The Prisoner's Dilemma and Social Theory: An Overview of Some Issues

Philip Pettit

Would rational individuals co-operate, even if wholly self-interested? In the event of their not doing so, might there still be reason why they should? And in that case, would the enforcement of co-operation be justified; would it be something which the individuals ought to welcome? The predicament projected in such questions has been at or near the focus of social theory since at least the time of Hobbes. But recently it has become particularly salient. It has found a perspicuous and persuasive exemplification in the prisoner's dilemma.

The prisoner's dilemma gets its name from the following story. A district attorney confronts two prisoners who are held for a crime which she believes they jointly committed. Knowing that she lacks evidence to pin the crime on them in court she devises a predicament which ensures that they both confess.

She interviews each on his own and offers a deal. Confess and you go free if the other chap refuses whereas you get a three-quarters rap if he also confesses. Refuse to confess and you get a quarter rap — on some trumped-up charge or whatever — if the other prisoner also refuses while you get the full term if he confesses. The D.A. tells each that she is offering the same deal to the other and confidently waits for a confession.

Her confidence is based on the fact that she thinks each will rank the payoffs as follows on a scale between 1, the worst payoff, and 4, the best; the figures need have only ordinal significance:

Confess when the other confesses: 2 (the punishment payoff, P)
Refuse when the other confesses: 1 (the sucker's payoff, S)
Confess when the other refuses: 4 (the temptation payoff, T)
Refuse when the other refuses: 3 (the reward payoff, R)

If the D.A. is right about the prisoners' rankings then a standard assumption of rationality suggests that each will indeed confess. The strategy of confessing is dominant. It is better when the other man confesses, yielding 2 rather than 1; and it is better when he refuses, giving 4 rather than 3.

The dilemma which the D.A. manufactures for her prisoners is obvious once we construct a matrix, showing the payoffs for each under every combination of strategies. One prisoner chooses between the rows, the other between the columns. Row's payoffs are shown first in each box, Column's second:

	Refuse	Confess
Refuse	3, 3	1,4
Confess	4, 1	2, 2

The dilemma consists in the choice between refusing and confessing. Confess and you will end up with 2, refuse and you will end up with 1. And this despite the fact that if you could only both refuse at once you would each get 3.

More formally explicated, two conditions characterise the prisoner's dilemma. The first is that the strategy of confessing is dominant for each. The second is that the outcome of both parties following that strategy is Pareto-inferior to the outcome of each refusing to confess: it does worse for some — in our case all — and is better for none.

Commentaries on the dilemma often emphasise different features but these two capture all that is important. Sometimes the fact that confess-confess is an equilibrium receives more attention. But that means just that it is an outcome from which no party can unilaterally depart with advantage, a feature ensured by the fact that confessing is dominant for each. Sometimes too the essence of the dilemma is said to be that T is preferred to R, and R to P, and P to S. That is true, given the structure of the payoffs, but not surprisingly: it ensures after all that the two conditions mentioned are fulfilled.

In a standard taxonomy there is only one variety of two-party game satisfying these conditions (see Rapaport and Guyer). The importance of the prisoner's dilemma story is that it illustrates the sort of co-operative predicament in question. It shows that there are cases where the concern of two parties to maximise their own payoffs can lead them each to get a lower payoff than might otherwise have been attained. It demonstrates that there are instances where individualistically rational strategies collectively yield sub-optimal results.

If this sort of co-operative predicament occurs commonly in social life, then the prisoner's dilemma story points to an important lesson. It shows that we cannot rely on the invisible hand to distil collective good out of individual self-interest, to convert private vice into public virtue. The invisible foot can do as much harm as the invisible hand does good. The story suggests that there is another hand at work where co-operative predicaments are solved and that, if there is not, there ought to be. It has both explanatory and policy-making implications.

In this paper I wish to consider a number of questions bearing on the significance of the dilemma story. I do not have an original thesis to defend but under each heading I shall make some points that are not recognised, or at least not commonly recognised, in the literature.

My questions are:

1. Are the motivational assumptions of the prisoner's dilemma realistic?
2. Are the assumptions about the setting realistic?
3. Does the dilemma generalise to the many-party case?
4. Is the prisoner's dilemma model useful in explanation?
5. Is it useful in policy-making?

1. Are the motivational assumptions realistic?

There are two motivational assumptions which the District Attorney makes about the prisoners. The first is that the payoffs reflect the concern of each with just his own welfare. The second is that each will try to maximise the payoff he receives. How far can these be varied? And are they realistic, so far as they cannot?

Contrary to what is almost universally suggested, the first assumption can be dropped: the payoffs may reflect on other-regarding rather than a self-regarding disposition (contrast Hardin 1982, p.10, and Axelrod 1984, p.6-7). This appears in the fact that the matrix fits a variant story as well as the original.

Suppose that the prisoners are perfect altruists, each concerned only with the other and not at all with himself. Suppose further that the D.A. knows this. She can still get them each to confess by giving them the following line. Confess and the other chap goes free if he refuses, while he gets a three-quarters rap if he also confesses. Refuse and the other gets a quarter rap if he refuses too, while he gets the full term if he confesses.

Under this offer, the matrix for each remains as it was before, the figures now reflecting an altruistic disposition. Confessing will still be dominant since it will benefit the other more, whether the other confesses or refuses. Thus both will be altruistically led to confess and only the D.A. will gain.

