THE PHILOSOPHY OF SOCIAL SCIENCE

An introduction

MARTIN HOLLIS

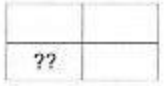
Professor of Philosophy at the School of Economic and Social Studies
University of East Anglia, Norwich

CAMBRIDGE
UNIVERSITY PRESS

1994

CHAPTER 6

Games with Rational Agents



Individualism, in a general and robust form, maintains that there are only particulars, with the methodological rider that, in the final analysis, reference to particulars can account for whatever seems to involve something more. But social scientists are as inquisitive as anyone about human individuals, what makes them tick and whether they are the creators or creatures of the social world. This curiosity, I hope, warrants a focus on versions of individualism, where individuals are human agents with desires and beliefs and act in ways which account for what happens. Admittedly, there are behaviourists, who contend that a properly scientific approach neither wants nor needs to credit individuals with a subjective point of view or, indeed, any mental states. Also we have noted that, in some theories usually deemed individualist, the individuals are firms, nations or other agents lacking flesh and blood. But, indulging my own curiosity, I propose to explore the thesis that 'history is the result of human action, not of human design'.

That quotation comes from the eighteenth-century Scottish philosopher Adam Ferguson and nudges us towards economics as the social science where a humane kind of individualism has been most thoroughly deployed. Lionel Robbins defined economics as 'the science which studies the relationship between ends and scarce means which have alternative uses', and this makes it potentially a very far-ranging science indeed (1932, p.15). Politics and sociology are only two of the social sciences which have lately become fascinated with 'economic' analyses of social interaction. The idea that we behave in all aspects of our lives as we typically do in market situations is not only intriguing in itself

but also prompts a potent individualist account of institutions, norms and practices, thus robbing holism of its trumps.

'Humane' is perhaps the wrong word. The economic theory of Rational Choice treats us as rational, self-interested individuals, each intent on maximising our own utility. Game Theory, which goes on to analyse interaction, rests on Rational Choice theory. This hardly sounds like humane treatment; but the approach can be made more generous than it seems at first. The chapter will start by defining a rational agent with the standard notion of rationality in economics. It will then sketch the elements of Game Theory. That will be followed by a discussion of the emergence of norms and whether we now have an analysis which can account for them.

RATIONAL AGENTS

'The first principle of economics is that every agent is actuated solely by self-interest.' Edgeworth's remark remains a good starting point, not least because his title, *Mathematical Psychics*, so well captures the spirit of what is afoot (1881, p.16). But it needs caution. Not all economists will accept it as their first principle, although neo-Classicists commonly do and although others, like Keynesians and Marxists, usually employ the stock notion of rationality when analysing individual action. Also there is still the problem, left unresolved when we introduced it on p.57, of deciding whether Edgeworth's dictum offers a falsifiable hypothesis or has some other status. For the moment, however, let us concentrate on rational agents and the sense in which they are actuated by 'self-interest'.

Rational Choice theory begins with a single, ideally rational individual, classically Robinson Crusoe alone on his desert island. He has three components: *fully ordered preferences*, *complete information* and *a perfect internal computer*. He acts rationally in as much as he chooses the action which he correctly calculates to be most instrumental in satisfying his preferences.

Suppose, for instance, that Crusoe is trying to decide whether it is worth the trouble of making a net, so as to catch more fish than he now can with a pointed stick. The consequences of making a net are not only (probably) more fish tomorrow but also the effort of making it and less fish today. Think of consequences as possible states of the world, brought about by his choice of action. If he was sure of one fish a day without a net and sure of four fish with a net after a fishless and tiring first day, his rational choice would depend simply on how he ranks these two outcomes. He simply chooses the action whose consequences he prefers.

Where outcomes are not certain, matters are more complex. The theory assumes that he has a complete ranking of all possible outcomes, regardless of their probability. It then assumes that his information is complete, in that he knows how likely each outcome is. (To be exact, he has a 'subjective probability distribution' which is complete and consistent; and thus has a subjective view of the chances of, say, two fish or three or four and so on, which will not land him in contradictions.) Since he also has a perfect internal computer, he can now calculate the *expected utility* of making a net and compare it with that of not making one. The expected utility is the sum of the utility of each possible outcome, discounted for the probability that it will not in fact occur. (To grasp the idea, think of a £1 bet on a card drawn at random from a full pack, with generous payoffs of £5 for a Spade and £3 for any red card. Assuming that utilities are reflected in cash profits or losses, a bet on a Spade has an expected utility of 0.25 (= £5/4 − £1) and a bet on red of 0.50 (= £3/2 − £1).) A rational agent always rejects actions with lower expected utility and is indifferent between those with the same expected utility. Notice that the calculation may need to allow for varying costs as well as varying probabilities.

The complexities make it plain that the ideally rational agent is very idealised indeed. None of us ever has such a complete and consistent set of preferences over the range even of the likelier outcomes. We have nothing close to complete information and the device of working with subjective probabilities in a world of uncertainty is distinctly artificial. We are not blessed with perfect internal computers. Nevertheless, this is the ideal-type case of a simple and potent notion of rationality. We act rationally, when we know what we want, have a shrewd idea how likely each

course of action is to satisfy us at what cost, and choose the action which is thus the most effective means to our ends.

Rational action is thus *instrumentally* rational action. It does not matter whether people prefer oranges to apples, guns to butter or virtue to vice. Rational agents can have any (consistent) preferences, and are rational if and only if their choices maximise their expected utility accordingly. There is no further question of the rationality of their ends. Questions of whether preferences cause actions or merely derive from what is chosen can wait.

