to believe. But by hypothesis the only good reason for me to think that the assumption is true is that I am sensitive to the rational features of options. And so an evidential circle looms.

The circle, otherwise drawn, is this. I may believe – I am entitled to believe – that I am rational, in particular rational in this decision, only if I may believe that I am sensitive, in particular sensitive in this decision, to the rational features of options. But I may believe that I am sensitive in this decision to the rational features of options only if I may believe in the premises of the argument, and so only if I may believe that I am rational. Thus I may believe that I am rational in this decision only if I may believe that I am sensitive in the decision to the rational features of options, and I may believe that I am sensitive in this way only if I may believe that I am rational.

Stage 2

The upshot is that in the scenario described I would not be in a position to avail myself of the argument under examination. This result is not just of speculative interest, for it can be shown that the scenario actually obtains.

However we differ in our theories of rationality, we must all agree that rationality among options is the first thing to define and that it is this which enables us to define rationality among agents. We know who the rational agents are by knowing who are disposed to choose

rational options and we know what the rational options are independently. We may claim to know that an option is rational, depending on our theory, through knowing that it maximises conditional expected utility or some other Bayesian function, or that it maximaxes, maximins or whatever. But however we go, we can tell that an agent is rational only through knowing that he is disposed to select options that relate in such a manner to his beliefs and desires. Rationality, as we may put it, is an option-based property. It belongs to options in the first place, agents in the second.

The fact that rationality is option-based in this sense means that the only sort of evidence there is for me to believe that I am rational must ultimately come down to evidence that I am sensitive to those relational features of options in virtue of which they, the options, are rational; rationality now consists in such sensitivity. Assuming that there is indeed good reason for me to believe that I am rational, this means that the scenario described actually obtains. There is at bottom one sort of good reason for me to think that I am rational, and one sort only: viz., the reason provided by the fact that I am sensitive to the rational features of options.

Conclusion

What is true of the scenario is true then of the actual world. I am not entitled to believe that I am rational if I ever let that belief figure in the derivation of what I should do. Rationality fails to satisfy the last of the three conditions mentioned earlier.

I conclude that in a prisoner's dilemma I cannot avail myself of the argument from rationality. So far as I am prepared to use the argument in determining what to do, I deprive myself of any good reason to believe in one of its premises: the proposition that I am rational. The argument promises a short-cut to decision, in particular a short-cut to a cooperative decision, but the path promised is a dead-end.[11]

This conclusion means that the argument from rationality does not establish that the prisoner's dilemma is an exploitable Newcomb problem. But there remains a sense in which it is nevertheless a problem of that kind: a sense in which the Lewis result still stands. The reason is that if someone else, or even the agent himself, views the predicament from the outside, casting the rationality argument in the

third person, then he may be able to endorse all of the premises. Just rehearsing the argument will not cause problems for the acceptance of any of its premises in the way that using it for decision-making purposes would do so. The observer can reckon that the agent does indeed stand to maximise conditional expected utility if he cooperates. But he can reckon this only because he enjoys a privileged lack of access to the agent's decision-making; it is crucial that the thought to which he subscribes is not liable to affect what the agent does.

4. THESIS 2

The argument from mutually recognised rationality, abortive though it is, may encourage the investigation of parallels. The argument would turn in the parallels, not on the known rationality of the participants and their choices, but rather on some other salient characteristic: their envy, greed, or desire for power; even perhaps, as Lewis suggests, their irrationality in certain respects. What is required of such a characteristic is that it satisfies three conditions parallel to those imposed on rationality: it is a property of agents; it requires particular choices of them, at least in prisoner's dilemmas; and, most important of all, an agent can assume that he exemplifies the characteristic and still let that assumption figure in the derivation of what he should do.

Let the mutually recognised characteristic be predicated of me and my partner by 'Fi' and 'Fy' respectively. The parallel argument will follow the same lines then as our original.