Given that the prisoners need not be self-regarding, must they at least seek to maximise their payoffs? Again, no: this assumption is as variable as the other. Suppose that we have a game described, with self-regarding payoffs, in the following matrix:

	Refuse	Confess
Refuse	3,3	4,1
Confess	1,4	2,2

This matrix answers to the situation of our altruistic prisoners in the story just told, but with the crucial feature that the payoffs represent their respective interests, not their altruistic dispositions.

Consider now what happens if the prisoners are assumed to maximise, not their own payoffs, but those of their partner. Each clearly will be driven to confess, since confessing in such a case will be better for his partner, regardless of what the partner does. Equally clearly, however, it will lead the two into an outcome which they must rank below the outcome of refuse-refuse.

I have shown that the twin assumptions of self-regarding payoffs and maximising strategy can be independently relaxed. Obviously they can also be relaxed together. We can easily imagine a story for which the game characterisation is the matrix just given but with the payoffs reflecting a disposition that is not self-regarding. In such a case the policy of maximising the payoff to the other will still produce the Pareto-inferior outcome.[1]

What these considerations show is that the co-operative predicament illustrated in the prisoner's dilemma does not arise just for self-interested agents. The payoffs which agents maximise need not be self-regarding and they need not even be their own. But what if the agents are not maximisers? Does that tell against the possibility of the co-operative predicament in question?

If the payoffs to each agent reflect his preferences all things considered — things regarding others perhaps as well as himself — then it scarcely makes sense to think that he does not maximise them (see Pettit, 1984a, pp.172-173). But in the story about the altruists we assumed that payoffs need not be of this encompassing character. Might the predicament arise then for agents who satisfice on the achievement of such payoffs, or take some other non-maximising line?

There is no reason why it should not. Given appropriate stories and matrices, even policies like minimising the payoff to oneself or to the other, or just realising one less than the best, will lead the agents into the Pareto-inferior outcome. There is no prophylactic available here against the dangers of the co-operative predicament.

There remains one motivational assumption in the prisoner's dilemma which we have not yet questioned. Whether payoffs are self-regarding or not, whether agents are concerned with their own payoffs or those of others, whether the concern is to maximise or to effect a different result, we have assumed so far that each agent is focussed on distinct goals. The parties may not be self-interested but they are interested in divergent goods. This even holds of the altruists, each of whom cares only for the other.

We have here the core motivational assumption in the two-person prisoner's dilemma (see Parfit 1984, Chapter 4; also Moore 1984). That is clear from the fact that once it is relaxed, the predicament is resolved. Consider what happens if just one prisoner is an altruist, so that both maximise the other's payoffs; or if both compare respective payoffs in some way and maximise the average return; or if they are subject to any twist of motivation which gives them a common goal. The predicament in many such cases disappears. The pair achieve the outcome that best answers to their desires.

We can see this if we go back to the original matrix, viz.

	Refuse	Confess
Refuse	3,3	1,4
Confess	4,1	2,2

If Row is a perfect altruist and Column a perfect egoist then both will seek the refuse-confess outcome. If they maximise average returns — assuming the figures have cardinal significance — then they will each go for refuse-refuse. And so on.

This line of reflection suggests that the motivational assumptions of the prisoner's dilemma are quite realistic. Self-interested egoists may be just images of the economics textbook but agents with divergent interests are a common or garden sort. We each look after not just ourselves, but our families, our friends and our countries, our hobbies, our jobs and our causes. Furthermore, as stockholders we appoint boards to further the company's interests; as workers we subscribe to the union to promote the workforce's welfare; as citizens we elect governments to see to the country's good. We are divergently motivated ourselves and we countenance and encourage such motivation in the institutional agents that we license.

It is no surprise therefore to find that two-party dilemmas, or at least potential dilemmas, are the stuff of everyday life: neighbours face them over fencing, draining and noise; friends over restaurants and films and other friends; companies over prices and markets; governments over pollution and armaments; and so on in a familiar pattern. The world is rife with co-operative predicaments.

2. Are the assumptions about the setting realistic?

Just as the prisoner's dilemma is often thought to be atypical because it requires self-interested agents, so it is sometimes cast as marginal on the grounds that the participants are cut off from communication with each other. This response is likely to be reinforced by the fact that allowing subjects in prisoner's dilemma experiments to communicate, even about irrelevant matters, makes it more likely that they will achieve the co-operative outcome (see Rapaport 1974).

I do not believe that the absence of communication is a significant assumption in the prisoner's dilemma. The point has often been made that, so long as the payoffs remain as they are, communication will not help the prisoners get to the refuse-refuse outcome. The only way in which it might help is by causing the payoffs to shift so that there no longer is a predicament. Having promised to refuse to confess each might find the reward payoff R more attractive than the temptation T; this, because it matters to him that he keeps his promises. That effect aside however, communication would not help to resolve the predicament. Each might promise to refuse but if he pays attention to the payoffs, he is bound to break his word.

This point holds even for our perfectly altruistic prisoners. Provided neither puts a premium on keeping his word, each will break any promise to refuse and try to further the interests of the other by confessing.

The possible effect that I put aside in the above discussion points us towards a deeper assumption about setting than the absence of communication itself. This is that the payoffs given in the matrix for the dilemma pick up everything that matters to each in his choice. He must not care about breaking or being seen to break a promise, about behaving or being seen to behave in a unilateral fashion, about signalling an individualistic conception of the relationship with his partner, or whatever. The payoffs that he considers, whether his own or those of the other, must be the only factors relevant to his choice.

This assumption is not very plausible. It would be likely to hold only under two conditions: first, that the parties do not mind having a reputation for rugged individualism or unilateralism; and secondly, that they do not meet in indefinitely extended sequences of dilemmas and other interactions.