We can now be clearer about 'self-interest'. Edgeworth took people to be so self-regarding and selfish, at least in the realm of commerce, that they were indeed 'self-interested' in the everyday sense of the term. More generally, economics has been dubbed 'the dismal science' partly because economists often take a similar view of agents in economic or even social life at large. But, strictly, the standard first principle assumes only that agents are guided by their own preferences. In this sense saints are as 'self-interested' as sinners and the theory of Rational Choice is not committed to any view about how saintly or sinful we are. Although we shall need to ask later quite what, philosophically, it is committed to, we begin by assuming only that agents seek to maximise their own expected utility.

THE THEORY OF GAMES

Rational Choice theory opens with a single agent in an independent environment. Having given a basic definition of rational choice, it then explores the implications making the agent no longer certain about the consequences of action. The environment sets *parameters*, within which choices are to be made. A static environment is not required but any dynamics are independent of the agent's decisions. It is not as if the god of the sea were trying to anticipate Crusoe's moves. Call choices made in an independent environment *parametric*. As soon as Man Friday enters the scene, however, Crusoe's rational choice may depend on what Crusoe will choose. Each may need a *strategy* which takes account of the

other's strategy. Call choices which are interdependent in this way *strategic*. Game Theory starts here. It analyses strategic rational choices in an ideal-type setting where each rational agent knows, among other things, that other agents are rational in the sense already defined.

That sounds daunting but the basic idea is still simple. The basic scenario requires two agents, each with a choice of two actions. Since it saves confusion if we can refer to the agents as 'he' and 'she', let us replace Crusoe and Friday from now on with Jack and Jill. Suppose that Jack and Jill are motorists who meet at opposite ends of a narrow bridge with room for only one car. Each must choose whether to advance or wait. There are four possible outcomes: (Stop, Stop), (Go, Go), (Stop, Go), (Go, Stop). The situation is a 'game', with the schematic form shown in Figure 6.1.

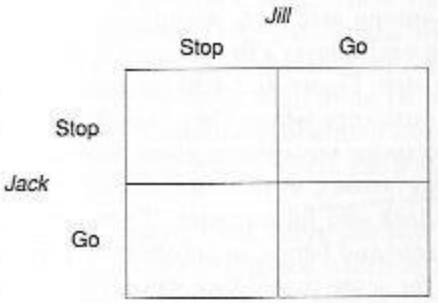


Figure 6.1

What happens depends partly on how each player ranks the four outcomes. Crucially, however, what *payoff* each receives may depend on which choice of action the other player makes. So each must also allow for how the other ranks the outcomes and also for what the other is thinking. Thus Jack may need to know what Jill expects him to do and *vice versa*. We can best deploy the idea by considering four basic games.

FOUR BASIC GAMES

(1) Coordination

Suppose initially that neither minds who waits and that there are thus two outcomes which each ranks equal best (and two inferior ones). This makes the game a *coordination game*, as shown in Figure 6.2.

The pairs of numbers in the boxes give the *payoff* to each player with the left (or row) player's first and then the top (or column) player's. Thus (1,1) in the bottom left box tells us that, if Jack goes and Jill stops, Jack receives '1' and Jill receives '1'. The payoff numbers sometimes represent specific 'goods' like sums of money (or 'bads', if prefaced by a minus sign). Sometimes they are to be thought of as 'utils', the famous units of Jeremy Bentham's 'felicific calculus' which he hoped would one day enable utilitarians to calculate 'the greater happiness of the greater number', when choosing among actions or social policies. Sometimes they merely represent each player's preferences, rather than quantities of anything, so that Figure 6.2 tells us only that both players prefer the two outcomes where they coordinate to those where they do not. I postpone any queries about whether it matters how we read the *utility numbers*, as they are usually called.

In Figure 6.2 Jack and Jill are stuck. There are, in a sense, two *solutions* to the game and hence, in another sense, none. (Stop, Go) is a solution, in the sense that if Jack stops, Jill's rational choice is

		Jill: Stop	Jill: Go
Jack	Stop	0, 0	1, 1
	Go	1, 1	0, 0

Figure 6.2 Coordination game I

to go; and if Jill goes, Jack's rational choice is to stop. That gives us the crucial notion of a *Nash-equilibrium*, a pair of strategies, one for each player, where each is a best reply to the other. The pair forms an equilibrium in that it is a stable outcome, since neither has a better strategy, given the other's strategy. (It is called a 'Nash-equilibrium' after the game theorist John Nash.) (Stop, Stop) is not an equilibrium, because, if Jack stops, Jill's rational choice is 'Go'.

If (Stop, Go) were the only equilibrium, we might conclude that the players would know what to do. But (Go, Stop) is another. Neither Jack nor Jill can deduce which the other will aim for. So perhaps each considers tossing a coin. This introduces a further idea, that of a *mixed strategy*, which I shall mention briefly and then set aside, because it only spoils the elegance and impact of the basic analysis. A mixed strategy is one arrived at by, so to speak, using weighted dice to decide. Here Jack might play Stop with a probability of one half; if Jill knew this, she could not improve on doing likewise; nor could Jack then do better by changing his strategy. The pair of strategies would form a *mixed strategy equilibrium*. (The example should not be read as suggesting that one half is the only suitable probability value or that there is a mixed strategy equilibrium only where the players are otherwise stuck.)

