Free Riding and Foul Dealing

Philip Pettit

The Journal of Philosophy, Vol. 83, No. 7. (Jul., 1986), pp. 361-379.

Stable URL:
http://links.jstor.org/sici?sici=0022-362X%28198607%2983%3A7%3C361%3AFRAFD%3E2.0.CO%3B2-4

The Journal of Philosophy is currently published by Journal of Philosophy, Inc..

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXIII, NO. 7, JULY 1986

FREE RIDING AND FOUL DEALING*

IT has been recognized for some time that the free-rider problem is usually an *N*-party prisoner's dilemma. What has not been so clear, however, is whether it is the only species of that dilemma, and, if not, how it differs from other members of the genus. This paper is addressed to those taxonomic problems.

I wish to argue that there are two significantly different types of many-party prisoner's dilemma, one of which is exemplified by the paradigm free-rider problem and the other by a common predicament which I describe as the *foul dealer*. The significance of the distinction appears in a number of ways. It is a formally salient divide; it marks a difference in the possibilities of strategically resolving a many-party dilemma; and it models, more or less accurately, the distinction between the free-rider and foul-dealer problems.

The free-rider problem is the predicament, familiar from informal social theory, of productive cooperation. It arises when everyone is better off if each contributes to a certain cause—by effort, finance, restraint, or whatever—than if no one does so, but when no one's contribution is likely to make a difference sufficient to repay him for the cost involved. The problem is how to persuade people to contribute when each may argue that others will succeed in furthering the cause, or fail to further it, regardless of what he does, and that, therefore, he may as well save himself the trouble of contributing.

* I was moved to think about the distinction discussed here by a conversation with Richard Tuck; I am unclear about what is his property, what mine. I am also grateful for helpful discussions with Geoffrey Brennan; for remarks received when the paper was read to economists at the University of Adelaide and the University of Melbourne and to philosophers at the 1984 Australasian Association meeting at the University of New South Wales; and for comments sent by a number of people who read the manuscript: mainly, Bob Goodin, Julian Le Grand, Jean Hampton, Margaret Levi, Max Neutze, David Papineau, Fred Schick, Jack Smart, and Christie Slade.

0022-362X/86/8307/0361$01.90

The foul-dealer problem is also familiar from informal social theory, though not under that name: it is the predicament, in a banal phrase, of peaceful coexistence. This is the problem of how to persuade people individually to disarm themselves of some instrument of offence when, albeit they are better off under universal disarmament, still they are each exposed by self-disarmament to the worst prospect of all—that of being a defenseless victim of another's aggression; and, equally, if they refuse to disarm they are each eligible for the best possibility of all—that of being an unopposed aggressor.

Within formal social theory, where such predicaments are usually modeled as games, in particular as prisoner's dilemmas, the difference between these problems has failed to appear.[1] My paper will serve to rectify this failure, identifying a salient formal analogue for the informal distinction. But the paper should do more than recover lost ground; for, as already suggested, the formal analogue reveals something which is of independent interest and which is not obvious from the original contrast. This is that there is a significant difference in how far the two predicaments are strategically resoluble. That will be a salutary lesson for those who assume that what solves one prisoner's dilemma solves all.[2]

In summary then, I have two aims: (1) to show that the formal distinction has material significance, answering to the difference between free-rider and foul-dealer problems, and (2) to demonstrate that it also has strategic significance, so far as the two types of dilemma lend themselves to different sorts of resolution.

My paper is in four sections. I begin with the formal distinction between two mutually exclusive and jointly exhaustive types of dilemma, A and B (section I). I plot the correspondence between type A and the free-rider problem (section II) and between type B and the foul-dealer problem (section III). And then, finally, I explain the strategic significance of the distinction between dilemmas A and B (section IV).

Before passing on to the main business, it will be convenient to mention three standard empirical assumptions that I make in my

[1] This homogenization is implicit in the classic formulation of collective predicaments: Mancur Olson, *The Logic of Collective Action* (Cambridge, Mass.: Harvard, 1965). It is sustained in the more recent tradition of casting Olson's problem as a prisoner's dilemma. See, for example, Russell Hardin, "Collective Action as an Agreeable *N*-Prisoner's Dilemma," *Behavioural Science,* XVI, 5 (September 1971): 472–481, and *Collective Action* (Baltimore: Johns Hopkins UP: 1982).

[2] The assumption is implicit in Hardin when, having shown that Olson's collective-action problem is a prisoner's dilemma, he writes: "The significance of the result stated above is that any analysis prescribing a solution for Prisoner's Dilemma must prescribe a similar solution for the game of collective action" (*Collective Action,* p. 28).

presentation. I mention them without further comment, since their significance should be clear in the light of later discussion.

The first is that agents care sufficiently about the future to be attracted by the gains of cooperation in repeated prisoner's dilemmas. The second is that they prefer a sequence of cooperative payoffs in such dilemmas to an alternation between cooperating when others defect and defecting when others cooperate. And the third is that the damage done to cooperators by a lone defection is not undone by extra defections.