The first condition is necessary because if the parties are concerned about having individualistic reputations, they will often be led to ignore the immediate payoffs in favour of their long-term standing. The condition is not fulfilled in the ordinary run of interpersonal exchanges but there are circumstances where it may be satisfied. These are settings in which an individualistic ethos is cultivated: among competing businessmen, among diplomacy and monopoly players, among the legal representatives of opposing clients, and the like.

The second condition is necessary because even individualistic parties will be concerned with more than the payoffs of a particular dilemma if similar dilemmas, or any interactions with reciprocal benefits, are likely to engage those parties in the future.

The iteration of two-party prisoner's dilemmas into an indefinite future undermines the assumption about setting involved in characterising each as properly a dilemma.[2] It means that each is only a potential dilemma, not an actual one.

The reason is intuitive. If two parties are going to be caught in a series of co-operative predicaments, then the strategy which will best serve each is to co-operate so long as the other does so. The strategy makes no sense in a one-off dilemma, since the agents make their choices independently. It is possible in an iteration of predicaments, because the other's choice in the previous interaction can be made the basis of one's own choice in the present one. The two parties can signal a willingness to co-operate and can punish one another for any failure to come into line. Each may co-operate for example except where the other has defected on the previous play: this is the so-called tit-for-tat strategy.

In the one-shot dilemma, only categorical strategies are available: co-operate or defect.[3] Once we have indefinite iteration, we are into a different game. Each is now in a position to choose a strategy for the sequence in prospect — the supergame, as it has been called — and the strategies on offer include conditional ones like tit-for-tat: I co-operate in the first play and afterwards I co-operate if the other has just co-operated and defect if he has just defected. In such a supergame there are equilibria with conditional strategies and these include ones that are Pareto-superior to the equilibrium constituted by each permanently defecting (see Taylor 1976, pp. 31-43). Thus we may expect rational participants to reach some such equilibrium. Probably the most salient occurs when each party uses tit-for-tat (see Axelrod 1984 and Goodin 1984).

It is extremely uncommon to find this second condition fulfilled for real-life two-party dilemmas (see Parfit 1984, Chapter 2). Most of us, and most of the institutional agents which we create, are involved in potentially long-term relationships with any other parties that we encounter in a co-operative predicament. Even if we have no worries about gaining an individualistic reputation, we will find ourselves concerned with more than the payoffs that characterise the dilemma. True dilemmas as distinct from potential ones probably only arise then in very exceptional circumstances.

3. Does the prisoner's dilemma generalise to the many-party case?

In the second section I took away with one hand what I conceded in the first section with the other. Having argued earlier that the core motivational assumption in the two-party prisoner's dilemma is realistic, I went on to show that its main assumption about setting is not. But we should not lose interest too quickly for, as we shall see in this section, the assumption about setting regains credibility in the many-party version of the dilemma. More than that indeed, it turns out that under the more general version the core motivational assumption may also be capable of being relaxed. Given large numbers, the invisible foot strikes back with a vengeance.

The definition of the two-party dilemma mentions two conditions: that the defect (ie. confess) strategy is dominant for both participants and that the outcome of defect-defect is Pareto-inferior to the outcome of co-operate-co-operate (ie. refuse-refuse). The straightforward generalisation for N persons, where N can be greater than 2, will require in parallel: that the defect strategy is dominant for each but that the outcome of universal defection is Pareto-inferior to that of universal co-operation (see Taylor 1976, Chapter 3 and Sen 1969).

The prisoner's dilemma comes into its own with large numbers because, given this definition, it is clear that free rider problems are all examples of it. The free rider problem is a familiar sort of predicament. It arises when there is a certain sort of good which a group can obtain for itself and from which no individual can be excluded; when the way to achieve the good is for members to take on a certain burden — of effort or restraint or finance; and when the good can be realised at a level sufficient to compensate contributors short of the point when all contribute. The problem is how to prevent people from letting others carry the burden, in the hope that they will enjoy the good nevertheless. Their own contribution might increase the level of the good available but it is assumed that it would not do by a sufficient margin to compensate them personally for the burden involved.

Examples of the predicament abound. Consider the situation of a team in tug-o-war; of a community which wishes to keep the local park free of litter; or of a group of television users who want to invest in equipment to boost certain signals in their area. In each case, the free rider threatens to strike. He will seek to enjoy the fruit of the labours of others, evading the burden of effort or restraint or finance which they have to bear.

The free rider problem is a prisoner's dilemma because for practical purposes it meets both of the conditions mentioned. Universal free-riding is Pareto-inferior to universal co-operation, since the cost to each in the latter event is outweighed by the benefit collectively enjoyed. Free riding however is a dominant strategy for every party. Each will reason, first, that if sufficient others contribute, then the extra benefit gained by his contributing will be less than the cost of doing so; and secondly, that if not enough others contribute, then the cost of his doing so will be in vain, since his contribution is unlikely to make the difference between provision and non-provision of the collective good. In either case free riding is the preferable strategy: it dominates co-operation.

I say that the free rider problem meets our two conditions for practical purposes. The reason for the qualification is that there is one case where free riding would not dominate co-operation. This occurs if the individual knows that he is the crucial contributor: he is the one who will make the difference between there being a sufficient and an insufficient number of contributors.

For practical purposes each can ignore this possibility in most free rider problems. It will be too unlikely to deserve consideration. Or it may be that the good to be achieved is vaguely defined and that the contribution of no individual can be held to make the difference between its realisation and its non-realisation. Grains of sand make a heap but no grain marks the divide between a heap and a non-heap. Similarly, as Richard Tuck has observed, individual contributions can make for a collective good without any one contribution being crucial for the achievement of the good (see Tuck 1979 and Fishkin 1982). This happens with goods such as having a reasonably clean park, having an adequate television signal, having a good community ethos, and the like.