In a coordination game, then, there are two (or more) equilibria and the players must each decide which to aim for. If Figure 6.2 captures the nub of the situation, why are narrow bridges not a source of chaos? How, for that matter, do motorists manage to pass on roads, where each is, in the abstract, indifferent between both keeping left and both keeping right? How, in general, do strangers manage to coordinate a hundred times a day? One answer starts by pointing out that Figure 6.2 represents a *one-shot* game, taken in isolation. Matters would be different, if it belonged to a series or *supergame*. If Jack and Jill were regularly to arrive at the bridge at the same moment, a convention might emerge, for instance that Jill should go first. Or, even if this is their only encounter, they might be able to take advantage of conventions which have emerged in other games, like 'ladies first' or 'the driver travelling uphill has right of way'. Or they might talk it

over and agree to a solution. There seem to be many possibilities. But we need to tread gingerly. If Game Theory is to be a potent tool for analysing social life, it must not simply assume the existence of conventions. In its ambitious versions, at any rate, it needs to show how rational agents can arrive at them and exactly how they can make one outcome a salient or focal point to aim at. Nor may it assume that the players can make agreements with the aid of language, if, as game theorists commonly presume, language is grounded in conventions. There are depths here, which we shall return to.

Meanwhile, not all coordination games have the symmetry of Figure 6.2, where the two equilibria are equally ranked by both players. Figure 6.3 shows a coordination game where Jack and Jill would both rather she went first.

There is still a second equilibrium, where Jack chooses 'Go', and Jill's best reply is 'Stop', in which case Jack indeed does best to choose 'Go'. But its payoff is (1, 1) which is worse for both than (2, 2). Where an outcome is superior for all players, it is natural to assume that each player is rational to play the strategy which contributes to it. (Such an outcome is dubbed 'Pareto-superior', after Vilfredo Pareto, an outcome being Pareto-superior to another if at least one player does better and no one does worse). In that case the coordination game in Figure 6.3 would have a unique solution, although even this seemingly irresistible thought will be challenged later.

		Jill	
		Stop	Go
Jack	Stop	0, 0	2, 2
	Go	1, 1	0, 0

Figure 6.3 Coordination game II

The key point about coordination games is that the players have a mutual interest in coordinating. Since there is no conflict of interest, it would be odd if rational agents failed to find mutually beneficial solutions, at least with repeated play. Superficially at any rate, this thought offers a simple key to the existence of a society where 'every agent is actuated solely by self-interest'. Provided that individual interests are mutually served, it is not surprising that individuals form associations. There is no puzzle about the emergence of a society which improves on the state of nature for everyone. Furthermore, conventions which benefit everyone need no enforcing among rational agents. So, if interests were never in conflict, civil society could be analysed as a set of coordination games and anarchists might be right to maintain that a society without government is as possible as it is desirable.

(2) The Prisoner's Dilemma

Since we are analysing an ideal-type world where everyone is wholly rational, it may not be an objection to this anarchist's utopia that social life is not utopian in practice. But interests can conflict, perhaps even in utopia. Where there is pure conflict, games may have no solutions short of war. Where interests overlap without coinciding, however, Game Theory becomes fascinating.

Suppose that Jack and Jill have become better acquainted and discovered a mutual penchant for crime. Soon they have committed several robberies and have graduated to murder. Sadly for them, however, they have just been arrested by a vigilant police force. The police can prove that they have done a robbery and know, but cannot prove without a confession, that they have also committed a murder. The cunning police chief has them placed in separate cells and makes each an offer. 'If you care to confess to the murder,' he assures each, 'and your partner does not, you will go entirely free. Your partner will be charged and executed but you will go free; and, of course, *vice versa*, if your partner confesses and you do not. If you both confess, you will both be charged with murder and, as a reward for being helpful, will receive a ten-year

sentence. If neither confesses, you will get two years apiece for robbery. This offer is being made to each of you.' Assuming that this can all be taken at face value, and Jack and Jill know it, what is each player's rational strategy?

The game is given in Figure 6.4, with the utility numbers representing the preference orders over the possible outcomes. Thus both rank (Silent, Silent) above (Confess, Confess) but are sharply at odds for the rest. Jack will be best suited if he confesses and she sits tight and worst suited contrariwise; for Jill it is the other way about. It may seem at first that each player's rational strategy depends on what the other is likely to do. After all, silence is a better strategy for both than confession. But Game Theory bids Jack reflect that, if Jill confesses, he does better to confess (ten years beats execution), and that, if Jill stays silent, he again does better to confess (freedom beats two years in gaol). Hence confession is better for him, *whatever she does*; and, by parity of reasoning, better for her, *whatever he does*. So both confess regardless and a contented chief of police has them both sent down for a decade.

To arrive at this conclusion abstractly, look at the utility numbers first from Jack's point of view and then from Jill's. Jack notes that 'Confess' scores him 4 and 'Silent' scores him 3, if Jill chooses the left-hand column; and 'Confess' scores him 2 and 'Silent' scores him 1, if she chooses the right hand column. Thus 'Confess' is his *dominant* strategy, in that it scores higher than 'Silent' (the *dominated* strategy), whatever Jill does. Similarly, for Jill's payoffs, 4 beats 3 in the top row and 2 beats 1 in the bottom

		Jill	
		Silent	Confess
Jack	Silent	3, 3	1, 4
	Confess	4, 1	2, 2

Figure 6.4 The Prisoner's Dilemma

row, making 'Confess' her dominant strategy. A rational agent never chooses a dominated strategy.

