Draft 9 October 24 2000

THEORY OF GAMES AS POLITICAL THEORY:

A NON-MATHEMATICAL EXPLORATION

Brian Barry

Columbia University

For Keith Dowding, Anne Gelling and
my ungodson Jonathan, with fond memories
of the past and high expectations of the
future.

THEORY OF GAMES AS POLITICAL THEORY:

A NON-MATHEMATICAL EXPLORATION

I GAME THEORY, POWER AND COLLECTIVE ACTION

II THE ANALYSIS OF POLITICAL POWER

PART 1

GAME THEORY, POWER AND COLLECTIVE ACTION

1. INTRODUCTION

Game theory is an analytical tool which helps to clarify thought about strategic interactions – that is to say, situations in which the optimal choice of an action or strategy (rule for a series of actions) for one actor depends on the choice made by another actor (2-person game theory) or the choices made by more than one other actor (n-person game theory). Since virtually all of politics falls under this description, the potential relevance of game theory is unlimited.

The actual utility of game-theoretical analysis is another question, but I would not have written this book if I did not believe that it can contribute to the resolution of some problems that cannot even be clearly stated without it. I shall discuss in what follows some uses (and some abuses) of game theory in political theory. My own suspicion is, however, that there is much more room still left for the application of game theory to political theory, and that the main reason for its limited use is that few political theorists have even an elementary grasp of game theory. This book is not a substitute for a systematic study of game theory. It provides an elementary exposition of basic ideas in game theory and develops these ideas only in directions that bear on issues central to political theory.

Other books that expound game theory and its applications to politics tend to combine sophisticated game theory with rather crude political theory. In this book, by way of contrast, it is the game theory that is rudimentary. There is no algebra and no symbolic logic. Moreover, the examples that are worked through take the simplest form possible. The apparatus used consists only of payoff matrixes and decision trees, both of which are explained with a degree of redundancy that may make some readers impatient. My objective is, however, to break down the resistance of even those most subject to mathematics phobia – provided they actually want to find out about game theory, of course. And my experience has been that there is no substitute for going very slowly. The other side of the picture is that I presuppose a certain familiarity with political theory. In particular, I rework some ideas from Hobbes and Hume in terms of game theory without pausing to explain the broader context within which these ideas occur in Leviathan and the Treatise.

The book is divided into two parts. Part I provides some basic analysis of zero-sum (ch. 2) and non-zero-sum (ch. 3) games, then uses the framework of non-zero-sum game theory to discuss power (ch. 4) and collective action (chs. 5 and 6). Part II is entirely devoted to the topic of political power. The first four chapters (chs. 7 – 10) take up in turn four important forms of political power. These are often run together, but I hope to show that they require very different analytical approaches in order to see how they operate. Chapter 11 is devoted to a much-discussed (and poorly-discussed) question: Is power zero-sum? Chapter 12 consists of a few concluding remarks.

Although much of the book is concerned to explain as lucidly as possible ideas that are the common currency of game theory, I should like to believe that the chapters in Part I on power and collective action offer a fresh view of some familiar topics. In Part II, I have sharpened up some debates by deploying the formal analysis that I develop to challenge some widely-held views. In chapter 7, I explain why obedience to governments cannot rest on power. In chapter 8, I argue against the apparently obvious idea that a political party holding the balance of power has more power than the others. Chapter 9 contains, among other things, a deductive argument showing the extraordinarily demanding conditions that have to obtain for the theory of consociational democracy to be valid. In chapter 10, I suggest that owners of capital have a great deal of power, and are not merely fortunate in that governments have to take their interests into consideration. Finally, in chapter 11, I claim that anyone who simply answers 'yes' or 'no' to the question 'Is power zero-sum?' must be talking nonsense. I hope, therefore, that even those who are already familiar with both game theory and political theory will find something of interest in this book.

2 ZERO-SUM GAMES

1. Pure Strategies

Although the zero-sum game has only a very limited application in politics, it is worth starting with because it provides the simplest introduction to game-theoretical analysis. Also, as a matter of intellectual history, the solution concept for two-person zero-sum games was worked out first by John von Neumann long before he and Oskar Morgenstern launched game theory in The Theory of Games and Economic Behavior (1944). It remains a classic example of the way in which our intuitions can be organized and clarified by an appropriate formal analysis.

Without further ado, let me introduce the example of a zero-sum game that I shall be discussing throughout this chapter. Let us suppose that, during the Gold Rush, A is employed by Wells Fargo to drive the wagon containing sacks of gold dust from the gold fields to San Francisco, and that the notorious bandit B is intent of robbing him of the gold. Despite other bits of local colour, A and B will be making frequent appearances throughout this book, sometimes accompanied by C and occasionally by D. In the present case, remember that A is the (potentially) ambushed and B the bandit. It will be noticed, incidentally, that all these characters are referred to as 'he'. While nothing in this book is difficult, some of it is complicated, and I have abandoned the attempt to make the language 'non-sexist' because it made the exposition just too clumsy. Please accept my apologies right at the start.

To set up a choice of strategies for each party, let us say that A has a choice of two routes: one goes through a wide valley and one through a narrow defile. B likewise has a choice: he can lie in wait where the road goes through the wide valley or where it goes through the narrow defile. Each has to commit himself to one course or another in advance. That is to say, neither can make his own choice depend on knowledge of the other's choice. The sketch map constituting Figure 2.I. illustrates the set-up. Tables and figures are collected at the end with all the tables preceding all the figures.) What is the nature of the strategic problem facing A and B? Clearly, if A goes in one direction and B goes in the other, A gets through every time with the gold. This is the outcome most preferred by A and least by B. Now compare the two outcomes that have A and B converging on the same spot. First, if A goes through the wide valley and B lies in wait there, B has some chance of making off with the gold but not a very good one; second, if both converge on the narrow defile, B has a very good chance of making off with the gold. If we assume that the only thing either of the parties cares about is what happens to the gold, we can deduce that the first of these alternatives is preferred by A to the second, whereas the second is preferred by B to the first.

The rankings of the four possible outcomes are shown in Table 2.I. It will be seen that A's ordering of outcomes is exactly the reverse of B's. (In statistical terms, there is a perfect inverse correlation between them, giving a Spearman coefficient of –1.) This is what defines a zero-sum game. To put it another way, a zero-sum game is one with the maximum possible amount of conflict built into it, because the better an outcome is for one player the worse it is for the other.

It is important to recognize that this is <u>all</u> that is meant by calling a game zero-sum. The term is therefore misleading in as far as it suggests that there is some stuff whose quantity is conserved and that the question is how it is to be divided. It is true that, as I have told the story so far, there is a fixed amount of gold that either gets through to its destination or is appropriated by the bandit. But this is not an essential feature. I have already said that the cargo takes the form of sacks of gold dust. It would therefore be a plausible addendum to the story to say that a short sharp struggle where the road runs through the narrow defile will result in some gold dust getting lost, while a more protracted fight where the road runs through the wide valley will result in more getting lost. (Bogart fans will be reminded of <u>The Treasure of Sierra Madre</u>.) None of this makes any difference to the strategic problem for A and B, as long as it leaves the rank ordering of outcomes for both the same.

If A prefers outcome x to outcome y, we can say (following the standard terminology) that x gives A more ordinal utility than y does. We can then assign numbers to A's utilities, but with the proviso that the only thing that matters about these numbers is whether they are bigger or smaller than one another: this is precisely what is meant by saying that they are only ordinal utilities. Table 2.2

uses the simplest array of whole numbers possible to lay out the strategic set-up so far described.

What we have in Table 2.2 is a payoff matrix, and such matrixes are an essential tool of game theory. I shall be making extensive use of them in the rest of this book. It is therefore worth taking a little time to understand how to read a payoff matrix. In Figure 2.2, A is the Row player: his choice determines whether the outcome lies in the upper or lower row. B is the Column player: his choice determines whether the outcome lies in the right or the left column. Putting together their choices determines a single cell in the payoff matrix as the outcome. For example, if A picks the lower row and B picks the right column, A and B both go to the wide valley, which produces –1 for A and +1 for B. (In a payoff matrix, Row's payoffs are shown before Column's in each cell, which can be memorized as the Roman Catholic convention.)

The payoffs assigned in Table 2.2 are the simplest possible, and those in each cell do sum to zero. This may, despite all warnings, tempt the reader to think that there is still some stuff – this time utility – whose quantity is being conserved. To see that this really is not so, look at Table 2.3. This represents exactly the same situation, because the rank order of each player's payoffs remain the same. Having made the point, I shall revert to the simpler numbers in Table 2.2. What should A and B do to give themselves the best attainable payoff? Obviously, A would most prefer the outcome in which he goes by one route and B lies in wait at the other (and he does not mind which route it is), but he has no way of bringing this about single-handedly. Of the two outcomes in which he is intercepted by B, he prefers the one in which it happens in the wide valley, giving him a payoff of –1 as against the –2 he gets if it happens in the narrow defile. It is therefore rational for A to go via the wide valley, since this cuts his maximum loss to –1. This is his maximin strategy: it maximizes the minimum he can get in the worst case scenario.

(A brief excursus on maximin. In his 1958 article, 'Justice as Fairness', John Rawls justified the choice of maximin as the decision-rule in the Original Position – or its functional equivalent in that article – by saying that each agent should choose principles 'as if his enemy were to assign his place'. Clearly if your enemy is to assign you your place, you had better maximize the minimum payoff, since you know in advance that that is the one you will get. However, Rawls provided no reason for saying that the choice should be made on this assumption, and subsequently dropped the claim. He has kept the notion that the appropriate decision-rule in the OP is maximin, but has never succeeded in motivating it adequately, as numerous critics have pointed out.)

What about B? He would most like to encounter A at the narrow defile, but, again, he has no way of bringing that about. His best bet is to track A's calculations, anticipate that A will go through the wide valley and take up his own position there. Thus, the outcome that follows if each follows the best strategy is the one represented in the lower right hand cell in Table 2.2 , in which both converge in the wide valley, with a payoff of –1 to A and 1 to B.

2. Mixed Strategies

This analysis, however, assumes that the two players have only pure strategies available to them, so that they must for some reason commit themselves to always doing the same thing. We could give this a concrete interpretation if we said that there is currently no way of getting a wagon from the gold fields to San Francisco, so Wells Fargo has to construct a road. Assume that the costs of building a road along either route are the same, so the only consideration affecting the choice is the risk of a hold-up. Then, pretty clearly, Wells Fargo will pick the broad valley and B will make plans accordingly.

In general, there is no reason for the parties to be stuck with pure strategies. Why shouldn't A think along the following lines? 'If B is always going to go to the broad valley, I can go to the narrow defile and be certain of escaping his clutches.' But if A starts making a habit of this, B, missing him at the broad valley, will arrange an ambush at the narrow defile, thus getting his highest payoff and giving A his lowest. All this presupposes, of course, that the parties have to take independent decisions. If it is fairly quick for B to shift from one position to the other, he can wait and see which road A takes and respond accordingly. Under the circumstances, A is in the same position as Wells Fargo building the road. He will minimize his losses by always taking the road through the wide valley, on the natural assumption that B will always be waiting for him, whichever route he takes. Notice, however, that this is not A's unconditionally best move. It remains true that he is better off if he goes to the narrow defile while B goes to the wide valley, or if he goes to the wide valley and B goes to the narrow defile. The point is simply that A has no reason for expecting B to be so obliging as to bring about either of these outcomes if B has the option of always going to the same place as A does. The advice given by game theory is predicated on the presumption that the other player is rational, which in this context means capable of figuring out the utility-maximizing strategy. I shall return to this point in the next section.

If the two parties have to decide independently where to go, we can conclude that the two pure strategies picked out before are not in equilibrium against one another. If B sticks to his best reply to A's best pure strategy, A has an incentive to depart from it; and, if A departs from his best pure strategy, B has an incentive to depart from his best reply to A's best pure strategy. This sort of 'if I do this and he does that . . . ' thinking is precisely what game theory is designed to help with. Von Neumann's conjecture, which led to the creation of game theory, was that, if we allow for mixed strategies, every zero-sum game has an equilibrium. In other words, there is always a pair of strategies, defined over randomized choices, such that each is the best reply to the other, in the sense that neither party can do better against the other party's optimal strategy than play his own optimal strategy. (Thomas Schelling, The Strategy of Conflict, p. 175n., quotes Von Neumann as writing: 'As far as I can see, there could be no theory of games . . . without that theorem.')

To explain how these mixed strategies can be calculated, we have to introduce stronger measurements of individual utility than the ordinal ones so far used. We need to be able to say by how much A prefers one outcome to another, and the same for B. In the rest of this book I shall for the most part need only ordinal utilities, and the numbers that I assign to outcomes will simply be chosen to make it easy to see that the story being told is internally consistent. Every now and then, I shall need to be able to say that, of three outcomes (x, y and z), the distance between x and y is much less than that between y and z. For example, compared to the status quo (x), an interruption in of supplies of strawberry jam for a week (y) is far less of a deprivation than (z) an interruption for the same period in the supply of electricity and (because the pumps stop working) the supply of water.

I do not believe that there is anything very arcane about the interpretation of such a statement, and I cannot imagine that anybody can seriously pretend not to understand it. However, long after philosophers had abandoned it, economists were trapped in a naive form of verificationism, according to which assertions were meaningless unless they could be cashed out in terms of actual or hypothetical observations. In the case of human beings this meant (actual or hypothetical) behaviour. Von Neumann and Morgenstern developed a measure of utility along these lines, treating assertions about intensities of preference as assertions about hypothetical choices between gambles. Only adherence to a discredited dogma could lead somebody to say that the meaning of an assertion about preference intensity is carried by statements about hypothetical choices between gambles. But for some purposes it is useful to be able to say that intensities of preference may be thought of as being expressed by hypothetical choices among gambles – on condition that the parties have neutral attitudes to gambling as such, so that they are neither risk-averse or risk-seeking.

In the previous section, all that was necessary to set up the analysis was that the parties agreed on the proposition that, if both converged on the narrow defile, B was more likely to make off with the gold than if they converged on the wide valley. What we now need is that A and B assign probabilities to each outcome. These probabilities could be different: A might have one estimate and B another, and (to make things even more complicated) each might form an incorrect estimate of the other's estimate. This would make for a very messy game. I am interested only in presenting the simplest possible case, so I shall suppose that the probabilities are common knowledge. That is to say, each has an estimate, these estimates are the same, and each knows that they are the same. Only in the past couple of decades have game theorists put knowledge conditions in the forefront of their analyses. Although recent developments are too complex for my purposes, I shall nevertheless have occasion to point out a number of times later in this book how important to the way a game is to be analysed is the presence or absence of common knowledge. I shall, in particular, draw attention to the frequency with which common knowledge assumptions (of a strong and implausible form) are relied on without being explicitly stated or a fortiori defended.

I shall stipulate specifically that both A and B believe that B's chances of catching A are twice as great if they both go to the narrow defile as they are if they both go to the wide valley. These ratios are all we need. To make it more concrete, however, we may as well add that in the first case B has four chances in five of getting away with the gold and in the second case only two chances in five. If A and B have neutral attitudes to gambling, this implies that A and B would both be indifferent between (1) both going to the wide valley and (2) a fifty-fifty chance of (a) both going to the narrow defile and (b) each going to a different place. (Because they have a neutral attitude to gambling, it follows also that they would be indifferent between (1) two chances in a hundred that both go to the wide valley plus ninety-eight chances that they both go different ways and (2) one chance in a hundred that both go to the narrow valley plus ninety-nine chances that they go different ways: this illustrates that the assumption of neutral attitudes to gambling is hardly trivial.) In terms of Von Neumann and Morgenstern utilities, this means that each attaches twice the utility/disutility to the outcome where both go to the narrow defile as each does to the outcome where both go to the wide valley.

Table 2.2 can be brought back to illustrate this, since the outcomes have this ratio for both players. However, to make the computation more perspicuous, let us say that A gets 10 units of disutility from losing the gold and B gets 10 units of utility from capturing it. Then we have the payoffs set out in Table 2.4. Let me once again emphasize that this ascription of utilities is arbitrary. Any set of numbers that retained the right ratios would be equally good. We could if we liked say that A loses 10 units of utility from losing the gold while B gains 100 units of utility from getting it. B's numbers would then be 80 and 40 instead of 8 and 4, but this would make no difference. Thus, although these are cardinal utilities, in as far as the ratios have to be preserved to keep the same game, they are not interpersonally comparable cardinal utilities of the sort that the utilitarian calculus requires. In fact, I cannot emphasize sufficiently strongly the following: nothing in game theory turns on interpersonal comparisons of utility. All we ever need to do is ask how each player rates outcomes – which may or may not include preference intensities for that player. We never have any occasion for comparing intensities across players. Perhaps B does care a lot more about the outcomes than A: after all, it isn't A's gold, whereas if B gets it he has all the benefit. It may then be that the ascription of 100 units to B against 10 to A for the utility/disutility of gaining/losing the gold is accurate in terms of interpersonally comparable cardinal utility. But nothing in the analysis of strategies follows from this.

(A qualification that is not really a qualification: if one party's utility depends on that of another, interpersonally comparable utilities could go into the utility measures entering the payoff matrix. But this is not really a qualification because the strategic analysis turns on the entries in the payoff matrix, whatever lies behind them. I shall occasionally call attention to the problems raised by interdependent utilities of this kind, but I shall not try to incorporate them formally into the analysis at any point. The idea behind this book is to provide the most elementary treatment of every issue, with complexities beyond its scope indicated with sporadic hand-waving like this.)

