II

PD NORMS

1. *The Prisoners' Dilemma*

1.1 *Presentation*

THE well-known story of the prisoners' dilemma is the following:

Two guilty prisoners, against whom there is not enough incriminating evidence, are interrogated separately. Each faces two alternative ways of acting: to confess the crime, or to keep silent. They both know that if neither confesses, they will be convicted of some minor offence, concerning which there is sufficient evidence against them, and will be sentenced to a year in prison. If both confess, each will be sentenced to five years in prison. However, if only one confesses, he thereby turns king's evidence and is thus set free, whereas the other receives a heavy term of ten years.

Matrix 2.1 depicts the situation in terms of years in prison awaiting the prisoners in each of the possible combinations of their actions.

		B: Not confess	B: Confess
A	Not confess	1 / 1	10 / 0
	Confess	0 / 10	5 / 5

2.1

Matrix 2.2 depicts the situation in terms of the desirability of each of the possible combinations of their actions for each prisoner. That is to say, the numbers, or 'pay-offs', indicate

roughly the participants' relative strength of preferences over the possible outcomes, on some ordinal scale. (In both matrices prisoner *A* is Row-Chooser and prisoner *B* is Column-Chooser, their gain-and-loss measures being indicated in the upper-left and bottom-right corners of each cell, respectively.)

		B: Not confess	B: Confess
A	Not confess	1 / 1	−2 / 2
	Confess	2 / −2	−1 / −1

2.2

Two points are to be noted about matrix 2.2. First, the action 'confess' *dominates* the action 'not confess', for both *A* and *B*. That is to say, if *A* confesses, his pay-off is higher than it would have been had he decided not to confess, regardless of *B*'s choice of action. And the same goes for *B*. In terms of the matrix this finds its expression in the fact that, for Row-Chooser, each entry in the second row is higher than the corresponding one in the first row ($2 > 1$ and $-1 > -2$), and similarly for Column-Chooser with respect to the dominance of the second column over the first.

Secondly, the state of affairs obtained when both choose their dominant actions (i.e. the state represented by the bottom-right cell of the matrix) is an *equilibrium* in the sense that each stands to lose were he alone to deviate from it (as the pay-off for unilateral non-confession is -2, which is worse than -1). This implies that this state of affairs is *stable*. (It is, moreover, stable in the strong sense, since each stands actually to lose by unilateral deviation from it, and not just not to gain from it.)

Note, finally, that any matrix of the form shown in 2.3, where $T > R > P > S$, represents a situation embodying a dilemma the same as the prisoners'. (The letters stand, almost traditionally by now, for the following: 'R' for 'reward', 'T' for 'temptation', 'S' for 'sucker', 'P' for 'punishment', 'C' for 'co-operation', and 'D' for 'defection'.)

	C	D
C	R / R	S / T
D	T / S	P / P

2.3

1.2 *The Dilemma*

Let us examine the prisoners' considerations as to which course of action they are to opt for. The lines of reasoning to be presented are of course well known, indeed they almost belong to folklore by now. I consider it worth while, however, to elaborate them here since this is the basis for most of what follows.

Suppose, first, that there is no way for them to discuss their situation, and in particular no possibility of coming to an

explicit) agreement between themselves as to how they are to act. Each prisoner will realize that the worst he is to expect from choosing to confess is a term of five years, whereas from choosing not to confess it is ten years. So it is reasonable for each to regard confession as a *safer* course of action than non-confession. If, as a result, they both opt for confession, their choice is said to reflect the consideration of *maximizing security level*.[1]

There is, moreover, a stronger argument for preferring confession to non-confession. Each prisoner will realize that whatever the other does, he himself stands to gain more (or actually to lose less) by confessing than by not confessing. The consideration is based on what is known as the *dominance principle*: the choice of confession dominates (as was shown on p. 19 above) the choice of non-confession, regardless of what the other does.

(I have said that this is a stronger argument for preferring confession than was the former. The reason is that any combination of the participants' chosen actions which is an intersection of their dominant strategies is, in particular, also an intersection of their maximum security level strategies.)

The state of affairs arrived at on the basis of these forceful and convincing considerations, viz. double confession, is, however, jointly undesirable. It implies a sentence of five years to each. And it must be felt to be all the more frustrating once the prisoners realize that there is in fact an alternative outcome which is mutually desirable: they could have ended up with a relatively light term of only a year in prison had they co-ordinated their choices of actions appropriately.

