THE EMERGENCE OF NORMS

EDNA ULLMANN-MARGALIT

OXFORD
AT THE CLARENDON PRESS

CONTENTS

I. AIM AND METHOD *page*

1. The Subject Matter 1
   1.1 'Rational Reconstruction', 1
   1.2 'Formal Features', 5
   1.3 'Social Interaction', 7
   1.4 'Generation of Norms', 7
   1.5 Norms, 12
2. Methodological Individualism 14

II. PD NORMS

1. The Prisoners' Dilemma 18
   1.1 Presentation, 18
   1.2 The Dilemma, 19
2. The Main Proposition—PD Norms 21
3. A Generalized PS-structured Situation 22
   3.1 First Approximation, 23
   3.2 Refinement, 25
   3.3 The Condition of Individual Insignificance, 27
4. The Solution as a Stabilizing Device 29
5. The Mortarmen's Dilemma 30
   5.1 Presentation, 30
   5.2 Minelaying, 32
   5.3 Discipline, 33
   5.4 Honour, 36
6. An Assessment of the Proposed Solutions to the Mortarmen's Dilemma and Their Relation to PD Norms 37
7. PD and Morality 41
   7.1 Are PD Norms Norms of Morality? 41
   7.2 Norms for Keeping People in a PD Situation, 44
8. Functionally Equivalent Solutions to PD Problems 45
   8.1 Division of Labour in PD Situations, 48
9. Public Goods 49
10. The Generalization Argument and PD Norms 53
    10.1 The Modified Generalization Argument, 59

11. Hobbes and the Prisoners' Dilemma 62
11.1 The Original Situation of Mankind as PD-structured, 63
11.2 Hobbes's Solution, 65
11.3 An Answer to Gauthier, 67

III. CO-ORDINATION NORMS

Introduction 74

1. Co-ordination Problems 77

2. Co-ordination Problems and Norms 83
2.1 An Established Solution Turned Norm, 83
2.2 Norms as Solutions to Novel Co-ordination Problems, 89
2.3 On Conformism, 93

3. The Distinguishing Features of Co-ordination Norms 96
3.1 General, 96
3.2 The Source of Effectiveness, 98
3.3 Whose Interests Are Served by the Conformity to Co-ordination Norms? 101

4. The Social Role of Co-ordination Norms 104
4.1 The So-called Integrative Function, 104
4.2 An Instrument for Social Choice, 106
4.3 The Problem of Rationality, 110

5. Co-ordination Norms and PD Norms 114
5.1 Structural Differences, 114
5.2 Implications of Structural Differences—General, 115
5.3 Implications of Structural Differences—Norms, 117
5.4 An Intermediary Case—The Stag Hunt, 121
5.4.1. An Example, 124
5.5 The Oligopolists' Problem, 127

6. Co-ordination and Co-operation 129

IV. NORMS OF PARTIALITY

1. *Status Quo* of Inequality 134
1.1 Presentation, 134
1.2 Dahrendorf's Derivation of Social Inequality, 135

2. The Disfavourably Placed Party 139

3. Improvement of Absolute Position 141

4. Improvement of Relative Position 146
4.1 Presentation, 146
4.2 Relative Position Improvement in the Light of the Postulates of Game Theory and Pareto Optimality, 148
4.3 Envy and Rationality, 150

5. Improvement of Relative Position at the Expense of Absolute Position 155
6. Interpersonal Comparisons 157
7. The Favourably Placed Party 159
8. The Paradigmatic Situation 162
8.1 Characterization, 162
8.2 The Disfavourably Placed Party's Strategic Move, 164
9. Fortifying Methods 168
9.1 The Possible Methods, 168
9.2 Force *v.* Norms, 170
10. Norms of Partiality 173
11. Functional Explanation 176
12. Latent Functions and Latent Interests 180
12.1 Latent Functions and the Conspiracy Fallacy, 180
12.2 Latent Interests and Dahrendorf's Model, 182
13. Norms of Partiality and Coercion 187
13.1 The Coercer and the Impersonality of Norms, 188
13.2 The Impartiality of Norms and Self-coercion, 189
13.3 Is Coercion a 'Zero-Sum' Concept? 193

BIBLIOGRAPHY 198

INDEX 203

II

PD NORMS

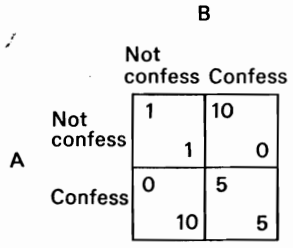
1. *The Prisoners' Dilemma*

1.1 *Presentation*

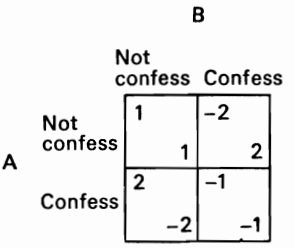
THE well-known story of the prisoners' dilemma is the following:

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.