We can now, finally, ask what is the optimum strategy for A and for B. Eschewing algebra, as promised, I can state the condition for an optimum strategy for each player as follows: it is one such that the expected minimum payoff for a player is the same, regardless of the choice that that player makes. Thus, if we focus for a moment on A, he should face the same minimum expected payoff if he goes to the wide valley or if he goes to the narrow defile. His minimum payoff when he goes to the wide valley is – 4, and his minimum payoff when he goes to the narrow defile is – 8. To equalize his minimum expected payoffs by adopting a mixed strategy, he needs to go to the wide valley two-thirds of the time and to the narrow defile the remaining one-third of the time. (He could implement this by throwing a die before setting out on each occasion and taking the route through the narrow defile if a five or six came up, otherwise the wide valley.) One-third of – 8 is 2 2/3, and so is two-thirds of – 4, so his minimum expected payoff either way is – 2 2/3. This is a significant improvement on the – 4 that he gets by following the pure strategy of going to the wide valley every time.

B similarly needs to find a mixed strategy that equalizes his minimum expected payoffs from the two choices open to him. If he goes to the wide valley two-thirds of the time and the narrow defile the rest of the time, he can guarantee himself an expected payoff of 2 2/3 either way. (Two-thirds of 4 is 2 2/3 and so is one-third of 8.) This is, of course, worse for B than the 4 that he got from intercepting A every time in the wide valley. But the whole point is that there was nothing to make A go to the wide valley all the time, and if A did elude him B would come away with 0, as we can see in Table 2.4. The recommended mixed strategy of 2 2/3 is the highest minimum expectation he can hope for. (Of course, on any given occasion, he will get 8, 4 or 0; but _ex ante_ or in the long run – depending on how we construe probability – he cannot get less than 2 2/3.)

A's mixed strategy is his maximin strategy: it maximizes his minimum expected payoff. It is at the same time his minimax strategy: it minimizes B's maximum expected payoff. In other words, by playing the recommended strategy, A can hold B down to an expected payoff of 2 2/3. No other mixed strategy could hold B to a lower one. Indeed, any alternative would give B a chance to do better. Similarly, B's recommended mixed strategy holds A down to an expectation of – 2 2/3, and any other mixed strategy gives A a chance to do better. That the maximin and minimax mixed strategies of any given player are the same is the key result established by Von Neumann. It holds only for zero-sum games: what makes non-zero-sum games interesting is precisely the fact that one outcome can be better than another one for me without being worse than the other one for you. (R.B. Braithwaite's 'solution' to the problem of the trumpeter and the pianist, in his Theory of Games as a Tool for the Moral Philosopher, discussed in my Theories of Justice, turned on the difference between the maximin and minimax strategies of the two parties.)

What is the significance of the equivalence of maximin and minimax in zero-sum games with mixed strategies? It means that, if A plays the one-third/two-thirds strategy, he cannot do worse than get – 2 2/3, and B cannot do better than get 2 2/3, regardless of what B does. If B goes to the wide valley all the time, the payoffs are – 4 for A and 4 for B two-thirds of the time and 0 the rest of the time, for an expectation of – 2 2/3 for A and 2 2/3 for B. If B goes to the narrow defile all the time, the payoffs are – 8 for A and 8 for B one-third of the time and 0 for the rest of the time, for an expectation of – 2 2/3 for A and 2 2/3 for B. If B follows his own optimal mixed strategy, the payoffs are – 4 for A and 4 for B four-ninths of the time (2/3 x 2/3), - 8 for A and 8 for B one-ninth of the time (1/3 x 1/3), and 0 the rest of the time. Since 4/9 x 4 = 16/9 and 1/9 x 8 = 8/9, B's expectation is 24/9, which reduces to or 2 2/3. A's expectation is, of course, – 2 2/3. The expected payoffs come out, it can be calculated, at – 2 2/3 and 2 2/3 for A and B respectively for any mixed strategy that B plays. A parallel analysis can be carried out for the case in which B plays his two-thirds/one-third mixed strategy, illustrating the way in which he gets 2 2/3 and A – 2 2/3 whatever A does.

It may be thought that, if the expected payoffs are the same whatever B does, as long as A plays his optimum strategy, there is no advantage to B in playing his own optimum strategy. This is true as far as it goes. But the point is that it remains true only as long as A does continue to play his optimum mixed strategy. As soon as B departs from his own optimum mixed strategy, he gives A an incentive to depart from his, giving A the chance to do better and make him worse off. Let us suppose that B goes to the narrow defile half the time instead of one third. Then A would be better off going through the wide valley all the time, improving his expectation from –2 2/3 to –2 and lowering B's from 2 2/3 to 2. Suppose instead that B goes to the narrow defile only a quarter of the time instead of one third. Then A can do better by going to the narrow defile all the time, again raising his expectation to – 2 and lowering B's to 2.In exactly the same way, if A departs from his optimum mixed strategy, he leaves himself open to exploitation by B. Rational players would therefore (by the definition of rationality as expected utility-maximization) both play their optimum mixed strategies.

3. Judging Game Theory

Let me repeat a point made earlier. The strategy recommended by a game-theoretic analysis is not recommended unconditionally: it tells you what is the best move against the other player's best move. Imagine that A, in defiance of common sense as well as game theory, insisted on taking the road through the narrow defile every time. Then, obviously, B would do best by always lying there in wait. Analogously, a strong chess player playing a weak one may be able to win quickly with a series of moves that would lose against another strong player. Books containing advice about how to play chess tend not to offer this kind of advice because they are concerned with asking what is the best move in anticipation of the best counter-move.

Ian Shapiro and Donald Green, in their book Pathologies of Rational Choice Theory (Yale University Press, 1994), criticize rational choice theory (of which game theory is a large part) because it fails to predict. Hardly any social science predicts, however, in the sense of prophesying the future. The most that social scientists aspire to is to retrodict (that is to say, 'predict' past events in terms of other things in the past), and this is what positivistic accounts of 'explanation' amount to. We know how to maximize the explained variance, in this sense of 'explanation': throw everything in as an independent variable that you can measure and tweak the slopes and intercepts until you get the biggest correlation coefficient with the dependent variable. (I simplify, of course, but this is close enough for the present purpose.) That the results defy interpretation, as they sometimes do, does not impugn the status of an explanation of this kind.

If you are in the prediction business, in this sense, there is no point in wasting time on game theory. Game theory predicts outcomes if its premises are all met: the players must be 'rational' (which means that they understand the theory intuitively if not formally), the payoffs to each player must be as postulated, and the structure of the game (knowledge conditions, moves available, etc.) must also be as postulated. This is, of course, as much of a tautology as the assertion that somebody whose arithmetic was impeccable would always add up a column of numbers correctly. But in both cases we have a real, though conditional, prediction.

Consider the requirement of 'rationality', understood as following the utility-maximizing strategy as deduced from the theory. Clearly, if game theory were supposed to produce predictions beyond the narrowly-specified ones I have talked about, it would be an objection that most people do not know any game theory and may well not be all that good at intuiting it. But is this an objection to the theory of games? If it is, it is equally an objection to the theory of chess. I doubt if one in a thousand recreational chess players very often makes a move (in the middle of a game, anyway) that would be recommended by a grandmaster, or even the highest level setting on a reputable chess program operating on a PC. If we thought that the theory of chess was supposed to be predictive, we would have to pronounce it a catastrophic failure. But that is not its point. What it can do is illuminate the strategic problems thrown up by chess, and the theory of games can do the same in a more abstract way.

If I were asked to 'predict' (i.e. retrodict) what would actually happen in the example discussed in this chapter, I would be inclined to put my money on A and B converging every time on the wide valley. There are two reasons for this. The first is that, although A might see that in principle he could do better by sometimes going the other way, it was hard to know what to do with this idea until Von Neumann came along. (Almost uniquely in the history of ideas, nobody has, to the best of my knowledge, managed to dredge up an anticipation of Von Neumann's fundamental theorem.) The second reason is that, even if the wagon-driver had the right insight, he might reasonably doubt that his bosses would be able to share it, in the absence of Von Neumann's theorem. He might well therefore fear that, if he went to the narrow defile a few times and got away with it, he would still lose his job if he went that way and was caught: his bosses might say he was 'just lucky' the other times and hold him responsible for losing the gold the last time. None of this in my view undermines the claim of game theory to have told us something interesting that we didn't know before.

Let me conclude this discussion by taking up the problem of ascribing utilities, because it is absolutely critical in everything that follows. Green and Shapiro, as I will note several times in subsequent chapters, seem to assume that, if game theory is to be any use, we must be able to read off the players' utilities from the mere description of a situation, where this simply describes what each gets in the way of money, injuries, jail sentences and so on. But this is an absurd demand. Clearly, if utilities are incorrectly ascribed to the players, then game theory will not predict correctly, even if the players are 'rational' in the sense defined. Thus, in the example I have been using, I have assumed throughout that the utilities of the two players do take a zero-sum form. This is actually quite implausible. If we were to look at the payoffs of organizations behind them (Wells Fargo and, let's say, Bandits Incorporated), they might well be zero-sum in that they are entirely predicated on whether or not the gold gets through or is captured. But the individual players are likely to be intensely interested in something else: their chances of getting killed or injured in the struggle over the gold.

Wars are archetypically zero-sum at the level of states (more or less) in that the more one wins the more the other loses. At the level of individual combatants, however, this is highly implausible. Even if they would all prefer their side to win, they can all see that the chance that they will personally alter the outcome is so small that it can be disregarded. Each, therefore, has an incentive to avoid getting killed or injured, even if doing so is infinitesimally bad for his side. Game theory does here yield a prediction: it predicts that, provided soldiers on opposite sides can somehow communicate a willingness to engage in mutual restraint in prosecuting the war, a reciprocal scaling down of hostilities can occur. Robert Axelrod, in The Evolution of Co-operation, points out that exactly such informal truces did occur in the First World War where opposing forces were in continuous occupancy of the same positions. For example, crossroads on the other side would be shelled, to satisfy the top brass – but only on a set timetable, so that nobody got hurt.

By analogy, the courier and the bandit might (explicitly or tacitly) co-ordinate on a strategy in which the courier hands over the gold peacefully with a frequency that is not so high as to arouse suspicion and lose him his job. An obvious frequency to pick is the two times in five that, we have stipulated, he loses it in a fight if both go to the wide valley. (We might suppose that, before they cut the deal, they have fought often enough to make this probability common knowledge.) The point is, of course, that we can analyse this revised statement of the situation only as a non-zero-sum game, because the parties do not now have strictly opposing preferences for outcomes. I shall now turn to non-zero-sum games.

3. NON-ZERO-SUM GAMES

1. Degrees of Conflict and Co-operation

A zero-sum game is, as I said in the previous chapter, a game that maximizes conflict, because the better an outcome is for one player the worse it is for the other. All other games are non-zero-sum games, which immediately suggests (what is in fact the case) that they provide a far richer field of study. The rest of this book will be concerned with non-zero-sum games. If a zero-sum game is, by definition, one of pure conflict, it follows that a non-zero-sum game cannot be one of pure conflict. To put the same thing positively, a non-zero-sum game must be one in which there is at least one outcome (defined by the choices of A and B) that both judge to be preferable to some other outcome (also defined by their choices). It is easy to overlook this. A Prisoner's Dilemma – to be discussed in a moment – is certainly conflictual, but what gives it its point is that there is an outcome (neither confess) that both prisoners prefer to another one (both confess). Similarly, if I threaten you with some sanction unless you comply with my demand, this is certainly setting up a conflict. But it is a pointless exercise unless there is some outcome (you comply, I don't carry out the threat) that we both prefer to another one (you don't comply, I carry out the threat). I shall discuss power relations of this kind in the next chapter. I simply want for the moment to emphasize the general point that every non-zero-sum game must have embedded in it somewhere a possibility of co-operation in at least the limited sense I have specified. If it does not, it is not a non-zero-sum game.

Robert Axelrod put forward in his (1970) book Conflict of Interest a way of measuring the degree of conflict between the players in a game. For my present purpose, something more informal will do. Let me simply suggest, then, that non-zero-sum games can be categorized as more or less conflictual (or, conversely, less or more co-operative) according to the degree to which the preference orderings of the parties over outcomes diverge. If there is no divergence, there is no conflict. Imagine two drivers approaching one another on a straight road. If they are both indifferent between both driving on the left and both driving on the right, we have the payoff matrix shown in Table 3.1. What they care about is avoiding a collision. The rule of the road for any given place is a norm (or a law, but one that does not need enforcement) prescribing which of the two equally good choices each driver should make. This rule or norm creates a convention.

In a two-person case, the payoff matrix given in Table 3.2 should present even less of a problem, since both players prefer the same unique outcome – both drive on the right – perhaps because both vehicles are constructed with left hand drive. If we consider an n-person version of the same problem, however, it may not be so easy to arrive at the mutually preferred outcome. For an n-person game, we can still continue to deploy the matrix in Table 3.2 if we interpret A as a single person deciding which side of the road to drive on and take B as 'how everybody else in this society drives'. From A's point of view, the 'choice' made by B is parametric: that is to say, it will not be affected by what A does. If 'everybody else' drives on the right, A is better off driving on the right, but if 'everybody else' drives on the left A is better off driving on the left.

Since the payoffs tell us that they are all better off driving on the right rather than on the left, why might they not? A society in a 'state of nature' would have no difficulty in adopting the preferable rule of the road. But if the existing rule of the road prescribes driving on the left, switching it presents obvious difficulties. The Swedes made the switch overnight, after a lot of preparation, at a time when the main requisite was changing traffic signs. Britain could have done it too before it started building motorways. Now that the rule of road is given (literally) concrete embodiment in the arrangement of cloverleaf connections between motorways, a change would be very expensive and also extremely disruptive. The QWERTY keyboard is another example, familiar to everyone, of an inefficient equilibrium that everybody is locked into. It was designed with the deliberate intention of being awkward, so as to prevent the early users of typewriters from going too fast and jamming the keys. This constraint has not existed for many years, but changing to a more efficient keyboard would require an effort of co-ordination that is apparently infeasible.

We can now introduce an element of conflict by moving on to the payoff matrix represented in Table 3.3. Here, two friends prefer eating together to eating separately, but one would prefer to eat at a Mexican restaurant and the other at a Chinese restaurant. (Let us ignore the possibility that they live on the Upper West Side and thus have the cultural advantage of being able to eat at a restaurant that offers both Mexican and Chinese cuisine.) We can see from Table 3.4 that their first and second preference orderings are opposed: A prefers Mexican/Mexican to Chinese/Chinese while B has the reverse preference ordering. However, both prefer either of these to eating alone. Of the two remaining outcomes, A obviously prefers eating at the Mexican restaurant (which implies that B eats at the Chinese one) while B prefers eating at the Chinese restaurant (which implies that A eats at the Mexican one). Thus, they agree on the order of the two eat-alone outcomes. But what really matters is that for each person the best eating alone option is worse than the worst eating together option. We can describe this as a case in which the parties gain by co-ordinating their actions, but in which one party does better than the other. Another reminder: saying this does not require us to commit ourselves to statements about interpersonally comparable cardinal utilities. All we mean here is that, if the two people eat together, there are two possibilities: one is that A gets the outcome that he most prefers while B gets only his second-best outcome, and the other is that B gets the outcome he most prefers while A gets only his second-best outcome.

If it is simply a matter of two people choosing a restaurant, we can imagine them flipping a coin or (if this is a regular event) alternating the right to choose. But once we think about the n-person version, we can see that the equilibrium may very well get locked into one of the two possible outcomes that are preferred by everybody to the non-cooperative ones. For example, once norms about the relations between the sexes or the etiquette of race relations are in place, it is advantageous to everybody to observe them. But these norms may well be (and usually have been) strongly biased in favour of one group. (For a discussion of biased norms, see chapter 8 of my Justice as Impartiality and Edna Ullman-Margalit's The Emergence of Norms.) Social norms can be changed if enough members of the disadvantaged group are prepared to accept the costs of breaking them in the hope that this will destabilize the norm and pave the way for the emergence of a more advantageous norm. Commonly, however, a change in the norms requires legal intervention.

A constitution may itself be thought of as a convention for co-ordinating the system for taking decisions binding on a society. (See Russell Hardin, Liberalism, Constitutionalism and Democracy.) Any constitution, however, works to the advantage of some interests and against the advantage of others. Constitutions can survive despite this bias as long as even those who could expect to do better under some alternative set of arrangements prefer continuing to live under it to the kind of civil disorder that is typically involved in shifting to a new equilibrium.

Another large-scale phenomenon illustrating biased co-ordination is capitalism. (It may be surmised that there is in fact some connection between capitalism and constitutionalism, though I shall not explore it here.) Capitalism results in extraordinarily unequal payoffs for individuals, but its productivity depends on expectations of stability (especially among owners of capital), with the consequence that serious attempts to remove the bias are liable to make most people worse off. I shall return to this issue in chapter 10, where I shall suggest that an analysis in terms of the power of owners of capital is also valid.

We can inject more conflict into the situation if we make more of A's payoffs follow an order that is the reverse of B's payoffs. Table 3.5 presents a situation that may be familiar to some readers. Here, A and B share an apartment. As before, conflict is moderated by the fact that they both put the same outcome last: here, the worst outcome for both is living in total filth and squalor. Rather than endure this, each would rather clean the apartment all by himself. (We see the ordering of outcomes for each in Table 3.6.) However, cleaning the apartment by himself is for each only his third preference. Both would rather they both share the job of cleaning (the second preference of both), but best of all for each would be to have the other do all the work. Thus, while A and B both put the not clean/not clean outcome last, their preferences over the other three are a mirror image of one another: A most prefers that B cleans, next that both clean, and third that he cleans himself, while B's preferences are the reverse of this.