But prisoner *A* will not decide not to confess unless he is absolutely sure that *B* will not confess, for if *B* confesses while *A* keeps silent, *B* goes off free and *A* is condemned to the heaviest sentence. The same goes for *B* with respect to *A*. But of course there is no way of being sure what the other will do if there is no possibility of communication before taking their decisions. And the risk involved in assuming, or hoping, that the other will not

[1] This, basically, is the idea behind the famous game-theoretical principle of mini-max, which instructs a player to examine all the (pure) courses of action available to him, find out what is the maximum loss he may expect to suffer in each, and finally choose that action which minimizes that loss.

Strictly speaking, however, the mini-max principle applies to games of *pure* conflict only (i.e. only to zero-sum games, in which one person's gain is the other's loss)—a class of which the Prisoners' Dilemma is not a member.

confess is, under the circumstances, too great for either of them to take. So once again they are locked in on the jointly destructive course of double confession.

Let us now remove, therefore, the restriction on communication, and suppose that they are allowed to discuss their situation. Would this solve their dilemma?

Clearly, they will come to a quick agreement to co-ordinate their choice of actions so as to achieve the mutually desirable outcome. That is to say, they will agree not to confess. But this is not the end of the story. Having come to this agreement and having persuaded each other that it guarantees the best they may hope for under the circumstances, each will have good reasons to walk out on it. Reasonably expecting *B* to keep the agreement, *A* will find that he stands to gain a lot—his freedom—by treacherously choosing to confess. While hesitating whether to resist this temptation he may further reflect that this consideration is likely to occur to *B* too, which means that perhaps *B* may after all *not* be expected to keep the agreement. And this, obviously, is all the more reason for *A* not to keep his own part of the agreement. So either way, whatever he expects *B* to do, *A* will decide to confess. The same, of course, applies to *B* with respect to *A*. So once again the available-yet-unattainable attractive outcome produced by joint non-confession eludes them. Their having come to an agreement, provided it is not enforceable, does not prevent their being stuck with the harsh consequences of having both confessed.

The dilemma of the prisoners is thus clear. The most rational choice for each leads to a state of affairs which is jointly destructive and at the same time stable. The jointly beneficial outcome, on the other hand, although available in principle, is highly unstable and hence is all but unattainable in practice.

2. *The Main Proposition—PD Norms*

What the prisoners need, quite clearly, is to find—or to develop—some means of stabilizing the jointly desirable state of affairs brought about by their both keeping silent. To be sure, the prisoners' problem is not in principle unsolvable; there might be instances of the dilemma where such means *are* present. For instance, the existence of trust, solidarity, or friendship between them might directly resolve their dilemma (on this issue see

section 8 below). So also would an agreement between them, once it is enforceable; that is, provided that its breach is certain to be punished and, furthermore, that the threat of this punishment outweighs the temptation of breach.

However, my concern in the present chapter is not the particular problem of the two prisoners as presented thus far. Rather, I shall be interested in situations the structure of which is revealed on analysis to be similar, in its main features, to that of the Prisoners' Dilemma (henceforth to be referred to as PD) situations, but which is a generalization, in a sense yet to be explored, thereof. It can be said, therefore, that I shall be interested in situations which are generalized PD-like ones.

Regarding such situations it might already be noted by way of a preliminary conjecture that the larger the number of participants in them, the less the likelihood that there will exist trust, solidarity, or friendship among them, and also the more problematic (and possibly costly) the possibility of their coming to an explicit and enforceable agreement capable of co-ordinating their choice of actions in a mutually satisfactory way.

With the intention of providing an early guideline to the arguments in this chapter, I shall now state its main proposition. It is that generalized PD-structured situations constitute a type of contexts which are prone to generate norms. Unfolding this contention somewhat, the idea is the following: A situation of the generalized PD variety poses a problem to the participants involved. The problem is that of protecting an unstable yet jointly beneficial state of affairs from deteriorating, so to speak, into a stable yet jointly destructive one. My contention concerning such a situation is that a norm, backed by appropriate sanctions, could solve this problem. In this sense it can be said that such situations 'call for' norms. It can further be said that a norm solving the problem inherent in a situation of this type is *generated* by it.

Such norms I shall call PD norms.

3. *A Generalized PD-structured Situation*

Having said that I shall here be concerned with generalized PD-structured situations, I have yet to state more precisely first, what 'essential features' a situation has to possess in order to qualify as a PD-structured one, and, second, in what sense

such a situation is considered to be a generalization of the prisoners' predicament.