For all that has been said, free rider problems may be potential prisoner's dilemmas rather than actual ones. The crucial questions are whether they fulfil the assumptions about motivation and setting mentioned in previous sections.

The motivational assumption is that each party has a distinct aim. That is satisfied in most free rider predicaments because, while each desires the collective good at issue, each wants it at least cost to himself. Each is in pursuit therefore of a different goal.

In passing, it is worth noting that by most accounts free rider problems can arise even for convergently orientated agents such as utilitarians. Suppose that N persons seek to maximise overall utility and are confronted with the task of providing a collective non-excludable good. They may each still free ride, on the grounds that since the marginal benefit of his contributing is less than the cost, the way to achieve maximal utility overall is for him to leave the creation of the good to others.

Some writers contend that such utilitarians are simply making a mistake in calculation; we shall discuss this view in the final section. The alternative story is that they are victims of their limited perspective. Each seeks to maximise overall utility but each sees that goal from a different point of view. He has to decide what intervention he will make, given that it is independently settled whether enough others will make their contributions or not. If this story is accepted, then we see that the motivational assumption of divergent interests can be slightly relaxed without undermining certain many-party prisoner's dilemmas.

Do free rider problems satisfy the assumption about setting as well? It appears that they often do. Free riders may gain individualistic reputations — if they are noticed — but we know that in some settings at any rate that won't matter. More importantly, they will not be concerned, at least in the case of large groups, with punishment in future dilemmas for past defections. One reason is relative anonymity and the fact that there will not be a single recurrent partner to mete out retribution; another we will come to later. Thus it seems likely that in many free rider predicaments the parties will be attentive just to the payoffs that characterise it as a prisoner's dilemma. We will have a genuine dilemma on our hands, not merely a potential one.

Given free rider problems, it is clear that prisoner's dilemmas generalise to the many-person case and indeed come into their own there. In concluding this section I would like to turn to a topic not much discussed in the literature. This is the source of the dilemma in the free rider case: the reason why the payoffs there assume the pattern characteristic of the dilemma.

Suppose that all the members of a group are required to provide a non-excludable good. It is clear in that case that no-one will be tempted to free ride. Everyone will be deprived of his share of the good if he defects and we assume that this benefit outweighs the cost of contributing. A similar point holds in the case where a subgroup of K members is capable of providing the collective good, if only at some minimal threshold (see Olson 1965 and Schelling 1978). If anyone knows that K-1 other parties, and only K-1, have contributed or will contribute, then he is no longer tempted to free ride. This shows that the reason why the free rider problem exists, the reason why the payoffs there assume the profile of a dilemma, is that everyone must deem it extremely unlikely that he will be the decisive contributor.

We can now set up a contrast — never, surprisingly, made in the literature — between free rider problems and another species of prisoner's dilemma: these I will call foul dealer problems. The free rider defects only because he does not think he will be decisive for the minimal achievement of the good in question. The foul dealer may also not believe that he will be decisive but that is not his reason for defecting. He will have reason to defect — defecting will remain dominant for him — even if he knows that N-1, let alone K-1, have already co-operated.

Both free rider and foul dealer problems display the structure of costs and benefits which makes them prisoner's dilemmas: both meet the two conditions mentioned. The difference comes in the source of those costs and benefits. The free rider balances the marginal benefit of his contribution to a collective good against the absolute cost of that contribution, psychological, physical or financial. The foul dealer — at least when he knows he is decisive — balances his share in the benefit procured by co-operation against the comparative cost of a lost opportunity: the opportunity to take advantage of the co-operation of others, scoring at their expense.

Another way of making this contrast is to consider the attitude which N-1 contributing members ought to take to a defector. In a free rider problem, where N-1 is greater than K, they may feel resentment at the defector's costlessly enjoying the fruit of their labour. They are marginally worse off after all than if he joined in. However it will generally make no sense for them to think of retaliating by massive defection; this matter will come up again in the final section. That is because the defector does not make them worse off than they would have been if they had not co-operated and had not thereby exposed themselves to free riding. He may put them below the universal co-operation baseline but he does not force them beneath the baseline of universal non-co-operation.

In a foul dealer problem things are quite different. The N-1 contributors are not just free ridden in a costless enjoyment of their labour. They have their efforts nullified and their work undermined. The foul dealer reduces them to a level below the baseline of universal defection. Thus it will be rational for them, not just to feel resentment, but to defect themselves in retaliation.

There is a weakened and more general version of the foul dealer problem that should also be mentioned. In this predicament, the lone defector reduces some party or parties, but not all, below the baseline of universal defection. Here it will be rational for co-operators to threaten to defect in response to any single defection if they are of a maximin mentality, wishing at all costs to avoid the worst possibility.

Foul dealer problems are clearly a purer form of prisoner's dilemma than free rider ones. Two-party dilemmas are all of the foul dealer sort, since the lone defector does not share in the benefit of the other's co-operative effort but takes damaging advantage of him. Many party dilemmas too can fit the bill. One example might be the predicament of a community of armed and violent people. All are better off if each destroys his weapons at a prearranged time. But the knowledge that he will make the difference between co-operation by N and N-1 or by some threshold K and K-1 does not give anyone a reason not to defect. On the contrary, it promises maximal benefit.