A reasonable prediction is that each will indicate an unwillingness to clean the apartment unless the other co-operates, so that they finish up sharing the work. But if A could somehow convince B that under no circumstances will he do any cleaning, B is left with only the options of cleaning the apartment himself or living in squalor. (Suppose, for example, that for some reason only Saturdays are possible for cleaning the apartment and A makes an irrevocable arrangement to be away every Saturday, or A tells B that, as a Seventh Day Adventist, it would be violating his religious beliefs to work on a Saturday.) If B is a rational utility-maximizer, he appears to be condemned to clean the flat himself, since that gives him the outcome he prefers out of those available. There is, on the face of it, something peculiar about this conclusion. To explain what it is, however, we need to be able to think about long-run strategic calculations, and I shall postpone that topic until later in this chapter.

This payoff structure also defines what is known as a game of 'Chicken', after one possible interpretation of it. 'Chicken' is (or was) a game allegedly played by youths in the American Southwest. Two participants drive cars towards one another along the dividing line in the middle of a straight road. If, at the last minute, one pulls over to the right, while the other keeps going straight ahead, the one who swerves is 'chicken' and the other gets the glory. If both swerve, there is a slight loss of face by both, and if neither swerves there is a high speed collision in which both die. As will be seen in Table 3.7, the payoffs can be represented with a payoff matrix identical with that of the apartment-cleaning game. The ordering of payoffs is thus the same, though we may suppose that the gap between the other payoffs and the worst (in terms of intrapersonally comparable cardinal utilities) is greater in this case. Once again, however, I must emphasize that the strategic situation flows from the ordering of the payoffs and nothing but the ordering of the payoffs.

According to the standard analysis of a 'chicken' game, you should be able to 'win' by ostentatiously lashing the steering wheel (so that the car cannot do anything except go straight ahead) with a chain secured by a padlock and then throwing away the key. However, the fact that you have done so has to be communicated to the other driver and he has to believe it. And if by any chance he has already done the same it is a recipe for catastrophe whether or not he knows what you have done. I shall return to this topic of precommitment in the next chapter, since it is critical to the analysis of power.

J.F. Nash introduced a now canonical version of what is in essence the same problem in his early (1950) paper 'The Bargaining Problem'. In this, he supposed that two people have to divide a certain fixed sum of money. If they can agree on any particular division of it, that is what they both get. If they cannot agree, they get nothing. Let us make the sum a thousand dollars and stipulate that any division must be expressed in whole dollars. On the assumption that neither player has any reason for agreeing to an outcome that is just as disadvantageous as the non-agreement outcome, neither will offer the other a split that gives him $1,000 and the other nothing. However, if A offers B $1, B is better off taking it than getting nothing, so it is a possible offer by A. Table 3.8 lays out the preference orderings for this game, on the (crucial) assumption that both players prefer more money to less, regardless of any other aspect of the outcome. This does not mean that they both care equally about the money (an interpersonal comparison) or that money is linear with Von Neumann/Morgenstern intrapersonal cardinal utility, so that (for example) both are indifferent between $300 for sure and a 50-50 chance of $600. All it means is what it says: each prefers more dollars to fewer dollars. There is, if you like the lingo, a monotonic relation between money and utility.

As in the cleaning-the-apartment and 'chicken' examples, both players agree on the worst outcome – non-agreement with a payoff of zero dollars each. But, also as in those cases, all their other preferences are exactly the reverse of each other's. Nash proposed a solution to this game which made it turn on the parties' Von Neumann/Morgenstern utilities. With linear utilities of the kind I just illustrated, for example, his theory advised the parties to take $500 each. I have discussed this kind of game – now known as a Nash bargaining problem – and Nash's own solution to it in Part I of Theories of Justice, and I shall not say anything about it here. The only point to make is that there is nothing compelling the parties to accept Nash's (or anybody else's) solution. If A could convince B that he will accept nothing except the outcome that gives him $999 and B $1, B will, as a rational utility maximizer, have to accept it, since it is still better than the non-agreement payoff of O. (A might, for example, write down this offer and leave immediately, arranging to be incommunicado until the expiry of the period within which the agreement has to be reached if it is to be valid.)

Both of the conflictual payoff structures discussed so far have in common the feature that both parties put the same outcome last. The significance of this can be appreciated if we observe that eliminating it gives us the famous (or infamous) Prisoner's Dilemma (PD) payoff structure, illustrated in Table 3.9.

(William Poundstone's Prisoner's Dilemma is an entertaining account of the invention and application of the PD, interspersed with a sketch of the life of John Von Neumann, the father of game theory.) As I have already said, any non-zero-sum game must by definition have some congruence between the players' preferences, and here it lies in their both having the same second and third preferences. Both prefer the payoff if both play the Co-operate move than the payoff if both play the Defect move. What gives the PD its bite is, however, that the first preference of A is the last preference of B, and vice versa.

The order of payoffs displayed in Table 3.10 is what makes a game a PD. Notice that in its general form we can state it so that the two alternatives 'Co-operate' and 'Defect' are simply defined abstractly as whatever actions A and B might take that jointly bring about payoffs that for each player have the ordering given in Table 3.10. The concrete embodiments of 'Co-operate' and 'Defect' can be anything that makes the payoffs come out like this: what A does and what B does need not even be the same kind of action.

2. The Simultaneous One-Shot Prisoner's Dilemma

Why is this structure of payoffs called a Prisoner's Dilemma? The answer (which can be found in Poundstone and elsewhere) is that A.W. Tucker made up a story to exemplify the payoff structure that involved two prisoners, and the name has stuck. Since it will be helpful to give a concrete example of a PD at this point, I may as well give Tucker's. We are told that the District Attorney is holding two prisoners in different cells. He tells each that he has enough evidence to get them both jailed for a year. However, if the prisoner he is talking to confesses to some larger crime, he will be released and his confession will be used to send the other prisoner down for ten years. If both confess, he promises to seek only five years jail for each. It is not at all apparent why either should believe any of this. (What's to stop the D.A. from using one confession to convict both, or going for ten years if they both confess?) Thus, the critical problem of the assumption of common knowledge of the structure of the game rears its head again. Let us ignore that problem, however, and construct the payoff matrix as in Table 3.11.

Notice that the payoffs in Table 3.11 are stated in terms of jail sentences. We cannot therefore assume that the preferences of the players have a PD structure. To arrive at that conclusion we have to stipulate (along the same lines as the stipulation that the players in the Nash bargaining game preferred more money to less) that each of the two prisoners cares only about his own time in jail and prefers less to more. The stipulation may fail: in the Nash case, the parallel stipulation performs poorly as a predictor of behaviour. We can build into the structure of a Nash bargaining problem the position that arises if A writes down his offer and leaves by turning the original set-up formally into an ultimatum game. To do this we simply give A the power to propose a division to B, and specify that B has to take it or leave it. Experimental evidence suggests that the players in the B position would sooner leave it than take it if the proposed division is too grossly unequal, and that the players in the A position anticipate this by not proposing grossly unequal divisions. Within the present framework, this implies that (at any rate within the range of outcomes that social psychologists can afford to offer their subjects) the B players do not in fact prefer more money to less unconditionally. Similarly, experiments with monetary payoffs of a PD form quite often result in the subjects both playing the co-operative move. (I shall say more about this later.) Within the present framework, this again means that objective payoffs do not correspond to the subjects' utilities.

Let us assume for now that both prisoners unconditionally prefer less time in jail to more. Then the ordering of their payoffs is as shown in Table 3.12. If we look at A's ordering among outcomes, we can see that he is better off confessing regardless of what B does. Of course, he would prefer it if B does not confess, but he has no control over what B does. The point to be emphasized is that, if B does not confess, A is better off confessing; and, if B does confess, A is still better off confessing. Confessing is for A a dominant strategy: it is his best choice regardless of what B does. By a similar analysis, B is better off confessing than not confessing regardless of what A does. It therefore pays both to confess, yet the consequence of that is that they both finish up with their third preference (confess/confess), whereas if neither had confessed they would have finished up with their second preference (not confess/not confess)

Although the whole point of the PD is that each player has a dominant strategy, I am amazed sometimes to see in print even now the idea that what drives the prisoners towards confession is uncertainty about what the other will do. This is quite irrelevant to the logic of the PD. As long as the utility payoffs are captured by the payoff matrix constituted by years in jail, confessing is a dominant strategy. If the other prisoner confesses, you get 5 years instead of 10 years by confessing yourself. If the other prisoner does not confess, you go free instead of being jailed for a year. Either way, you are ahead by confessing.

If we define rationality as maximizing your payoff, and assume that jail sentences correspond to payoffs, there is no question that confessing is rational in the PD. Attempts have been made to get round this conclusion, but it is easy to see that they must be doomed. It is, of course, true (1) that both prisoners would do better if neither confessed than if both confess and (2) the outcome in which they both confess is the outcome of rational choices by both prisoners. It is this feature of the situation that led to its being called a dilemma. But it is not really a dilemma in as far as confessing is the dominant strategy. It has also been called a paradox. The phenomenon of cycling among group preferences (to be discussed in chapter 8) is also often referred to as a paradox. But the most we can say in either case is that the analysis has potentially disconcerting implications. There is nothing paradoxical about the fact that the pursuit of individual self-interest may be collectively disadvantageous, or the fact that consistent individual preferences over policies may aggregate to inconsistent collective preferences. (See Russell Hardin, Collective Action, ch. 9.)

Let me again emphasize that the PD can easily be 'solved' (i.e. the con-confess/non-confess 'right answer' achieved) if the jail payoffs in the matrix are not the utility payoffs. If each prisoner adheres to the maxim of 'honour among thieves', for example, both may prefer to behave honourably. This could be a motive for not confessing even in the absence of an expectation that the other will not confess (so-called 'Kantian motivation'). Alternatively, a prisoner could be motivated by the expectation that the other prisoner will not confess plus the thought that he would then be a dirty rat to confess.

It may be noted that, if the payoffs really do correspond to jail sentences, and nothing but jail sentences, it makes no difference to the outcome whether or not the prisoners can talk before they make their decisions: the optimal strategy is to pledge non-confession, encourage the other not to confess, and then confess. However, experiments by social psychologists using a PD game (defined in terms of monetary payoffs) have found that allowing subjects to talk before they make their decisions increases the probability that they will play the co-operative move (i.e. the one corresponding to non-confession). The increase is striking: standardly, the rate of co-operation more than doubles. What this suggests is that the utility payoffs may diverge from the monetarily-induced ones under the influence of such factors as trust.

Green and Shapiro treat this kind of result (for which they set out the evidence) as an 'anomaly' (p. 90). But there is nothing in the least anomalous about the fact that monetary payoffs do not necessarily induce corresponding utility payoffs. Nothing in the theory of games commits anybody to the assumption that people must unconditionally prefer less jail to more or more money to less. If they do not, the game is no longer a PD – that's all. One would get the impression from Green and Shapiro that to say this is to indulge in some kind of shady manoeuvre. But no other theory is expected to work if the conditions it stipulates do not obtain. Galileo's theory of falling bodies does not predict that a feather and a lead weight released together from the roof of the World Trade Center will hit the ground at the same moment. Its prediction holds only conditionally and a condition which is not met here is the absence of air resistance. Nobody seems to think that saying this constitutes a desperate piece of fudging to cope with the failure of the theory to predict the phenomena. If it isn't a problem for Galileo that his theory works only where its premises hold, it isn't any more of one for game theory.

What sort of a game do we get it both A and B would prefer to co-operate provided the other does? Turn to Table 3.9, and substitute for '1,1' in the upper left cell '3,3'. We then get, of course, a change in the ordering of payoffs in Table 3.10. A and B both now put 'A plays C, B plays C' on top. Being an exploiter ('A plays D, B plays C' for A and 'A plays C, B plays D' for B) moves down to second place. The other two outcomes remain in third and fourth places. This game is a hybrid between the co-operative game depicted in Table 3.2 and a PD. It shares with that co-operative game the feature that both players like the same outcome most. But it departs from it by otherwise having PD payoffs, which means that there is still the possibility of getting the 'sucker' payoff (I co-operate, you defect). This has no analogy in Figure 3.2, where both players lose equally from failure to co-ordinate on the best outcome.

What shall we call this? There is a usage of the term 'Assurance game' according to which all games with a single outcome best for everybody are 'Assurance games': everybody has a motive for picking the move corresponding to that outcome as long as he is assured of the others' doing so. However, this means that a lot of games with very different payoff structures are all lumped together. (Michael Taylor, in The Possibility of Cooperation, p. 38, gives three different payoff matrixes with divergent strategic implications.) I shall, for the purposes of this book, reserve the term 'Assurance game' for the particular payoff matrix in question here.

Since we are dealing here with a one-shot game, the transformation of PD payoffs defined over jail or money into Assurance game payoffs defined over utility must arise from a norm of reciprocity: the desire to do good in return for good, the desire not to abuse another's trust, and so on. As we shall see, however, Assurance payoffs can arise in iterated games out of calculations of long-run self-interest. The Assurance game (understood as I have defined it) is thus a crucial aspect of applied game theory – arguably more important than the PD.

The PD is not necessarily an undesirable phenomenon, from the larger point of view. It may be in the public interest that the prisoners confess. More significantly, the structure of a competitive market is an n-person PD, as far as the sellers are concerned. Consider a cartel, which keeps up prices by collusion above the level that would generate normal profits. In the absence of an effective mechanism for enforcing fixed prices, each seller has an incentive to shave the price, thus producing almost as much profit per item while getting a larger market share at the expense of the others. Hence the instability of cartels. Once the number of sellers becomes large enough, the possibility of maintaining collusion to keep up prices disappears and the rationale of Defect all round comes through.

At the same time, it is worth noticing that there is an application of the case of the findings of the social psychologists. The hypothesis advanced to explain these findings is that talking to one another fosters trust. On condition (and this, of course, already implies that the monetary payoffs are not the real payoffs) that each subject is willing to cooperate if the other does, each is then more likely than otherwise to make the co-operative move in the game. Analogously, Adam Smith said that businessmen rarely meet, even for purposes of recreation, without some scheme contrary to the public interest (i.e. for some kind of restraint of trade) coming out of it. Even formal price-fixing cartels typically cannot apply actual sanctions against defectors: the price-fixing is maintained through a common sense of its advantage and trust that the others will not cheat. (It is not accidental that cartels in the USA were called 'trusts'.) However, price-fixing (whether it arises by formal agreement or informal collusion) is a game that is played out over time, so that each party can respond to what the others do. It is thus best analysed as an iterated PD, and I shall discuss this in section 4.

3. The Sequential One-Shot Prisoner's Dilemma

The classic PD involves simultaneous decisions, as I have said. But I have also pointed out that it still pays to Defect even if you know that the other player will Co-operate. Because of this, the logic of the PD still operates when it is played sequentially. (N.B. we are still talking about only one play.) Hobbes, for example, thought that contracts with simultaneous performance (I hand you the potatoes, you give me the apples) are feasible even in the state of nature. But he added, famously, that 'covenants without the sword are but words, and of no help to secure a man at all'. By a covenant he meant an agreement in which one party performs first and the other undertakes to perform later (I give you the potatoes now, you promise to hand over the apples when they have ripened and been picked). The obvious problem with this is that, in the absence of an enforcement agency, you gain no advantage by doing your part when the time comes, at any rate if we model the interaction as a one-shot play. (I shall come back to Hobbes and covenants later in this chapter, when I come on to discuss iterated games.)

Hume provided a carefully-stated illustration of the problem of obtaining co-operation in a sequential PD in the absence of mutual trust (Treatise, pp. 520 – 1 in the Selby-Bigge/Nidditch edition). 'Your corn is ripe today; mine will be so to-morrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon your account; and should I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security.' Notice that Hume takes the trouble to specify that neither farmer's utility is a function of the other's: they have 'no kindness' for one another. Each one's payoff depends entirely, therefore, on two things: the amount of his crop that is harvested (either by him alone or with the help of the other), and the total amount of labour that each expends (either just on getting in his own crop or on that and on helping his neighbour as well).

We can represent the set-up by a decision tree. (Such trees have the curious property of growing downward.) Let us posit that the value to each farmer of the whole of his own crop is 12 units of utility, and that if both farmers collaborate the whole crop is harvested. Let us also for simplicity say that by himself each can harvest only half his crop, for a value of 6 units. We also have to factor in the cost of the effort involved in harvesting. Say that harvesting your own crop by yourself costs 2 units. Then, since that yields half a crop and collaboration yields a whole crop, we can reasonably say that collaborating with the other farmer also costs 2 units. (We assume, therefore, no advantages from collaboration as such – we can suppose them each working half the land simultaneously but independently.) We now have the materials for the decision tree in Figure 3.1. Once again, let me add the caution that the numbers themselves play no part in the analysis, and that only the ordering of payoffs counts. The numbers are there simply to make it easy to check that the story is internally consistent.