2.1



2.2

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.)

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).

In game theory, a single participant is not necessarily envisaged as a single individual. It might be an aggregate of individuals (e.g. a firm, a political party, etc.). So that in fact a two-sided game might involve a (possibly large) number of persons as participants. However, it is not the number of persons involved in a game situation which is of significance to the game theorist; it is, rather, the number of decision-makers. Thus, when one considers a game between two firms, one might take them as two single players—even though they might each be composed of many individuals—once it is understood that each firm (regardless of how the processes of arriving at decisons are carried out *within* it) speaks with just one voice, and that in this respect its 'moves' are analogous to those of a single individual.

The generalization intended here, on the other hand, is genuine in that it refers not just to a large class of participants, but to a class of participants each of whom is a unit in terms of decision and action. Of course, a generalized situation *cannot* be represented by a 2×2 matrix of the familiar type. (See, however, section 9 below.)

The second axis, which does not find an explicit expression in the above list of features, concerns the factor of time. The original Prisoners' Dilemma is ordinarily conceived as a 'one-shot' game; i.e. as a situation involving a dilemma which is presented to the (two) participants just once, and simultaneously. A generalized PD-structured situation, on the other hand, is one in which the dilemma faced by the (multiple) participants is recurrent, or even continuous.

As a result, in addition to yielding a situation involving a large number of people among whom the dilemma is recurrent, the generalization should also yield situations in which the class of participants is not only large, but also indeterminate. That is, situations of particular interest to us here will be such that involve a community which is neither fixed nor closed, but which admits initiation of the young, of newcomers, etc. In such a situation, therefore, where people come and go, the recurrent dilemma is not in general faced, in each of its occurrences in the community concerned, by exactly the same people, but rather by a community with somewhat fuzzy borderlines.

Moreover, the participants involved in such a situation *know* this to be the case, and have to take the indeterminacy of the

class, as well as the anonymity of its members, into account in their considerations as to what they might expect the others to do, and ultimately as to how they themselves are to act. Consequently, a solution to such a generalized dilemma would have to be one that is capable of being extended to cover the many instances of the problem, possibly among the anonymous members of an indeterminate class, and not one that solves, *ad hoc*, just a particular occurrence thereof, among specific participants.

Each of these generalization axes is of crucial importance to the central contention of this chapter. It might be said, somewhat schematically, that the smaller and the more determinate the class of participants in a generalized PD-structured situation, and the more isolated the occurrence of the dilemma among them, the more likely it is that there might be solutions other than (PD) norms to the pertinent problem (see section 8 below). And conversely, the larger and the more indeterminate the class of participants, and the more frequent the occurrence of the dilemma among them, the more likely it is that a solution, if any, would be in the form of a PD norm.

3.2 *Refinement*

When the class of participants is large, the essential features of a generalized PD-structured situation as stated above are stronger than need be and can be somewhat relaxed. It is a little over-dramatic to formulate the conditions in terms of 'all' and 'all but one'. It should obviously be possible to substitute 'most' for 'all' in some appropriate way. Also, the fact that the dilemma is recurrent should find some expression in any reformulation of the features, possibly by means of time-terms like 'always' and 'sometimes'.

I propose, therefore, the following refined characterization.

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

(i) If, in any occurrence of the dilemma among them, most of them do A the outcome is (and is known to them to be) mutually harmful;

(ii) If, in any occurrence of the dilemma among them, most 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 when most of them do A;

(iii) Each of the persons involved obtains, at least in some occurrences of the dilemma among them, the highest possible payoff in the situation when he himself does A while most of the others do non-A;

(iv) If, in any occurrence of the dilemma among them, some do A, the outcome to the non-A doers is less beneficial than it would have been had everyone done non-A.

Note that in feature (iii) it is not required that in *each* occurrence of the dilemma each participant stands to gain most by unilaterally doing the 'bad' action. Rather, it states the weaker requirement that each participant is at least *sometimes* tempted to do so. The generalized PD case is, consequently, not to be regarded as an amalgam of discernibly isolated proper (two-person) PD cases. It is, rather, a situation of complex social interaction which, when taken in its entirety, displays the four features just stated. These features clearly constitute a somewhat weakened generalization of the features characterizing the proper two-person PD case.