Is there no honour among thieves? Not if, as in Figure 6.4, there is an unique Nash-equilibrium, however disastrous. Confession is the best reply to confession; silence is not the best reply to silence. How exactly has the chief of police trapped them? An obvious suggestion is that the trick was to put them in separate cells, where they could not communicate. Very well, give them a few minutes together to plan a joint strategy. No doubt they will agree to keep silent. But will they keep their agreement? Each must now choose whether to keep it or break it; so substitute 'Keep' for 'Silent' and 'Break' for 'Confess' in Figure 6.4. That seems fair enough, since, by the measure of the resulting sentences, each presumably ranks the outcomes as in the table.

Outcome			*Ranking*
Self	+	*Other*	
Break		Keep	1st
Keep		Keep	2nd
Break		Break	3rd
Keep		Break	4th

If Other plays 'Keep', Self does better to play 'Break'; if Other plays 'Break', Self again does better to play 'Break'. As in Figure 6.4, there is a dominant strategy for both players and it sums to a Pareto-inferior outcome. In other words, utility numbers being unaffected, communication and an apparent agreement leave the game essentially as before. Jack and Jill have a mutual interest in avoiding confession by both but cannot yet escape their dilemma by agreeing to cooperate. Words are cheap talk, as game theorists are wont to say.

What happens if the game is repeated several times? Here it may seem that each player will perceive the folly of pegging away at an inferior outcome. A strategy of 'tit for tat' looks more promising: play obligingly in the first round and then keep it up, provided that the other player does likewise. If Jack and Jill

both play tit for tat in a ten-round game with the payoffs as in Figure 6.4, each will emerge with a score of thirty, instead of the twenty which repeating the unsociable strategy of the one-shot game would produce.

But this thought hits a dramatic snag. In a ten-round game, there is nothing to gain from cooperative play in the final round. Since there is no further round, where uncooperative play can be punished, the final round is effectively a one-shot game with the old strategy dominant. But in that case there is nothing to gain from cooperative play in the ninth round, since rational players are going to play their dominant strategy in the tenth regardless. The same therefore goes for the eighth round, the seventh, the sixth and so on. The whole cooperative scheme unravels back to the opening round, where Jack and Jill will thus find that each is rational to play the equilibrium strategy whatever the other does. How sad!

That chain of reasoning is typical of some quirky-looking results apparently implicit in Game Theory, and presents a significant crux, to be explored later. It does not hold, if the game has infinitely many rounds or if players do not know which round is the last. But, where it does hold, it holds however many rounds the game has. The case for cooperation unravels logically as easily for one hundred rounds as for ten, even though one might suppose that the trouble starts only as the game nears its end. Intriguingly, however, when the game is played with student volunteers in the economics research laboratory at my university, the (known) length of the series does matter, with defections starting only when the end comes close and economics students defecting earlier than others. This is consistent with similar American reports (including the point about economics students) and, more loosely, with what seems to happen in everyday life. Is there more honour even among thieves than Game Theory implies? If so, is it rational or irrational to act on it?

Before we broach such questions, here are two more games. Since I cite them for what they have in common, one might have been enough. But, since they are distinct and tend to turn up in different realms of discussion, there is room for both.

(3) Chicken

When Jack and Jill emerge from prison, they are angry enough to fight a duel. Like Texas teenagers of the 1950s (whose behaviour gave the game its name), they each acquire a car and line up facing each other a distance apart on an empty lonely road. In a moment they will drive down the narrow highway at speed towards a head-on collision. Whoever swerves first will be 'Chicken'. For each, loss of face would be terrible but a crash would be worse.

Chicken differs from the Prisoner's Dilemma in having two equilibria with pure (unmixed) strategies: (Swerve, Middle) and (Middle, Swerve). Jack thus has no clear choice of strategy, since, by parity of reasoning, neither has Jill. That might seem to make for a happy outcome, at least in a Chicken supergame, since the players each lack a dominant strategy leading to collective disaster. But, whereas there may be some pressure (so far mysterious) to cooperate in the repeated Prisoner's Dilemma, players of Chicken are less disposed to be amiable. Since Jack knows that it is rational for Jill to swerve if she expects him not to, it pays him to acquire a reputation for bravado. One way to acquire it, moreover, is to play even a one-shot game by visibly setting himself on a collision course from which she knows that he cannot back down. These strategies of 'reputation-building' and 'commitment', as game theorists dub them, enrich the game in

		Jill	
		Swerve	Middle
Jack	Swerve	3, 3	2, 4
	Middle	4, 2	1, 1

Figure 6.5 Chicken

ways which do not tend to benefit both players, as they do in the Prisoner's Dilemma.

That makes it disquieting to find Chicken analysed in depth by those studying arms races. There has been much debate whether the game which reveals most about an arms race is Chicken, the Prisoner's Dilemma or some other game. The belief that it is Chicken inclines policy-makers to nuclear strategies which petrify the rest of us. But complexities soon set in. For instance, the chilling policy of Mutual Assured Destruction (MAD) recommends arming to the teeth with nuclear weapons on the grounds that no one will then dare to start the Chicken game. Even if this were to make for nuclear peace, however, it might thereby encourage Chicken games with lesser weapons, since no one will dare retaliate with a nuclear strike.