Clearly, both farmers are better off if both reap each time than if each only reaps his own field, just as both prisoners were better off not confessing than confessing. But, just as in the simultaneous-choice PD, if the payoffs assigned really reflect the farmers' complete payoffs, they will not co-operate. Although this may seem obvious, it is worth spelling it out so as to introduce the important idea of backward induction, according to which the players start by looking at the outcomes from the last move and then working back from there. Suppose, then, that we look at the left hand node after A and B have both worked on A's fields (marked 2 in the figure). B can see that A does better by not reaping B's fields when the time comes than by reaping, since he loses two units of utility by doing the work and gains nothing. B therefore predicts that, if A is rational (i.e. maximizes utility as defined by the decision tree), A will not help him with his crop when the time comes. Hence, the predicted final outcome for B of 'A reaps, B reaps' is +2, which is the payoff he gets if A does not reap. He compares this with the end-point that occurs by taking the rightward path from the initial starting point (the node marked 1 in the figure). Because he does not now lose 2 units by helping A, this gives him +4, and is thus superior. The upshot is that B will prefer not to accept A's proposal that they collaborate in reaping their crops. For by backward induction he arrives at the answer that he will end up worse off by accepting the offer than rejecting it. Of course, he could be wrong about this. Perhaps A would have helped him get his crop in when the time came. But this could come about only if the payoffs ascribed to A have been incorrectly assigned (e.g. A would feel that not helping B after B had helped him would make him a dirty rat) or A is not rational in the sense defined (i.e. utility-maximizing).

4. The Iterated Prisoner's Dilemma

It may be said, correctly, that we err in modelling the two farmers' problem as a single play. Presumably, the same situation can be anticipated to come up again next year. A may therefore have a motive for helping B this year after B has helped him, in the hope that they will co-operate again next year. Hume made precisely this point in the paragraph following the one I have just quoted. 'Moralists or politicians', he says, cannot hope to make people altruistic (so that the two farmers would help one another out of 'kindness') but they can 'give a new direction' to 'natural passions' such as 'selfishness and ingratitude'. They can 'teach us that we can better satisfy our appetites in an oblique and artificial manner, than by their headlong and impetuous motion' (p. 521). Hume now formulates the principle of tit-for-tat (i.e. conditional co-operation) as the best way of playing an iterated PD. (See Robert Axelrod, The Evolution of Co-operation.) 'Hence I learn to do a service for another, without bearing him any real kindness; because I forsee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv'd him, and he is in possession of the advantage arising from my action, he is induc'd to perform his part, as foreseeing the consequences of my refusal.' (ibid.)

(There is a subtlety in this passage which I shall for now leave aside. This is the point that I may reciprocate a good turn done by you not so much because I hope you will co-operate with me again as because I hope that others, seeing that I am a trustworthy type, will co-operate with me in future. I shall take up this kind of reputational effect in chapter 6, section 3.)

An annual play of the game between the two farmers, it may still be said, leaves a lot of room for uncertainty about whether they will still be in the same situation next year. If we speed up the rate of iteration, we can show more clearly the case for co-operation. Suppose farmers A and B have to cross one another's land every day to get to their flocks and herds. Closing the gates in the other's farm is slightly burdensome, but having the gates left open allows the sheep and cows to get out of their pastures and is highly inconvenient to the farmer concerned. In these circumstances, I don't think any of us will find it hard to imagine that both farmers will be meticulous about closing one another's gates.

Even here, however, backward induction raises its ugly head. Suppose the lease of farmer A runs out in five years time and is non-renewable, and both farmers know this. Then on the last day, if neither farmer has any 'kindness' for the other, neither has any incentive to shut the other's gates: we are back to the one-shot case. But since the gates will be left open on the last day, regardless of what has gone before, there is no incentive for either to shut them on the next to last day. And so by backward induction we can conclude that, even when the expiration of the lease is still five years away, neither has any incentive to close the other's gates. I believe this to be a genuine paradox in that the premise of backward induction seems sound in principle, yet it generates absurd results. (Not everybody agrees the results are absurd – Jon Elster for one – but I think they are.) Numerous attempts have been made to 'solve' the problem, but I have yet to see one that seems to me satisfactory. What we do know (see Axelrod, The Evolution of Co-operation), is that most people do co-operate over most of a finite (e.g. hundred-play) PD in experiments.

Fortunately, in any case, real life rarely presents us with known finite numbers of iterations, and so the problem generated by backward induction seldom occurs. It is worth noting that Hobbes understood the logic of the iterated PDs and argued that one of the laws of nature, which held even in the state of nature, was 'that men perform their covenants made'. So, if the other person has performed, you should do your part, as long as you can do so safely. For that conduces to peace, which is an overriding interest of everybody. This raises the question, however, why a sovereign is necessary after all: if people could see the logic of the laws of nature, wouldn't they co-operate without sanctions to guarantee peace? Hobbes gives three answers. One is that each person in the state of nature is the final judge of what his own safety demands which leads to insecurity. A second is that you were not obliged to be the first to perform a covenant, only to perform second if the other has performed first, so there may be few covenants. The third is that some people are driven by the desire for glory or eminence. For them, in game-theoretic terms, life is not a non-zero-sum game in which we can all gain by co-operating. Rather it has a zero-sum element defined by status: I can only be up if others are down. (As Gore Vidal once said, 'It is not enough to succeed – others must be seen to fail.') Pursuers of eminence are bound to upset the apple cart because Hobbes believes most people will settle for equal status but not for inferior status. Those who pursue superior status will therefore create conflict.

(Michael Taylor points out in The Possibility of Cooperation (pp. 142-3) that it is not necessary in order to turn a co-operative game into one with a zero-sum payoff structure that agents should be solely concerned with their relative eminence. It is enough if this enters into their utility functions. Taylor, however, models the pursuit of relative eminence as a 'difference game' in which people's (second-order) utilities depend, positively or negatively, on the difference between their (first-order) utilities and those of others. This would , clearly, require the first-order utilities to be interpersonally comparable cardinal utilities. It seems to me, however, that Hobbes believes the relevant difference to be not in 'felicity' as such but in possession of the external marks of superiority themselves: 'the eminence of the Faculties of Body, or Mind' and 'Riches, Reputation, Friends, and the secret working of God, which men call Good Luck'.)

Let me now return to the cartel case, as promised, looking at it formally now as an iterated PD. The point here is, again, that over time each participant has an opportunity to respond to the actions of others. Suppose that prices can be changed only every three months. Whether a cartel is explicit or tacit, what keeps it going is the expectation that any price-cutting by one member will be met by price cuts by the others. The largest amount of time during which a firm can have a price advantage over the others is three months. This is a system of deterrence not by a threat of punishment (to be analysed in the next chapter) but by a threat to wipe out the advantage. And, of course, after the advantage has been wiped out, everybody – including the price-cutter – is worse off than before.

The conditions most favourable to the success of a cartel are (1) high barriers to entry, so that existing firms are not too worried that high profits will attract new entrants, (2) small numbers, (3) an expectation that the existing firms will be around a long time in the future, (4) a tendency to take long views and (5) a market which is not extremely price-sensitive. All of these conditions, it may be noted, are met by Ivy League universities. Until they were stopped by an anti-trust suit, they used to get together for a tuition fee fixing meeting each year. But even in its absence, it is easy to see how, with all the conditions that favour it, tuition fees can be pushed up in concert. Even if fees do not rise in perfect synchronicity, a university that pushes a little ahead of the others (a 'price-leader') can afford a small drop in applicants, confident that others will catch up or push ahead further next year.

Hume relied on following suit to explain how justice (by which he meant respecting others' property) could arise out of self-interest in the absence of enforcement (Treatise p. 498). 'Every member of society is sensible of this interest [in maintaining a system of justice] : Every one expresses this sense to his fellows, along with the resolution he has taken of squaring his actions by it, on condition that others will do the same. No more is requisite to induce any of them to perform an act of justice, who has the first opportunity. And thus justice establishes itself by a kind of convention or agreement; that is, by a sense of interest, suppos'd to be common to all, and where every single act is perform'd in expectation that others are to perform the like.' In the same way as the members of a cartel are kept in line (if it works) by the fear that any individual failure to stick to the pricing convention will be worse for everybody, so, according to Hume, the members of a society are kept in line by a fear that failing to adhere to the justice convention will bring about results so bad that the losses from it will swamp any short-run gains to be made by breaching it. For, 'without justice, society must immediately dissolve, and every one must fall into that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be suppos'd in society' (p. 497). (Although Hume is careful never to mention his great predecessor, it is obvious that this is precisely Hobbes's case for keeping covenants shorn of its contractarian trappings. It is also noticeable how close Hume comes to repeating Hobbes's description of the 'state of nature'.)

5. Breaking Down the Prisoner's Dilemma into Small Steps

To conclude this chapter, it is worth noticing that there is an important lesson to be drawn from the analysis of the iterated PD about the way in which to induce co-operation. This is that the more a project can be broken down into small steps the more easy it will be to gain co-operation for it. Thus, if the members of the cartel could react to a price change the next day, the advantage of stealing a march on the others would be one day's sales instead of three months' sales.

What may be regarded from one point of view as a single project can sometimes be broken down into a series of discrete moves, leaving each party free to react to the choices made on the preceding move. Thus, one way of looking at the reduction in American and Soviet nuclear arsenals was to think in terms of a big treaty such as the Strategic Arms Limitation Treaty. But such treaties have the drawback that neither side is prepared to trust the other to carry it out second if it carries it out first, and there is not (and cannot be, since the parties are the two superpowers) any third-party enforcement of the treaty, applying significant sanctions against whichever side fails to comply. The obvious alternative is that neither side reduces its arsenal in one fell swoop. Each scraps a few weapons and waits to see the other side's response. (In the nature of nuclear weapons, small reductions on either side do not affect the strategic balance, since each side can still annihilate the other many times over. At very low levels, this might have been a problem, but that is, unfortunately, a problem that is still a long way off.) For this incremental reduction in nuclear arms, no treaty is necessary in the first place. In fact, it may be suggested that the treaty is, at best, simply a way of co-ordinating expectations. In the absence of a treaty, the United States and the Soviet Union did indeed succeed in reducing their arsenals in some periods by precisely this method of matching arms reductions.

Hume illustrated the point with another story, also involving two farmers (Treatise p. 538). 'Two neighbours may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each others mind; and each must perceive, that the immediate consequence of his failing in his part is the abandoning of the whole project.' The point about this is that each keeps an eye on the other's contribution, and stops work if the other does. Thus, as Russell Hardin says (Collective Action, p. 132) 'although it seems like a one-shot problem, Hume's meadow will take time to drain, and the two neighbours will either both participate or both not participate . . . , so that the Prisoner's Dilemma interaction will be ongoing for a while, not strictly one-shot.' Let us even suppose that the neighbours have one mechanical digger between them, so that they cannot usefully work at the same time. One may then work with the machine digging ditches in the morning and the other in the afternoon. Provided each day's work is small enough in relation to the total, we can expect the logic of the iterated PD to come into play and induce cooperation.

There is, however, a condition for each agreeing to undertake the project, which is that each somehow makes it credible to the other that he will stop work if the other does, even if it would actually be advantageous to him to finish it himself at that point. To see why this is crucial, let us stipulate that the meadow will not show any improvement unless the job is finished. (Drains that are not completed will just fill with water.) And let us say that the whole job requires 200 units of disutility (and that each of the neighbours experiences the same amount of disutility for the same contribution), and that a half share in the drained meadow is worth 150 units of utility to each. Ex ante, this looks like a good bet, since, if they do an equal amount of work on the project, each will gain 150 units of utility for an expenditure of 100.

Suppose, however, that, when each has contributed 30 units of disutility to the job, A declares that he has no intention of doing more. If B is a non-strategic utility-maximizer, he will calculate as follows. 'We have between us already incurred a sunk cost of 60 units, leaving 140 units to do. If I finish the job myself, it costs me 140 units and I get 150, since the meadow is then drained.' At this point, the two neighbours are in exactly the same position as the two people sharing an apartment whose situation I analysed earlier in the chapter. Either would be better off finishing the job all by himself than leaving it unfinished.

Where this case differs from the apartment-cleaning case is, however, that in this case B would have been better off not entering into the agreement in the first place if it results in his finishing the job himself. For he ends up by making an investment of 170 units of his own disutility in return for a gain of 150 units: the 30 units he expended jointly with A at the beginning and the 140 he expended by himself after A pulled out of further co-operation.

A, of course, does very well out of this trick, getting 150 units of utility for an expenditure of only 30. Since the two neighbours are symmetrically placed, so each can see how the other could play this trick – in other words, each can put himself equally well in the shoes of A or B. By backward induction, each will therefore refuse to agree to enter into an agreement to co-operate. The only way round this is for the neighbours to convince each other that they will not behave as non-strategic utility maximizers. Thus, precommitment of some sort is critical. Each has to say that he is prepared to cut off his nose to spite his face. Perhaps the most persuasive line is for each to say that he would so dislike being taken for a sucker that he would sooner let the meadow stay undrained than allow the other the satisfaction of exploiting him. It is worth noting in passing that this connects up with Allan Gibbard's sociobiological explanation of norms of fairness in Wise Choices, Apt Feelings: common norms of fairness, whose violation arouses indignation, help to co-ordinate mutually advantageous deals on one out of the (usually numerous) alternatives that are all superior for both parties to no deal at all. (In the Nash bargaining game, for example, the obvious common norm was push the parties to a split giving them $500 each.)

4. POWER OVER OTHERS

1. Why Carry Out Costly Threats?

Having said something in chapter 3 about non-zero-sum games, let me now move on to what is, one way or another, the main topic of this book, the analysis of power. Power, as I have already remarked, can arise only in a non-zero-sum game. To see why this is so, we have only to ask under what conditions it is worth the while of A to attempt to exercise power over B. The minimum conditions are as follows. First, there must be some possible act of B (call it b2) that is better for A than the one B would choose if left to his own devices (call it b1). Otherwise, A has no incentive for trying to change B's behaviour. The second condition is that this same act (b2) must be better for B than doing the thing he would prefer to do (considered in itself) (b 1) and subtracting from that payoff the negative payoff constituted by the sanction threatened by A. This implies a non-zero-sum game.

The maximum amount of power in the hands of A occurs when the cost of carrying out the threat against B is zero. (This would be true, for example, of a country with some rockets approaching their use-by date, but real life examples are fairly rare.) If we call A's doing nothing a1, and A's carrying out the threatened sanction against B a2, then the condition of maximum power is that A's payoffs from a1 and a2 are the same. This is illustrated in Table 4.I. The payoffs in the cell constituted by B complying and A punishing are in parentheses because this outcome should not occur with rational play. It is of the essence of an effective threat that the bad thing you threaten in the event of non-compliance will be withheld in the event of compliance: if B expects to suffer the sanction whether he complies or not, he may as well not comply. A topical illustration of this point is as follows: if the Israeli government's idea of 'restraint' is shooting dead more than a hundred Palestinians and injuring thousands more, it is hard to know what anybody is to make of a threat not to show restraint. (As can be seen from the right-hand column from Table 4.I, the sanction carried out by A is assumed to reduce B's utility level by 2; subtracting that from the payoff arising from the compliant act produces a total of –1, and subtracting it from the payoff arising from the non-compliant act produces a total of 0.) As usual, all these numbers are merely a way of telling the story, and what actually matters is the ordering of the outcomes.

What makes the analysis of power interesting is that the normal case is one in which it costs A something to apply sanctions against B. This means that if B plays b1 (i.e. does not comply), A's payoff from a2 (apply the sanction) is actually lower than his payoff from doing nothing (a1). This is illustrated in Table 4.2, in which it costs A one unit of utility to punish B for non-compliance. (From now on I shall assume that the lower left cell is void.) The implication is, however, that A now has no incentive to punish B for non-compliance, because the payoff to A from doing so is –1, whereas the payoff from doing nothing is 0. If we invoke backward induction, we can see that B will realize this, so B has no reason to comply with A's demand. And, by further backward induction, we can see that A will realize this in turn and conclude that it is a waste of time even to issue the demand and make the threat.

Since this is a game played sequentially, its structure is most perspicuously displayed in a decision tree. The data from Table 4.2 underlie Figure 4.I. Here, however, we start at a further stage back. Table 4.2 was based on the supposition that A had already made the demand, and the question was whether B would comply or not and what A would do in the event of B's non-compliance. In Figure 4.1, we start with A's decision to either issue the demand or not. If the demand is not issued (the right hand branch from node 1) the payoffs are simply those in the upper right cell in Table 4.2: B does what he most likes and A does nothing about it. The left hand branch of the decision tree follows the possible outcomes if A does issue the demand. We can clearly see here the crucial point that A is worse off by choosing the right hand branch from node 3 (carry out the sanction) than by choosing the left one (do nothing). But the left branch from node 3 leads to the same outcome as that which occurs if A does nothing in the first place (the right branch from node 1), so there is no point in A's making the demand.

This appears to be a paradox. If the analysis is right, why does anybody ever attempt to exercise power – and why does anybody ever actually carry out the threat if it is unsuccessful in modifying behaviour in the desired direction? There is no paradox here, however, any more than there is in the case of the one-shot prisoner's dilemma. If we focus on one play (as we have implicitly been doing), it really is true that there is a loss to A from carrying out the threat, which undermines its credibility in the first place. So far from being a theorist's fantasy, however, it is precisely this logic that haunted the 'wizards of armageddon': the inventors of the notion of nuclear deterrence at the Rand Corporation and the Hudson Institute.

2. The Problem of Nuclear Deterrence

As a preliminary to this, let us analyse deterrence in general. So far, we have looked at a case in which A attempts to change B's behaviour from what it is now. In this alternative, A fears that B is going to change his behaviour in a way deleterious to A's interests. In Figure 4.2, the right hand branch from node 1 represents A's choice to do nothing in face of the risk that B may change the status quo. If B's payoffs are correctly characterized, we shall expect B to change the status quo under these conditions (i.e. go down the right hand branch from node 3), since he is free to do what he prefers without any fear of incurring a penalty. If A goes down the left hand branch from node 1, this means that we have an attempt to deter B from changing the status quo. If this attempt is successful, the outcome is 1 for A and 1 for B (the left hand branch from node 2), which is the same as the payoffs from the status quo in Figure 4.2. Suppose, however, that B does not comply. This means that, if A does nothing, the status quo has changed in a way that leaves A with 0 and B with 2, as shown at the bottom of the left hand branch from node 4. (This is, of course, the same as the scenario if A does not make any threat in the first place and B makes the change that A dislikes.)