Foul dealer problems are unlikely to be more than potential predicaments. This is because every defector will attract retaliation in further exchanges, at least if the problem is of the non-weakened sort. One-shot dilemmas therefore will give way to a supergame in which equilibria other than universal defection will be available and Pareto-superior. Foul dealer problems, like two-party dilemmas generally, will be short of an appropriate setting.

Finally, a note in redemption of a promise. I said earlier that there was a further reason besides anonymity why parties in a free rider problem need not be concerned with factors other than the characteristic payoffs. The reason should now be obvious. Resentment may be a rational response towards a free rider but retaliation will not be. Thus the free rider, unlike the foul dealer, does not have reason to fear the punishment of massive defection in further instances of the predicament. He can defect without much fear of the consequences.

4. Is the prisoner's dilemma model useful in explanation?

Suppose I am asked to explain why there are trees planted on one of the university lawns. I will not know what sort of explanation is sought, let alone provide it, unless two further matters are clear to me. First, I must be aware of the contrasting possibility or possibilities envisaged in the question. Am I asked to explain why there are trees rather than flowers; why there are trees rather than nothing; why there are trees planted rather than trees freestanding in boxes; why there are trees planted on that lawn rather than some other; and so on?

Secondly, I have to see what the background suppositions are in the light of which the actual situation is to be explained. Is the question asked within everyday psychology, on the assumption that gardeners plant trees and the like for aesthetic effect? Or is the answer to that taken as given, and the question posed within Freudian psychology, on the assumption that aesthetic tastes spring from deeper sources?

These observations on contrasting possibility and background supposition are relevant, not just in the case of our trivial example, but for any form of explanation (see van Frassen 1980, Chapter 5). What they show is that the object which an explanation is meant to make intelligible is not the simple state of affairs expressed by the sentence 'q' which occurs in the explanation 'That p explains why q'. That object, the *explanandum,* is something expressed by the more complex sentence: 'q rather than r, given s', where r is the relevant contrast (s) and s the relevant supposition (s).

Returning now to the prisoner's dilemma model, we are in a position to notice its explanatory benefit straightaway. This is that the model generates fruitful *explananda* for social theory. It does so by highlighting social predicaments which are at least potential dilemmas and by providing them with a background of appropriate contrasts and suppositions.

The highlighting of predicaments occurs when, with the model in mind, we begin to discover, as M. Jourdain discovered that he had been speaking prose, that we have been dealing with prisoner's dilemmas all our lives. Am I to contribute to the emergency fund for Ethiopia? Am I to make a correct tax return? Am I to put the anti-pollution device on my car? Is my union to do its bit for advancing the cause of the unemployed? Is my city or state to play its part in maintaining national highways? Is my country to pay the cost of upholding a boycott against South Africa? Social life begins to bristle with potential co-operative predicaments, once we look at it from the viewpoint of the prisoner's dilemma.

Suppose that a predicament is one that is resolved by or for the parties involved. In that case the model will cause us to look for the factor or factors that make a difference between it and the obvious contrasts: analogous predicaments in which the participants fail to co-operate. Suppose on the other hand that it is not resolved. In that event the contrasts selected will be counterpart situations in which co-operation is achieved and we will be invited to search for the factors that are lacking in the predicament on hand.

What the model does in both of these cases is to give us good questions to ask. The questions are good on a number of grounds. They do not make 'unrealistic' motivational assumptions which would deprive the answers of a connection with the real world. They allow answers which are comparable across very different instances and which may accumulate in a body of general knowledge. And the answers which they invite engage with our interests as social participants: they show us what we are doing right when we succeed in co-operating and what we are doing wrong in other cases. In short, the questions generated by the model make possible a social theory that is at once relevant, general and practical.

But does the prisoner's dilemma model suggest background suppositions as well as contrasts for the social explanatory enterprise? I believe that it does, though this point is not so manifest. The suggestion which I find in the model is that where co-operation is achieved, that is to be psychologically explained by reference to the effects of iteration or other external forces on the opportunities or payoffs or dispositions of the agent. In other words, the background supposition is that people, and the institutional agents which people create, behave in a rationally interpretable manner: they act intelligibly, in the light of independently intelligible attitudes (see Macdonald and Pettit 1981, Chapter 2). That this is the supposition suggested may not be so obvious however, given that there is an alternative candidate in the offing.

The alternative proposal is that society conforms to the picture projected in the writings of some functionalists, being such that the prerequisites of survival, stability, cohesion or whatever are more or less deterministically fulfilled. The prisoner's dilemma model might just be taken to suggest that there is a function-satisfying mechanism at work, the role of the model being to highlight putative functions in need of fulfilment. Where the predicaments highlighted are resolved, the functionalist story would run smoothly. Where they are not, we would be invited to see why their resolution is not after all important or to identify some block to social functioning, some spanner in the works of society (see Stinchcombe 1980).

The functionalist supposition of a self-adjusting society is highly dubious, since it is not clear why or how society should come to adjust in this way (see Macdonald and Pettit 1981, Chapter 3). In any case, however, I think that it is foreign to the way of thinking associated with the prisoner's dilemma model. In that model we are invited to consider the psychological payoffs in virtue of which certain co-operative predicaments obtain. Those payoffs make for a predicament only so far as the parties are expected to act in the light of their payoffs. But to be expected to act in that way is to be supposed to be rationally interpretable. Thus the background supposition is as I described it earlier.

In summary then, the explanatory use of the prisoner's dilemma model is that it gives us good questions to ask in pursuit of rational as distinct from functional social theory. But before leaving the topic there are three further comments I would like to make.