1′. If Fi & Fy & Sw, then $p(Cy/Ci) > p(Dy/Ci)$ and $p(Dy/Di) > p(Cy/Di)$.
2′. Fi & Fy & Sw.
3. Therefore $p(Cy/Ci) > p(Dy/Ci)$ and $p(Dy/Di) > p(Cy/Di)$.
4. And therefore, it may be that cooperation maximises conditional expected utility: viz., that

$$p(Cy/Ci)d(Ci \,\&\, Cy) + p(Dy/Ci)d(Ci \,\&\, Dy) > p(Cy/Di)d(Di \,\&\, Cy) + p(Dy/Di)d(Di \,\&\, Dy).$$

We can deal briefly with this argument. I assume that envy, greed, power-hunger and other such salient candidates for the role of the F-property are all option-based. F-agents are defined as agents who are reliably disposed to select F-options, as we may call them, not vice

versa. And an option is an F-option in virtue of features fixed independently of who are liable to choose it.[12] Like the features which make an option rational, F-making features may be relational rather than intrinsic; indeed they may even relate to the attitudes of the agent in question. The important point is that you cannot tell whether someone is an F-agent until you know whether the options he is disposed to choose are F-options, whereas you can tell what an F-option is without knowing who is an F-agent.

The fact that an F-property is option-based means that whether or not it satisfies the other two, it cannot meet the third of the constraints that parallel the conditions on rationality. No one is entitled to assume that he satisfies the property if he lets that assumption figure in the derivation of what he does.

The reason is that if I am F so far and only so far as I am sensitive to the F-properties of options, then at bottom the only good reason there is for me to believe that I am F is that I display such sensitivity. But if I decide what to do on the basis of the assumption that I am F, then there is good reason to believe that I am sensitive in this way only if there is good reason to accept that assumption: I can hardly be sensitive through arguing from a premise which there is no good reason to believe. And so another evidential circle looms.

The circle, more carefully described, is this. I may believe that I am F, in particular that I am F in the decision before me, only if I may believe that I am sensitive to the F-features of options, in particular sensitive to the F-features of the options on hand. But I may believe that I am sensitive to those features only if I may believe in the premises of the argument from which I take my guidance, including the premise that I am F; I can hardly expect such sensitivity to be generated by relying on a premise which I am not entitled to believe. Thus I may believe that I am F in this decision only if I may believe that I am F-sensitive in this decision and I may believe that I am F-sensitive in the decision only if I may believe that I am F.

We are in a position to conclude that the possession of no option-based similarity can provide participants in a prisoner's dilemma with reasons for thinking that cooperation may maximise conditional expected utility. It may be possible for outside observers to see the dilemma as a Newcomb problem, arguing that the similarity establishes a probabilistic correlation between the participants' choices. But this perception of the dilemma cannot be internalised by

the participants themselves. It is bound essentially to an outside perspective.

5. THESIS 3

The discussion of Thesis 2 suggests that an argument deploying a property that is not of an option-based kind may play the role for which rationality and F-characteristics in general are not fitted. Suppose that we could identify a property that belongs, like rationality, both to agents and options but which invites quite a different sort of definition. Suppose in particular that we could only define options exemplifying the property as options which agents who possess the property are reliably disposed to select.

Let 'Gi' and 'Gy' ascribe such an agent-based property to the participants in a prisoner's dilemma. This property will be expected to satisfy our three familiar constraints: it must be a property of agents; it must require particular choices of agents, at least in prisoner's dilemmas; and an agent must be entitled to assume that he exemplifies the property, even if he lets that assumption dictate one of his choices.

The argument involving the G-property will run on familiar lines.

1″. If Gi & GY & Sw, then $p(Cy/Ci) > p(Dy/Ci)$ and $p(Dy/Di) > p(Cy/Di)$.
2″. Gi & Gy & Sw.
3. Therefore $p(Cy/Ci) > p(Dy/Ci)$ and $p(Dy/Di) > p(Cy/Di)$.
4. And therefore, it may be that cooperation maximises conditional expected utility: viz., that

$$p(Cy/Ci)d(Ci \& Cy) + p(Dy/Ci)d(Ci \& Dy) > p(Cy/Di)d(Di \& Cy) + p(Dy/Di)d(Di \& Dy).$$

Given that the G-property is agent-based, this argument does not suffer the weakness of the other two. The agent-basis means that no matter how I make up my mind, I can always rely, if I know myself to be a G-agent, on deciding so that Gi. In particular, I can rely on this even when I decide what to do, as here, on the assumption that Gi. The assumption is robust under its own practical application. G-ness satisfies the condition which the option-based properties fail.

