Newcomb's Problem and Repeated Prisoners' Dilemmas

Christoph Schmidt-Petri†‡

I present a game-theoretic way to understand the situation describing Newcomb's Problem (NP) which helps to explain the intuition of both one-boxers and two-boxers. David Lewis has shown that the NP may be modelled as a Prisoner's Dilemma game (PD) in which 'cooperating' corresponds to 'taking one box'. Adopting relevant results from game theory, this means that one should take just one box if the NP is repeated an indefinite number of times, but both boxes if it is a one-shot game. Causal decision theorists thus give the right answer for the one-shot situation, whereas the one-boxers' solution applies to the indefinitely iterated case. Because Nozick's set-up of the NP is ambiguous between a one-shot and a repeated game, both of these solutions may appear plausible—depending on whether one conceives of the situation as one-off or repeated. If the players' aim is to maximize their payoffs, the symmetric structure of the PD implies that the two players will behave alike both when the game is one-shot and when it is played repeatedly. Therefore neither the observed outcome of both players selecting the same strategy (in the PD) nor, correspondingly, the predictor's accurate prediction of this outcome (in the NP) is at all surprising. There is no need for a supernatural predictor to explain the NP phenomena.

†To contact the author, please write to: Faculty of Economics and Management, Witten/Herdecke University, Alfred-Herrhausen-Str. 50, 58448 Witten, Germany; e-mail: christoph.schmidt-petri@uni-wh.de.

‡I am grateful for helpful comments and criticisms by Jason Alexander, Richard Arneson, Thomas Baldwin, Jossi Berkovitz, Richard Bradley, Luc Bovens, Juliana Cardinale, Nancy Cartwright, John Collins, Peter Dietsch, Adam Elga, Branden Fitelson, Till Grüne-Yanoff, Jonathan Halvorson, Jim Joyce, Isaac Levi, Ned McClennen, Paul Schweinzer, Dana Tulodziecki, Bruno Verbeek, Ioannis Votsis, and Jo Wolff. Special thanks to Ulrich K. Müller. I gratefully acknowledge financial support from the AHRB, the Aristotelian Society, the Department of Philosophy, Logic and Scientific Method of the London School of Economics, as well as the Philosophy, Probability and Modeling Group at Konstanz, which was supported by the Alexander von Humboldt Foundation, the Federal Ministry of Education and Research, and the Program for the Investment in the Future (ZIP) of the German Government through a Sofja Kovalevskaja Award to Luc Bovens.

Philosophy of Science, 72 (December 2005) pp. 1160–1173. 0031-8248/2005/7205-0042$10.00

1. Introduction. Newcomb's Problem (hereafter NP), originally presented in Nozick 1969, runs as follows. In front of you there are two boxes, and you are offered the choice of either taking both or just box 2. You know that box 1 contains $1,000. You also know that some allegedly supernatural being (the 'predictor') may have put money in box 2 as well, following this rule: if he predicts you to take both boxes he will put nothing. If he predicts you to take just box 2 he will put $1m. A further fact is crucial: you know that the predictor has almost always been right in his predictions. Assuming that you only care about the money and prefer more of it to less, should you take both boxes, or just box 2?

Those who recommend taking just box 2 (so-called 'one-boxers') are typically represented as 'evidential' decision theorists along the lines of Jeffrey (1983). The one-boxers' reasoning is said to be that, *by assumption*, the probability of the predictor predicting correctly is very high—say he is right in 99% of cases, whatever you choose. Hence, taking one box would seem to give you extremely good evidence that the predictor also predicted you would take one box, that is, that there is $1m in it. Learning that you are almost certainly just about to get this million (by taking this box) would be good news; you think that it is quite desirable to have $1m. On the other hand, if you opted for both boxes, the predictor will likewise have predicted this and will have left box 2 empty. Yet, in the present context, bringing about the piece of news that you are extremely likely to get $1,000 is not quite as desirable. Taking one box brings about a more desirable piece of news than taking both boxes, so you should take box 2 only.