The first is that the dilemma is just one, albeit probably the most important one, of a number of social predicaments. Others are represented in games like the assurance game or the game of chicken, as they are known in the literature (see Taylor and Ward 1982). These are likely to serve similar explanatory roles to that which we have assigned to the dilemma. Using the dilemma as a model does not mean disavowing the use of such other games. The point scarcely needs emphasis.

The second comment I would like to make is that when we explain a social response or institution as a solution to a prisoner's dilemma, or indeed to any social predicament, we are required to show how it could have emerged among rationally interpretable agents. Otherwise our explanation is no more helpful than the functionalist's observation that the phenomenon is functional. It is not surprising that as-if aetiologies have an important place in the sort of social theory associated with the dilemma model. Such *a priori* and counterfactual derivations are our only guarantee that it is sensible to claim that the response or institution occurs because it solves the predicament.

My final comment is of a more general character. It is that the use of the prisoner's dilemma model is part of a broader explanatory enterprise and that the invisible hand, even in the sense in which it may work like an invisible foot, is only one mechanism which may be invoked there. The enterprise begins with presumptively rational human beings and calls upon a variety of mechanisms to explain how the unseeing or half-seeing initiatives of such agents give rise to aggregate and organisational phenomena: phenomena which, from the viewpoint of the agents' interests, may be benign or malign or mixed.

The mechanisms which such theory calls upon are usefully characterised in the image of guiding hands. The hand which guides may combine the efforts of individuals, or select from among their products, or simply pre-empt what they do. Each hand may operate in a hidden or overt manner. Thus the combinative hand may be *the invisible hand* with which we are familiar or *the helping hand* of public co-operation. The selectional hand may be *the dealing,* and perhaps double-dealing, *hand* of the covert manipulator, or *the even hand* of the open poll or contest. Finally, the pre-emptive hand may be *the hidden hand* of backstage control or *the heavy hand* of dictatorship (see Pettit 1984b).

The prisoner's dilemma model commits a theorist to the belief that it is mechanisms such as these which move the social world, meshing individual behaviour into institutional outcome. It is not *the iron hand* of history or sociology which does that work, or at least not in a manner that undercuts rational agency.[4]

5. Is the prisoner's dilemma model useful in policy-making?

Because it captures the structure of a recurring sort of social predicament, there can be no doubt but that the prisoner's dilemma model is bound to be of use to policy-makers. I do not mean just to policy-makers in the narrow public service sense. I mean to all of us, when we consider how society is best organised under its civil, economic and political aspects (see Pettit 1980, Chapters 1-3).

The model serves as a discipline for the imagination when we examine the question of how co-operative predicaments are best resolved. It enables us to distinguish the various sorts of resolutions possible and to debate their merits in neglect of distracting detail.

The resolutions can be roughly classified as strategic, psychological and political. The strategic gets parties over a potential co-operative predicament by embedding it in a sequence of predicaments and generating a supergame in which it is rational to co-operate. The psychological relieves them of their problem by inducing attitudes or principles of such a character, usually of such a moral character, that co-operation becomes desirable: the troublesome payoffs are reformed or made irrelevant. Finally, the political resolution introduces such restraints or threats or promises as alter the opportunities or payoffs of the participants and, once again, rationalise the co-operative option.

In the assessment of these various sorts of resolution, it is plausible to rank them according to degree of interference with the participants involved. On such a ranking, strategic solutions would be best, psychological second best, and political worst. This is not the place to debate such an ordering principle or to consider other grounds of assessment. What I would like to do instead is to take the principle as given and mention some problems that stand in the way of three non-political solutions which have recently been touted.

The most radical of these is a resolution — or perhaps a dissolution — which Richard Tuck has proposed for the free rider problem, or at least for certain instances of that problem (see Tuck 1979). Tuck's line is that not even strategic factors are necessary to resolve this predicament; it withers under clear scrutiny, and indeed under the scrutiny of regular participants.

Tuck argues that the potential free rider finds himself in the same sort of predicament as the potential procrastinator and that just as procrastination is a problem which our ordinary common sense gets us over, so we can rely on everyday wisdom to get us over the free riding difficulty. In each case, as Tuck constructs them , our task is the achievement of a vague outcome such that each individual contribution is incapable of making the difference between its realisation and its non-realisation. The procrastinator is a shepherd, in Tuck's example, who wants to build a cairn by adding a stone a day. The free rider is a member of a group which wishes to obtain a parallel, vaguely defined goal by having each individual perform the counterpart of adding a stone.

We may grant that at least certain sorts of free rider problems are of this character. The question is whether we should go along with Tuck's conclusion that just as the shepherd's problem — no single stone makes the difference between a cairn and a non-cairn — is solved *ambulando,* so we may expect the free rider difficulty to be overcome; that just as people have enough sense not to procrastinate, so they ought to be sufficiently intelligent not to free ride.

Unfortunately, I do not think that we can indulge ourselves in such optimism. There is an important asymmetry between procrastination and free riding which Tuck's parallel covers up. The reason it is not sensible of me to put off adding today's stone may be that such an omission is likely to set up a disposition in me to postpone every addition and that would be fatal for the cairn. Such a reason does not obtain in the free riding predicament. My free riding is not likely to cause others also to free ride and so the consideration that militates against procrastination does not carry any force here.

A less radical proposal than Tuck's has been suggested recently by Russell Hardin (see Hardin 1982, pp.28-29). He is primarily concerned also with the free rider problem. The argument is that there are potential resolutions of some many-party dilemmas which strategic considerations should cause to emerge; and that such a resolution must be available for the free rider problem, since it is an instance of a prisoner's dilemma.