I am prepared to endorse the argument involving the G property. The difficulty in pursuing the line suggested, however, is that it is hard

to come up with examples of appropriate G-features. Still, it is not impossible.

Suppose that neurophysiology were sufficiently developed for the participants in a dilemma to be able to tell that they are neurophysiologically indiscernible. And suppose that such indiscernibility means that they are bound to make the same choices in symmetrical situations. In that event, the participants could invoke the agent-based property of having such and such a neurophysiological profile in the role ascribed here to the G-feature. They could argue from their common neurophysiological character and from the inevitability of their each choosing an option in that character.

In the sort of situation postulated the evidence I have on my partner's disposition to choose is that he is a perfect neurophysiological replica, destined to mimic my every choice. There may be difficulty in this thought, particularly as I consider myself to be in interaction with my replica, but the difficulty is of a familiar kind: it is the tension involved in seeing myself simultaneously as a chooser and as a physically determined system.

If I have to see my partner as a replica of this kind, then I will certainly see our dilemma as a Newcomb problem. It will be clear that if I cooperate, and this is in my neurophysiological nature, then the other will cooperate too. Since I cannot help but act in my neurophysiological nature, it will be clear, more simply, that if I cooperate, the other will cooperate also. And that is enough for me to see that the dilemma is a Newcomb problem in which cooperation may maximise conditional expected utility.

We must concede that this kind of situation, and appropriate probabilistic variants, constitute dilemmas in which evidential decision theory may counsel cooperation. But the concession takes little from the brunt of our earlier conclusions. The realm of neurophysiological replicas, perfect or probabilistic, is an imaginary world. Possibilities which are only vindicated there are of vanishing significance.[13]

NOTES

[1] See Lewis: 1979, 'Prisoner's Dilemma is a Newcomb Problem', *Philosophy and Public Affairs*, vol. **8**. One reason for wanting to oppose Lewis is related to the line I run in Pettit: 1986, 'Free Riding and Foul Dealing', *Journal of Philosophy*, vol. **83**; reprinted in 1987, *The Philosopher's Annual*, vol. **9**: see Note 12 of that article. I find some consolation, but not enough, in J. H. Sobel: 1985, 'Not Every Prisoner's Dilemma is a Newcomb Problem' in Richmond Campbell and Lanning Sowden (eds.), *Paradoxes of Rationality and Cooperation*, University of British Columbia Press, Vancouver.

[2] On the evidential decision theory which supports such maximisation see Richard Jeffrey: 1983, *The Logic of Decision*, 2nd edition, University of Chicago Press. For a survey of this and other Bayesian theories see Ellery Eells: 1982, *Rational Decision and Causality*, Cambridge University Press, Chapters 1–3.

[3] See Lewis: 1981, 'Causal Decision Theory', *Australasian Journal of Philosophy*, vol. **59**.

[4] The standard presentation is Robert Nozick: 1969, 'Newcomb's Problem and Two Principles of Choice', in Nicholas Rescher (ed.), *Essays in Honour of Carl G. Hempel*, Reidel, Dordrecht. The usual story-line is that a predictively reliable genie has determined whether *r* or not *r*; specifically, that he has determined that *r* if and only if he has predicted that I will choose A.

[5] This feature, as Hugh Mellor reminded me, is the crucial one, not the feature from which I derive it in the text. But nothing is affected by my deriving it as I do and it happens that the derivation motivates a convenient way of presenting the arguments discussed in the sections following.

[6] Lewis employs an unusually wide conception of a Newcomb problem and I follow his practice. He does not require that the dominated option should actually maximise conditional expected utility, only that it may do so: this, because the choices of the participants, however causally independent, are not probabilistically so. See 'Prisoner's Dilemma is a Newcomb Problem', page 239, Note 6.

[7] These two features define the prisoner's dilemma under a strict conception of the predicament. See Pettit: 1985, 'The Prisoner's Dilemma and Social Theory', *Politics*, vol. **20**.

[8] Lewis, 'Prisoner's Dilemma is a Newcomb Problem', page 239. I skip the precise details of the case which Lewis makes because I do not think that those details affect the line which I argue.