Even without pursuing the example, we can notice a further crux. The real-world games of war and peace are played not among ideally rational agents but among less abstract Jacks and Jills. Jack needs to know not whether the game truly is Chicken but whether Jill thinks it is. To complicate matters, however, there is a sense in which the game cannot fail to be what the players think it is: the game is theirs. Also, neither would quite think it a 'game' of any sort, were there not intense interaction between academic game theorists and policy-makers (who pay for much of the research). Here is a case where the ideas which social scientists put into the heads of agents shape the very world which the social scientists are trying to analyse. If the case is as typical as I suspect, it gives reason for thinking not only that the agents' understanding is relevant to the social scientists' explanations but also that, being the stuff of the social world, it sets them a profound methodological challenge.

(4) Battle of the Sexes

Having at last cooled off, Jack and Jill resume their partnership. But their problems are not over. Soon they find themselves playing Battle of the Sexes. This game gets its Thurberesque or mildly Freudian name from the following scenario. Jack and Jill have agreed to spend the evening together at an event, which is to be

		Jill: Bullfight	Jill: Concert
Jack	Bullfight	4, 3	2, 1
	Concert	1, 2	3, 4

Figure 6.6 Battle of the Sexes

either a bullfight or a concert. But they have forgotten to agree which and it is too late to communicate. Each prefers an event in the other's company to going alone but Jack likes bullfights and Jill likes concerts. Each must therefore choose where to go, given the preferences shown in Figure 6.6. How should each choose?

Here (4, 3) and (3, 4) are both equilibria but, for a one-shot game, there is no pure strategy. If a convention were to emerge, for instance that women defer to the wishes of men, then the supergame might be determinate and so might even the one-shot game in a society where the convention was well known. For, as soon as Jack believes that Jill expects him to go to the bullfight and Jill knows this, the bullfight becomes the rational choice for both. In a repeated game, these expectations serve to lock Jill into a series of bullfights – a symbolic point about any society where even one-shot encounters are influenced by a presumption that the Jacks call the shots and the Jills oblige. Potentially there is a highly instructive lesson here about the nature of power and why losers are 'rational' to respect a distribution of power which works against them, when out-of-equilibrium strategies would suit them worse. It bears, for instance, on whether those who engage in free-market transactions thereby show themselves to do so freely and with consent, or, on the contrary, may merely be assisting in their own alienation.

These four games are not the only ones worth knowing about but I hope that they are enough to give a feel for game-theoretic analysis, to indicate some instructive applications and, in

a moment, to raise questions about its basis. At any rate, we can now address its claim to offer a thoroughly individualist analysis of social institutions and of society at large. I shall make some general remarks about social theories which rest on a notion of contract and then raise a crux about the analysis of social norms.

THE SOCIAL CONTRACT

Ambitious claims are often made for Game Theory as a tool of social analysis, especially by those who see it as the cutting edge of a general individualism. We have had strong assurances from J. S. Mill that 'human beings in society have no properties but those which are derived from, and may be resolved into, the laws of nature of individual man' (p.10), and from Jon Elster that 'the elementary unit of social life is the individual human action' (p.19). Yet we are born into a world of institutions which socialise us, shape our goals and values, condition our options and outlast our demise. So what could render plausible Elster's claim that 'to explain social institutions and social change is to show how they arise as a result of the action and interaction of individuals'?

The broadest question is why societies exist at all and a simple answer might be that they embody a social contract, being associations of individuals who find it rational to cooperate. Coordination games illustrate this answer readily and offer the plausible suggestion that primary 'institutions' are simply the sort of conventions which emerge with repeated play as guidance where there are multiple equilibria. Wherever there is no conflict of interest, individuals have nothing to lose and plenty to gain by hitting on rules which it suits everyone to follow. Although not every institution can be analysed as a deposit resulting from previous games of coordination, it is not implausible to suggest that societies rest ultimately on mutual self-interest, so that their very existence can be analysed as a solution to a basic problem of coordination. To strengthen this ingenuous thought, it is especially plausible to think of language as a set of conventions which serve a mutual interest in coordination. It does not matter what we call a spade, provided that we all give it the same name. Different languages are different solutions to the same ultimate

coordination game, and so, perhaps, are the various schemes of moral or normative concepts through which societies ensure stability in thought, word and deed.

If this were all there was to it, society could, in theory, arise through a sort of social contract which needs no enforcement and exist without government. This cheerful anarchism is not refuted by the obvious prevalence of enforcements, because the process of coordination can go wrong. One enjoyable anarchist theme is that government is indeed needed, but only to remedy the ills caused by government, without which there would be no ills to remedy. Similar thinking may inspire libertarian ideas of what a completely free market would achieve, although this is not to belittle the strenuous intellectual efforts by economists which yield the complexities of general equilibrium theory. Enforcement is needed only to correct distortions of what could be better achieved by unenforced cooperation among fully rational individuals.

A contrary view of the social contract makes the Prisoner's Dilemma the crucial game. This view has a distinguished pedigree, usually traced to *Leviathan* by Thomas Hobbes, published in 1651 in the aftermath of the English civil war. Contemplating that grim episode, Hobbes produced an individualist analysis of human nature and society intended to explain 'the art of making and maintaining commonwealths'. *Leviathan* opens with several chapters exhibiting human beings as mechanically-driven creatures, each bent on securing his own 'felicity'. 'The felicity of this life,' Hobbes remarks in Chapter 9, 'consisteth not in the repose of a mind satisfied' but in a continual progress of desire from one object to another, not 'to enjoy once only, and for one instant of time; but to assure for ever the way of his future desire'. This being the human condition, all mankind has 'a perpetual and restless desire of power after power, that ceaseth only in death'. Here is a classic statement of the idea that all rational action is aimed at maximising the agent's expected utility, with the Hobbesian rider that, 'because life itself is but motion' (Chapter 6), the aim can never remain satisfied for long.