I. TWO TYPES OF DILEMMA, A AND B

The two-party prisoner's dilemma is characterized by two conditions: first, that defecting is the dominant strategy for each person; and, secondly, that the defect-defect outcome is Pareto-inferior to the cooperate-cooperate outcome: it is worse for some—in fact, for all—and better for none.[3]

The natural generalization of the dilemma to the many-party case prescribes a corresponding set of constraints: first, that defecting is a dominant strategy for each party; and, secondly, that the outcome of universal defection is Pareto-inferior to that of universal cooperation. Other generalizations are possible, but this is the one with which I shall work.[4]

The most important feature of the generalization belongs to other candidates as well. This is that it introduces an indeterminacy which is without counterpart in the two-party case. The definition stipulates that the outcome of universal cooperation is better for each than universal defection and that every outcome is better for defectors than for cooperators. But it says nothing about how partial-cooperation outcomes compare with one another, or indeed with the uniform outcomes.

We can formulate this point more perspicuously with the help of some symbols. Let 'C' stand for "cooperate" and 'D' for "defect." Let (C, D) designate the outcome of a (two-party) dilemma in which

[3] This definition is somewhat stronger than necessary. In place of the first condition, we might stipulate just that defect-defect is an equilibrium: that is, that no one gains—though he may not actually do worse—by unilaterally departing from it. The stronger definition is satisfied by the classic examples. Also, it is required to ensure the truth of what is often presented as an alternative definition: that for each party, defecting while the other cooperates is preferred to double cooperation, which is preferred to double defection, which is preferred to cooperating when the other defects. See Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984), pp. 9/10. See also below.

[4] This is the generalization in Michael Taylor, *Anarchy and Cooperation* (London: Wiley, 1976), ch. 3. See his fn 7, p. 62, for references to other generalizations. Taylor follows Amartya Sen, "A Game-theoretic Analysis of Theories of Collectivism in Allocation," in T. Majumdar, ed., *Growth and Choice* (New York: Oxford, 1969).

the first prisoner cooperates and the second defects. And let '>' mean "preferred by all"; '$>_i$' mean "preferred by person *i*."

With these symbols in hand, we can reformulate the two conditions defining the two-party dilemma. Dominance means that, for the first party,

$$(D_1, C_2) >_1 (C_1, C_2) \qquad \text{and} \qquad (D_1, D_2) >_1 (C_1, D_2)$$

Pareto inferiority means that

$$(C_1, C_2) > (D_1, D_2)$$

Putting these conditions together, then, we see that they define a complete ordering of outcomes for the first party, as they also do, of course, for the second. We have:

$$(D_1, C_2) >_1 (C_1, C_2) >_1 (D_1, D_2) >_1 (C_1, D_2)$$

There is no indeterminacy allowed.

No such complete ordering of outcomes is fixed by the two defining conditions for the many-party dilemma. Dominance means that, from the point of view of person *i*, an outcome in which he defects is always preferable to the corresponding outcome in which he cooperates. We can express that as follows:

$$(C_1, C_2 \cdots D_i \cdots C_{n-1}, C_n) >_i (C_1, C_2 \cdots C_i \cdots C_{n-1}, C_n)$$

$$(D_1, D_2 \cdots D_i \cdots D_{n-1}, D_n) >_i (D_1, D_2 \cdots C_i \cdots D_{n-1}, D_n)$$

$$(D_1, C_2 \cdots D_i \cdots C_{n-1}, C_n) >_i (D_1, C_2 \cdots C_i \cdots C_{n-1}, C_n)$$

and so on

The Pareto-inferiority condition, on the other hand, means that the outcome in which all cooperate is better from the point of view of everyone than that in which all defect. That is,

$$(C_1, C_2 \cdots C_i \cdots C_{n-1}, C_n) > (D_1, D_2 \cdots D_i \cdots D_{n-1}, D_n)$$

Together these conditions define a complete ordering of the four outcomes that correspond to the outcomes ordered in the two-party case:

$$\begin{aligned} &(C_1, C_2 \cdots D_i \cdots C_{n-1}, C_n) \\ >_i\ &(C_1, C_2 \cdots C_i \cdots C_{n-1}, C_n) \\ >_i\ &(D_1, D_2 \cdots D_i \cdots D_{n-1}, D_n) \\ >_i\ &(D_1, D_2 \cdots C_i \cdots D_{n-1}, D_n) \end{aligned}$$

But the conditions leave an important indeterminacy, because they give us no information on how person *i* should rank outcomes in which two or more cooperate but one or more defect. All the conditions tell us is that, when two such outcomes differ only in how person i chooses, i should prefer that in which he defects. Thus they leave open the question of where to place an outcome like $(D_1, C_2 \cdots C_i \cdots C_{n-1}, C_n)$ in i's ranking. We do not know whether it comes above universal cooperation, below universal defection, or in between the two.

The indeterminacy allowed by these conditions enables us to introduce any number of distinctions between different types of many-party dilemma. I wish to avail myself of this opportunity and to press the case for a distinction between what I shall call *type A dilemmas* and *type B dilemmas.* I do not want to prejudge against other distinctions, but, for the record, I believe that none will prove to be as important as the A-B divide.