[9] This line of argument is related to one that I examine and criticize in Pettit: 1986, 'Preserving the Prisoner's Dilemma', *Synthese*, vol. **86**. See too Lawrence Davis: 1977, 'Prisoners, Paradox and Rationality', *American Philosophical Quarterly*, vol. **14**; and Lanning Sowden: 1983, 'That There is a Dilemma in the Prisoner's Dilemma', *Synthese*, vol. **55**.

[10] In order to avoid any possible misunderstanding, I should stress that I deny this with regard only to a narrow decision-theoretic sense of rationality. I freely admit that rationality may meet this third condition, for example, in the sense in which rationality is what enables people in a coordination problem to recognise the salient solution; in that sense of course rationality will not meet condition *b*.

[11] Notice, however, that under the assumption that Ry and Sw there is a different short-cut available. The short-cut offers no consolation for opponents, however, since it leads to the conclusion that only defection can be rational, not cooperation. The argument is this. Were I to believe that RC, as we may say, then I would conclude that Cy, in which case I would see that defection was the rational option for me: i.e., that RD. Thus it must be the case that RD. There is no coherent alternative.

[12] The intention of the notation is obvious in view of the rationality parallel but some

care is needed. The R-agent chooses an option because it has those features in virtue of which we call it rational. The F-agent likewise chooses an option because it has certain features in virtue of which we describe it as the envious or greedy or power-hungry choice. But while we may shorten the remark about the R-agent, saying that he chooses an option because it is rational, we will be misleading if we make the corresponding abbreviation in the other case. The F-agent does not choose something because it is envious or greedy or power-hungry: those are our descriptions, not his. Rather he chooses it because it has properties that he would enunciate in other terms. It is important to be clear about this matter but once we are clear about it, there is no harm in saying that the F-agent makes his choice on the basis of recognising that FC or FD. [13] This paper benefitted from criticisms at seminars in the Research School of Social Sciences, Australian National University and the Centre for Public Choice, University of East Anglia. I owe a great debt for written comments received from David Braddon-Mitchell, Peter Forrest, Frank Jackson, Peter Menzies, Huw Price, Howard Sobel and Kim Sterelny, and also from two referees for this journal. I revised the paper during the tenure of an Overseas Fellowship at Churchill College, Cambridge and I am grateful to the college for the facilities put at my disposal.

Research School of Social Sciences
Australian National University
Canberra, A.C.T. 2601
Australia

STRATEGIC BEHAVIOR AND COUNTERFACTUALS*

ABSTRACT. The difficulty of defining rational behavior in game situations is that the players' strategies will depend on their expectations about other players' strategies. These expectations are beliefs the players come to the game with. Game theorists assume these beliefs to be rational in the very special sense of being *objectively correct* but no explanation is offered of the mechanism generating this property of the belief system. In many interesting cases, however, such a rationality requirement is not enough to guarantee that an equilibrium will be attained. In particular, I analyze the case of multiple equilibria, since in this case there exists a whole set of rational beliefs, so that no player can ever be certain that the others believe he has certain beliefs. In this case it becomes necessary to explicitly model the process of belief formation. This model attributes to the players a theory of counterfactuals which they use in restricting the set of possible equilibria. If it were possible to attribute to the players the same theory of counterfactuals, then the players' beliefs would eventually converge.

1. MUTUAL RATIONAL BELIEFS

In interactive contexts, such as those treated in game theory, what it is rational to do depends on what one expects that other agents will do. Could we take these expectations as given, the problem of strategy choice would be simple: an agent would choose the strategy which maximized his payoff under the assumption that all other agents act in accordance with his expectation. If the agents' reciprocal expectations are not given, then the problem arises of what expectation is to be entertained by a rational player who expects the other players to act rationally, and to entertain similar expectations about him and the other players. The problem has been appropriately termed by Harsanyi one of 'mutual rational beliefs' (Harsanyi 1965).

There are games in which this problem is absent, and what it is rational to do is straightforward; for example, there may exist a 'dominant strategy', which yields a better outcome than any other strategy, whatever the other players do. But in general this is not the case.

When an explicit treatment of expectations is required, game theorists assume that players are equipped with subjective probability distributions over the other players' choices and are *practically rational*

Synthese **76** (1988) 135–169.