'Two-boxers' (mostly causal decision theorists such as Gibbard and Harper (1978), Lewis (1981), Joyce (1999), but see also Eells (1982)) argue that since the predictor *pre*dicts your action, his decision whether to put the money in box 2 has already been made *before* you choose; at any rate there is nothing you can do to influence this prediction. Even though there is a *probabilistic* dependence between your action and the prediction (the correlation between one/two boxes actually being chosen when the predictor predicted that one/two boxes would be chosen is very strong), this probabilistic dependence is of no particular relevance for the 'real' desirability of the actions because it does not reflect a *causal* dependence; you can't *make* the predictor predict what you would most prefer him to predict just by acting one way or the other. If he predicted you to take just box 2, there will be $1m in it, if he predicted you to take both there will be nothing. Take both boxes and you get $1,000 plus whatever is in box 2. Take just box 2 and with some luck you will turn millionaire—but in that case you might just as well have taken the other box too and have an additional $1,000. Since you get $1,000 more for sure by taking

both boxes *whatever* the predictor has predicted it is best to take both boxes.

2. The NP and the Prisoners' Dilemma. Steven Brams (1975) and subsequently David Lewis (1979) have shown that a 'two-person' NP has the same structure as a Prisoners' Dilemma game (PD).[1] Recall that in a PD, the only Nash equilibrium is for both players to defect.[2] They argue that the payoff matrix for two people playing a PD could look like this (first yours, then his payoff, in $):

	C: he takes one = silent	D: he takes both = rats
C: you take one = silent	1m/1m	0/1,001,000
D: you take both = rat	1,001,000/0	1,000/1,000

An illustrative story along the lines of Lewis (1979) to bring life to this matrix might be helpful to bring out the intuitive similarity between PD and NP.[3] So suppose that, as in the original PD, you are one of two prisoners separately being questioned about a crime you have committed together. The police has set up a clever reward system to help the course of justice. If you rat on your accomplice, you'll get a thousand dollars (the money is already in front of you, in an enticingly transparent box). But if you don't rat, *your partner* gets rewarded with a million dollars. The same offer has been made to him too.

So if you rat (i.e., if you 'defect'), then, if your partner does not rat, you get $1,001,000 (and he gets nothing), whereas if he also rats you just get $1,000 (and so does he). Conversely, if you don't rat (i.e., if you 'cooperate'), then if neither does your partner, you get $1m as a reward for your partner's silence (and of course so does he, as a reward for your silence). However, if he does rat while you don't, you get nothing (and he gets $1,001,000).

Finally, the 'Newcomb twist': the detective tells you that comprehensive

1. That there is some similarity between the NP and the PD has also been pointed out by Gibbard and Harper (1978) and Horgan (1981).

2. A Nash equilibrium is a profile of strategies such that neither player wishes to revise his strategy choice, given the strategy choice of the other player. Each player's strategy is said to be a 'best reply' to the other player's choice. For instance, suppose you cooperate. My best reply to this would be defection. But *your* best reply to my defection is not cooperation, since you are better off defecting when I defect. Hence 'cooperate, defect' is not a Nash equilibrium. For an excellent introduction to some of the game theoretic results used in this paper see Routledge 1998.

3. But whether you find this particular story plausible ought not to influence your assessment of the general argument.

empirical studies in a number of states have robustly shown that in situations such as this one the prisoners practically always decide for the same option. Furthermore, the other prisoner has already made his choice. Actually, now you notice the black box in front of you! This box might contain a million Dollars! As you know that you will get the content of the black box anyway, should you take the transparent box too?

It should be clear how this is both PD *and* NP for you. Characteristic for a PD is the existence of a strictly dominant strategy for both players, and the Pareto-inferiority of the unique Nash equilibrium.[4] Here the strictly dominant strategy for both you and your accomplice is to rat since that always gives an *additional* thousand Dollars. However, both of you would do better if you both cooperated by staying silent. Characteristic for a NP is the option of getting $1,000 for sure by taking both boxes (that is, ratting), and causal independence but strong probabilistic correlation between your choice and the content of the second box. All of this is satisfied in this example: whether you find the million in the second box will depend not on what you have causal influence about, but on what the *other* prisoner does (on what the predictor predicts). The one thing you know about the behaviour of your fellow prisoner is that in situations such as this, the two prisoners practically always behave alike (the predictor predicts accurately), and that he has already made his choice (prediction).