If a many-party dilemma is one in which no cooperator is made worse off by a lone defector than he would be under universal defection, then we have a type A dilemma. If it is one in which the lone defector plunges some cooperator or cooperators below that baseline, then it is a type B dilemma. Take the outcome described above, where only the first party defects: $(D_1, C_2 \cdots C_i \cdots C_{n-1}, C_n)$. We have a type A dilemma if none of the cooperators is worse off in this outcome than he is under universal defection: that is, under $(D_1, D_1 \cdots D_i \cdots D_{n-1}, D_n)$. Otherwise we have a type B dilemma.

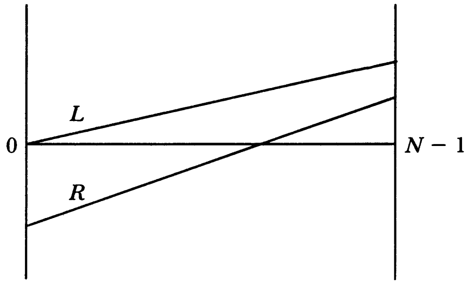
The A-B distinction is formally unambiguous, and what I want to show is that it has material and strategic significance. The next two sections deal with the question of material significance, the final section with that of strategic.

Before leaving this section however, I would just like to mention that the A-B distinction is hardly ever noted in the literature on prisoner's dilemmas and that one influential writer, Thomas C. Schelling, actually elides the possibility of a type B dilemma.

Schelling defines a many-party dilemma in a manner which preserves dominance but broadens the Pareto-inferiority condition.[5] Not only is universal defection Pareto-inferior to universal cooperation. It is inferior to the cooperation of some number K, where K may be less than the total population, N. So far I have no objection.

[5] See "Hockey Helmets, Daylight Saving and Other Binary Choices," reprinted in *Micromotives and Macrobehaviour* (New York: Norton, 1978). Notice that for Schelling the returns to all cooperators are the same, so the lone defector in a type B dilemma would have to plunge all cooperators below the baseline of universal defection. Perhaps it is the unlikelihood of such an event which leads him to discount the type B case.

But then, in the course of introducing an influential mode of representing payoffs in the prisoner's dilemma, Schelling goes on to assume that K will never be greater than $N - 1$: see the figure below. Schelling's diagram has a horizontal axis measured from 0 to $N - 1$. A curve labeled R represents, on the vertical axis, the return to cooperators for each number cooperating between 0 and $N - 1$. Another curve labeled L depicts, for each such number of cooperators, the return to the defectors at that level of cooperation. Schelling assumes that the R curve will cross the baseline of universal defection somewhere between 0 and $N - 1$. More importantly, since less remediably, his mode of representation enforces the assumption that the Nth cooperator cannot be the one who causes the curve to cross the baseline. This assumption defines the type B dilemma out of existence.



Schelling's Diagram

II. TYPE A AND THE FREE-RIDER PROBLEM

There is not much challenge in the task of showing that the type A dilemma is exemplified in social life; after all, much contemporary analysis of collective action, as in the Schelling case, is devoted to revealing the presence of the A structure. What I hope to do in this section, however, goes some way beyond this goal. I will set out a definition of what I call *the paradigm free-rider problem,* and I will show, first, that every such problem, so far as it is a prisoner's dilemma, is a type A dilemma; and, secondly, that under normal circumstances every type A dilemma constitutes a paradigm free-rider problem. For practical purposes this will establish a coincidence between the type A dilemma and the free-rider problem.

Before presenting my definition of the problem, it is worth mentioning three ways in which instances may vary: this will keep us alive to the range of the predicament. One concerns the nature of contributions, which may vary, not just in requiring effort or restraint or finance, but also in more formally significant ways. Each party may be presented with the same contributory demand or a different one. And in either case the demand may be to select a single contributory option or to select any of a number of such options: it may be to pay rather than not, or it may be to pay this or that or the other amount rather than not. I assume in my analysis that problems can always be broken down so that each person faces just a binary choice, and moreover a choice between the same alternatives.[6]

The second way in which free-rider problems may vary concerns the costs that contributors have to bear. These may be the same from party to party, or they may differ: this, because their individual contributions are more onerous for some parties than for others or because the good they collectively seek does not have the same attraction for each. In the analysis that follows I leave open the question of whether costs are the same or different; the account is designed to apply to both cases.

The third way in which free-rider problems may vary is in the nature of the goods on offer. Goods may be lumpy or smooth. The *lumpy* good, whether it lumps at one level or many, has this characteristic: that an individual's contribution need not benefit anyone in any degree; it will do so if and only if it comes at the threshold where a lump is achieved. *Smooth* goods differ in regard to that feature. They are such that everyone's contribution makes some beneficial difference, however small. In what follows I make no presumption as to whether goods are lumpy or smooth; again, I believe that my analysis applies in both cases.