In that case it is far from plain how society is possible and Hobbes turns a sharp eye to 'those qualities of mankind concerning their living together in peace and unity'. The mordant

Chapter 13 is titled 'Of the Natural Condition of Mankind as Concerning their Felicity, and Misery'. It takes the crux to be that 'if any two men desire the same thing, which nevertheless they cannot both enjoy, they become enemies' and opines that 'in the nature of man there are three principal causes of quarrel'. These are 'first, competition; secondly, diffidence; thirdly, glory'. Competition makes men invade one another for gain. Diffidence, or, as we would now say, distrust, makes for preemptive strikes. Glory, which equates loosely with what we now call status, makes men aggressive whenever they sense that they are undervalued. Such inbuilt causes of quarrel make mere associations fragile, to say the least. 'Hence it is manifest that during the time men live without a common power to keep them all in awe, they are in that condition which is called war; and such a war as is of everyman against everyman'. This condition is Hobbes' celebrated and bleak state of nature, where 'there is continual fear, and danger of violent death; and the life of man is solitary, poor, nasty, brutish and short'.

So how do we come to live together in peace and unity? Hobbes reasons that men are inclined to peace by 'fear of death; desire of such things as are necessary to commodious living; and a hope by their industry to obtain them'. These passions incline us to peace, but are not enough to overcome the causes of quarrel, however, unless there is 'a common power to keep all in awe'. Otherwise we shall each continue to invade one another, because that remains our dominant strategy, whatever others do. We are still 'enemies', wanting things which we cannot all have and, although each will apparently subscribe to articles of peace, each will break them if he can. Since 'the weakest has strength enough to kill the strongest, either by secret machination, or in confederacy with others, that are in the same danger as himself', no one is safe.

There is some dispute among game theorists as to which game best illuminates the theme of *Leviathan*. But there is an evident case for deeming it the Prisoner's Dilemma, although with more players than Jack and Jill – an *n*-person version, setting what game theorists term the *free-rider problem*. Everyone is better served by peace than by war; so we might suppose that peace will emerge

spontaneously. But, even if it did, each player is still better served by being a free-rider who takes the benefits but fails to contribute. For instance, if a peaceable convention of promise-keeping emerges, the free-rider makes promises, gets something in exchange and then does not keep them. If all behave like this, society reverts, of course, to war (whose nature 'consisteth not in actual fighting; but in the known disposition thereto'). The bones of the argument were contained in the discussion of the vain agreement to keep silent which Jack and Jill made on p.125, when allowed to communicate in the Prisoner's Dilemma.

In Hobbes' words 'Covenants without the sword are but words, and of no strength to secure a man at all' (Chapter 17). Accordingly he argues that the only escape is to create 'a power to keep all in awe' and to arm this sovereign with the sword. That is the 'Leviathan' of the book's title, a sovereign authority created by a social contract to protect us from invasion, domestic and foreign, and to secure us in our covenants with one another. The Introduction describes Leviathan as 'an artificial man', shown in the original frontispiece as a crowned king armed with the weapons of church and state. This sovereign figure seems to be wearing chain mail, but look closer and it is in fact made up of tiny human individuals. That captures Hobbes' theme precisely. Society is an artifice which lets rational individuals escape the Prisoner's Dilemma.

NORMS AND COOPERATION

That gives us two individualist ways of analysing social norms, both fundamental enough to be called theories of the social contract. They are not the only such theories but they point conveniently to a broad division of social models into those premised on consensus and those premised on conflict. Broadly, consensus models start with coordination and then need to account for norms which are not merely self-enforcing; conflict models insist that our basic interests may overlap but emphatically do not coincide, and thus make the root problem how cooperation is possible. Either way, we are owed a theoretical account of the

role of norms in a world where all agents are rational individuals guided by their own preferences.

The shared individualism makes the basic game in both analyses *non-cooperative* in Game Theory parlance. This may sound a perverse way to describe a coordination game, where the players have a mutual interest in cooperating. But game theorists classify a game as *cooperative* only if players can rely on any agreement being kept, and conventions arising even in coordination games do not have the strongly binding character here envisaged. Whereas non-cooperative games presuppose a solution to the problem of how norms arise and why they persist, coordination games seem to need make no assumption of norms and institutions are thus deemed non-cooperative. The problem, in essence, is how and whether non-cooperative games can give rise to cooperative ones.

Hobbes' answer is that they cannot do so directly but that rational individuals can agree to create a power which they cannot then escape. With Leviathan in place and equipped with a sword, everyday contracts can be made and bargains struck in the knowledge that defections will be punished. In the state of nature everyone arms to the teeth because being armed is better than not being armed, whatever others do. But once there is a sovereign claiming the monopoly of legitimate force, the payoffs for going armed change, since peace and 'commodious living' become possible. This strongly suggests that norms, like promise-keeping, truth-telling and respect for moral obligations in general, work only in so far as there are sanctions. We are good when it pays to be good; and it pays only when we are sure of punishment for being bad.