Thus, to consider the example of promise-breaking, in order for it to be a generalized PD-like situation we need not suppose that every one of us is always tempted to break his/her promises; it suffices that each of us sometimes finds it to his/her advantage to do so. Similarly, the situation involving a decision between truthfulness and mendacity can be construed as a generalized PD-structured one, in which feature (iii) holds in its weak version.

It might, furthermore, be possible, in some specific situation, to make more precise just what is meant by 'most' and 'some'. One might think, for instance, of the example of the lawn: if only a few tread it each day, the damage to it might not be significant. So they themselves gain a short cut and the others can still derive pleasure from contemplating the still beautiful lawn. However, once a few more tread it each day, the lawn might by now be appreciably damaged, a state of affairs which is assumed, for the sake of the argument, to be repugnant to all. Similarly, one might think of the case of petty shop-lifting in a supermarket which, once it transgresses certain limits,

inevitably leads to the collectively undesirable outcome of price-raising due to a decline in profits, extra security precautions, etc.

In general, this further refinement is conceived as being along the following lines:

Let the community of participants be denoted by C. Then, with regard to any occurrences of the dilemma in C, there is some maximal sub-set of C, say D (for defectors, dodgers, deviators, or what have you), which is in general rather small relative to C, such that if all members of D do A while the members of C—D do non-A, the A doers obtain for themselves the highest possible pay-off in the situation, and yet the outcome of the non-A doers is still not significantly affected by the deviators and is in general rather satisfactory. However, if more members of C defect and join the sub-set D, then deterioration takes place. That is, for any D′ such that $D' \subseteq C$ and $D \subset D'$, if the members of D′ do A then the outcome for all—both A doers and non-A doers—is already harmful.

This is still quite crude. In order to refine it further one would have, for instance, to consider the point that it may not be a certain crucial *number* (or *proportion*) of people who deviate that brings about the jointly harmful consequences; it might also be the *frequency* in which these deviations occur. Also, attention would have to be paid to the fact that in some cases it might be the addition of just one deviator to D which upsets the balance, while in other cases (perhaps in most) it takes a certain *range* of additional deviations to produce that effect. Such further refinements, however, although quite interesting in themselves, are immaterial for the issues in hand.[3]

3.3 *The Condition of Individual Insignificance*

Special attention should now be directed to feature (iv) (pp. 23, 26 above)—that is, to the extent to which one's doing the 'bad' action directly harms the doers of the 'good' action. In the case of the original story about the prisoners, unilateral confession not only brings about freedom for the confessor, but

[3] I believe that David Lyons's carefully worked-out notions of the *density* of a practice, of general *v.* universal practices, and in particular of *thresholds* and *threshold effects* are directly applicable to any attempted refinement along the lines suggested above. (Lyons, 1965, pp. 63–75.)

sends his mate to the longest term in prison. That is, in the case of just two participants, unilateral deviation (defection) is directly at the expense of the other, and results in a state of affairs where the former obtains the highest possible pay-off in the situation and the latter the lowest.

However, it stands to reason that the more participants in the situation, the less the damage to the majority by the minority of deviators (provided it is small enough). If, for instance, there are in a large community only very few tax-evaders, the system will go on functioning and no appreciable decrease in the provision of goods and services by the state will occur. To be sure, *some* harm, at least in principle, is always done by such a minority: it would surely be better, even if it is only very slightly better, for the community as a whole were there no tax-evaders at all and as a result were there somewhat more resources in the state treasury. For this feature of the generalized PD-structured situations to obtain, there is one necessary condition which must be fulfilled. It is that the contribution of each participant taken alone be sufficiently small and untraceable as to be practically insignificant to the over-all outcome. The whole point, of course, is that the composition of such near-insignificant contributions is, as from a certain point (or range), quite significant; indeed it is capable of swaying the final outcome.

This condition—call it the *condition of individual insignificance*—implies (but is not implied by) another condition, which we might call the *condition of homogeneity*. It requires that the contribution of each participant has more or less the same weight as that of any other participant relative to the over-all outcome. Jointly they ensure that there be no single participant (or a small group of participants) whose decision whether to do A or non-A tips the scale as to the over-all outcome, regardless of what the others do. (In economics the fulfilment of these two conditions results in a state of *perfect competition* where a large number of small producers compete for the market and there is no leading firm which dominates it.)