You might now be tempted by roughly the following reasoning (let's call it 'evidentialist'). Given that in the past the two prisoners have practically always behaved alike, your cooperating (taking one box) will be evidence for the other player cooperating too, a situation in which, if realized, you would net $1m both and therefore only be convicted of the lesser crime. If you ratted (took both boxes), ideally he would not follow suit (the predictor gets it wrong and you would do even better than cooperating since you get the million plus the $1,000 from the transparent box 1); but by the assumption of similarity in behaviour most probably he would rat on you too, and you would end up with just $1,000 (the black box 2 is empty as he accurately predicted your taking both boxes). So, as due to the above-mentioned similarity in behaviour between you and your fellow prisoner you are very likely to both make the same choice *you* should cooperate since if *both of you* cooperate you will get more money than if both of you defected.

But there is, as in the NP, the contrary reasoning that you are better

4. A strictly dominant strategy is a strategy such that, in equilibrium, you always do better using this strategy than using any other strategy. An outcome is Pareto-inferior with respect to some other outcome if at least one agent would be better off, and no one worse off on the latter.

off defecting whether the other player cooperates or not (you prefer \$1,000 over \$0 in case he defects, and \$1,001,000 over \$1m in case he cooperates), hence cooperating is strongly dominated and you should rat. Your choice has no 'causal' influence on what is rational for the other prisoner to do, even if there is a strong correlation. As the situation is symmetric actually you should both defect rather than cooperate.

3. Analysis. This analogy between PD and NP provides the background for my comment on the NP. From the above exposition it emerges that by adopting the evidential decision theorist's reasoning about the NP to the PD, he seems to prescribe to play 'cooperate' in the PD. But such a recommendation would clearly be mistaken.[5] We have uncontroversial tools in place to solve problems like the PD which dictate otherwise. Specifically, game theory tells us that to cooperate in the PD is irrational (cooperating is strictly dominated). More accurately, game theory tells us that to cooperate in the PD is irrational *provided that it is a one-shot game*. But in indefinitely iterated games (that is, in games of finite but unknown duration and in infinitely repeated games) playing a cooperative strategy may be the right thing to do.[6] So how about *repeated NPs*?

I suggest that (corresponding to PDs) the best thing to do is to choose both boxes in a one-shot NP, and to choose one box iff the NP is repeated an indefinite number of times. The argument is by analogy, given the structural similarity between NP and PD which has been demonstrated by Lewis. Instead of importing the paradoxes surrounding the NP into the analysis of the PD, my suggestion is to export the clear-cut game theoretical results about the PD into the analysis of the NP: in one-shot PDs, defecting is rational. Since NPs are PDs, and defecting corresponds to taking both boxes, taking both boxes is rational in one-shot NPs.

In indefinitely iterated PDs, both players' always cooperating is one (of many) possible equilibrium outcomes, but it is a particularly salient outcome since it is the Pareto-efficient and symmetric outcome—it maximizes payoffs for both players (see Harsanyi and Selten (1988) for extensions to the concept of Nash equilibrium). Assuming that players maximize

5. The above reasoning is also known as the 'fallacy of the twins', see Bicchieri and Green 1997 and Binmore 1994, 203ff.

6. See Fudenberg and Tirole 1991, 110ff. With a 'cooperative strategy' I mean a strategy that, in equilibrium, results in both players playing 'cooperate' in all stage games. For instance, a 'grim trigger' strategy is cooperative in this sense: cooperate until the other player defects, then defect ever after. In equilibrium, the other player will never defect for fear of being too severely punished provided he cares enough about the future.

payoffs, they will cooperate.[7] Again, since NPs are PDs,[8] and cooperating corresponds to taking one box, taking one box is the best thing to do in indefinitely iterated NPs.[9]

7. The 'folk theorem' tells us that there are many possible equilibria in the indefinitely repeated PD if you are a bit patient, including the pair of strategies 'always defect, always defect'. The 'problem' with indefinitely iterated games is that game theory does not prescribe any particular equilibrium of these as more *rational* than any other. For instance, the strategy 'always defect', is not ideal despite the fact that 'always defect, always defect' is an equilibrium in *both* the one-shot and the indefinitely iterated case (unlike 'always cooperate, always cooperate', for the best reply to a 'always cooperate' strategy is 'always defect') since in equilibrium you would do better by *cooperating* if the game is repeated. 'Cooperate, cooperate' is the equilibrium outcome with the highest payoffs, which seems to render it at least *prudent* to aim at, but to date there is no compelling 'rationality' argument in its favour. See Harsanyi and Selten 1988. Note that the deficiency is in the theory; everyone agrees in practice that the outcome with the highest payoffs is privileged.