Interestingly, however, Hobbes clearly thinks that, although there are no obligations in the state of nature ('the notions of right and wrong, justice and injustice have there no place'), the creation of Leviathan makes possible obligations which are indeed morally binding. Even prisoners of war, released upon a promise to pay a ransom, have an obligation to pay up, even though beyond reach of reprisal. This goes with his finely balanced reply to 'the fool', who 'hath said in his heart that there is no such thing as justice' (Chapter 15) and remains critical for current

contractarian theories of ethics and justice, like David Gauthier's *Morals by Agreement* (1986). Whether Hobbes can maintain this, given his own analysis, is disputable and I shall not discuss *Leviathan* itself further here. But the issue is absolutely crucial for the whole attempt to analyse institutions as conventions and to regard conventions as emerging from interaction among rational agents to their mutual benefit.

No society can function without trust. Our account of games among instrumentally rational agents leaves it unclear how far ideally rational agents can be trusted. That is partly, I think, because the exact sense of 'trust' is not yet plain and partly because too little has yet been said about how rational agents are motivated.

There is a weak sense of 'trust' in which Jack can be trusted to do whatever it is predictable that he will do, rather as a reliable alarm clock can be trusted to ring at the set time. In this sense, Jack can be trusted to do what maximises his expected utility. But, to get at the problem of trust, we need to distinguish this usage from two others. To say that no society can function without trust is usually either to say that it needs social norms to do with truth-telling, promise-keeping and the honouring of agreements, which operate even on occasions when one could break them without penalty, or to say that it needs members who recognise and respect moral obligations. Whether social norms and moral obligations are finally distinct is for dispute and more will be said in Chapter 10. To bring out what they have in common, consider the game theorist's apothegm that 'words are cheap talk'. The thought behind it is that, since Jack will do only what suits his preferences, Jill does well to attend to his preferences and not to his words. For example, his offer to keep silent in a one-shot game is to be trusted only if he would keep silent without it. By contrast, for persons bound by social norms or moral principles, to give one's word creates an effective reason for keeping it.

Since strategies of reputation-building and commitment are available to rational agents, it may seem that Game Theory can incorporate binding agreements readily enough. But the motivation of a rational agent is solely forward-looking. All the game diagrams indicate clearly that actions are motivated solely by their

resulting payoffs. Jack's preferences can, of course, be influenced by sanctions and it may be that some of these sanctions are internal. Thus it may be that he feels rotten, if he breaks a promise; but, if so, this shows up as a disutility in his mathematical psychics. Reasons for action are never backward-looking, in the sense of operating solely because of a past event. To make that possible, we would need a notion of rationality distinct from the instrumental one governing this chapter.

There may be scope, however, for making agents more complex. Comparison with utilitarianism is instructive. Its initial version, that we should always do the *act* with the best consequences, arguably could, if adopted by everyone, lead to a society without a reliable institution of promise-keeping. This consequence would be worse than what would happen if we always acted in accordance with the *rules* which would lead to the best consequences if followed by all. Rule-utilitarianism can thus seem the better version, with its forward-looking reasons for acting on backward-looking reasons. But critics complain that it does the job only if so Kantian that it is no longer a version of utilitarianism. Similarly, rational agents who were more reflective and distanced from their own preferences might each do better for themselves. But it is not plain that the change can be introduced without destroying the basis of Rational Choice theory. Whether it can will become clearer in Chapter 9.

Short of this, there remains a strong case for holding that cooperation sometimes requires the *prior* acceptance of social norms or moral obligations which the theory therefore cannot account for from scratch. Yet I do not want to press too hard. The examples offered have been ones where out-of-equilibrium outcomes are Pareto-superior. But what might rationally prompt honour among thieves in the Prisoner's Dilemma or mutual non-aggression in Chicken is only a higher prudence. It has yet to be proved that suitably prudent agents need be radically different from the rational seekers of marginal advantage with whom we began.

Even if some conventions are proving elusive to analyse, thus giving trouble for the individualist account of rules norms and practices, others still seem straightforward. So let me just mention

very briefly two further problems, both arising from the existence of multiple equilibria in many games.

The more obvious one is illustrated by Battle of the Sexes, where each equilibrium distributes benefits unequally. If a convention emerges, for instance that Jills do what best pleases Jacks, it is easy to see why it may persist. But how does one particular equilibrium emerge as salient or focal? Chance is a possible answer. But, contemplating plausible examples of the game, one is more inclined to point to the distribution of power. Nothing in the chapter affords any clue to the nature and origins of power, since vague remarks about social evolution merely veil the problem and the fiction of an explicit social contract is only a fiction.

The subtler problem concerns the whole idea of convention. A convention is being analysed as a set of mutual expectations which reinforce one another to make a particular equilibrium salient. On the highway, for instance, Jack is rational to keep left if he expects Jill to keep left and *vice versa*. Yes; but he is also rational to keep right if he expects Jill to keep right and *vice versa*. Both pairs of expectations are mutually self-reinforcing. How and why exactly does an established habit of keeping left (in Britain) give Jack a sufficient reason to keep left next time? To reply that left has become salient begs the question and, surprisingly, nothing yet said offers more. Left is Jack's rational choice only if he expects that Jill expects . . . (that he expects that she expects . . .) that he will choose left. Equally, right is his rational choice if he expects that she expects . . . (that he expects that she expects . . .) that he will choose right. Each of these infinite hypotheticals is a tautology and nothing shows why previous behaviour renders one categorical and the other irrelevant.

Even more astonishingly, perhaps, the same point arises with the version in Figure 6.3, where both players prefer the same equilibrium.

Stop is indeed Jack's uniquely rational choice *if* he expects that she expects . . . But, equally, 'Go' is still his uniquely rational choice, *if* he expects that she expects . . . Again something more is wanted than has yet been offered to determine which hypothetical is to guide action.