Now consider the case in which A does carry out the threat to punish A. This is costly to A, so A gains no advantage from doing so and in fact becomes worse off, as we can see from the payoffs at the bottom of the right hand branch from node 4. By backward induction, we again get results parallel to those in the previous case: B can predict that A will not carry out the threat, so B has no reason for complying with it, so there is no point in A's making it in the first place. All A can do is hope that, if B has not changed the status quo up to now, he or she does not really prefer a change (in other words, the value ascribed to B from a change from the status quo is incorrect) and that the status quo will therefore continue.

It may be observed that the whole of the criminal law rests ultimately on the possibility of deterrence, if we make the assumption that in the absence of a threat of punishment some people would prefer to break the law. The analysis of deterrence is therefore at least as important for political theorists as the case of attempts to change behaviour in order to get people to do something, as against refraining from doing something. I shall return to this later, but let us for now pursue the issues raised by the idea of nuclear deterrence.

The key to the theory of nuclear deterrence was that it was essential for a country to have a second-strike capacity, that is to say the capacity to endure a first strike by the other side and still be left with enough nuclear weapons to launch a devastating strike against the other side. It might be thought that the problem of cost does not arise in launching this second strike, which functions as the punishment threatened if deterrence has failed. For there is in some obvious sense no cost in launching the missiles that have survived in their hardened silos: this is, after all, what they were built for and there is no alternative use for them. However, we can extend the concept of the cost of carrying out a threatened punishment in a plausible way. If we count in the cost of suffering retaliation by the other side in the cost of punishing, we can reintroduce the problem.

Concretely, assume that the Soviet Union, though it has always officially rejected the theory of nuclear deterrence, has absorbed the lesson that it is important to have a significant second-strike capacity. It therefore ensures that it still has plenty of missiles left after launching a first strike that will be capable of surviving a counter-strike by America's (ex hypothesi depleted) nuclear arsenal. Say that the Soviet Union does launch a nuclear attack. Let us assume (maybe counterfactually) that the 'hot line' is still intact after the Soviet first strike and that the President of the Soviet Union calls the American President to say that, if the United States bombs the Soviet Union, the Soviet Union will retaliate with a devastating further attack. (This presupposes, of course, that the Soviet first strike leaves enough of the USA standing to make what is left worth saving: having a second strike capacity should at least guarantee that, but it is pretty cold comfort.)

If we apply Figure 4.2 to this case, the USA is A and the Soviet Union B. The USA has taken the left hand branch from node 1, by threatening the Soviet Union with sanctions if it changes the status quo (i.e. bombs the USA). At node 2, the Soviet Union has taken the right hand branch, and gone ahead anyway with bombing. The USA, after the telephone call, is at node 4. This implies that its bluff has been called, since carrying out its threat is worse at this point than doing nothing. How can the United States avoid getting into this situation? If we look again at Figure 4.2, we can see that it would do so if it could somehow eliminate the choice that is to be made at node 4, so that it can only go down the right hand branch. By backward induction, the Soviet government will then realize that, if it chooses to go down the right hand branch at node 2, this will inevitably lead to the outcome with payoffs A – 1, B 0. Since this is worse for the Soviet Union than the status quo for it (which give it 1), it does not therefore launch a first strike, and, by further backward induction, the United States gains by making the threat.

How could the United States give plausibility to this meta-threat – that its threat will be carried out, even though it would be better off not carrying it out? Clearly, by persuading the Soviet Union that there will be no room for discretionary decision-making at node 4: a first strike by the Soviet Union will automatically trigger a counterstrike and that's all there is about it. As the 'wizards of armageddon' recognized, this could be implemented by a 'Doomsday machine' – a device that triggered off a nuclear strike automatically under certain conditions and could not be overridden. (This device was, it may be recalled, incorporated in Stanley Kubrick's Dr Strangelove, in which it was supposed that the Soviet Union had one that would be triggered by a single nuclear bomb.) Alternatively, each team in charge of a weapon could simply be told to fire it under certain conditions and disregard any orders to the contrary. Indeed, it would be quite enough to create uncertainty to let it be known that each team had instructions to fire its missile unless it was told not to: given the doubts about the survivability of command and control systems, this would be enough to provide a reasonable fear of retaliation.

A different approach would be for the American President to convince his Soviet counterpart that he is crazy, and would launch a second strike however irrational it might be to do so. According to a recent biography of Richard Nixon, (Anthony Summers, The Arrogance of Power [New York: Vintage, 2000]), he encouraged Henry Kissinger to tell the Soviet leaders that he (Nixon) was totally out of control and might do anything if provoked. This was, as Summers shows in terrifying detail, very close to the truth, so it may well have been believed. Thomas Schelling, in The Strategy of Conflict (p. 22), draws attention to the effectiveness of a credible show of irrationality: 'If a man knocks at a door and says that he will stab himself in the porch unless given $10, he is more likely to get the $10 if his eyes are bloodshot.'

Undoubtedly, if nuclear deterrence really prevented war during the height of the Cold War, it operated via this kind of existential uncertainty. However, it seems to me far more likely that deterrence was irrelevant because neither side would have preferred the other to be destroyed. (In other words, the payoffs attributed to B in the end-point of the right hand branch from node 3 were not as shown, whether the United States or the Soviet Union was regarded as B.) The United States, after all, could have used its nuclear weapons with impunity before the Soviet Union developed bombs and delivery systems, and it would require a good deal of paranoia to imagine that the leaders of the Soviet Union were so unrelievedly depraved as to prefer the destruction of the United States to its continued existence. The general point here is one that I shall have occasion to return to in my discussion of political power in Part II of this book. There is no advantage in trying to exercise power unless there is some feasible change in somebody else's behaviour that would be beneficial to you or some feasible change in somebody else's behaviour that would be deleterious to you. If what they will do anyway is what it best suits you for them to do, you have no occasion for trying to exercise power. Indeed, even if you have no power at all, you may be fortunate enough to get a lot of the outcomes you want if those who do have power use it in ways that benefit you.

3. The Indirect Exercise of Power

Given the frailties that attempts to exercise power are liable to, how does it ever come about that people are got to change their behaviour or to refrain from changing it by threats? Mechanical devices are one answer, and have some uses beyond the (fortunately fanciful) Doomsday Machine. Once, trying to find a parking space so as to go to a political philosophy seminar at the University of Southern California, and frustrated by a series of car parks equipped with spikes that (as the notice advised) inflict SEVERE TIRE DAMAGE to deter people from driving in through the exit (thus circumventing the barrier at the entrance), my colleague remarked that this was somehow more persuasive than stationing an attendant with a pitchfork at the exit and warning everybody that he would plunge it into their tyres if they tried to drive into the car park past him.

Alternatively, things may be set up so as to make the infliction of the sanction for non-compliance almost or completely costless. Through most of history, a frequently used way of doing this was the taking of hostages. Rodin's well-known group statue of the Six Burghers of Calais commemorates an example: in 1346, England captured Calais and took six prominent citizens as a surety for good behaviour (as defined by the English) by the citizens. In earlier times, a defeated ruler might be required to hand over his eldest son (or maybe more family members) to the conqueror, to be kept in his court and killed in case the defeated ruler breached the terms of peace imposed on him. The Soviet

Union deterred prominent citizens who left the country from defecting to the West by not allowing their spouses or children to be out of the country at the same time. Thus, although it might have been quite hard to apply sanctions to those who defected, it would have been quite easy to inflict penalties on those left behind.

Sometimes, where the parties were equally strong, power was exercised by each party over the other and hostages were exchanged. This gave each side the opportunity to deploy low-cost sanctions against the other. On related lines, anthropologists explain the peace-inducing effects of elaborate systems of cross-cousin marriage as a kind of mutual hostage-taking by different (and potentially hostile) lineages. The nuclear stalemate was sometimes described as one in which each government held the other's inhabitants as hostages.

Hostage-taking brings up a point that is of general significance and will come up later again. This is that sanctions threatened in an attempt to exercise power do not have to fall directly on the person or body to whom the demand is addressed. The English hoped that the citizens of Calais cared enough about their six leading members to be deterred from causing trouble. The ancient rulers and those of the Soviet Union hoped that the recipients of their (explicit or implicit) demands cared enough about the hostages in the hands of the party making the demands to comply. Similarly, if we analyse nuclear deterrence in terms of an exchange of hostages, we are in effect saying that the government of each side (who might well be safe in bomb-proof shelters) were motivated not by their own personal survival but by that of their citizens.

Again, if the child of wealthy parents is abducted, and the safe return of the child is made conditional on the payment of a ransom, the object is to secure compliance with a demand made on the parents by threatening to harm the child. Although it is the child that stands at risk from the kidnappers, there is nothing that the child could do that would be of advantage to them. (This is not quite true in that they might ask the child to plead with the parents to pay up, on a tape or in a handwritten letter; but this is incidental to the objective of extracting money from the parents.) Thus, we have to analyse the case as one in which the kidnappers seek to exercise power over the parents through their ability to kill or injure the child.

This gives rise to a slightly subtle point that will prove to be important in chapter 10. If you can lower somebody's utility, this means that you have power over them. For you can threaten to lower their utility unless they comply with some demand, and thus attempt to exercise power over them. But the ability to lower this person's utility is of no value to you if there is nothing they could do that you would want them to do. The ability to lower their utility can nevertheless be used so as to attempt to exercise power over some third party, as we have seen. In this case, the victim is in your power in that you have the ability to lower his utility. By making the threat to a third party, however, you are not attempting to exercise power over the victim.

There is a more general conceptual point about power to be made. Attempting to exercise power is threatening to do something that will lower the utility of the addressee of a demand. This can be done directly or (as we have just seen) indirectly by threatening to lower the utility of somebody on whose utility the utility of the addressee depends. But the ability to lower somebody's utility in the event of noncompliance is also the ability to lower it unconditionally. This ability is power because it could be the basis of a threat. But if it is simply used to lower the victim's utility, that is not an exercise of power, though it could be said to be a demonstration that power existed. Once the victim's utility has been lowered, however, the possibility of threatening to lower it obviously disappears and with it the power.

Suppose, for example, that you owe me a large sum of money and that if I call it in immediately you will be bankrupted. That gives me power over you, because I might demand that you comply with some demand or face bankruptcy. But I might simply decide one day to call in the money anyway. That is not an exercise of power, because there is no demand involved. It is, as I have said, a demonstration that I did have the power, since it shows that, if I had made a demand, the threat backing it up was credible. In the circumstances, however, what happened was that I chose to give up the power and get the money (or however much you had) instead.

All this depends, let me concede, on a stipulative definition of 'power over'. But I believe that the phenomenon of trying to gain compliance by the use of threats is distinctive and important enough that we should reserve a special form of words for it. Thus, parents can beat or lock up children (at any rate until they are strong enough to resist). But I want to deny that simply beating or locking up children is an exercise of power over them: in itself it is simply an application of superior strength or brute force. The exercise of power over somebody entails some modification of his will. The threat of punishment by parents, where this is contingent on non-compliance with some demand, is an attempt to exercise power over their children. Hurting them or depriving them of liberty in itself is not.

4. The Rationale of Punishment

Mechanical devices ('doomsday machines', spiked barriers) can remove discretion. Social devices such as hostage-taking can lower or eliminate the cost of carrying out threatened sanctions. Neither has wide enough applicability, however, to get rid of the basic problem with power that I have identified. This is the problem that, if the threat fails and carrying it out is costly, it is disadvantageous to carry it out. Assuming that this is common knowledge, it appears to undermine the credibility of the threat and (by a further operation of backward induction) render it pointless even to make it in the first place.

It has been argued that a version of this problem afflicts any act-utilitarian account of the rationale of punishment. For act utilitarianism is purely future-orientated: whatever has happened up to now, we should always do the thing anticipated to have the best overall consequences. According to D.H. Hodgson, in his book Consequence of Utilitarianism, act utilitarians could not operate a system of deterrent punishments, because, once the crime has been committed, utility will be lower if it is punished than if it is not. The criminal is, by design, worse off, and in addition there may well be a cost of carrying out the punishment: keeping somebody in prison is very expensive. Even if the punishment is not itself costly to carry out, it is still likely that there will be a net loss of utility from punishing as compared to not punishing. Hodgson therefore deduced that punishment as a practice could be sustained only in a society of rule utilitarians, who simply acted in every case on the rule with the best consequences, or a society of non-utilitarians who thought (as Kant did for example) that there was a duty to punish criminals, regardless of any utility to be expected from it. (Kant notoriously said that, if a society were about the dissolve, it should execute anybody on death row first.)

The obvious counter to Hodgson's argument is that, even if there is a net loss of utility from punishing this particular law-breaker, there is still a prospective gain in utility from punishing him, as an example to others. We have to start somewhere, it may be said. Hodgson, of course, anticipates this common-sense response and finds it wanting. According to him, an act utilitarian can never say 'We have to start somewhere', because each case is a fresh start and there is never a good act utilitarian reason for doing something that reduces utility. Thus, on Hodgson's analysis, an act utilitarian judge is in an analogous position to that of the neighbour in the meadow example (in the previous chapter) who is condemned by non-strategic utility-maximization to finish the job of draining the field all by himself.

To the best of my knowledge, nobody has found Hodgson's argument compelling. For, despite everything he says to the contrary, it still seems that the judge can say 'The announced policy is to punish those found guilty of breaking the law. Although there is a loss of utility that inevitably arises from our punishing you, there is also a gain in utility from our making an example of you and thus deterring others from breaking the law. Since the expected utility from having a law-abiding society is so great, the loss entailed in carrying out the punishment is greatly outweighed by its deterrent effect.' To check the reasonableness of this, imagine that the two neighbours owned hundreds of meadows. If A stopped work on the first field after expending a mere 30 units of utility, B could surely say 'It's worth writing off this meadow if necessary rather than set a precedent in which I'm left in the lurch at this point in every meadow and have to finish it by myself, with a net loss of utility each time.' What gave the meadow example its punch was that, even if the process of draining it was broken down into small steps, the project of draining the meadow was itself conceived of as a one-off undertaking and was thus in the end still a one-shot PD. The revised version eliminates that feature and makes co-operation more secure.

I am not especially concerned with utilitarianism here – though utilitarianism does throw up many problems involving strategic interaction – but I am concerned with the strategic analogue of Hodgson's objection to the coherence of act utilitarianism. Here, we do not have to factor in the loss of utility to the person who is punished. All we need to ask is whether or not, taking the long view, it is advantageous to the government to announce a set of penalties for law-breaking and set up an apparatus (police, judges, prison officers etc.) to catch law-breakers, try them, and punish them if found guilty. Although it remains true that any given punishment reflects a failure of deterrence in that individual case, it contributes to what in the literature of criminology is called 'general deterrence'. It is not hard to see why carrying out the punishment can be presented as worth doing, bearing in mind that this is the only way in which anybody else (or the same person, indeed) can be led to believe that future law-breaking will be punished.

The problem of infirmity of purpose is, once again, to be overcome by creating a machine. This time, however, it is a machine constituted by people playing assigned roles – in other words, an organization. Once they have been instructed that their job is to fight fires, catch criminals, wage war or whatever it may be, ideal Weberian bureaucrats will not ask themselves if they would be better employed doing something else. As in the Charge of the Light brigade, 'Theirs not to reason why/Theirs but to do or die.' So, if the job description requires it, the executioner will indeed execute the last people on death row before society dissolves, and an Eichman will pride himself on a well-run death camp. As this illustrates, bureaucracy can be pathological in certain circumstances, but the point that is relevant here is simply that (for better and for worse) it is certainly possible to set up a piece of machinery that will act according to rules, regardless of the short-run utility to the society of doing so.

Even without such machinery, a strong enough sense among the members of a society of the advantage of enforcing laws should, in principle, lead to co-operation in their enforcement. It has been suggested that there is some deep problem for Hobbes's sovereign by institution in getting started. (See, for example, Gregory Kavka, Hobbesian Moral and Political Theory, ch. 6.) But on Hobbes's premises there should be no problem. First, we should never forget that keeping covenants tends to peace (the surest way to individual self-preservation), and that includes the covenant among the people that authorizes the sovereign to act on their behalf. Therefore, each subject is obliged to obey the sovereign's laws – as long as they do not put his own survival seriously at risk – for exactly the same reason as people are obliged to perform their covenants in the state of nature, as long as they can do so safely: the gain in security from improved prospects of peace outweighs whatever immediate profit there may be in reneging. Thus, if everybody understood Hobbes's theory and acted on it, enforcement would be necessary only when the sovereign commanded somebody to do something that posed a serious threat to his life. This proviso poses notorious problems for Hobbes in trying to explain why a soldier is obliged not to run away from the battle field if his life is seriously at risk from fighting. But with that exception it seems reasonable to think that a prudent sovereign should be able to avoid demanding self-sacrifice. Of course, the sovereign cannot (as Hobbes acknowledges) oblige anybody to give himself up for punishment, because that would violate his 'right of nature'. But if everybody understood and acted on Hobbes's theory this contingency would not arise, because nobody would break the law in the first place.