The *paradigm free-rider problem,* as I shall call it, arises under four conditions. They are:

1. There is a nonexcludable good attainable for a group: that is, a good enjoyed by all, though perhaps with different intensities, if enjoyed by any.[7]

[6] What if each faces a formally similar set of alternatives, such as that of paying or refusing a certain percentage of income? This is excluded by my assumption. It would mean that K, in the definition that follows, may have different values for different contributors.

[7] Although nonexcludable, the good need not be joint of supply and, therefore, need not strictly be a public good. On the definition of the latter, see Taylor, *Anarchy and Cooperation,* pp. 14/15.

2. It can be attained, and rationally attained—whether at one of many levels, however smoothly related, or at the only level possible—by K members of the group, where K is less than all: the reward to each contributor in a subgroup K will exceed the cost of his contribution, though not necessarily by the same margin.
3. It cannot be attained, or at least not rationally attained, by just one member of the group: the reward to the lone contributor will not cover the cost he has to bear.
4. The fear of contributing when the good is not produced, and the hope of not contributing when it is, make it rational for each person not to contribute to the production of the good: specifically, it means that the strategy of not contributing maximizes expected utility.

We get from the paradigm free-rider problem to various peripheral versions by relaxing any one or any combination of the last three clauses. The last clause may be relaxed, for example, so that what is required is just that not contributing is the maximin or maximax strategy.[8] The second and third clauses may be relaxed so that K may be 1 or N, where N is the total number in the group.

My definition of the paradigm problem is not arbitrary: I believe that it identifies the main sort of collective-action predicament with which recent social theorists have been concerned. The three variable clauses all have the backing of contemporary authority.[9] And not without reason, since they serve to define a homogeneous and common class of social predicaments. In any case, what I shall mean, henceforth, when I speak of the free-rider problem is the paradigm just defined.

Examples of the free-rider paradigm abound. Consider the situation of a team in tug of war, of a community that 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

[8] Such a conception of the free-rider problem appears in Michael Taylor and Hugh Ward, "Chickens, Whales, and Lumpy Goods: Alternative Models of Public-goods Provision," *Political Studies,* XXX (1982), and in Jean Hampton, "Free-rider Problems in the Production of Collective Goods," unpublished typescript.

[9] The second clause would naturally be supported by Schelling, for reasons already noted. The third has the support of Olson, since he thinks that if K is equal to 1 then "there is a presumption that the collective good will be provided" (*op. cit.,* 34). The last clause is supported by the reading of Olson implicit in Brian Barry, *Sociologists, Economists and Democracy* (London: Collier-Macmillan, 1970), p. 24.

of the labors of others, evading the burden of effort or restraint or finance which they have to bear. He will do this, despite the risk that there will not be enough contributors, because that is the way for him to maximize expected utility.

What I want to show is that any free-rider problem that constitutes a prisoner's dilemma will be a dilemma of type A and that any type A dilemma that meets certain plausible assumptions will be a free-rider problem. The first part of this claim is stronger than it sounds, for it is normally and reasonably supposed that most free-rider problems constitute prisoner's dilemmas. I will begin by defending that assumption.

The free-rider problem will be a prisoner's dilemma if and only if universal defection is Pareto-inferior to universal cooperation and defection is a dominant strategy. The Pareto-inferiority condition is clearly fulfilled in the free-rider predicament. But what of the dominance condition? Does the fact that defection has a higher expected value than cooperation, as clause 4 stipulates, mean that it dominates cooperation? And this, not just for the case covered by clause 3, where no one else cooperates, but in the different cases defined by different numbers of cooperators?

Not strictly, since there is one situation under which anyone will prefer to cooperate. This is when K is a determinate number; when he knows that $K - 1$, and only $K - 1$, others have contributed or are going to contribute; and when the good is such that no one receives any benefit until K contributors have done their bit: that is to say, the good is lumpy at level K.[10] But this situation need not inhibit the assimilation of the free-rider problem to the prisoner's dilemma. The reason is that, even with lumpy goods, the case envisaged is usually so improbable that it can be discounted or else is downright impossible.

It will be improbable under a variety of circumstances: for example, if K is a large number, or if it is a small number relative to the total population. It will be impossible if K is not so much a number as an indeterminate range: in that case 'K' will be not a numeral but an expression like 'a lot' or 'a good many'. The first sort of case is familiar. The second arises when the collectively produced good is characterized by a vague predicate which, as Richard Tuck has observed, generates something like the paradox of the heap.[11] Just as a

[10] On lumpy goods, see Hardin, *Collective Action,* ch. 4. Notice that this sort of case gives rise to game-theoretic problems other than prisoner's dilemmas. See Taylor and Ward, *op. cit.,* and Hampton, *op. cit.*

[11] See Richard Tuck, "Is There a Free-rider Problem?" in Ross Harrison, ed.,

pebble never makes the difference between a heap and a nonheap of pebbles, so one contribution in such a case cannot make the difference between the realization and nonrealization of the good: say, between a littered park and an unlittered park.