Hardin's premise is true and interesting. He shows that in the many-party dilemma there are conditional strategies which are not just Pareto-optimal equilibria, but Pareto-optimal co-ordination equilibria. The equilibrium is an outcome such that no one can do better through his own unilateral departure from it. The co-ordination equilibrium is more demanding: here no one can do better through any unilateral departure, his own or someone else's. The availability of such an equilibrium does suggest that the predicament may be strategically resoluble. Let a group get to such an outcome — Hardin describes some possible routes — and all will prefer that each conform, thereby having a motive to keep everyone in line by whatever sanctions are available.

I do not believe however that Hardin's conclusion follows. The reason is that some prisoner's dilemmas are foul dealer problems, others free rider predicaments, and resolutions of the one sort may not be available for the other. More particularly, I believe that the co-ordination equilibrium solution is appropriate for foul dealer cases but not for the free rider problems to which Hardin applies it.

The Pareto-optimal co-ordination equilibria which Hardin hails involve conditional strategies such as tit-for-tat: I co-operate unless someone has just defected. If everyone followed such a strategy in a prisoner's dilemma, we would certainly have a co-ordination equilibrium. By switching, each would invite punishment and do worse than he does with tit-for-tat. By anyone else's switching, he would be required to mete out punishment, i.e. defect next time around, and he would also do worse himself.

The problem with the co-ordination equilibrium envisaged however is that the conditional strategies involved are not generally credible in a free rider predicament, only in a foul dealer one. The foul dealer's defection puts all others beneath the baseline of universal defection and so he must believe them when they threaten to defect for his every defection. The free rider's defection does not do this and he would find the corresponding threat quite hollow. No group of parties in a free rider problem could persuasively aspire to converge on an equilibrium of conditional strategies.

The weakest predicament in which such strategies would be generally credible is the weakened foul dealer problem. This arises when the defector is bound to plunge somebody, though not necessarily everybody, below the baseline of universal defection. Here it would be credible of everyone to threaten defection for a defection by another, if it is credible that everyone maximins, protecting himself against the worst eventuality: viz. being the person reduced to misery. That could just be believed, at least under certain sorts of payoffs.

Despite suggesting the argument mentioned, Hardin does at one point concede that the credibility problem generally blocks the strategic resolution of free rider predicaments; it allows it only under very special circumstances (see Hardin 1982, p.194). Those circumstances occur when parties are able to pre-commit themselves to a strategy like tit-for-tat. Given a small group where everyone knows everyone else for example, the parties could pre-commit themselves by staking a bet, or just their honour, on retaliation when tit-for-tat requires retaliation. Such circumstances are not typical, unfortunately, of free rider situations.

Where Tuck believes that to free ride is a failure like procrastination, and Hardin thinks that it is often a strategic error, Derek Parfit argues a less radical but still surprising thesis. This is that free riding is a moral mistake, at least for someone who is impartially benevolent: someone who has an equal concern for all, including himself (see Parfit 1984, Chapter 3).

Parfit makes a number of points, the brunt of which is that many apparent examples of free rider problems that arise for utilitarians are indeed just apparent. But when he considers the really troublesome cases, he has to call on extremely controversial intuitions. These cases can be exemplified in a trivial instance: all the members of a community are interested in maintaining the cricket oval in good repair and can do so only if enough of them do not walk across it for a short-cut. The problem is, why should a utilitarian who is in a position to save himself considerable trouble by taking the short-cut, not do so when he gets the chance. His deviation will not make any perceptible difference to people's enjoyment of the ground and we may assume that it will not be noticed by others and will not cause widespread defection.

Such a utilitarian weighs, on the one hand, the trouble of not taking the short-cut and, on the other, the imperceptible harm of doing so. One way in which he might be persuaded not to go the shorter route is by being brought to realise that the harm is real, though imperceptible, and is liable to outweigh his own convenience, being suffered by many. The other is by being led to think that he ought to consider, not just the effect of his own compliance or defection, but the joint effect of all those complying or defecting: this will constitute a perceptible benefit or harm.

Parfit urges the attraction of both lines of thought but not, to my mind, persuasively. His argument for the first is that if an imperceptible harm is not a real harm then nothing is (Parfit 1984, p.78-79). The imperceptible harm will be at least as bad as a harm that differs from it by an imperceptible increase; as bad as a harm that differs from that harm in turn by such an increase; and eventually as bad as the very substantial harm to which such increases can lead. Thus if it is not a real harm, neither will that substantial injury be one. This argument turns however on the assumption that 'at least as bad as' is a transitive relation, so that if

X is at least as bad as Y, and Y as Z, then X is at least as bad as Z. In the sort of case on hand that assumption is false. An injury X can feel at least as bad as Y, and Y as Z, but X not feel as bad as Z.

Parfit's argument for the second line — that the utilitarian ought to consider more than the effects of his own action — is less explicit than that for the first. But it is no more compelling. It does not offer independent grounds for the conclusion, coming ultimately to the claim that unless utilitarians consider the joint effects of all actions like their own, they will fall prey to prisoner's dilemmas. This appears in the remark with which he concludes his discussion. 'We must accept this view if our concern for others is to yield solutions to most of the many prisoner's dilemmas that we face' (Parfit 1984, p.86).

It may yet seem that Parfit has reason on his side. If so, let me make some points in clarification of where I stand. I do not deny that the right thing for the utilitarian to praise is compliance, even if free riding is the right thing to do. I do not even deny that the right thing for utilitarians to do as a group is to try to ensure that they each comply. All I say is that if the utilitarian is in the enterprise of unilaterally promoting maximum happiness, then the only thing for him to do in cases like that described will be to take the free ride.