Suppose, nevertheless, that somebody does break the law. There is still no great problem. We should remember that professional police forces are an invention of the nineteenth century. Before that, the 'hue and cry' was an important instrument for the apprehension of criminals: the victim would raise the alarm and public-spirited citizens would go in pursuit. The Hobbesian sovereign could avail itself of this. We must again bear in mind that you are obliged to keep your covenants even at the cost of some inconvenience because the greater good of contributing to peace outweighs the inconvenience. The covenant setting up the sovereign authorizes the sovereign to demand assistance in enforcing the law, and every able-bodied person is obliged to help, as long as he can do so safely. If it is said that there is always an element of risk in pursuing somebody, so nobody is ever obliged to help, Hobbes's whole system collapses, since paying somebody to do it cannot make any difference: the 'right of nature' still remains. (This is precisely Hobbes's difficulty with soldiers.) It seems to me, however, that Hobbes could reasonably assume that in many cases the risk of grave injury or death from joining in a collective pursuit would be low enough not to trigger the 'right of nature'.

Like the argument for the obligation to keep covenants in the state of nature discussed in the previous chapter, this all turns on accepting Hobbes's premises, and I shall discuss their plausibility in chapter 6.

5. Power and Strategy

Because there is one state facing a large number of subjects, the logic of general deterrence is strong. Even if experience suggests that a certain individual is not prevented from pursuing a criminal career by the risk of punishment, it is still worth punishing him. There would clearly be significant weakening of general deterrence if it became known that you would go scot free provided you could build up a reputation for incorrigibility. But this kind of general deterrence is a special case – albeit an important one – and in this section I shall take up at a more abstract level the case of power relations when the actors take account of future payoffs as well as those inherent in a particular attempt to exercise power.

The basic points are simple. What is difficult is knowing what to do with them analytically. Thus, if you are the party making the demand, your threat to carry out some sanction if it is not met is more credible the more convincing you can make it that a great deal rides on its being widely perceived that you do carry out your threats. Thus, as Diego Gambetta points out in The Sicilian Mafia, the entire stock in trade of a mafioso is a reputation for ruthless and effective application of threatened sanctions in the event of non-compliance with a demand. In a sense, the fragility of the mafioso's position – that it rests on nothing but reputation – is precisely what gives him power. If he threatens to blow up your restaurant unless you pay him protection money, you know that he cannot stay in business unless he carries out his threat.

What is not as readily recognized, however, is that exactly the same can be said about the recipient of a demand. If you can make it convincing that a lot rides on your maintaining a reputation for refusing to give in to threats, whatever the short term cost, you make it less likely that you will be threatened. Thus, some governments have an inflexible policy of never negotiating with hijackers, and being prepared to accept whatever loss of life (among the hostages or its own troops) may arise from a policy of always trying to get the hostages out without conceding demands. Other things being equal, we may surmise on the basis of this analysis that governments with an inflexible policy will be less prone to hijacking than others. Of course, things are very far from equal, in that some governments' policies are of more interest to potential hijackers than those of other governments. It may well therefore be that there is in fact a positive correlation between the number of hijackings a government suffers and the inflexibility of its policy, because governments subjected to a lot of hijacking tend to be more inflexible as a result. All we can say is that, if the analysis is correct, a government with an inflexible policy will be subject to fewer hijackings than it would have been otherwise.

Even in the case of law enforcement, where the logic of general deterrence is strongest, a group that builds a reputation for obstinacy may be able to face down law enforcement agencies. An excellent example of such a group is provided by the Amish in the United States. In the past, Amish elders have demonstrated their willingness to go to jail rather than comply with laws that they dislike. Putting them in jail generates mainly negative publicity for those who have undertaken the prosecution. This reputation for accepting sanctions rather than complying therefore makes public officials very reluctant to attempt to enforce laws that the Amish have indicated that they are not prepared to accept.

A corollary of this line of analysis is that both those who make demands and those who are the target of them will increase their chances of success if they can consolidate disparate demands. Thus, companies turn over to collection agencies whose stock in trade is (like that of the Mafia) a reputation for carrying out threats. Conversely, kidnappers who demand a ransom for the release of their victim are helped by the fact that the demand is likely to be seen as a one-off by the family of the victim. Clearly, the more families pay up, the more attractive kidnapping becomes as a profession. But effects of this kind are not likely to weigh much with the victim's family. The only future-orientated thought that may give them pause is the fear that giving in will encourage future kidnapping aimed at them. (Elizabeth Barrett Browning's dog, Flush, was apparently being held to ransom constantly.) This thought is more likely to motivate victims of blackmail, where it may be impossible for the blackmailer to prove that he has destroyed all the damaging evidence in return for payment.

To deal with the one-off problem, Thomas Schelling suggests in The Strategy of Conflict (n. 11, p. 39) that there could be 'a law that required the immediate confinement of all interested friends and relatives when a kidnapping occurred'. The state would, in effect, consolidate the cases of all targets of kidnapping and adopt an inflexible policy on behalf of the families, whether they wanted it or not. Schelling says that this proposal is 'perhaps impractical' and it is, indeed, hard to imagine a government's getting popular support for such a policy, however rational it might be in the long run. There is, in any case, an implementation problem because families that were disposed to settle would not inform the authorities of the demand either at the time (since they would be prevented from paying) or afterwards (since they would be admitting to the crime of not reporting the demand). The attractiveness of kidnapping as a profession might therefore actually increase as a result of such a law.

6. Power and Success

Ultimately the game defined by a demand backed by a threat is a battle of wills. We can quite easily say what makes for a larger rather than a smaller amount of power: the less it costs to carry out the threat backing the demand, and the more damaging the sanction will be to the addressee of the demand, the more power the agent making the demand has. (See Keith Dowding, Rational Choice and Political Power, pp. 74 – 5.) But we still cannot simply read off the outcome from a knowledge of the power relations of the parties, even if it is common knowledge what the payoffs are. The case is not analogous with that of poker: in poker, common knowledge of everybody's hands would make the result a foregone conclusion, since everybody would know who had the best hand. Power games, in contrast, are in general indeterminate even if both the parties know everything there is to know. There is still the question left open of what they will do in the light of the information. Over the long run, we should find that, for a demand of any given size, the more power the person making the demand has in relation to the person to whom the demand is addressed, the more often the demand will be successful. But there is no mechanical relationship.

We can, in one respect, say rather more than this about the prospects of a demand's being successful. For it is an important fact about power that the size of the threat sets an upper bound to the size of the possible demand that can be made successfully, backed by that threat. ('Size', here, is measured in terms entirely of the addressee's utilities.) If you purloin my $10 watch and threaten to smash it in front of me unless I pay you $100, I will obviously prefer you to carry out the threat. If a smashed watch is no use to you and it takes effort to smash the watch, I may even get it back intact if I can convince you that it would be completely irrational for me to comply with your demand. (I am assuming for the sake of the example that smashing the watch and returning it are the only options open to you.) If the watch is worth $1,000, by contrast, I may well fork out $100 to get it back, though the outcome now becomes indeterminate in the usual way: if I can somehow convince you that I will pay nothing for the return of the watch, and it costs you an effort to destroy it, I may still get it back. We are back at the 'battle of wills' scenario.

Although the point that the size of the threat limits the size of any successful demand may seem obvious once stated, neglect of it – or miscalculation of the size of the demand – is surprisingly common and has often had momentous consequences in prolonging wars and attempts by states to suppress insurrections. Clausewitz's crucial insight was that the point of fighting is not to beat the other side's armed forces but to change the minds of its leaders about the acceptability of some demand. This is the significance of his famous assertion that war is diplomacy carried out by other means. Now that so much diplomacy is backed by the manipulation of economic offers and threats (trade and aid), the relevance of Clausewitz's point that force is merely one of the tools of diplomacy should be even clearer than it was in the time in which he wrote, when states were far less enmeshed in complex economic relationships.

If the only use of warfare (or the threat of war) is to secure compliance with some demand, it follows that the demand must be precisely formulated in advance, and must not be larger than can be secured at an acceptable cost. Thus, Clausewitz insisted that a country that starts (or enters) a war must have precise war aims and stick to them. A large part of the task of the military profession was, he argued, to prevent the politicians from engaging in what would now be called 'mission creep': the gradual extension during a war of the objectives. Democratic countries are especially liable to this pathology, because, as public opinion becomes inflamed in the course of the conflict, politicians tend to respond by expanding the demands on the other side.

Thus, for example, the demand for ' unconditional surrender' maximized the size of the demand made by the Allies in the Second World War. Quite likely nothing that the Allies would have been prepared to concede in advance could have shortened the war against Germany. But perhaps a guarantee of the Emperor's immunity (which was granted anyway) might have led to an earlier surrender by Japan. This presupposes, of course, that the object was to end the war as quickly as possible. But the same point can be turned round to satisfy anyone who believes that the American government wished to impress Stalin by demonstrating atom bombs on a couple of cities. In that case, we can say that insisting on unconditional surrender was the surest way of prolonging the war until these weapons were ready for use.

The United States seems particularly liable to underestimate the willingness of leaders and people in other countries to put up with privation, destruction and death rather than yield to demands to constraining their political independence. A trivial example that is nevertheless telling in illustrating the quality of information being fed to the government: many years ago somebody connected with the British diplomatic corps mentioned to me that, before the Berlin Wall was erected, he was assured by a CIA agent in West Berlin that the East Germans would never do such a thing because of the disruption to the subway system it would cause! Similarly, the President and the Secretary of State seem to have been genuinely surprised that Milosovic did not cave in over Kosovo in a few days after a few bombing sorties. (It seems to me doubtful that he would have capitulated even when he did in the absence of the incipiently credible threat of invasion.)

The point, formally put, is that an overwhelming superiority of power – defined in the usual way as the capacity to inflict large costs on others at low cost to oneself – does not turn into success if the demand backed by the threat is too big in relation to it. It is even less likely to be successful, especially in the case of protracted conflict, if the gain to the side with superior power from getting its way is relatively small. For, even if the cost per month is not high, it may well come to be seen as too high to be worth paying, as the cumulative costs mount up with no end in sight.

The classic postwar illustration of this is, of course, the Vietnam War. The American stake in this amounted to no more than the entirely speculative 'domino theory' – which was actually proved false when no other countries followed Vietnam into the communist camp after American withdrawal. (In fact, au contraire, the United States bombing of Cambodia put Pol Pot in power.) On the other side, the Vietnamese independence movement had been accepting heavy losses to get rid of imperial power since 1947. In that year, Ho Chi Minh said to a French friend 'You will kill ten of my men while we kill one of yours, but you will be the ones to end up exhausted.' (Frances Fitzgerald, review in Sunday New York Times Book Review, October 15 2000, pp. 14 – 15, p. 15 of William J. Duicker, Ho Chi Minh [New York, Hyperion, 2000].) Exactly the same thing happened in turn to the United States.

Similarly, overwhelming Israeli military force cannot be turned into Palestinian submission to a peace treaty that is more advantageous to Israel than a certain extent that is still to be determined but apparently exceeds the concessions (in relation to the status quo on the ground) so far offered by Israel.

The American administration has persistently failed to recognize this, perhaps misled by the belief (which may well be correct) that the old and ailing Arafat would personally be prepared to settle for the trappings of being the tinpot president of a Middle-Eastern Bantustan on almost any terms. But though his signature would be binding on anything he signed, he could not carry his people.

This is not, once again, to say that the possession of power is irrelevant to the outcome. Compared to the peace supported in numerous United nations votes by every government in the world except those of Israel and the United States (now that South Africa has changed sides) – complete Israeli withdrawal from the West Bank, and a Palestinian state with its capital in East Jerusalem and the return of all refugees and their descendants – there can be little doubt that Israel can parlay its power into something that gives it more of what it wants, as a permanent settlement of the Palestinian issue. The point is, simply, the one made by Ho Chi Minh: the 'revealed preference' of Palestinians is for losing ten (or more) of their number for every Israeli rather than accept a permanent settlement that fails to satisfy their national aspirations to an adequate degree – whatever precisely that degree may be. (For an argument that grossly unjust settlements resting on extremely unequal bargaining power are 'no more than truces', see my Justice as Impartiality, pp. 31 – 9).

Game theory, like the neoclassical economics of which it is an offshoot, has nothing to say about the preferences over outcomes that define payoffs.

These are simply taken as given. To the extent that advice can be extracted from game theory, it is impartially available to jailed criminals, gold-robbing bandits or firms operating a cartel as well as actors with more socially valuable ends. In terms of Kant's distinction, it gives rise only to a system of hypothetical imperatives of the form 'If you want this, do that.' The major piece of advice that can be distilled from the discussion in this section is that it is an error to focus exclusively on power at the expense of careful attention to utilities. Before making a demand backed by a threat (especially where the process may be protracted), the actor with the power needs to ask two questions. First, is getting this demand satisfied worth enough to me that I am prepared to make the sacrifices necessary to carry out the threat? And, second, is the demand so obnoxious to the other party that my threat is unlikely to make it seem worthwhile to it to concede the demand rather than put up with the sanction threatened?

The entire history of independence movements since the end of the Second World War (from India in 1947 to East Timor most recently) illustrates how often overwhelming power fails to produce success. Whenever the occupying state cares substantially less about keeping possession of a territory than the inhabitants of that territory do about gaining control over it, all the power in the world cannot stop the independence movement from succeeding – short of killing most of the inhabitants. (Even driving them out will work only if they settle somewhere else permanently.)

Among some territorial animals (for example, the males in some species of birds such as red grouse), it has been found that , in the vast majority of cases, the possessor of a territory will drive off a marauder as long as they are anything like equally matched physically. Thus, equal power gives rise to systematically asymmetrical outcomes. There is presumably some physiological basis for this behaviour: the defending animal secretes more testosterone or suchlike, one supposes. It is easy to see how such traits, which secure stability of territory, might have been selected for. We might be anthropomorphizing if we were to talk here about utilities, though economists who have established production and consumption functions for animals do so without qualms. But we can at any rate say that the outcome is as if the possessor cares more about keeping the territory than the marauder cares about seizing it. Human beings are more complicated than grouse: unlike some enthusiasts for sociobiology I am happy to acknowledge that. Nevertheless, especially in the era of nationalism, it looks as if something analogous may be going on.

5. COLLECTIVE ACTION AND THE SOCIAL CONTRACT

1.Collective Action Problems

Throughout chapters 3 and 4, I was dealing with collective action problems, but deliberately refrained from offering a formal analysis of the issues raised in terms of the concept of collective action so as to avoid trying to do too much at once. There is a direct connection, however, because the classic collective action problem is an n-person Prisoner's Dilemma.

An early and well-known statement of the problem of collective action was Garrett Hardin's essay 'The Tragedy of the Commons'. Suppose a mediaeval village has a commons (containing, say, pasture and woodlands, as they typically did) and every villager has an unrestricted right to exploit it. Then each will, if motivated purely by self-interest, put beasts on the commons to graze up to the point at which it is not worth it to him to have more. Similarly, each will cut firewood up to the point at which it is not worth it for him to cut more. The result, if all the villagers do the same, is that the pasture will be overgrazed to the point that it is of little use to anyone and the woods will be destroyed. The structure, it may be seen, is that of a PD: all would be better off if all showed restraint, but it is disadvantageous to any one of them to do so unilaterally. If I put only the number of beasts on the commons that would be sustainable if everybody did the same, this simply means that somebody else will put more on instead. Even if this did not happen, and there were a small reduction in the total, it would still be true that the improvement to the commons would be minuscule in relation to my loss from having fewer beasts on it, because the gain from the small reduction in overgrazing that I bring about is spread over all the villagers.

In fact, the 'tragedy of the commons' did not occur in mediaeval Europe, though its grim logic appears to be at work in Africa and the Indian subcontinent, resulting in the destruction of grazing land and an appalling rate of deforestation – which, in addition to making firewood less available, also creates far worse flooding than was previously the case, because there is nothing to impede the run-off of heavy rains. However, the oceans – 'that still remaining commons of mankind', as Locke called them – illustrate the 'tragedy of the commons' well, with the destruction of whales so that they were no longer an economic proposition and more recently the depletion of fish stocks in the North Sea. In the absence of an enforceable agreement to restrict hunting or fishing, each boat catches as much as it is worth catching, even though eventually this will put all the boats out of business.

In the absence of restrictive legislation, the air is a commons and so are the roads. As a consequence, each person is free to decide whether to drive or not drive, as he or she sees fit. If the result is air pollution and traffic congestion, that is a 'tragedy of the commons', because for each person the reduction in air pollution and congestion from not driving is too small to change the calculus of advantage from driving. Air pollution and traffic congestion are 'public bads' (the opposite of public goods) in that everybody 'consumes' whatever amount there is of them. It may well be that we would all be better off driving less and having less of these 'public bads', but we are trapped in an n-person PD. We need the state to change the payoffs by adding sanctions to the anti-social choice.

How did the mediaeval villages avoid the 'tragedy of the commons'? The answer is that use of the commons was governed by an elaborate set of rules specifying who could graze what, who could cut turf, who could pick up fallen branches, and so on. These rules were nor normally enforced by criminal law. But the inhabitants of a village interacted with one another in multiplex ways and depended on one another for so much that there was no difficulty in making informal sanctions work.

Michael Taylor, a political theorist with a penchant for anarchism, asked in his book Community, Anarchy and Liberty what would be the conditions under which collective action problems could be solved in the absence of a state. He concluded that a high degree of interdependency among a fairly small group of people interacting in multiplex ways could provide the conditions under which informal social sanctions could work to constrain behaviour in the appropriate ways. He conceded, however, that most of us would hate to live in such a society. If we are really concerned with individual liberty (and not merely opposed to states), the implication is that we need a state which has built-in mechanisms to avoid oppression. For the conditions necessary in order to solve collective action problems without a state are strongly inimical to individual liberty.