8. But which NPs are PDs? Sobel (1985, 1991) thinks that while there are some NPs which are PDs, there are both NPs and PDs which are not PDs and NPs, respectively. He argues that *PDs are only NPs* if the conviction that the other player behaves alike is primitive (i.e., not derived from the structure of the situation); in other words, for a PD to be a NP Sobel has to allow the possibility of cooperation in a one-shot PD (e.g., if the players share some peculiar mental characteristic which makes them cooperate). I think that such a situation is either not a PD, or it is (in a somewhat deviant sense) and the players are cooperating irrationally—which makes the analysis irrelevant to rational choice theory. Indeed, Sobel (1991) argues, in a nutshell, that *NPs are only PDs* in case the structure of the NP is such that the predictor is nothing but the other player in a standard PD. In Sobel's exposition this is obscured because the accompanying narrative is very different, and not symmetric for the two players: "the predictor . . . cares mainly about the money so that being right is not worth $1,000 to him . . . in addition, *he dislikes Takers, and would prefer that the predictee not be one*" (1991, 202, italics as in original). Of course this would imply that *any* ordinary PD may be a NP, contrary to what Sobel (1985) claimed.

Sobel's confusion derives, I think, from his assumption that the characteristic mark of a NP is that the reliability of the predictor is primitive—which I deny (see below). In consequence, Sobel's analyses fail in explaining the reliability of the predictor: e.g., the hypothesis (1991, 200ff) that 'the predictor wants only to be right' does not entail that the predictor *will* be right, a problem Sobel bypasses by circularly *assuming* that the player will definitely take both boxes. Of course if the predictor knows *this* he won't have much difficulty in predicting it either.

9. John Mackie (1977, 154) also recommends taking one box in a specially constructed repeated case ("the [predictor] thinks that the player's actions are determined by his character . . . and the player knows that the [predictor] believes this") and argues that "these assumptions—in particular that of repeated playing of the game—reverse the direction of causation and enable the player's choice to determine the [predictor's] move". But if the NP is a PD we do not have to talk about causation at all. Sorensen (1985) only considers finitely iterated PDs and NPs of known duration and does not reach my conclusions.

Of course, just as there are specific conditions under which cooperation in finitely

4. The Disagreement. What does this argument add to our understanding of the NP? It is only by considering the rational solutions to both one-shot and repeated NPs that allows us to realize that *both* one-boxing *and* two-boxing intuitions are entirely accurate in some cases. In particular, the causal decision theorist captures the right intuition for one-shot situations, whereas the evidentialist is correct for the selection of a strategy for indefinitely iterated games.

But this approach to the NP is only helpful if it also explains more thoroughly why causal and evidential decision theorists give the 'wrong' answer for indefinitely iterated and one-shot NPs, respectively. I think that it does. Ultimately the defect will lie in the fact that the NP is a problem of *game*- (rather than *decision*-) theory, which uses somewhat different solution concepts, but it is nonetheless of interest to see how these competing intuitions may be accounted for once the game-theoretic background is in place. I do not attempt to give a full explanation here, but another look at the PD might provide some clues since there is a very similar puzzle about the PD.

A common worry is how to explain the apparent divergence between the recommendations of 'economic rationality' (defect) and 'common-sense intuition' (cooperate) in one-shot PDs. The problem with the 'rational' solution is taken to be that it delivers a suboptimal outcome. Proponents of the 'rational' solution usually claim that the unconscious focus on less abstract 'real life' situations provides an explanation of the appeal of the 'intuitive' solution. They say that because in real life we play PDs over and over again with the same people (at least we hardly ever know for sure that the current interaction will be the last one with this player), to cooperate is relatively more likely to be the game-theoretically rational solution than to defect. One might also argue, with some plausibility, that there is no such thing as a one-shot PD situation in real life in the first place. All of this explains why to cooperate, practically always the rational solution in real life, seems more natural even in highly theoretical situations.