Assuming that usually collective goods are nonlumpy or that K is of a kind to generate one or other of the cases just described, I shall take it that the dominance condition as well as the Pareto-inferiority condition is normally satisfied by the free-rider problem. That the conditions are satisfied means that the problem is an instance of the prisoner's dilemma, and the question then is whether my specific claims hold. Is every free-rider problem a type A dilemma in particular? And is every type A dilemma a free-rider problem?

It is demonstrable that the problem is a type A dilemma, because the lone free rider has exactly the effect of the lone defector in that sort of dilemma. As he is alone in his defection, $N - 1$ others must have contributed to the good. But $N - 1$ is the upper bound on K; so K or more must then have contributed. That means that the collective good must have been achieved and that everyone is better off than under universal defection. Thus the lone free rider does not plunge anyone below the baseline of universal defection. The free-rider problem is a type A prisoner's dilemma.

The reverse claim is substantiated by showing that, in any type A dilemma, the clauses defining the free-rider problem are more or less certain to be fulfilled. I shall show this for each clause in turn.

The $N - 1$ cooperators in a type A dilemma must have created a good that they enjoy regardless of the lone defection, since they are better off than if they had all defected too. More strongly, this good must be available to all, including the defector, since otherwise his deprivation would presumably motivate him to cooperate too. Thus the first clause in the definition of the free-rider problem is sure to be satisfied.

The second and third clauses put limits on K: that is, on the minimum number of cooperators who can rationally provide the good. K must not be more than $N - 1$, where N is the total population, but it must be more than 1. These limits are bound to obtain in a type A dilemma. There certainly are no more than $N - 1$ cooperators required to produce the good in such a dilemma, since the lone defector does not make cooperators worse off than under universal

Rational Action (New York: Cambridge, 1979). See also James S. Fishkin, *The Limits of Obligation* (New Haven, Conn.: Yale, 1982), ch. 7, and Derek Parfit, *Reasons and Persons* (New York: Oxford, 1984), sec. 29.

defection. And there must be more than one required, for if a single cooperator does better than he would do under universal defection, then defecting does not dominate cooperation.

The last clause is also almost certain to be fulfilled. In a type A dilemma defecting dominates cooperating, and under normal circumstances that means that the expected value of defecting will be higher than that of cooperating. Normal circumstances obtain so long as no one can think that if he cooperates it is more probable that others cooperate, and sufficiently more probable to raise the expected value of cooperating above that of defecting. No one can think this on the grounds of a causal connection with others, since agents in a one-shot dilemma act independently. Nor can anyone plausibly argue such a case on the grounds that others are similarly constituted.[12] At least in real-world type A dilemmas, the fact that defecting dominates cooperating means that defecting maximizes expected utility.

I conclude that the free-rider problem and the type A dilemma can be taken as effectively one and the same predicament. The paradigm free-rider problem is usually a type A dilemma, and every plausible type A dilemma is a paradigm free-rider problem.

Since the free-rider problem is the standard illustration of the many-party prisoner's dilemma, the question remaining is whether the type B dilemma has any substantive significance. Can we find interesting examples, or is it just a formal possibility?

III. TYPE B AND THE FOUL-DEALER PROBLEM

The type B dilemma contrasts with the other variety in regard to the significance of the lone defector. Here the lone defector makes others—or at least one other—worse off than under universal defection; there he does not make others worse off at all. Here he is seriously injurious, there he is merely an irritant.

What sort of examples are likely to illustrate the type B dilemma? It must be that by cooperating all can make themselves better off than if all defect. And it must be that the extra return to the lone defector causes at least one other to do worse than under universal defection.

[12] David Lewis argues the case for some two-party dilemmas which are distinguished by relatively high cooperative rewards: see his "Prisoner's Dilemma Is a Newcomb Problem," *Philosophy and Public Affairs,* VIII, 3 (Spring 1979): 235–240. But the corresponding case in the many-party dilemma will be rather harder to establish: there will have to be a higher probability that if one cooperates, others do so. On related matters, see *ibid.*, pp. 238/9. In any event, I have doubts about the argument which are related to the considerations raised in my article "Preserving the Prisoner's Dilemma," *Synthese,* forthcoming.

That benefit therefore must be enjoyed, not just by the efforts of others, as in the free-rider case; it must be enjoyed directly at some other's expense.

One situation that would fit these constraints is a limit case, so to speak, of the free-rider problem. It is an instance of that peripheral version of the problem which arises when the second clause in our original definition is removed: that is, when K can be N. Suppose that the collective good on offer is smooth, so that each individual's contribution benefits each member of the group to some degree. And suppose that all members are compensated for their individual efforts only when all have contributed. In such a case the lone defector will hold at least some cooperators below the baseline of universal defection, since it is only at the point where he contributes too that each makes a net gain.

On its own, this sort of case would hardly establish the material significance of the type B dilemma, for it is not clearly more than a theoretical possibility. Any genuinely smooth good that pays dividends for some only at the point of universal cooperation is not going to be much of a good for those parties. It is difficult to envisage examples, and it is not certain that there are any.