Hume's case of the neighbours who agree to drain a meadow (which is, we may note, stipulated to be a 2-person commons) was in fact designed to make the point that a state is needed to solve collective action problems. Thus, having said that two neighbours would easily be able to carry out an agreement to drain their meadow, Hume goes on: 'But 'tis very difficult, and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expence, and wou'd lay the whole burden on others. Political society easily remedies both of these inconveniences' (Treatise, p. 538).

In terms that have become familiar even outside the circle of game theory adepts, each of the thousand potential beneficiaries has a strong temptation to be a 'free rider' and 'lay the whole burden on others'. Unpacking the concept of the 'free rider' literally, we envisage a system of public transport. The train or bus will still go if one of the passengers does not pay the fare, but if enough fail to pay the service will deteriorate or cease altogether due to lack of funds. Each passenger, however, can reason that the amount of deterioration in the service or the increased probability of its ceasing that will result from his own failure to pay the fare will be so small that it is greatly outweighed for him by the cost of paying it. Analogously, each of the thousand people who stands to benefit from the draining of the field can figure that the field will either be drained or not by the efforts of others, and that the chance of his making the difference to the outcome's being draining rather than not draining is so small as to be completely outweighed by the advantage of not expending any effort.

The best-known statement of the problem of collective action in its general form is Mancur Olson's The Logic of Collective Action. Olson's central point was simply that it is often assumed that, if some outcome is beneficial to all the members of a group, and is worth more to each than his share in the cost of contributing to it, self-interested individual members of the group will have an adequate motive for contributing to the good. But in any case in which there is no way of excluding the non-contributors from the enjoyment of the good, this is not so. Thus, for example, we cannot explain membership in trade unions on the basis of their beneficial effect on wages nor can we explain membership in pressure groups on the basis of the advantage of getting more favourable public policy outcomes.

Olson's analysis of Hume's example would run as follows. Suppose that draining the meadow is an extremely good bargain for me in that I stand to gain five times as much from its being drained as it costs me in labour to do my share of the work. (We can cash out these ratios in terms of the von Neumann/Morgenstern utilities discussed in chapter 2.) Even under these favourable conditions, it still does not pay me to contribute to the job unless I believe that my participating has a better than one in five chance of making a difference between the meadow's being drained and its staying the way it is. Given that there are a thousand neighbours who all (we assume for the present purpose) have the same stake in the draining of the meadow, it seems implausible that I would single-handedly make more than a one in five difference to the probability of its being drained.

Since a collective action problem is an n-person prisoner's dilemma, it follows that the Defect option is a dominant strategy if the loss from contributing outweighs the following amount: the expected increase in the probability of getting the public good multiplied by the value of the public good. In other words, it does not make any difference to the logic of the argument what I expect the others to do. If they would complete it without my help, I'd be a damned fool to incur the effort; and if it won't be completed even if I do help, I'd be even more of a damned fool to incur the (ex hypothesi wasted) effort. (This, of course, presupposes that I am moved neither by altruism nor a norm of reciprocity.)

As I have already said of all game theory, Olson's 'logic of collective action' is not straightforwardly predictive. If you like, it says that, other things being equal, large groups will find it harder to provide themselves with public goods than small ones. Even this, however, is true mainly because small groups are more hospitable to conditional co-operation (which I will discuss in section 3), and this already falls outside Olson's 'logic'. In any case, other things are not equal, and a variety of factors may play a role in determining what happens. One important possibility is that the payoff structure may not really be that of a PD. As we saw in chapter 3, people may actually prefer (for normatively-driven reasons or future-orientated self-interested ones) to co-operate as long as others do. The implication is that the game form is really that of an Assurance game, as I defined that in chapter 3.

What, then, is the use of Olson's 'logic'? The answer is strictly negative: it rules out one possible answer to the question: how does a public good get supplied? Since, however, this answer has often been thought of as the obviously right one, this is important. For it creates a question needing an answer where there may not have appeared to be one before. Thus, suppose we were to ask why blacks in the American South took part in the Civil Rights movement of the 1960s. The answer ruled out by Olson is that it was in the interest of blacks collectively to have civil rights so it was in the interest of blacks individually to take part in the Civil Rights movement. That there was, nevertheless, a Civil Rights movement is not, even prima facie, a refutation of Olson's 'logic', as Green and Shapiro suppose. It simply sets up a research programme: if the move from collective self-interest to individual self-interest is invalid, what is the explanation?

This was the challenge accepted by Dennis Chong and answered in his book Collective Action and the Civil Rights Movement (Chicago: University of Chicago Press, 1991). Green and Shapiro (p. 88) describe this as a 'post hoc' account in which 'collective action is transformed from a Prisoner's Dilemma into an Assurance game, whereby participation becomes more rewarding than abstention'. This suggests that Chong simply fudged up some utilities to 'save the phenomena', but in fact he mined memoirs and diaries written by participants in order to establish independently what were the motivations of the actors. And these did indeed show that, under certain conditions of mutual assurance, people would prefer participating to not participating: conditional co-operation was the most preferred move among all the members of a group.

In the final paragraph of Collective Action, Hardin writes: 'Olson's logic of collective action is based on a strictly static analysis of the costs and benefits of any given collective action uncoupled from other exchange relationships' (p. 229). Where 'enduring movements' create a network of conditional co-operators, he goes on, this 'may motivate a high level of activity, especially . . . if the activity can partly be localised, as in the American civil rights, antiwar, women's and environmental movements' (ibid). Hardin concludes as follows: 'Alas, to understand such activity in a particular case might require such intensive observation as would try the patience of an anthropologist and such attention to nuance as would frustrate a philologist.' Here follows a footnote with the obligatory reference to Clifford Geertz and the remark: 'To understand common instances of political activity in rational terms would require "microscopic" descriptions of relevant individuals' behaviors and intentions.' Hardin's final sentence is 'At this point, I take comfort in being neither [an anthropologist nor a philologist] and therefore in leaving the field to others' (pp. 229 – 30). Work such as Chong's is, I would suggest, precisely the kind that was made possible by Hardin's book and other game-theoretic studies of collective action.

2. The Social Contract Solution

Olson himself suggested that the way in which organizations induce people to participate on a self-interested basis is by offering them 'selective incentives' – that is to say advantages that people get in return for contributing and cannot get without doing so. For example, even if trade unions cannot have the gains from collective bargaining paid only to their members, they can confine to their own members their role of representing workers in individual conflicts with management. Lobbying organizations can offer their members magazines or specific help on an individual basis, as well as lobbying on public policies whose applicationcannot be limited to those who contributed. And so on.

States offer selective incentives for compliance with their laws in the negative form of freedom from the sanctions they would otherwise impose for noncompliance. Hume's explanation of the way in which 'political society easily remedies' the 'inconveniences' of getting people to plan and take part in collective projects fits in with this. Left to our own inclinations, each of us may find it advantageous to disobey the law, even though we are all better off if we all obey it than in a 'state of nature'. A system of laws with enforcement mechanisms helps to narrow the gap between the behaviour that is individually beneficial and the behaviour that is collectively beneficial. Hume says (pp. 538-9) 'Magistrates find an immediate interest in the interest of any considerable part of their subjects'. Governments, he goes on, can deal with the problems of planning and execution that would foil a thousand independently-acting individuals. 'Thus bridges are built; harbours open'd; ramparts raiz'd; canals form'd; fleets eqiip'd; and armies disciplin'd; every where by the care of government, which tho' composed of men subject to all human infirmities, becomes, by one of the finest and most subtle inventions imaginable, a composition that is, in some measure, exempted from all these infirmities' (p. 539).

The theory of the social contract is an attempt to embody the point that everybody gains from government, as against a 'state of nature'. The essence of a contract is that A agrees to do something that he would sooner not do (call it x), in exchange for B's doing something he would sooner not do (call it y), because they both prefer the situation which x and y are both done to the one in which neither is done. Thus, in the original PD story, the prisoners would like to be able to make an enforceable agreement not to confess. This is preferable to the other symmetrical possibility of both confessing. The contract rules out for each prisoner his most preferred alternative – he confesses, the other does not confess – but each realizes that this is not attainable anyway. By giving up the option of confessing, each obtains a better outcome than he could hope for by retaining the option.

Similarly, an enforceable contract would solve the problem of the two farmers in Figure 3.1. The problem, it may be recalled, was that, if B helps A with his crop, he has no assurance that A will help him with his crop in turn when it ripens. In fact, A's best payoff comes from not helping B (A + 10, B + 2). As a result, by backward induction, B does not help A and each loses half his crop. Clearly, if A could enter into an enforceable undertaking to help B provided B has helped A, we could regard the right hand branch from node 2 as ruled out. (This could be done if a penalty of 10 units, say, was attached to A's non-performance.) The path from 'A reaps, B reaps' (on A's crop) now goes only down the left hand branch to 'A reaps, B reaps' (on B's crop). Both have an incentive to make a contract binding the one whose crops ripen first (whichever it is) to help the other, on condition that the other has helped him. This would be, in Hobbbes's terms, a 'covenant with the sword.' Notice, again, that A gains by giving up the option that is most advantageous to him: the option of not helping B after B has helped him. He does this on order to induce B to help him, which B will not do if A retains the option of not helping in turn when the time comes.

The contract of the social contract theorists is (as Hume called it) an 'Original Contract': it creates the possibility of having, among other things, enforceable contracts of the ordinary kind. The trouble with it is (as Hume again observed) that there is no such contract. Hobbes, indeed, acknowledged this – though he did not make a point of it – in his depiction of the way in which sovereignty by institution comes about. It is, he says, as if every man should say to every other man . . . . But, as Ronald Dworkin pointed out in his early paper on Rawls's A Theory of Justice, a hypothetical contract cannot give rise to actual obligations: even if I would have agreed to something, that does not make things the same as if I had actually agreed. This is not a valid objection to Rawls, for whom the hypothetical agreement establishes what is just and is not itself intended as a source of political obligation. (That is covered by a 'natural duty'.) But it is a valid objection to Hobbes. Locke, of course, tries to weasel out of the difficulty by invoking 'tacit consent' but there is no need here to go into what is wrong with that. Hume, as we have already seen, appeals directly to the thought that underlies social contract theory: the idea that we all do better by giving up some liberty in return for security and the provision of public goods. (These were, it may be recalled, the two things that Hobbes said would be missing in the 'state of nature'.)

Unlike Socrates, but like Glaucon (who is often characterized as the first exponent of social contract theory), social contract theorists assume that at least some of us would be liable to commit injustice if we could get away with it. But we all see that we are all worse off with that option available to us (exactly as the prisoners and the farmers were worse off if the exploitative option was available) than if it were removed by an apparatus for the enforcement of laws. The logic of this is unassailable. Hume's point is simply that we can say all this without any talk of a social contract. Government can be justified directly as the solution to a collective action problem.

Where does this leave the discussion of Hobbes in the previous chapter ? I argued there that Hobbes can get his sovereign by institution off the ground by invoking the 'law of nature' to the effect 'that men perform their covenants made' as long as they can do so safely, and this includes the covenant that sets up the sovereign. This obliges everybody to obey the law and also obey the sovereign's orders to help catch those who do break it. But if this covenant is only hypothetical, and therefore not binding, where does that leave him? I would say, with Hume, that he can say everything he needs to say without the device of the social contract. He can appeal directly to the sense each person has of the advantages that flow from getting out of the 'state of nature'. If these advantages are as great as Hobbes maintains they are (and so does Hume), Hobbes can say that it is worth incurring a small cost to improve the prospects of peace. This may or may not be plausible but if it is not plausible then neither is Hobbes's claim that it is advantageous in the long run to keep covenants even at some short-term cost to oneself. The two propositions stand or fall together, since the obligation to keep covenants is derived from their role in bringing about peace.

To illustrate this point, let me briefly refer to Jean Hampton's book <u>Hobbes and the Social Contract Tradition</u>. She recognizes (p. 174) that, for Hobbes, 'the subject must oblige "himselfe, to assist him that hath the Sovereignty, in the Punishing of another" (Lev 28 –2); that is, the subject must actively assist the sovereign when ordered to do so in punishment and enforcement activities involving others.' But she argues that there looks to be a free-rider problem here (pp. 175 – 6), because each member of the society either gets the public good (peace) or not, and each can reason that helping the sovereign to capture somebody will not make enough difference to the probability of that outcome to make it worth it. As with the meadow, each can think: 'If enough others will follow the sovereign's orders to go after somebody, I'm better off leaving them to it; and if not enough others will, things will go to hell anyway, and I certainly don't want to be one of the few to stick my neck out.' Hampton proposes a solution, which I shall discuss in the next chapter. But the point I want to make now is that Hobbes would rule out this kind of case-by-case calculation. If he is right, the advantage of keeping covenants is so great (because of their contribution to peace) that it is always in your long-term interest to do so, as long as you can do so safely. (Obviously, if you would be risking your own life by keeping a covenant – for example, giving yourself up to the sovereign for punishment – the gain to personal security from peace is outweighed by the immediate gain from not keeping the covenant. But this is not a serious weakness in Hobbes's theory. I am constantly surprised that so many people (e.g. Hampton) seem to think this proviso is a serious difficulty in Hobbes's theory. I do not believe that any legal system in the world relies on individuals voluntarily giving themselves up to be punished. There have been parts of the world in which fear of collectively-suffered sanctions leads a group to hand over one of its number, but that is clearly a different matter.)

I shall take up at length in the next chapter (sections 3 and 4) the validity of Hobbes's basic claim. Before that, I need to discuss two possible ways of escaping collective action problems without resorting to external sanctions such as punishment imposed by a state.

6. ALTERNATIVES TO COERCION IN COLLECTIVE ACTION PROBLEMS

1. Privileged Groups

In the previous chapter, I focused on coercion as the solution to collective action problems. This could be coercion by a state or it could be coercion applied in the context of a dense network of social interactions, as in the case of the mediaeval villagers and their rules about the use of the commons. In this chapter, I shall discuss two other ways in which collective action problems can be solved. As we shall see, they are of rather limited applicability, so the conclusions of the previous chapter can still be allowed to stand as true for most cases.

The first solution turns on two concepts introduced by Olson in The Logic of Collective Action. Suppose there is some good that, if it were produced, would benefit more than one person and from whose enjoyment nobody can be excluded. (It is thus, in at any rate one sense of that notoriously slippery concept, a 'public good'.) There are now two possibilities. One is that one person (or more than one person) would individually find it advantageous to supply (at least some of) the good. This is, in Olson's terms, a 'privileged' group. The other possibility is that it is not in anybody's individual interest to supply the good, though there is some possible system of contributions such that all would benefit from making the contribution and having the good supplied. This is, in Olson's terms, a 'latent' group.

Olson, as Russell Hardin has complained, muddled together the privileged/latent distinction and the large/small distinction (Collective Action, chapter 3). So, it may be noted, did Hume. Certainly, large groups will tend to be latent while small groups are more likely to be privileged. But the logic of the distinction turns not on size but on the structure of payoffs. We can illustrate this point by going back to the meadow-draining example. The payoffs envisaged to illustrate it had the implication that, even though only two people were involved, they formed a latent group. For the cost of draining the meadow was assumed to be 150 units of utility and the gain to each neighbour from its being drained 100.

Using this new vocabulary, we can add that, once they had put in 30 units of work each, leaving only 140 needed to finish the job, the group composed of the two farmers became a privileged group. For, given this sunk cost, it would now pay either neighbour to complete the draining himself. Since the game was symmetrical, however, it did not point towards either as the 'natural' person to do all the rest of the work, and thus set the stage for a good deal of acrimony. Now consider the case where a thousand people stand to benefit from draining the field. Hume says it would be *difficult* to co-ordinate on a common plan and organize the work. But pretty clearly what leads him to say it would be *impossible* actually to get the job done (as against the mere difficulty of planning it) is the motivational problem that he identifies.

Notice, however, the assumption underlying this analysis, which leads to the conclusion that the thousand people constitute a latent group. This is that each member of the group stands to gain only one thousandth of the total benefit from the drainage of the field. Suppose instead, however, that one person has the right to graze animals on the meadow and cut the grass to make hay for winter feed, while the other 999 have only the right to walk in the field. It could be that the one person with the major rights would gain 200 units of utility from draining the field, which means that it would be worth his while to drain it all by himself if draining it takes 150 units. Let's imagine that each of the other 999 stands to gain one unit of utility from the draining of the field. The total gain is thus (roughly) 1,200 units. Equity might suggest that the person with major rights in the field should have to contribute only one-sixth of the cost. But we can predict that, if the field gets drained at all, the major beneficiary will have to do it. Similarly, suppose that there were only two neighbours but that one of them would gain (and it was common knowledge that he would gain) 200 units from draining the field, while the other stood to gain only, say, 20 because he did not wish to use it for grazing and only for walking on. The neighbour who stood to gain the most might find it very hard to persuade his neighbour not to take a free ride on his efforts.

Notice that to be the party that makes a group 'privileged' is not itself by any means an unequivocal advantage. It does mean that the collective action problem is solved (at least partially) because it pays somebody to carry out (at least part of) the project. But it also means that this party is liable to carry the whole burden of producing however much of the public good is produced. The 999 other beneficiaries in the 'privileged' group of neighbours gained a strategic advantage by being disorganized: if they had had a common authority with the power to tax them, the major beneficiary would have had a much better chance of negotiating a more favourable deal by threatening not to do anything unless the others contributed.