Be that as it may, you might now think, why would specifically the

iterated PDs is rational (see e.g., Kreps et al. 1982; cf. Routledge 1998, 106ff), there will be conditions under which it is rational to take one box in the finitely iterated NP, rather than two boxes, as the standard backward induction argument would suggest. Finitely iterated games do not seem to have any explanatory relevance here.

Discussing the relation between PD and NP, Hurley (1991, 181) mentions *en passant* that in an indefinitely repeated PD cooperation may be individually rational; but she thinks that even though adopting what she calls 'cooperative' reasoning from the PD to the NP is a plausible explanation for the one-boxers' intuition, it is quite obviously mistaken, given that the NP is based on a case of a common cause (1991, 175), which she takes to be the best explanation for the NP's structure. Consequently she argues that in a repeated NP, one should take both boxes too (1991, 190).

evidentialist think of iterated games when approaching the NP? Obviously, if NPs are nothing but PDs then being a one-boxer is just an *instantiation* of the just mentioned intuition to 'cooperate' in the one-shot PD and should be amenable to the same explanation—the overall frequency and success of that strategy in real life. But there is also a good reason intrinsic to the set-up of the NP. It is the evidentialist's adoption of the *probability as frequency* of the predictor being right in the past as his *probability as subjective degree of belief* in the accuracy of the next prediction which (unconsciously) directs his attention to the repeated case. *Only* the evidentialist requires such a probability to make his decision (the causal decision theorist only uses 'dominance' reasoning in this case), but the only explicit probability to fix expectations in this example is the past accuracy of the predictions.[10] The evidentialist uses the only data available, which is about the long run, to have anything at all to go by. It would be a clear fallacy, however, to believe that since past predictions were 99% accurate, the *next* prediction must be accurate with a probability of 99%, too.[11] Whereas the probability the evidentialist *requires* is epistemological, a subjective degree of belief, the one he *relies* on is statistical. For the one-shot scenario, this is not a good basis, but of course such long-run data would be more reliable for *long-run* prediction.

Since causal decision theory was developed in reaction to the NP, and arguably similar common-cause scenarios where the evidentialists seemingly went wrong[12] we do not have to analyze why it recommends *taking both boxes*—a major aim of the project was to provide us with a theory in which taking both boxes falls out as the uniquely prescribed outcome. In fact we knew the theory's desired prescription before the theory itself even existed. The corresponding question would be why the causal decision theorist focuses on the *causal* properties of the situation to yield this choice. It seems reasonable to presume that the focus on causation in the NP derives by simple analogy from the field of probabilistic cau-

10. Hubin and Ross (1985, 440) have similarly suggested that "controversy between one-boxers and two-boxers persists [because of] selective attention to certain constraints on the puzzle." Also see Levi 1978.

11. It has been argued elsewhere that it does not *follow* from the mere fact that the predictor predicted accurately in 99% of past cases that the probability of his *next* prediction being accurate is 99% too, irrespective of what you do. See Levi 1978, 1982, and 2000. Jeffrey's theory does not require one to mechanically adopt statistical data in such a way—this would indeed be a "bizarre theory" (Pearl 2000, 108). Also see Price 1986, 197; 1991.

12. Whether Bayesian decision theory is unable to deliver the intuitively right solution in *any* of these cases (including the NP) is heavily disputed. See, e.g., Levi 1978 and 2000; Eels 1982; Price 1986, 1991.

sation, the debate about the extent to which probabilistic correlation provides evidence for causal connection.

I am not claiming that philosophers writing on the NP are confused as to whether the decision problem is one-shot or indefinitely iterated. My claim is that, as it is described by Nozick, the NP itself is ambiguous. If this is so, one should expect this ambiguity to be reflected more or less perspicuously in the proposed solutions. It is quite common in the literature on the NP to claim that the problem is not well specified. Some think it is overspecified in some respects (e.g., Hubin and Ross 1985), some think it is underspecified (e.g., Cargile 1975; Mackie 1977; Levi 1978), or even outright inconsistent (e.g., Bar-Hillel and Margalit 1972). My analysis falls into this heterogeneous 'no-boxing' tradition. But rather than rejecting the NP as mis-specified, I go beyond this literature by showing *how* this mis-specification actually explains the observed disagreement between one-boxers and two-boxers: the disagreement arises because the standard descriptions of the NP mix factors which lead one to think of a repeated situation with some which emphasise the one-off character of the interaction, *and* one-boxing is rational if the situation is repeated, while two-boxing is rational in the one-off case, given the NP-PD analogy.