Note that in situations where the condition of individual insignificance is satisfied the temptation to do the 'bad' action might persuasively be argued to be greater than it is in PD-structured situations in which it is not satisfied (such as in the

original case of the prisoners). The argument is that in situations of the former kind unilateral defection to the tempting, 'bad' action is not at the personal expense of any of the other participants, and is moreover likely to go entirely unnoticed by them. It is an action which, if unilaterally done, gratifies its doer—possibly even highly gratifies him—while not directly and appreciably harming the others. Thus a potential inhibition which might exist in situations of the latter kind, regarding the infliction of direct and personal damage on one's partners, is removed in the cases where the condition of individual insignificance is satisfied. (Regardless of whether the situation is one in which the condition at hand is satisfied or not, though, it holds true that the tempting, 'bad' option dominates the 'good' one, in the sense explained on p. 19.)

Let me finally recall briefly the statement of the main proposition of the present chapter (as put forward in Section 2 above), that generalized PD-structured situations are prone to be norm-generating. The last two sections aimed at clarifying the notion of a 'generalized PD-structured situation'. I would like to add at this point the conjecture that generalized PD-structured situations which satisfy the condition of individual insignificance—and hence the condition of homogeneity too—are more likely to constitute recruiting fields for norm-generating contexts than are those which do not. In generalized PD-structured situations in which the condition of individual insignificance is not satisfied (and the condition of homogeneity is either satisfied or not, as the case may be), I believe that it is more likely that solutions other than (PD) norms might exist, or be devised. This point will be taken up again later (Sections 8 and 8.1 below).

4. *The Solution as a Stabilizing Device*

The problem posed by a situation of the generalized PD variety is clear: it is how to protect a state of affairs which is jointly beneficial yet unstable from deteriorating—via the yielding to temptation by each (or by most) of the participants involved—into a state of affairs which is not only jointly harmful but also highly stable so that the possibility of redemption from it seems at best remote. That is, what is needed is some stabilizing device.

A stabilizing device in the type of situations under consideration would be anything which successfully eliminates the temptation to deviate from what is obviously the 'good' action. In terms of the matrix representation (matrix 2.3, p. 19 above) this means that an adequate stabilizing device would be one which would change the pay-off structure of the situation in such a way that the second action—referred to by 'D'—will no longer dominate the first—'C', and that the jointly beneficial outcome (represented by the upper-left cell) will become an equilibrium from which it would pay no one to deviate alone.

All this amounts, very simply, to the statement that what is required is a reduction in the pay-offs referred to by 'T' at least to the level of the pay-offs referred to by 'R' but preferably to a considerably lower level than that. A prospective solution characterized in this way is, it seems, just what T. C. Schelling had in mind when putting forward the following propositions (1964, p. 487):

> Game theory can help to identify solutions, that is, ways that one may change the situation so that choices will be made differently. . . . Game theory can also discover that reducing pay-offs to both players, if the reductions are selectively accomplished, to change a dominant strategy into one no longer dominant and to provide a new equilibrium point, can benefit both players. Though they might appear to be worse off from selective reductions in their pay-offs, in fact each gains by the resulting change in strategies chosen.

I propose now to construct an imaginary example of a PD-structured situation and to examine in some detail several ways out of the dilemma which all fall under the above characterization. I shall then proceed to assess these solutions, and to examine which of them, if any, is connected with PD norms, and in what way. The analysis of this example will prove instructive, I believe, for the understanding of the nature and function of PD norms in general.

5. *The Mortarmen's Dilemma*

5.1 *Presentation*

Let us imagine two mortarmen in two isolated outposts, facing an enemy attack. Each of them considers that he has two alternatives open to him: to remain at his post and fight, or to take to his heels.

Now suppose the possible outcomes are:

1. If both stay in their posts and shell the enemy, the attack is repelled.
2. If both run away there is nothing to hold the enemy up, so the enemy break through right away and manage to take both of them prisoner.
3. If one stays in the post and the other runs away, the one who stays behind manages to hold the enemy up by engaging them in fire just long enough for the other to escape safely, but then gets killed by the enemy who eventually break through.

Suppose too that both mortarmen know these to be the possible outcomes.

Let us further suppose that the mortarmen's preferences among the outcomes are represented by the following 'pay-offs':

To remain in the post and repel the attack: 1
To remain in the post and get killed: −2
To desert successfully: 2
To desert and be taken prisoner by the enemy: −1

The situation may now be summarized in the classical PD matrix (2.4).

	R	D
Remain	1 1	−2 2
Desert	2 −2	−1 −1

2.4

Each realizes that, whatever the other does, he is better off deserting: the D-strategy dominates the R-strategy. This leads them to double desertion (D, D), which results in them both being taken prisoner, whereas both would prefer the outcome (R, R) which entails their remaining alive and free. The trouble, typically, is that 'not only is a constant temptation [to desert] pressing on each player, but also the *knowledge* that the other player is tempted makes the defection practically compelling' (Rapoport, 1966, p. 129).