During the Cold War, it could be (and was) suggested that the same logic explained the disproportionate American contribution to the costs of NATO. Once the US government had adopted the view that its own security entailed the defence of Western Europe, the individual European countries could treat NATO as a 'privileged group', or at any rate close enough to one that the United States would be willing to do what was necessary as long as they made some modest contribution. The implication is, of course, once again that the Western European governments gained by not having a common defence policy: if they had had one, the United States could have argued that, with more population and at least as much GNP, Western Europe should assume an equal burden. That being so, why do we now find Western European governments agitating themselves about the need for a common defence policy? This can be explained by the same logic. During the Cold War, NATO policy was a public good for all NATO countries, since it consisted essentially of deterring the Soviet Union. Since there was a common interest, it did not matter if the USA dictated the policy on the basis of the maxim that 'he who pays the piper calls the tune'. (American policy elsewhere might be found obnoxious by many European governments, but they could not influence that anyway.) With the end of the Cold War, it is plausible that American and Western European interests diverge in relation to the rest of Europe, so that there is now an incentive for Western Europe to have some independent military capability. Conversely, if the United States government now sees its security as less bound up with what happens in Europe, NATO ceases to be (or even approximate) a 'privileged group' on the basis of United States involvement. Western Europe can therefore be expected to move towards a solution in which (like a taxing authority for the 999 neighbours) it overcomes its own internal collective action problem so as to provide a common defence policy to back up a common foreign policy.

2. Conditional Co-operation

An alternative non-coercive way of overcoming a collective action problem can be illustrated by returning to the original case, as stated by Hume, of the thousand neighbours who all stand to gain equally from draining a meadow. Suppose that it were feasible to monitor each neighbour's participation in the work. Then each might pledge himself to work only if all the others did. Notice that this commitment can be seen as a kind of threat, but of a new kind. To adapt Hobbes, we may say it is a threat of all against all. So far, I have been looking primarily at asymmetrical power relations, though mutual hostage-taking (in both the individual and the nuclear cases) was an exception. Here, we have a symmetrical n-person threat situation – in which all the parties have equal power.

The threat made by each is to stop work immediately if <u>any</u> of the others do. The effect of these threats – if they are carried out – is that anybody who stops work will trigger a mass walkout. The problem with this threat is the usual one of making it credible. If the job is largely completed and one person stops work, are the rest really going to leave it unfinished and waste all their labour so far?

Perhaps the neighbours all swear a mighty oath and believe that breaking it will imperil their immortal souls. Then the device of mutual threat should work. But this simply pushes the difficulty back a stage. So far it has been tacitly assumed that it is common knowledge that all thousand neighbours will benefit from the meadow's being drained to a degree that outweighs the disutility of doing one thousandth part of the work required to drain it. But this is an extraordinarily strong assumption. Quite likely it is not even true: we may plausibly expect that, for a variety of idiosyncratic reasons, a number of the neighbours will either not expect to gain much benefit or will (for health reasons, for example) suffer a lot from doing their share of the work. What is even more of a problem is that, even if it is true that all would make a net gain from doing their share of the work and having the job done, it seems impossible to imagine how everybody could be sure that this is true of everybody else.

Suppose a small group of the neighbours say that they would sooner the meadow stayed as it is than have it drained if the price is their participation in the draining. It may be God's honest truth or it may be a pure bluff. The point is that there is no way of being sure. (Notice that, if there were a monetary advantage to each from having the job done and each could be assessed a monetary contribution as their share of the cost, this problem would not arise – as long as nobody could convincingly say that their utilities did not track the money – 'It's not the money, it's the principle of the thing.') Are the remaining neighbours going to insist that they will call the whole project off unless everybody signs up? It would seem that the proposal should be amended to allow some to opt out, as long as not too many do. Suppose, for example, that it seems to be true that dividing the work nine hundred ways would still put nine hundred neighbours who joined in well ahead, compared to leaving the meadow undrained. Why not announce that the work will proceed with nine hundred pledges but not fewer?

This would work if everybody who would genuinely gain took the pledge and there were at least nine hundred of them. But if the neighbours were like that, the problem of free riding might not occur in the first place. If they are strategic utility-maximizers, each has an incentive to opt out, even if he really would gain from doing his share, hoping that there will still be nine hundred pledges forthcoming. The optimal move is to declare an opt-out early, so as to put pressure to stay in on whoever would push the number of opt-outs over a hundred. But there seems to be nothing to stop an ugly rush that is liable to overshoot. If there were some system for strictly sequential declarations (e.g. in alphabetical order) this could put the person who has to stay in or make the opt-outs come to a hundred and one in the same position as somebody in the earlier story who believes everybody will stop work if he does. But what if, when the number of opt-outs gets to a hundred, there are still some people left to declare who would genuinely prefer leaving the meadow undrained to contributing to its draining? (See Michael Taylor, The Possibility of Co-operation, chapters 2 – 4, for a mathematical analysis of threshold problems such as these.) There seems to be no way of avoiding this very real possibility.

The implication is that only in exceptional circumstances do we overcome collective action problems among self-interested people by introducing conditional co-operation. The exceptions are cases in which it is common knowledge that it is actually advantageous to everybody to play the Defect move into the indefinite future if anybody plays the Defect move. Under these circumstances, the problems raised by strategic withdrawal of co-operation do not arise, because everybody understands that they all have a direct (i.e. non-strategic) motive for their withdrawal of co-operation.

Where these special conditions hold, nobody can believe that the option exists in which he defects and others go on co-operating. We can represent this in a payoff matrix by collapsing all future plays of the game into a single set of possible payoffs. This payoff matrix is shown in Table 6.1. As will be seen, only the Co-operate/Co-operate cell and the Defect/Defect cell are available. If we wanted to be picky, we could say that there is a tiny advantage to be gained from being the first to defect, thus stealing a march on the others. Ex hypothesi, however, this gain is so minute in relation to the loss from the move to an All-Defect equilibrium that it should be treated as infinitesimal.

What might be a real world example of this payoff matrix? A cartel is about as good as it gets, with one firm in Table 6.1 as A and 'the rest' as B. If A lowers its price, the others cannot afford not to follow suit. And once the price has been reduced all round, it is at the least uncertain that it can be pushed up again. This is especially so if price-fixing agreements are illegal. For then a firm that raises the price unilaterally has to hope that the others will follow its lead, rather than simply appropriate its share of the market. Thus, the situation is, prospectively, not too far off one in which one play of Defect triggers a permanent shift to the Defect/Defect outcome.

3. Hobbes and Hume Revisited

I have promised to assess Hobbes's argument about keeping covenants and Hume's argument about respect for possessions. These arguments are structurally identical. Indeed, the only difference of any significance lies in their account of the causes of conflict. Hume apparently believes that conflicts over possessions are the only possible cause of descent into a 'war of all against all', which (as we have seen) Hume describes in precisely Hobbesian terms without, of course, mentioning Hobbes. In Hobbes's more profound understanding, fear of others can itself lead to war: people may think, for example, that their safety will be enhanced by disabling others while they are vulnerable – a 'pre-emptive strike' in the parlance of the 'wizards of armageddon'. He also, as we have seen, thought that the desire for eminence was a potential source of conflict. However, both men held that the prospect of reversion to a 'war of all against all' should be so appalling as to lead everybody to the conclusion that no temporary advantage could be worth taking if it increased the probability of a transition from peace to war. For Hume, this implied that everybody should recognize the overwhelming advantage of respecting the possessions of others. For Hobbes, it implied that everybody should be prepared to do everything they could with safety do to promote peace, and in particular keep their covenants.

Strictly speaking, the Hobbes/Hume analysis does not require the assumption that a single person's playing the Defect move must be anticipated to bring about a permanent (or at any rate indefinitely prolonged) All-Defect condition. The argument can afford to be probablistic in form: what it has to maintain is that the possibility of an individual violation precipitating the move to war is great enough to outweigh the advantage to the agent of the violation. Notice that, in terms of the earlier discussion, this makes every group of human beings living in a certain area into a privileged group. It is not, of course, that any one person can create the public good of peace all by himself. Nevertheless, it does mean that each person finds it in his own individual net interest to contribute to the public good of peace, as long as others do so as well, because the value of the public good to him (his own personal security) is worth more to him than the value of whatever he has to give up (taking others' possessions or not keeping covenants) in order to ensure that the public good continues to be supplied.

The obvious snag with this line of thought is that it is highly implausible. Hume is open to the counter-argument that his account of the thousand neighbours draining the meadow and his account of everybody's respecting others' possessions has a rather schizophrenic look. If it is, as Hume maintains, impossible that a thousand people could drain a meadow because each would seek to 'lay the expense' on the others, how can he be so sure that nobody could ever rationally believe that stealing is not in his interest if he can get away with it? Surely, if the problem of collective action is so malignant as to lead to inevitable cheating in the meadow-draining case, it may also infect the property-respecting case. Why could not somebody rationally believe that the probability of his theft's bringing about a 'war of all against all' is so low that it can be discounted in his calculus of the long-term pros and cons of stealing?

Hobbes does not expose himself to a similar accusation of internal consistency, but he is, if anything, even more open to a charge of implausibility. Thus, he even says that, if you are captured by pirates and released in return for a promise to pay a ransom when you get home, you are obliged to carry out your undertaking. For that conduces to peace in saving of the lives of future captives. It is doubtful that this is really an injunction well designed to save lives in the long run. As I argued earlier, paying ransom makes the trade of holding people to ransom much more attractive. Since there is bound to be some loss of life incidental to the conduct of piracy itself, adherence to Hobbes's proposal may well actually lead to greater loss of life in the long run.

Leaving that aside, however, what seems quite clear is that it cannot by any stretch of the imagination be shown to be to the long run interest of the victim to pay the ransom, which is what Hobbes has to maintain. Even if Hobbes is right, and non-payment of the ransom makes it a little bit less likely that future captives will be released on their own recognizance, this could hardly make it worth paying once you get home on the basis of your own self-interest, especially if you have no plans to go sailing in pirate-infested waters in future.

Hobbes backs up the argument for keeping covenants when you can do so safely by saying that it is contrary to your long-run interest to gain a reputation as the sort of person who does not keep covenants because then nobody will make them with you in the future. (It may be noted that this is David Gauthier's case for keeping promises, from an egoistic viewpoint, in Morals by Agreement.) However, this argument depends on an extremely strong publicity assumption: why should I noise it abroad gratuitously that I have failed to pay the pirates, if that will stop people from trusting me? In effect, Hobbes' argument would work only in the kind of small face-to-face community with multiplex interactions in which, as I have suggested in the 'commons' case, collective action problems can indeed be solved to some degree without external enforcement.

I mentioned earlier in this chapter that Jean Hampton had a proposed solution to the free-rider problem generated by Hobbes's account of the obligation to assist the sovereign in law enforcement activities. This (pp. 176 – 86) is nothing other than the conditional co-operation discussed in the previous section. If Joe the Desperado (her example) heads off into the desert with the loot, the sovereign (in this case, I suppose, the sheriff) informs six people that capturing Joe needs a minimum of six people and he has chosen which six it is to be. This, she maintains, gets rid of the free rider problem, because Joe will not be captured unless they all join in and if Joe gets away they will all lose security. (If it isn't true that they're all worse off, of course, Hobbes doesn't even get to first base.) However, this requires common knowledge that six is a number that is both necessary and sufficient to capture Joe. (Hampton's analysis rests on the assumption that if a posse of six is formed it will succeed.) It is unlikely that there is in fact a sharp threshold of this kind and even less likely that everybody will be inclined to believe it. Moreover, if (as Hampton stipulates) the sovereign's choice is arbitrary, why should not the free rider problem resurface a stage further back? Each of the six people designated can reason as follows: 'If I don't go, the sovereign can equally well tell somebody else to.' Hampton's method would work only if, in addition to the required (and implausible) common knowledge, there were some reason why, given that the critical number is six, there are only six feasible candidates for the job. Needless to say these conditions are so demanding as to make Hampton's supposed solution trivial.

Furthermore, if the logic of the free rider infects the obligation to obey the sovereign's commands to help enforce the laws, it just as much infects the obligation to obey the sovereign's laws oneself in the first place. The equivalent of Hampton's solution in the enforcement case has to be, in the primary case of obeying the laws, that one act of Defect is believed to lead (at least with sufficiently high probability to tip the scales) to All-Defect. But that, I have argued, is not a solution that stands up.

4. Precarious Peace: An Alternative Account

It is, I suggest, possible to find situations that satisfy one of the Hobbes/Hume conditions in that peace is precarious and could be replaced by warfare (of more or less intensity) as a result of one act of violence or at least a small enough number of them that any one has a significant probability of pushing the situation over the brink. Militants from either community in Northern Ireland could probably engineer a resumption of the low-level communal conflict characteristic of most of the past couple of decades by perpetrating a medium-sized atrocity, for example. For it is likely enough that militants from the other community would respond in kind and thus set off a spiral of communal violence. In Bosnia, a dozen youths with rifles travelling from village to village in the back of a truck could, in one day, by systematically killing the members of one ethnic group, permanently bring an end to peaceful interethnic relations that had in many cases persisted among neighbours for generations. Similar knife-edge situations are not at all uncommon. In the Israeli/Palestinian case, of course, the situation was so volatile that a provocative act (Sharon's visit to the al Aqsa mosque) was by itself enough to unleash escalating violence, though this very volatility suggests that it determined only the timing. (Russell Hardin's One for All contains a useful discussion of the incentives for intergroup violence.)

Does this provide support for the Hobbes/Hume analysis? Unfortunately not, for those who destabilize a fragile peace are motivated by the belief that doing so will further their political ends. What they foresee is not a 'war of all against all' but a war of one community against another (or more usually a violent minority of each community against the other community) which they expect to win. (We should recall from chapter 4 that 'winning' here simply means achieving your political aims: you can lose the war, in the sense that the other side inflicts more damage, but still win in the Clausewitzian sense if you prevail politically.) The object may be 'ethnic cleansing' or it may be some realignment of relations between the communities, which are expected to remain in situ. Either way, the point is that those who initiate the breakdown of peace are acting not as individuals but on behalf of a group, and their calculation is that, in the long run, the price of warfare will be worth paying, from the point of view of the members of the group.

Looking at things from the perspective given us by another (in Hume's case) two and a half centuries and (in Hobbes's case) three and a half centuries, we have to reach the verdict that Hobbes and Hume were touchingly naive about the sources of social conflict. Hume was almost unbelievably superficial: 'No-one can doubt, that the convention for the distinction of property, and for the stability of possession, is of all circumstances the most necessary to the establishment of human society, and that after the fixing and observing of this rule, there remains little or nothing to be done towards settling a perfect harmony and concord' (Treatise, p. 491). The passions other than those of 'interest', Hume says, are nowhere near as dangerous to civil peace. But he assumes that they are invariably directed at individuals, rather than groups. Thus, 'envy and revenge . . . are directed against particular persons, whom we consider as our superiors or enemies' (ibid).

Hobbes's darker vision could encompass more sources of serious conflict than could Hume's. Hobbes, after all, was only too well aware that civil war could arise from causes other than a struggle over individual bits of property. People could, he recognized, fight about politics and religion. But he claimed that this could have occurred only due to cognitive error: if people had understood properly the grounds of political allegiance and the (very limited) requisites for personal salvation, they could not have done so. All this, however, depended on the presupposition that suffering violent death is so much the worst evil that can happen to anybody that it must trump any countervailing value. Given this assumption, and Hobbes's ideas about the requisites for salvation, his conclusion – that political quietism is the best policy for everybody – seems to me to follow pretty well.

Hobbes did, as we have seen, ascribe a desire for eminence to human beings, and saw this as a potential source of conflict. But he was like Hume in thinking of it only in terms of one individual comparing himself invidiously with another. He did not see the huge destructive potential of the desire for <u>collective</u> eminence: the exaltation of my group (Aryans, whites, Croats) at the expense of another group (Jews, blacks, Serbs). Of course, Hobbes could still say that it is irrational to risk your life for a collective cause of this kind, even if millions of people have chosen to do so. Within the perspective of game theory, however, this is a senseless assertion. To repeat: rationality is to be predicated of means only, not of ends.

The deliberate fomenting of communal conflict in the hope that your side will come out in the end closer to achieving its political goals is, it seems to me, something that lay beyond the scope of the theories put forward by Hobbes and Hume. Hobbes, as we have seen, thought it would be irrational to provoke war, because personal security was necessarily paramount. Hume, it is notorious, said that it is not contrary to reason to prefer the destruction of the world to the itching of my little finger. But his discussion of respect for property shows that he attributed to human beings in practice a uniform set of preferences in which violent conflict came at the bottom a long way below anything else. Both, therefore, actually adhered to the same conception of human motivation, though they expressed it differently. If it were right, most of the large-scale violence of the past two centuries would be inexplicable.

The conclusion to be drawn is, therefore, that instances of peace resting on a knife-edge, and capable of being transformed into bloody conflict by one or a few acts of violence (or perhaps even provocation) do not conform to the Hobbes/Hume model. They do not arise from individually-orientated acts based on a wish to acquire others' possessions, a desire for individual eminence, or fear of attacks by other individuals. Rather, violence arises from the pursuit of collective goals that are so important to people that they are prepared to accept an increased chance of violent death to advance them. Hobbes could understand, as I have mentioned, that people might rise violent death in order to improve their chances of personal salvation, and he devoted a large part of Leviathan to trying to show that this was never really necessary. But I think that the idea of people risking violent death for its secular collective equivalent – the liberation or aggrandizement of the nation – would have been genuinely incomprehensible to him.