3.1 *First Approximation*

Stripping the story of the Prisoners' Dilemma down to its bare skeleton, we get, as a first approximation, the following characterization:

A PD-structured situation is any situation involving at least two persons each of whom is facing a decision as to whether to do A or non-A,[2] such that

(i) If all of them do A the outcome is (and is known to them to be) mutually harmful;
(ii) If all of them do non-A the outcome is (and is known to them to be) mutually beneficial—or at any rate *better* than the outcome produced by their all doing A;
(iii) Each of the persons involved stands to gain most by singly doing A. That is to say, one's highest pay-off is obtained when one does A while all the others do non-A;
(iv) One's doing A when the others do non-A is—at least to some extent—at their expense. That is, when all-minus-one do non-A, the outcome to the non-A doers is less beneficial than it would have been had *everyone* done non-A.

Examples of familiar situations which might be seen to answer this description are, say, those involving a decision between payment of income tax *vis-à-vis* tax evasion, conscription *vis-à-vis* draft dodging, voting in the general election *vis-à-vis* picnicking on that day, and—on a smaller scale—situations involving decisions such as whether to take the longer path or to short-cut through the well-tended lawn, whether to keep a promise or to break it, and the like.

The generalization involved in passing from the story of the prisoners to the types of situations described above is conceived as being carried out along two axes. The first obviously concerns the size of the class of participants, which is allowed to be indefinite. A clarification is in order here.

[2] It does not matter which of the two involves doing and which refraining from doing (or, for that matter, they might both involve doing: if A is voting, non-A might be construed either just as not voting, or as some alternative B—like going to the beach for the day).

It is these rules which Buchanan and Tullock offer as an explication of the rules mentioned in Dorman's (MGA). Or, more precisely, the norm affecting the transition from the non-optimal private adjustment to the Pareto-optimal state of affairs is offered by Buchanan and Tullock as the economic analogue to Dorman's 'rule against *x*'.

To sum up. Both my PD norms and Buchanan and Tullock's rules are proposed as predictive interpretations of the rules referred to in Dorman's (MGA). It is clear that these two sets of norms are not coextensive, and that each of them covers a good many of the norms (rules) encountered in ordinary human affairs. Is there, however, anything to choose between them as interpretations to Dorman? That is, is one of the proposed interpretations to be preferred over the other *as an interpretation* of Dorman's rules?

I submit that, as an interpretation, mine is to be preferred to Buchanan and Tullock's. The reason is, simply, that I believe I have shown that the situations in which the generalization argument is in fact applied and in which, moreover, its application tends to be successful and persuasive are generalized PD-structured ones—the ones which are prone to generate PD norms. I believe that they are *not* in general the type of situations Buchanan and Tullock are concerned with, in which there is need for public adjustment to replace a non-optimal state of affairs brought about by private adjustment.

11. *Hobbes and the Prisoners' Dilemma*

It is quite commonplace to regard Hobbes's original situation of mankind as a version of the Prisoners' Dilemma (e.g. Barry, 1965, pp. 253–4; Rawls, 1971, p. 269).[13] It has recently, however, been challenged (Gauthier, 1969) whether it is indeed warrantable to do so.

In this section I propose, first, to show that the original situation of mankind according to Hobbes does possess the essential features of a generalized PD-structured one; second, to examine Hobbes's solution to the pertinent problem and its relation—if any—to PD norms, and third, to offer a rebuttal to Gauthier's challenge.

[13] Rawls, as well as others, makes the mistake of taking Hobbes's *state of nature* to exemplify the Prisoners' Dilemma, whereas in truth the state of nature, as we shall presently see, is but *one* of the several states of affairs (represented, in the two-person case, by the four cells of the PD matrix) of which the entire PD-like situation is comprised. It is for this reason that I speak of Hobbes's 'original situation of mankind', rather than of his state of nature, as being a version of the Prisoners' Dilemma.

11.1 *The Original Situation of Mankind as PD-structured*

The state of nature is, according to Hobbes (1948), that of 'war of everyone against everyone'. It is a state which no one enjoys, which is uniformly bad for all: '. . . and the life of man, solitary, poor, nasty, brutish, and short' (p. 82). On the other hand we then learn that 'All other time [other than "that condition which is called war"] is *peace*' (p. 82), from which are implied its advantages. The state of peace is uniformly good for all, and indeed Hobbes states 'that every man ought to endeavour peace—to seek peace and follow it' (p. 85).