We now ask ourselves how can this dilemma be resolved. At the outset, however, we note that it is not only in the interest of the mortarmen's superiors, or their country for that matter, that they both remain in their posts and repel the enemy attack, but it is also in the interest of the mortarmen themselves to do so. The problem, then, is to find some means by which to deprive the desertion strategy of its tempting force; that is, to

change the situation—or the degrees of desirability of the outcomes—in such a way as to cancel the dominance of the D-choice over the R-choice.

I shall propose and examine three such ways.

5.2 *Minelaying*

The first way of solving the dilemma to be considered looks like a rather brutal one: mines are to be densely laid all around the mortarmen's posts.

The possible outcomes and the pay-offs connected with them are now:

To remain (when the other remains too) and repel the attack: 1

To remain (when the other deserts) and get killed by the enemy: −2

To try to desert but get killed by a mine: −2

The matrix of this situation is 2.5.

	R	D
Remain	1 1	−2 −2
Desert	−2 −2	−2 −2

2.5

A rough examination of the matrix shows clearly that the outcome (R, R) has all the 'good qualities' of a game-theoretical solution. It is by far the best outcome for both parties; there is no question whatever of even contemplating the D-choice. It is clear, then, that this minelaying solution, brutal as it may appear, turns out to work in the interest of both parties. It effectively eliminates the possibility of desertion, and consequently each is happy to stay by his mortar and fight, knowing that the other has in fact no alternative but to fight too, with the result that the attack is repelled and they both stay alive. The minelaying changes the situation—more specifically, the outcomes associated with the D-choice—in such a way as to endow the outcome preferred by both with the accessibility, stability, and dominance which it lacked before.

This minelaying device may call to one's mind the famous story that in World War I German soldiers were in fact chained to their machine guns, to prevent their running away. In this connection Schelling points out (1964, p. 476) that 'it is particu-

larly interesting that machine-gunners might volunteer to be chained, each willing to stay on condition his partners be made equally reliable.' In our case we may say that it is not irrational for the mortarmen themselves to ask for the laying of the mines around their posts (provided they are given no map of the location of the mines), thus assuring their solidarity in battle.

5.3 *Discipline*

The second variation on the mortarmen story is this: suppose the two mortarmen belong to a unit in which military discipline is harsh and effective. Whoever gives way under enemy attack is liable to be executed by his officers, if caught.

The possible outcomes and the pay-offs associated with them are now taken to be the following:

To stay with one's weapon (when the other stays too) and repel the attack: 1

To stay with one's weapon (when the other deserts) and get killed by the enemy: −2

To desert (when the other deserts too), get caught and be taken prisoner by the enemy, *and* be liable to execution if eventually handed over to one's own forces: −1·5
(This pay-off, it will be noted, is better than that for getting killed, which is −2, and worse than that for captivity, which—in the original mortarmen's matrix—was −1; the rationale being that this case involves certain captivity conjoined with probable execution.)

To desert (when the other stays behind) and be liable to execution: −1 or 0
(Here the pay-off depends upon the probability one assigns to one's being caught, and hence executed, by one's own forces. Note that if being caught, and hence executed, is taken to be certain, then the appropriate pay-off will be −2, and we shall be back, for all practical purposes, at the Minelaying variation of the story.)

The matrix of this situation is 2.6.

This matrix is an interesting one: it has two equilibrium points, viz. (R, R) (since −1 and 0 are worse than 1) and (D, D)

	R	D
Remain	1 1	−2 −1 0
Desert	−1 0 −2	−1·5 −1·5

2.6

(since −2 is worse than −1·5), the first of which is obviously preferred by both parties. It might seem, then, that the outcome of (R, R) is the natural and straightforward one, and that the desired solution of the dilemma is consequently guaranteed. However, there is a certain complication here. The (D, D) equilibrium, albeit being the worse of the two, is nevertheless the one contained in the plan maximizing both participants' security level (for the worst that may occur when choosing R is −2, whereas the worst that may occur when choosing D is −1·5, which is better). So that the situation is such that there are two equilibria: one good and in a sense risky, the other worse but safer.

This means that neither party will find remaining in the post the more rational course of action *whatever the other does*. If there is the slightest suspicion that the other might desert, desertion has to be considered the safe choice. But as long as there is no ground for suspicion, and since each stands to gain from remaining as long as the other remains, there is reason to assume that the choice of both mortarmen will fall on the R-strategy. At least this much is true, that if communication between them is at all possible they should have no difficulty in co-ordinating their choice of actions to bring about their remaining in the posts and repelling the enemy attack, to the advantage of both.