For instance, the long series of past reliable predictions, the predictor playing continuously, and the successful 'evidential' one-boxers all support the hypothesis that the NP is a repeated game. On the other hand, the uniqueness of *your* decision, which seems one-off, and the 'unsuccessful' two-boxers may lead one to believe that the NP is one-shot. Yet as the selection of the right strategy crucially depends on whether the game is one-off or indefinitely iterated, the current debate may well be accounted for as a debate about what structure actually is underlying the NP situation.[13] It just *isn't* clear what exactly that structure is (partly, too, because the predictor is said to be 'supernatural'). As soon as the possibility of a one-shot/repeated ambiguity is recognized, adopting Lewis's PD analogy and some results from game theory is a powerful way of clarifying intuitions.

5. Supernatural Predictors with Invisible Hands. In the preceding sections I have argued that by considering the distinction between one-shot and indefinitely iterated cases of the NP, together with the Lewisian argument

13. What rationality prescribes in other situations discussed in this context often seems much less problematic (e.g., some common cause of both lung cancer and the disposition to smoke). Because a common cause is usually *assumed*, these situations do not share the one-shot/indefinitely iterated ambiguity. See Hurley 1991, who bases her argument on this observation, and Eells 1982, Chapter 8.

about the analogy between NP and PD, the suggestions of both 'one-boxers' and 'two-boxers' can be rationalized. Besides doing analytical work, the similarity between PD and NP also served to illustrate this idea since essentially the same disagreement about what is the rational choice occurs with the PD too, and is often analyzed along the same lines.

But the PD can do even more work for the NP. I will argue in this section that if the NP is understood as 'half a PD' (that is, a PD seen from the perspective of just one of the players) not even the mystery about the predictor's reliability has to remain intractable. This is a significant advantage of the present analysis. Ultimately, both one-boxers and two-boxers have to claim that the predictor must be 'magic' since he is able to predict (what either party perceives to be) *ir*rational behaviour. But if both taking one box *and* taking two boxes is rational in some situations, and thus easily predictable in those situations, we only need to assume that the predictor is able to predict *rational* behaviour. Assuming that we are looking for a rational solution to the NP, this seems preferable.[14] The only challenge is to present a scenario consistent with what we know about the NP in which this is the case. This has essentially been done by David Lewis by drawing the analogy from NP to PD. I extend his argument to the iterated case to explain the NP in its entirety.

We said that the accuracy of the predictor's predictions in the NP corresponds to the two players' making the same choice in the PD. But in the PD there is no corresponding *mystery* about the two players choosing the same option as there is about the accuracy of the predictor in the NP. It's just a matter of incentives. Rather than being under the illusion that my choice 'causes' the other prisoner to behave as I do, the simple fact that *he has the same incentive structure as I do* provides an exhaustive explanation of the 'invisible hand' phenomenon of joint cooperation (or joint defection). I choose the strategy best for me, he chooses the strategy best for him, period. Since the game happens to be set up symmetrically this means we always do exactly the same thing: given rationality and desire to maximize payoffs, my fellow player defects, as I do, in one-shot games, and he cooperates, as I do, in indefinitely iterated games (and very rarely makes mistakes). But this startling similarity in our behaviour is a *consequence* of the symmetric structure of the game and not an additional 'magic' assumption. Any other rational person in this situation would do exactly the same thing. It is in this sense that, in Lewis's PD analogy, the other player's behaviour 'replicates' your behaviour and therefore predicts

14. It is to beg the question against my analysis to claim that the fact that the predictor has accurately predicted both one-boxers and two-boxers in the past, while taking one box/two boxes is in fact irrational, *shows* that the predictor *is* able to predict irrational behaviour.

it. An additional 'twin' assumption is not required here, neither for the PD, nor for the NP.[15]