But here emerges the problematic aspect of the situation: it is dangerous, and against the Hobbesian 'laws of nature', to be a lone peace-keeper: 'For he that should be modest, and tractable, and perform all his promises, in such time and place, where no man else should do so, should but make himself a prey to others, and procure his own certain ruin' (p. 103). This means that, once in the state of nature, no one will have an incentive to pursue a peaceful policy, if all the rest stick to their belligerent policy. And this in turn means that the state of nature, notwithstanding its 'incommodities' (p. 82), is stable. Furthermore, the purpose of every one in this war of all against all is, of course, to win: first and foremost to survive, and also to reap such fruits of winning as gain, safety, and reputation (p. 81). Hence it is clear that, whenever some 'make themselves a prey' to others, that is, unilaterally withdraw from the policy of war, the others thereby gain the upper hand over them and profit by it.

Hobbes recognizes the dilemma. When one has no hope of obtaining peace—and obtaining it on one's own is impossible—one may, says Hobbes, 'seek, and use all help, and advantages of war' (p. 85). This is in accordance with the fundamental right of nature, which is 'by all means we can, to defend ourselves' (p. 85). This right of nature is in fact closely akin to the principle of maximizing security level, viz.: make sure the worst, i.e., in our case, being defeated in war, does not happen to you;

that is, do not adopt a peaceful policy which will result in such a defeat if the others do not adopt it too.[14] Consequently, a possibility of achieving the peaceful state preferred by all, while existing, is conditional. Hobbes's 'second law of nature' requires that a man be willing to lay down his 'right to all things'—which inevitably leads to war—'*when others are so too*' (p. 85, Hobbes's emphasis). That is to say, Hobbes realizes that the nature of this 'game' is such that one would not choose to keep the peace unless one knows that the others' choice would be the same.

The essential features of generalized PD-structured situations are evidently present here. Notice, however, that whereas in ordinary situations of that type the emphasis is on the advantages implied by lone deviation from the mutually beneficial state of affairs (represented by the top-left cell in the PD matrix), the emphasis here is on the disadvantages implied by lone deviation from the mutually destructive state of affairs (represented by the bottom-right cell).

This is so since in ordinary PD-structured situations, like the ones which served as examples throughout this chapter, the assumption is that the *status quo* in which the participants are placed is the good state, their problem being to stabilize it so as to avert the danger of deterioration into the bad state. In the Hobbesian original situation of mankind, on the other hand, the assumption is that the *status quo* is the bad state of war of all against all. (I believe that it is only in games, as opposed to real-life situations, that the participants are extraneous, so to speak, to the situation. In a game, therefore, the problem of the participants is which course of action they are to choose from the two alternative ones. In the type of situation with which we are here concerned, on the other hand, the problem of the participants is whether to remain in the *status quo*—whether it be the good or the bad one—or to deviate from it.)

The problem inherent in the Hobbesian situation of mankind, then, is two-fold. In the first place it is to find some means of deliverance, i.e. some means of 'lifting' from the bad state to the good one, and in the second place it is the usual problem of how to protect and stabilize the latter.

[14] Notice, however, that this of course is no *guarantee* against defeat: in the state of nature resulting from the belligerence of all the strong will win, the weak will lose.

11.2 *Hobbes's Solution*

The gist of Hobbes's solution is that people get together and covenant with each other to keep the peace. This covenant is then presumed to provide the needed ground on which everyone may reasonably suppose the others to choose a peaceful policy. But this, as Hobbes recognizes, is still far from being an adequate solution to the problem at hand. It seems that in Hobbes's view the following point cannot be overemphasized: that 'the bonds of words are too weak to bridle man's ambition, avarice, anger, and other passions without the fear of some coercive power' (p. 89), and that 'covenants, without the sword, are but words and of no strength to secure a man at all' (p. 109).[15]

It is not enough, then, that people covenant with each other and promise each other to keep the peace. They have to keep their promises too (cf. p. 21 above). And if there is reason to suspect they (or some of them) would not, then they have to be made to keep them. To this end Hobbes proposes a procedure whereby a sovereign is installed over the people—through an act of authorization made by every one of them—who is the representative of each person and is responsible for all matters of peace and security among them. This is achieved by means of mutual contracts, between each and every other person, in which all commit themselves to transfer their natural right to defend themselves to the sovereign, this commitment being conditional upon its being mutual ('on this condition that thou give up thy right to him, and authorize all his actions in like manner'; p. 112). Thus the sovereign becomes possessed of unlimited right of sanction. His word is the law ('. . . whereas law, properly, is the word of him, that by right hath command over others'; p. 105), and they who disobey the law, thereby threatening the maintenance of peace, are liable to as severe a punishment as the sovereign 'shall think expedient'.