However, if, as we assume, the mortarmen belong to a strongly disciplined unit, they should be *in the habit* of obeying orders, knowing from experience that disobedience is not tolerated. So when they are stationed in their posts with the order to repel the enemy attack with their mortars, even if communication between them is not possible this habit of obedience will serve as the needed ground on which each will expect the other to remain in his post and fight, and therefore he himself will remain and fight too.

Thus we see that the contribution of discipline to the dissolution of the dilemma is two-fold. In the first place its existence and effectiveness change the pay-offs associated with the outcomes in such a way that flight no longer dominates fight, and that the outcome (R, R) becomes an equilibrium. In the second, the knowledge, common to both mortarmen, *that* their unit is disciplined constitutes the necessary ground on which each bases his expectation that the other will not desert. This

expectation is decisive at the point where each has to decide which of the two equilibria he is going to aim at by his choice of action, and it is likely to turn the scale in favour of the jointly beneficial outcome.

Note, finally, that discipline functions as a *deterrent device*. It is the threat of punishment which outweighs, and thus cancels out, the temptation to desert.

In a recent paper addressing the issues of morality and rationality with regard to the Prisoners' Dilemma, Amartya Sen urges us to consider a variation of the preference orderings of the two prisoners (1974, pp. 59 ff.) which in essence corresponds to my 'discipline matrix' (no. 2.6, p. 33 above). That is, he proposes to begin by leaving the facts of the original story about the two prisoners (as represented by matrix 2.1, p. 18 above) intact, and then to imagine that the prisoners' preference ranking of the four possible outcomes be such that it will be represented by a matrix like no. 2.6 rather than by the usual PD matrix (i.e. like no. 2.2, p. 18 above). Realizing that in this variant of the 'game' there are two equilibrium points, and that its structure is such that 'a contract of mutual non-confession [non-desertion] does not need any enforcement . . . whereas it is the crux of the matter in the [original] Prisoners' Dilemma' (p. 60), Sen proposes to call it the Assurance Game (or AG): 'Each prisoner [mortarman] will do the right thing if it is simply assured that the other is doing it too and there is no constant temptation to break the contract' (ibid.). Sen then proceeds to argue that a resolution of the prisoners' dilemma will come about if 'every one behaved *as if* they had AG-preferences and had the assurance of similar good behaviour by others' (ibid.), rather than the PD-preferences they actually have.

The idea of acts according to *as if* preferences is mysterious to me. What seems to me somewhat clearer, though, is just where Sen's approach to solving PD-type problems essentially differs from mine—and I would like to point that out. Sen's idea is to leave the reality of the dilemma-ridden situation as it is, and somehow to affect the participants' (moral) *perceptions* of and *attitudes* to it in such a way that they will act on 'as if' (AG) preferences rather than on their real (PD) preferences. What I maintain, on the other hand, is that in such cases certain devices (generally speaking, PD norms) will be generated such

that it is the *reality* of the situation which changes (e.g. to one of the discipline, or AG, type), and along with it, of course, the participants' preferences over the (new) outcomes.

(As to the relation between the 'discipline device' and PD norms, see below, Section 6.)

5.4 *Honour*

Suppose now that in the society to which the mortarmen belong honour is all in all. In their unit one is not so much in dread of death, provided it is heroic death, but of denunciation as a deserter and traitor. For these people glory sugars the pill of death, and a mark of disgrace as a traitor makes the prospect of saving one's life by escape, let alone an escape which ends up in captivity, gloomy, if not hideous.

Given this system of values, the pay-offs connected with the four possible outcomes are taken to be the following:

to stay and fight (when the other stays and fights too), and repel the attack: 1

to stay and fight (when the other deserts), get killed by the enemy and become a (dead) hero: −1

to desert (when the other stays behind and fights), be saved from death, retain freedom, but be denounced as a traitor and people's enemy: 0

to desert (when the other deserts too), be taken prisoner by the enemy, *and* be denounced as a traitor: −2

The essential point here is, then, that these people had rather die as heroes than live in the enemy's captivity, stigmatized as traitors. On the other hand we should not, and in fact for our purposes need not, go to extremes and impute to our mortarmen preference of being dead, even though as heroes, over being alive *and* free, even though as traitors. This is the rationale of the order of preference among the outcomes, as represented by the above pay-offs.

	R	D
Remain	1 1	−1 0
Desert	0 −1	−2 −2

2.7

The matrix of this situation is 2.7.