So, to take up a thread from the second section, in the PD we know perfectly well *why* the other prisoner mirrors my behaviour even without consulting psychologists or crystal balls. In the NP, however, we seem to lack any explanation for the success of the predictor. And it is precisely because his unbelievable reliability looks like an *assumption* of the story that the actual existence of a being like the predictor seems so hard to swallow (he is also said to be supernatural). But arguably the best explanation for the predictor being reliable in his predictions of your actions is still that you are being predictable in your actions. If we want to make sense of the NP we should ask whether there may be situations in which we would *anyway* get such reliable predictions without having to invoke supernatural magic, and whether the structure of such situations could be squared with the set-up of the NP. The PD is such a situation: in an ordinary PD the role of 'predictor' of your actions is automatically assumed by your fellow prisoner deliberating in order to get the best outcome for himself[16] (and it could be assumed by any agent observing the situation and having a minimal degree of knowledge of game theory). Since the other player is in exactly the same predicament as you are, his actions are a perfect guide to your actions. The other player is similar to you in that he simulates your *behaviour* rather than in that he shares your mental set-up, and he behaves similarly because he is in the same situation as you are. That is to say, if you are told that in a PD your fellow player behaves exactly like you, this does not add any information relevant to your rational decision making. Since you know that the PD is symmetric you can deduce from the structure of the PD already that he will behave exactly like you. In the NP the situation is exactly the same. If in the NP you are told that the predictor will accurately predict your choice, this

15. To make an additional twin assumption is not just redundant, but inconsistent with game theory. Binmore (1992, 311) says that "The [fallacy of the twins] is particularly inviting in the Prisoners' Dilemma because the game is symmetric." That 'cooperate, cooperate' is the solution prescribed by rationality in one-shot PDs has most prominently argued by Rapoport (1966) and Davis (1977, 1985). For an analysis and rebuttal of several such arguments see Bicchieri and Green 1997, for a more comprehensive analysis see Binmore 1994, Chapters 2, 3. Also compare Gibbard and Harper 1978, 157 and Horgan 1981, 180. It is not obvious that Lewis (1979) is not tempted by this fallacy since he talks of the other player being 'a more or less reliable replica' of you, despite the fact that in the one-shot PD, any rational player would defect without exception.

16. Binmore (1992, 249) points this out, but rejects the present solution to the NP because it does not account for the fact that the predictor correctly predicts one-boxers too. This worry is defused by considering indefinitely iterated games.

doesn't impinge on the rationality of your choice. It is because choosing two boxes in one-shot and taking one box in indefinitely iterated games are the best choices that accurate prediction is possible in the first place.

The accuracy of the predictor's predictions will only depend on his knowledge of the structure of the game, of your rationality and of your desire to maximize your payoffs.[17] It is therefore misleading to call such predictive abilities 'supernatural' as this renders the whole NP situation susceptible to criticisms along the lines of Jeffrey (1983, 25): "If cows had wings, we'd carry big umbrellas." Maybe; maybe not. But as Lewis (1979) has already argued, if some PDs are NPs, then the set-up of a NP should not in general considered to be less plausible than that of a PD. And this applies to the existence of the predictor too.

6. Conclusion. The NP is just a particularly picturesque version of a PD with half the story missing (as well as some added ambiguities). A major advantage of accepting this reductive solution is a sensible explanation of the reliability of the predictor which explains how the predictor both predicts one-boxers and two-boxers without having to be supernatural. And if the NP is just a PD then it is not such a difficult problem to solve after all. In one-shot situations the only rational thing to do is to take both boxes, but in indefinitely iterated games taking just box 2 provides you with a higher payoff. This approach also provides a rationalisation of the debate between causal and evidential decision theorists: one-boxers have focused on the one-shot situation, two-boxers on the iterated version.

REFERENCES

Bar-Hillel, M., and A. Margalit (1972), "Newcomb's Paradox Revisited", *British Journal for the Philosophy of Science* 23: 295–304.

Bicchieri, C., and M. S. Green (1997), "Symmetry Arguments for Cooperation in the Prisoner's Dilemma", in G. Holmstrom-Hintikka and R. Tuomela (eds.), *Contemporary Action Theory*, vol. II, *The Philosophy and Logic of Social Action*. Dordrecht: Kluwer, 229–249.

Binmore, K. (1992), *Fun and Games*. Lexington, MA: Heath.

——— (1994), *Playing Fair: Game Theory and the Social Contract*, vol. I. Cambridge, MA: MIT Press.