If we are to show that the type B dilemma is materially significant, then we must look beyond the sorts of cases that generate free-rider problems. A feature that characterizes all free-rider pr.oblems, paradigm and peripheral, is that the good on offer is one-dimensional, holding out the same kind of reward for each: viz., a share in the common benefit. So long as we remain with such problems, the only possibility of finding an example of the type B dilemma will be in the sort of peripheral case just considered. And there, as we have seen, the prospect is not promising.

In order to stand a chance of establishing the material significance of the type B dilemma, we must be able to identify collective goods that have a two-dimensional aspect. They must be such that, from the perspective of the lone defector, there is more to be enjoyed by defecting than just the share of the good procured by the efforts of others. As well as—or instead of—providing such a share of a common good, the efforts of others must make another particularly tempting sort of good available to him. They must present the lone defector with the opportunity to exploit one or more of the cooperators, letting him enjoy some predatory advantage.

What we have to look for then, if we are to find examples of the type B dilemma, is a predicament with the following features. A number of parties are in a position to make themselves better off than they are, by universally adopting a certain cooperative strategy. However, the very fact of cooperating creates the opportunity for a defector, or at least for a lone defector, to take advantage of cooperators, achieving a benefit which is superior to a share in the reward of universal cooperation. That reward constitutes only one dimension of the collective good on offer; the benefit available to the lone defector constitutes a second.

These remarks immediately suggest one sort of example. This is the state of nature, in more or less its Hobbesian form. Let everyone defect and we have the war of all against all. Let everyone cooperate, say by beating their swords into ploughshares, and we have the rather more desirable state of universal peace. But let just one party defect and his victims—all others, potentially—will be worse off even than in the war of all against all. He, on the other hand, will have the supreme prize: unopposed superiority.

In this example, the collective good on offer is the cessation of a pursuit in which one party's gain is another's loss. In a Hobbesian phrase, it is the eschewal of eminence, where eminence is something enjoyed at the direct expense of others. The lone defector's payoff is the attainment of such eminence in the absence of opposition from those who cooperate. It is not the costless enjoyment of the peace procured by others, as in a free-rider dilemma. It is the enjoyment of war under conditions where others, in their desire for peace, have left themselves undefended and vulnerable.

This is the predicament that I call the *foul-dealer problem,* since the lone defector does such mischief. Examples of the foul-dealer problem are as various as the different sorts of eminence that attract individual and institutional agents. The eminence pursued may consist in a certain sort of control, physical, psychological, financial, or whatever. Alternatively, it may be constituted by the achievement of some damaging advantage. The advantage may be immediate, as in queue jumping or price cutting, or it may come from repute in the eyes of others. This repute will involve the attainment of a higher standing on some comparative metric, whether a metric with absolute significance, like beauty or strength or virtue, or one of an intrinsically relative kind such as status.[13]

We find foul-dealer problems, then, in a great range of contexts. Among individuals and gangs and governments and businesses, since

[13] That price-cutting is a form of foul dealing, as it certainly is under certain circumstances, means that in those circumstances we may expect cartels to form and hold: this follows from section V below.

all of those are regularly bent on mutual control. Among the members of a panicking crowd, or among the parties in any situation where each is tempted to seize an opportunity in a way that damages others. And among prima donnas and politicians, advertisers and advocates, and any agents to whom success consists in making a superior impression.

It is a great oversight in the recent literature on collective action, that the foul-dealer problem has dropped out of view. It is a common sort of predicament; it is distinctive in exemplifying the type B dilemma; and it stands in stark contrast to every version of the free-rider problem. Where recent theory seems to assume that collective goods are one-dimensional, the sort of good involved here has a double aspect. At the point where the last party has to choose—and almost certainly before that—it shows a second face: not now the aspect of a good to be shared with others; rather that of an advantage that others cede.

The free rider aspires to a share in a one-dimensional collective good which the cooperation of others has created. The foul dealer aspires to enjoyment of an advantage over others which their cooperation in seeking a two-dimensional collective good allows him to take. The free rider seeks to benefit by the efforts of others, the foul dealer to benefit at their expense.

In the last section we began from an independent characterization of the free-rider problem, and we argued that, so characterized, the problem was more or less coextensive with the type A dilemma. Our approach in this section has been different, since the foul-dealer problem was introduced and characterized only in the process of identifying the sort of predicament required to exemplify the type B dilemma. The upshot, however, is very similar. It appears that, for practical purposes, the foul-dealer problem and the type B dilemma may be taken to be coextensive.

The importance of the two sections taken together is that they show that the A-B distinction has material significance. Each type of dilemma is exemplified, and is exemplified in a familiar sort of predicament. With the material significance of the distinction established, the next task is to show that it is strategically significant too. To that task I now turn.