The (R, R) outcome is here the obvious solution: it is the best outcome for both, it is an equilibrium (the only one), and it is contained in a dominant choice (and hence

also in the choice maximizing security level). In other words, one is here better off remaining and fighting *whatever the other does*. It is interesting to note that in this respect the 'honour device' of solving the Mortarmen's Dilemma turns out to be more effective than the discipline device, the choice of R in the latter case being dependent upon the expectation that the other chooses R too.

It is noteworthy that honour, on the foregoing analysis, functions not only as a device of deterrence, but also as something which is rather close to being a *compensative device*. This is so since we have here both *decrease* in pay-offs, where being denounced as a deserter is involved, and *increase* in pay-offs, in terms of esteem, reputation, prestige, and glory, where behaviour which is up to what is considered honourable is involved.

(The latter point might seem to echo, in a sense, Schopenhauer's observation—in *Parerga und Paralipomena*—that since the state cannot afford to pay its officers and officials a high salary, it pays them the second half of it in livery, decorations, and honorary titles.)

Note also that both disgrace and glory are conferred not only on the deserter or the hero (as the case may be) himself, but on his family too. This is of significance since it contributes to the fact that it might after all indeed make a difference to a man whether he dies as a traitor or as a hero.

6. *An Assessment of the Proposed Solutions to the Mortarmen's Dilemma and their Relation to PD Norms*

Three ways out of the Mortarmen's Dilemma have been examined in the last few sections. All of them consist, essentially, of the bringing about of restraints on choices, or of changes in the preference-order among the outcomes which are reflected in pay-off changes. The first way considered, the minelaying, is in fact a sort of technical device; for most practical as well as theoretical purposes it amounts to eliminating an alternative, that is, to a reduction of the matrix into a degenerate one-choice matrix. In this respect this device is similar, as we have seen, to the measure of chaining machine-gunners to their weapons.

Let us now turn to assess the other two means through which the dilemma was shown to be resolved. What is discipline? I

think it is better to ask: What is a disciplined unit? At any rate the answer to the second question is simpler, and more pertinent to our present purposes. Very generally, a disciplined military unit is one whose members regularly obey their commanders' orders. They might do so out of habit, fear of punishment, or some inner conviction about the value of obedience; probably out of some combination of these.

What is the object of military discipline? In order to approximate an answer to this question, let us start by assuming that the entire machinery of an army is ultimately geared towards an encounter of its infantry with the enemy's. When this takes place, the case is a generalized PD-structured one in which, moreover, the condition of individual insignificance is likely to be satisfied. Each soldier's contribution to the final outcome of the battle can probably be taken to be negligible;[4] each soldier might reasonably be envisaged as facing a temptation to desert, to 'disappear', to stay behind as soon as the fighting begins; if too many yield to this temptation the outcome is likely to be disastrous to all; their only chance of obtaining a collectively satisfactory outcome might lie in their all storming and fighting together.

This description is, of course, oversimplified. But the oversimplification helps me make my point. It is that what might be taken to be one of the most fundamental military norms, that a soldier ought never to desert in time of battle, can—at least in part—be construed as a PD norm. It is a norm which, when supported by sufficiently severe sanctions, is capable of solving, or indeed dissolving, potential generalized PD-structured problems of the type outlined above. General compliance with it, conjoined with common knowledge that it is generally complied with, guarantees the desired collective effort and solidarity in battle.[5]

[4] Unless there is an appropriately detailed division of roles and positions among the soldiers, in which case the picture is changed. On this issue, see below, Sec. 8.1.

[5] Within the context of the military, it might be contended, along the lines of Urmson (1958), that this PD norm is some kind of a 'basic duty', a minimal requirement of co-operation which prohibits behaviour that is intolerable if an army is to achieve its primary goal. Furthermore, this norm leaves plenty of room for heroic actions which go beyond this duty. Indeed, I think that a case can be made that what falls under Urmson's category of basic duties are in general PD norms.

There is, however, a further point to be considered here. Suppose that it is a long time since the efficacy of this norm has been put to the test. Suppose, that is, that there has not been a battle for a long time, so that at least as far as the younger soldiers are concerned, there are no previous cases to go by. The requirement that the PD norm concerned (prohibiting desertion in time of battle) be effective and be known to be effective is thus not satisfied in such a case. Does it follow that in these cases a state of general desertion would almost inevitably occur?