Brams, S. (1975), "Newcomb's Problem and Prisoners' Dilemma", *Journal of Conflict Resolution* 19: 596–612.

Cargile, J. (1975), "Newcomb's Paradox", *British Journal for the Philosophy of Science* 26: 234–239.

17. I am happy to accept that my solution does not apply to those NPs in which the predictor is *infallible in principle* (he *necessarily* predicts your actions correctly), but I doubt anyone would want to raise this as a serious objection. Even on my analysis the predictor may still be 100% reliable *de facto*—but this is contingently so. It is an advantage of my account that there is no discontinuity in such a case. For a discussion of infallible predictors see Sobel 1988.

Davis, L. H. ([1977] 1985), "Prisoners, Paradox, and Rationality", in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 45–58. Originally published in *American Philosophical Quarterly* 14: 319–327.

——— (1985), "Is the Symmetry Argument Valid?", in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 255–263.

Eells, E. (1982), *Rational Decision and Causality*. Cambridge: Cambridge University Press.

Fudenberg, D., and J. Tirole (1991), *Game Theory*. Cambridge, MA: MIT Press.

Gibbard, Allan, and W. Harper (1978), "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, vol I. Dordrecht: Reidel, 125–162.

Harsanyi, J., and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Horgan, T. ([1981] 1985), "Counterfactuals and Newcomb's Problem", in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 159–182. Originally published in *Journal of Philosophy* 78: 331–356.

Hubin, D., and G. Ross (1985), "Newcomb's Perfect Predictor", *Nous* 19: 439–446.

Hurley, S. L. (1991), "Newcomb's Problem, Prisoners' Dilemma, and Collective Action", *Synthese* 86: 173–196.

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

Joyce, James (1999), *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma", *Journal of Economic Theory* 27: 245–252.

Levi, I. (1978), "Newcomb's Many Problems", in C. Hooker, J. Leach and E. McClennen (eds.), *Foundations and Applications of Decision Theory*. Dordrecht: Reidel, 369–383.

——— (1982), "A Note on Newcombmania", *Journal of Philosophy* 79: 337–342.

——— (2000), "Review of James Joyce: The Foundations of Causal Decision Theory", *Journal of Philosophy* 97: 387–402.

Lewis, D. ([1979] 1885), "Prisoner's Dilemma Is a Newcomb's Problem", in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 251–255. Originally published in *Philosophy and Public Affairs* 8: 235–240.

——— ([1981] 1988), "Causal Decision Theory", in P. Gärdenfors and N.-E. Sahlin (eds.), *Decision, Probability and Utility*. Cambridge: Cambridge University Press, 377–405. Originally published in *Australasian Journal of Philosophy* 59: 5–30.

Mackie, J. L. (1977), "Newcomb's Problem and the Direction of Causation", in J. L. Mackie, *Logic and Knowledge, Selected Papers*, vol. I. Oxford: Clarendon, 145–158. Originally published in *Canadian Journal of Philosophy* 7: 213–225.

Nozick, R. ([1969] 1985), "Newcomb's Problem and Two Principles of Choice", abridged in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 107–133. Originally published in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel, 114–146.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Price, H. (1986), "Against Causal Decision Theory", *Synthese* 67: 195–212.

——— (1991), "Agency and Probabilistic Causality", *British Journal for the Philosophy of Science* 42: 157–176.

Rapoport, A. (1966), *Two-Person Game Theory*. Ann Arbor: University of Michigan Press.

Routledge, B. R. (1998), "Economics of the Prisoner's Dilemma: A Background", in P. A. Danielson (ed.), *Modelling Rationality, Morality, and Evolution*, Vancouver Studies in Cognitive Science, vol. 7. Oxford: Oxford University Press, 92–118.

Sobel, J. H. (1985), "Not Every Prisoner's Dilemma Is a Newcomb Problem", in R. Campbell and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 263–274.

——— (1988), "Infallible Predictors", *Philosophical Review* 97: 3–24.
——— (1991), "Some Versions of Newcomb's Problem Are Prisoners' Dilemmas", *Synthese* 86: 197–208.
Sorensen, R. A. (1985), "The Iterated Versions of Newcomb's Problem and the Prisoner's Dilemma", *Synthese* 63: 157–166.