IV. THE STRATEGIC SIGNIFICANCE OF THE DISTINCTION

Suppose that two people find themselves in recurrent prisoner's dilemmas. The iteration involved makes it possible for each to let his choice of whether to cooperate or defect depend on how the other chooses. That is to say, it makes it possible for each to choose strategically. In the supergame constituted by the iterated first-order predicaments, for example, they might each select the strategy of tit for tat. Under this strategy, each begins by cooperating and thenceforth does exactly as his partner did in the previous round.[14]

It is clear that in an indefinite iteration of two-party dilemmas, the participants would do better by each tit-for-tatting than by each permanently defecting.[15] Joint tit-for-tatting would give the cooperative result in each encounter and would yield a series of outcomes Pareto-superior to those of permanent defection.

Ought we then to expect joint tit-for-tatting among rational participants? Not if each assumes the worst of the other and expects him to defect permanently, since tit for tat will do worse than permanent defection against permanent defection: it will cause a loss in the first round. But suppose that each assumes that the other is rational. Ought we then to expect joint tit-for-tatting?

It would seem so. The outcome of joint tit-for-tatting is an equilibrium, and indeed a coordination equilibrium.[16] It is an outcome such that neither can do better by unilaterally departing from it and moreover—this is what makes it a coordination equilibrium—an outcome such that neither can do better by the other's unilaterally departing from it either. To depart unilaterally would be to invite the punishment of the other's defecting, and it would be to have to suffer the outcome of cooperating when the other punitively defects before getting back to joint cooperation. To have the other unilaterally defect would be to incur the obligation to defect in retaliation. Either way one is worse off than under continuing tit for tat.

Because it is Pareto-superior to the outcome of permanent defection and in particular because it is an equilibrium, indeed a coordination equilibrium, we may expect rational parties to be able to converge on the tit-for-tat solution. Or, if not on that solution, on some similar one. All that is necessary is that one can make it credible to the other, that he intends to tit-for-tat. The joint tit-for-tat outcome

[14] On tit for tat in the two-party case, see Hardin, *Collective Action,* pp. 165/6 and Axelrod, *The Evolution of Cooperation,* passim. Taylor, *Anarchy and Cooperation,* pp. 31–43, considers related strategies.

[15] Note the qualification that the iteration goes on indefinitely. If you and I set out to engage in what we know will 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 game $n - 1$. And so back by mathematical induction to game 1. Or so it seems, at least. For a different reading, see Frederic Schick, *Having Reasons* (Princeton, N.J.: University Press, 1984), pp. 71–73.

[16] The first result is stressed in Taylor, the second in Hardin.

is an equilibrium, and so, if either believes that the other tit-for-tats, then he knows that he had better do so too.[17]

In an iterated two-party dilemma, tit for tat ought to be a credible strategy. Either party can test that it is indeed the strategy of the other: he need only vary his behavior and watch for the response. And each party can see that it is rational for the other to tit-for-tat: the other's most realistic hope of reward is to match cooperation with cooperation and defection with defection.

So much for the solution of the iterated two-party prisoner's dilemma. The question to which we must now turn is whether we can equally expect the participants in an indefinite iteration of many-party dilemmas to reach a strategic resolution of their predicament: whether in particular we can expect them to converge on a strategy like tit for tat. I shall argue that our expectations ought to be different in the type A and type B cases. This difference constitutes the strategic significance of the distinction.

Given the iteration of a many-party dilemma, whatever the species, there will be conditional strategies of choice available such that universal compliance with them would be better for everyone than permanent defection all round. An example is generalized tit for tat, under which each begins by cooperating; cooperates in the second round if everyone else cooperated in the first, and defects otherwise; cooperates in the third round if everyone else other than those punishing earlier defection cooperated in the second, and defects otherwise; and so on. If everyone tit-for-tats in this sense, then each will get the cooperative result in every round and will be better off than under universal permanent defection.[18]

Ought we to expect participants in many-party dilemmas to reach the solution represented, for example, by universal tit for tat? Perhaps. It is true here, as in the two-party case, that the outcome of universal compliance with that strategy is an equilibrium, and indeed a coordination equilibrium. Each would do worse by unilateral departure, since he would then be punished next time around. And each would do worse by anyone else's unilateral departure, since he would have to punish next time around.

[17] Because it is also a coordination equilibrium, tit for tat will be the thing for the other to try to get him to do as well. The other is worse off, as he is worse off himself, through the first party's failure—if he does fail—to cooperate in tit-for-tatting.

[18] On this tit-for-tat solution, see Hardin, *Collective Action,* chs. 10–12. Taylor, *Anarchy and Cooperation,* pp. 43–61, considers related strategies and again stresses the significance of Pareto-superior equilibria rather than coordination equilibria.

The question remaining to be answered is whether tit for tat is a credible strategy in the many-party prisoner's dilemma. If it is, then, as in the two-party case, universal tit for tat is a potential solution of the predicament.

Here is where the distinction between the two types of dilemma is important. I shall argue that there is a great difference in how far participants in the two predicaments can make it credible that they intend to tit-for-tat.

In order credibly to avow tit for tat, each participant will have to be able credibly to cooperate initially, to offer cooperation for cooperation, and to threaten defection for defection. Everything turns on whether the threat is credible, for if it is then initial cooperation and the offer of reciprocal cooperation will be credible too. If he can believe the threat of others to defect for every defection, then each will be expected to cooperate initially and to cooperate in response to cooperation. But is the threat credible? That is the crucial question.