To be sure, this norm is—and it can safely be assumed that it is known to be—accompanied by the severest sanctions, possibly execution. So that the temptation to desert is considerably diminished. But still, as was shown earlier (p. 34 above), the equilibrium state of collective desertion, although less preferred by all to the equilibrium state of collective fighting, might be considered safer in the sense of being contained in a choice maximizing the security level of each. Consequently, the deterrent effect of the punishment for desertion might nevertheless be outweighed by the consideration of maximum security level. This would be so, provided that the systems of mutual expectations were such that each expected the others to desert rather than to stay on and fight.

But do we in fact have any reason to believe that this would be the nature of the mutual expectations involved?

Let us, at this point, go back to the opening statements of this Section and tie them in here. My argument is that if the units under consideration are, and are known to their members to be, disciplined, then deterioration into the state of general desertion would *not* take place even if there are no recent precedents of behaviour in time of battle, and so the problem would be overcome in these cases too.

The reasoning is this: that the units are disciplined means that they have the propensity to obey orders; and that they are known to their members to be disciplined means that each member thereof may reasonably expect each other to obey orders in any given case. Hence, in the case of the battle in question, even if there is no recent precedent of compliance with the particular (PD) norm not to desert, the fact that the units involved are, and are known to their members to be,

disciplined, suffices to make the pertinent systems of mutual expectations such that each will have good reason to expect the others to obey their officers' orders and fight rather than desert. Consequently, the equilibrium state to be aimed at by each would be the better one—that of collective fighting—and the problem would be overcome.

To sum up, then. The fundamental military norm prohibiting desertion in time of battle is, it has been argued, a PD norm. It emerges out of, and applies to new, situations which can be viewed as generalizations of the mortarmen's predicament, that is, situations which are generalized PD-structured ones. The efficacy of this norm, conjoined with the severe punishment for its breach, solves the problem inherent in this type of sitnation: the good outcome is stabilized, the danger of deterioration into the bad state of affairs is eliminated. Furthermore, in cases where the condition of the efficacy of this norm is not satisfied owing to lack of appropriate precedents, it has been argued that the alternative condition—that the units concerned be in general disciplined—is sufficient to guarantee that the pertinent problem would nevertheless be solved.

As to the issue of honour, finally, I think that it might be advisable to make a shift here, similar to the one made above from 'discipline' to 'a disciplined unit', and to proceed to deal with 'honourable behaviour' instead of with 'honour'. It seems that there are certain standards of behaviour associated with, and presumably following in some sense from, the idea (or ideal) of honour. Behaviour which meets these standards—compressed by such injunctions as 'Return from war either with your shield or on it'—is, accordingly, considered honourable. Be the precise content of the concepts of honour and honourable behaviour what it may, however, the essential point concerning the issues in hand is that, under specific circumstances (such as those considered in 5.4 above), they might give rise to certain values which are capable of affecting one's preferences over a given set of alternative outcomes in such a way that dilemmas like the mortarmen's practically dissolve. From the point of view of the military, therefore, it seems that it is worth while to foster and to cultivate the idea of honour among its soldiers, indeed to make them adopt the standards of behaviour connected with it as their own inner ones, for with

such soldiers the officers need have no fear of predicaments of the mortarmen's variety, still less of cases of general desertion actually occurring in time of battle.

7. *PD and Morality*

7.1 *Are PD Norms Norms of Morality?*

David Gauthier characterizes a moral system, and in a derivative way a moral man too, in terms of behaviour in a PD-structured situation (1967). Practically speaking, a moral system, according to Gauthier, is a system of principles which will make people involved in a PD-structured situation choose the co-operative (dominated) action rather than the alternative (dominating) action of defection. The reasoning is that a moral system must be one which 'requires that some persons perform acts genuinely disadvantageous to themselves as a means to greater mutual advantage'. And since in a PD not to make the defection choice, when one's partner (or opponent, as the case may be) is expected to make the co-operation choice, is indeed genuinely disadvantageous to one, it follows that a system of principles inducing one to choose the co-operative action notwithstanding, in order to increase the mutual advantage, counts as a moral system.

Now the derived behavioural characterization of a moral man is the following. A moral man is one who, in a PD-structured situation, chooses the co-operative action on the assumption that the other is also going to make the same choice, and who, moreover, does not deviate from this choice even if he be certain that the other cannot, for some reason, punish him later by deviating too. The reasoning which leads to this contention starts by noting that the choice of the co-operative action, on the expectation that the other is going to make the same choice is, according to Gauthier, out of *prudence*. It is a choice which involves a sacrifice of some greater personal advantage for a maximum of mutual advantage. When to this is added one's readiness to stick to this choice in spite of a postulated certainty that deviation from it would—with impunity—significantly improve one's own position, then, according to Gauthier, one is shown to be *trustworthy* too. And, finally, one's being both