Free rider problems are more resistant to resolution than any of our three authors suggest. That is an important lesson for policy-making. My own view is that many of our co-operative predicaments require a political initiative and in conclusion I would like just to mention a final consideration which supports this.

In the dilemmas envisaged so far there are always just two strategies available, co-operate or defect. But many of the social predicaments which resemble the prisoner's dilemma are more complicated: each party may defect, or may choose to co-operate in pursuit of solution 1, or in pursuit of solution 2, or in pursuit of any of a number of incompatible solutions. Co-operation is not a single option; it is many. If not recognising private property is a defect strategy in the state of nature, for example, then there are as many strategies of co-operation as there are possible systems of ownership.

This complexity in co-operative predicaments probably means that the state will often be necessary to select and then reinforce one of the available resolutions. However much we relish the invisible hand, we may still require the strong arm. That lesson is as old as Hobbes, but there is no reason here for surprise. So, after all, is an appreciation of the prisoner's dilemma (see Taylor 1976, Chapter 6).[5]

NOTES

1. Notice that there is a regress in the offing. Two altruists begin with certain payoffs. Learning about the other each neglects his original payoffs for payoffs that reflect his (now exclusive) concern with the other. But before acting he checks again, only to find that the other's payoffs have shifted in the way his own did. So on to another round,. and a regress (see Schick 1971).
2. Notice the qualification about an indefinite future. If you and I set out to engage in what we know will only be n dilemmas, then we each know that the rational thing to do in the nth game is to defect, since there is no punishment in prospect. But in that case there is no punishment in prospect either for defection in game n-1. And so back by mathematical induction to game 1.
3. Here and later the only strategies envisaged, categorical or conditional, are pure. Mixed strategies, involving gambles over pure ones, are not considered.
4. Functional explanation does not have to undercut rational agency; perhaps only the functionalist credo does that. Marx offers examples of non-undercutting functional accounts on the analysis of his theory in Cohen 1978.
5. I am very grateful to Ra Foxton and Christie Slade for research assistance and to Eveline Bancroft for typing the manuscript. I benefited greatly from discussion at a series of seminars on the prisoner's dilemma in which the paper was originally presented; the seminars were held in 1984 in the Research School of Social Sciences, Australian National University.

REFERENCES

Axelrod, Robert. *The Evolution of Cooperation*, Basic Books, New York, 1984.

Cohen, G.A. *Karl Marx's Theory of History*, Oxford University Press, 1978.

Fishkin, James S. *The Limits of Obligation*, Yale University Press, New Haven, 1982.

Frassen, B. van. *The Scientific Image*, Oxford University Press, 1980.

Goodin, R.E. 'Itinerants, Iterations and Something In-between', *British Journal of Political Science*, Vol. 14, 1984, pp. 567-70.

Hardin, Russell. *Collective Action*, Johns Hopkins University Press, Baltimore, 1982.

Lewis, David. 'A Prisoner's Dilemma is a Newcombe Problem', *Philosophy and Public Affairs*, Vol. 8, 1977, pp. 235-40.

Macdonald, Graham and Philip Pettit. *Semantics and Social Science*, Routledge and Kegan Paul, London, 1981.

Moore, F.C.T. 'The Martyr's Dilemma', unpublished ms. Department of Philosophy, University of Hong Kong, 1984.

Olson, Mancur. *The Logic of Collective Action*, Harvard University Press, Cambridge, Mass., 1965.

Parfit, Derek. *Reasons and Persons*, Oxford University Press, 1984.

Pettit, Philip. *Judging Justice*, Routledge and Kegan Paul, London, 1980.

Pettit, Philip. 'Satisficing Consequentialism', *Proceedings of the Aristotelian Society*, Supplementary Volume 58, 1984a.

Pettit, Philip. 'The Philosophies of Social Science' in R.J. Anderson and W.W. Sharrock, *Teaching Papers in Sociology*, Longmans, London, 1984b.

Rapaport, A. and M. Guyer. 'A Taxonomy of 2 x 2 Games', *General Systems*, Vol. 11, 1966.

Rapaport, A. 'Prisoner's Dilemma — Recollections and Observations' in A. Rapaport, (ed.), *Game Theory as a Theory of Conflict Resolution*, D. Reidel, Dordrecht, 1974.

Schelling, Thomas C. *Micromotives and Macrobehaviour*, Norton, New York, 1978.

Schick, F. 'Beyond Utilitarianism', *Journal of Philosophy*, Vol. 68, 1971, pp. 657-66.

Schick, F. *Having Reasons*, Princeton University Press, 1984.

Sen, A. 'A Game-Theoretic Analysis of Theories of Collectivism in Allocation' in T. Majumdar, (ed.), *Growth and Choice*, Oxford University Press, 1969.

Stinchcombe, Arthur. 'Is the Prisoner's Dilemma all of Sociology', *Inquiry*, Vol. 23, 1980, pp. 187-92.

Sylvan, R. 'Maximising, Satisficing, Satisizing', Discussion Papers in Environmental Philosophy, Research School of Social Sciences, ANU, Canberra, 1983.

Taylor, Michael and Hugh Ward. 'Chickens, Whales and Lumpy Goods: Alternative Models of Public-Goods Provision', *Political Studies*, Vol. 30, 1982, pp. 350-70.

Taylor, Michael. *Anarchy and Cooperation*, John Wiley and Sons, London, 1976.

Tuck, Richard. 'Is There a Free-Rider Problem?' in Ross Harrison, (ed.), *Rational Action*, Cambridge University Press, 1979.