There are two sorts of reason why a threat might be believed: one general, the other special. The general reason is that, under the circumstances specified in the threat, the threatened course of action would be rational, even if the threat had never been made. The special reason is that the threat is implicitly or explicitly tied to a precommitment in virtue of which the threatened course óf action would be rational under the circumstances given. The agent making the threat might have made a substantial bet that he would keep his word, or whatever.

Consider now the two sorts of dilemma. I wish to urge that, whereas the general sort of reason makes the threat credible in the type B case, only the special reason can obtain in that of type A.

In the type B predicament, the lone defector causes some cooperator or cooperators to drop below the baseline of universal defection. No cooperator can be sure of avoiding that fate, and so it will be rational of maximinning agents to respond to such a lone defector by all defecting next time around. If they each act so as to ensure against the worst possible outcome—that is, against the possibility of being one of the lone defector's victims—then they will each defect on the next play. But maximinning is a recognizably rational disposition, at least under many circumstances. It follows, therefore, that in the type B dilemma the course of action threatened in tit for tat is often a rational response to the circumstances specified. There is a general reason, then, why the threat involved in tit for tat should often be credible.

That an iterated type B dilemma is strategically resoluble should not be surprising. The two-party prisoner's dilemma is an instance of the species, after all, and we have seen that it can be resolved under iteration. In the two-party dilemma the lone defector makes one cooperator—the only cooperator there is—worse off than under universal defection. In both the two-party and many-party cases, the damage inflicted by the lone defector is substantial, and it is for that reason that the threat of retaliation is credible.

Let us turn now to the type A dilemma. Here the lone defector is not so much injurious as irritating. He puts cooperators below the scoreline of universal cooperation—and then perhaps not detectably—but he makes no one worse off than under the baseline of universal defection. Apart from the case of someone who is left on that baseline, it would seem irrational of cooperators to carry out the tit-for-tat threat in response to a lone defector, defecting next time around. To do so would be to sacrifice the collective good on that occasion and perhaps to put its future realization in jeopardy—and this, for what will normally be just a miniscule gain. But if the threatened course of action would be irrational, then there is no general reason why the threat should be found credible. And neither in that case is there any general reason why universal tit for tat should be thought attainable.

The only hope of marshaling a tit-for-tat strategy in the resolution of a type A dilemma lies in constructing special reasons why each party should carry out the threatened response to the free rider. This might be done by means of some recognized ceremony or on the basis of a public recognition of how important it is to each that he be seen to carry out his threats. In either case the recourse to special reasons would be effective only for certain specific sorts of groups. The explicit or implicit precommitment required of each party will be visible and impressive only, for example, in relatively small groups where individuals are known to one another or where the culture is so homogeneous that they are mutually predictable.[19]

I conclude that although there is a general possibility of resolving iterated type B dilemmas by recourse to tit for tat, the possibility of resolving type A dilemmas in this way is extremely restricted. This conclusion holds indeed, not just for tit for tat, but for any condi-

[19] This squares with Hardin's statement of his claims in *Collective action,* ch. 12, though in earlier chapters the claims sound more significant. What is particularly misleading in Hardin is his suggestion, p. 28, that a solution to one dilemma will be a solution for all.

tional strategy in which each party makes his cooperation dependent on the cooperation of all others. The same problem of credibility will recur with any such strategy.[20]

The upshot then is clear. The lone defector in the type B dilemma may trespass more grievously on cooperators than his counterpart in the other predicament, but, for that very reason, he is more easily resisted. The B-defector can be coerced into cooperation, because all parties can credibly avow a strategy like tit for tat. The A-defector is proof against such persuasions. He can rely on the rationality of those who avow tit for tat to ensure that, however resentful they may feel, they will not retaliate against his lone defection. This, at least, unless they can persuade him that they are subject to some retaliatory precommitment. In the one dilemma defection is strategically anomalous; in the other it is strategically compelling.

PHILIP PETTIT

Research School of Social Sciences
Australian National University, Canberra

[20] But what of the possibility of resolving the free-rider problem by means of a strategy conditional on how only a certain number of others behave? Might a group resolve the problem, for example, through a sufficient number cooperating in the first round, and then cooperating only if at least $K - 1$ others cooperated in the previous round? Taylor considers such strategies in *Anarchy and Cooperation.*

In free-rider problems with lumpy goods, the threat involved in such a strategy would be generally credible. If fewer than $K - 1$ others have just cooperated and K is the crucial threshold, then it will be rational for a party to defect. The problem with the strategy, however, is that, even if each believes that others are following it, this gives him no reason not to free-ride himself. He can hope to be one of the $N - K$ parties who get the benefit without shouldering the burden. And, of course, if everyone can entertain this hope, then there is no prospect for a strategic resolution of the predicament. On matters related to this problem, see Michael Taylor and Hugh Ward, *op. cit.*