The central point, however, in this procedure, is that the sovereign himself is not a party to any contract: 'the right of bearing the person of them all, is given to him they make sovereign, by covenant only of one to another, and not of him

[15] See also p. 115: 'understanding this easy truth, that covenants being but words and breath, have no force to oblige, contain, constrain, or protect any man'.

to any of them' (p. 114). He stands outside this network of mutual contracts. As a result there is no arguing with him, no threatening him, and no disposing of him. Moreover, all this is taken to be, and to be known by everyone to be, very much in the interest of every individual in the community, since it guarantees that the sovereign, who has the responsibility of keeping the peace, would also have 'untied hands' (p. 115) to maintain it. In practice this is achieved through punishments for peace-breaking which are severe enough to outweigh any prospective profit to be gained from it ('. . . some coercive power, to compel men equally to the performance of their covenants, by the terror of some punishment, greater than the benefit they expect by the breach of their covenant'; p. 94).

To sum up thus far: The original situation of mankind, as conceived by Hobbes, belongs to the PD variety. It is given that people are in the state of nature, as represented by the bottom-right cell in the PD matrix. Now a mechanism is proposed, consisting of a network of mutual contracts, which leads to the installation of a sovereign who is responsible for peace and who has—owing to the content of the contracts and to the fact that he himself is no party to them—the means of keeping it. Through this mechanism, when adopted, two things are achieved simultaneously: the much-desired 'lifting' from the state of nature (bottom-right cell) to the state of peace (top-left cell) takes place, and—at the same time—the situation of mankind (represented by the PD matrix as a whole) changes into a new and different situation.

In the new situation the state of peace, which is now the *status quo*, is good, as before, and the state of war is bad, as before. But deviation toward a belligerent policy is now going to be severely punished (whether it be unilateral or general: remember that the sovereign who inflicts the penalties is not a 'player' in this 'game' but controls it from without), so that there is now a deterrence against it, whereas the peaceful policy pays—even if it be unilateral, since its chooser enjoys the protection of the sovereign. (The last statement may perhaps be modified by taking into account the probability that the sovereign will in fact intervene to protect the peace-keeper before his belligerent neighbour manages to get the upper hand of him.)

That is to say, the installation of a Hobbesian sovereign resolves at once the two parts of the problem. In the first place it affects the needed transition from the bad *status quo*—the state of nature—to the good state of peace, thereby serving as a means of deliverance. In the second place, owing to the sovereign's unlimited right of sanction, the pay-off structure of the situation changes in such a way that once the transition into the state of peace takes place, the situation is no longer PD-structured. The belligerent policy no longer dominates the peaceful one; the state of peace is rendered stable.

Is Hobbes's two-fold solution related in any way to PD norms? I think that it is. Quite obviously it makes no sense to speak of the installation of a sovereign as a PD norm; nor, for that matter, can the mutual contracts be, in any straightforward sense, regarded as PD norms. However, I submit that it is because he is the originator of law in general, and of peace-promoting PD norms in particular, that a Hobbesian sovereign, empowered as he is by the mutual contracts, constitutes a solution to the pertinent double problem of transition and stabilization. To elaborate: the responsibility of such a sovereign is keeping peace and security; the law is his word. The principal norms he would issue, therefore, are such as would prohibit belligerence. These norms, accompanied by the threat of appropriate punishments—which the sovereign himself is entitled to administer—are those which in effect transform the situation into one no longer PD-structured: hence they qualify as PD norms.

Essentially, then, my point is that Hobbes's solution of the problem inherent in the original situation of mankind consists primarily of providing an originator of (peace promoting) PD norms.

11.3 *An Answer to Gauthier*

In his book *The Logic of Leviathan* (1969, pp. 77–85) David Gauthier raises the question of whether violation of a covenant can, according to Hobbes, be more rational (or, to use Hobbes's term, reasonable) than keeping it. In Gauthier's formulation: is it possible, within the Hobbesian system, that it is the case that R(A&B)*e*—which is read 'it is reasonable for A and B to enter the covenant', and yet it is not the case that R(A&B)*k*—which is read 'it is reasonable for A and B to keep the covenant'?