Jack \ Jill	Stop	Go
Stop	0, 0	2, 2
Go	1, 1	0, 0

Figure 6.3 Coordination game II

This may seem incredible. But the case for it is accepted by some, although not all, game theorists and stems directly from Thomas Schelling's respected *The Strategy of Conflict* (1960). If it is correct, it undermines David Lewis' *Convention* (1969), the classic statement of the view of conventions, linguistic and social, which is standard for Game Theory and has been assumed in this chapter. Incredible or not, however, it would not astonish two older philosophers. Hume would take it to reinforce his contention that all our reasonings rest in the end on custom and so cannot explain the basis of custom. Kant would take it to show that the instrumental rationality of self-interested prudence is subordinate to a higher rationality of practical reason where each person is directly motivated to do what is right by the test of fairness, justice or morality.

Since there is no room to flesh out these quick hints at deep philosophical perplexities, it is time to sum up.

CONCLUSION

Game Theory abstracts from the tangles of the social world to a pure realm peopled by ideally rational agents equipped with fully ordered preferences, complete information and a perfect internal computer. Their preferences can be completely and consistently represented by a ranking of the possible outcomes of interaction, an interaction being the sum of the consequences of separate individual actions. Their information includes 'common knowledge' that other players are rational agents, and is so complete that anything known to anyone is known to everyone. Their computers see to it that, taking probabilities into account, each can derive everyone's rational strategy, where there is one. Although we have considered only four of the games such people play, we can see the power of the theory and, at the same time, raise some awkward questions about its limitations.

Coordination games introduce the basic notion of strategic choice. Jack's rational choice depends on what Jill will choose and *vice versa*. With repeated play, it is easy to conjecture that the emergence of a convention can guide them to a mutually beneficial equilibrium. That makes an interesting suggestion about a sort of norm which needs no enforcement and, more grandly, about consensus as the basis of a theory of the social contract. On reflection, however, we can still wonder whether Game Theory itself contains enough to explain exactly how and why conventions guide choices.

The Prisoner's Dilemma makes the vital point that individually rational choices can sum to collectively inferior results. An 'Invisible Hand' often makes mischief for all. Interestingly, the kind of norm which might prevent this happening seems to need enforcement, because it is otherwise subject to free-riding. Hobbes thought so, and *Leviathan* remains crucial for theories of the social contract and much else. Meanwhile down-to-earth examples of the Dilemma are legion, if one examines the fragility of attempts to preserve rain forests, protect endangered species, save energy, achieve a voluntary incomes policy, stop the arms race, prevent global warming and more modestly, Keep Britain Tidy. Yet, if what is really needed is genuine trust and moral conduct, it is tantalisingly unclear whether one would want even a fully rational agent as one's neighbour. How far can such a rational fool ultimately be trusted?

Chicken sets the problem of what strategy is rational in a game without any pure equilibria, a problem also arising in many other games and indeed in everyday life. If Jack is unsure of his rational strategy because unsure of Jill's, then her uncertainties are increased by contemplating his. This makes Chicken games

dangerous, indeed lethal if played with weapons of destruction. Moreover players in real life may be uncertain whether their present game is truly Chicken and whether the other players also take it to be Chicken. Among the intriguing questions raised is what difference it makes when flesh-and-blood players, like American nuclear strategists or British Treasury mandarins, have had lessons in Game Theory and bear them in mind.

Battle of the Sexes was discussed only fleetingly but will crop up again. Jack and Jill both gain from coordination but the two ways of achieving it benefit them differentially. In a repeated game one player looks like getting locked in to an inferior equilibrium. Meanwhile the general question again arises about the limitations of Game Theory in analysing games where there are multiple equilibria.

Individualism, as purveyed by Rational Choice theory and Game Theory, deals with social norms in two ways. One is by showing how repeated interaction can generate them as solutions to problems arising in the games. But even if that works for truly consensual norms which suit everyone, it remains unclear that it does so for norms vulnerable to free-riding. The crux is trust and whether rational prudence can make us trustworthy even on occasions when we could escape reprisal. The other way is to tuck them into agents' preferences. Thus the good Samaritan had altruistic preferences which led him to rescue a stranger, while those with other preferences walked past on the other side of the road. When George Washington, as legend has it, owned up to felling a cherry tree with the words 'Father, I cannot tell a lie', he was acting on a strong ethical preference for honesty. The theory is silent about the sources of preference, which it treats as 'given', and the crux is whether, in failing to say more, it leaves individualism to be trumped by a holistic story about the social determination of preferences.

For instance, to look ahead, preferences are often aligned with roles. Parents tend to prefer outcomes which benefit their children. French industrial workers like M. Rouget tend to prefer the policies of the Left. Bureaucrats often make the interests of their bureaucracy their own. Military advisers to governments favour military solutions to political problems. This means that choices

which might seem inferior, if one considered them in abstraction from social positions, can become entirely rational if one connects preferences to roles, thus embedding norms in the analysis. It is disingenuous to claim that Rational Choice theory is thereby enabled to account for role-specific behaviour. If a structure of social relations is being tucked away in agents' preference orders and is doing the effective work, then we shall want to know more about it than Rational Choice theory attempts to tell us. A similar point can be made about the enriched psychology just envisaged, where it sounds no less disingenuous to present altruists as people who happen to derive utility from raising the utility of others. But this is to anticipate Chapter 8 and I end by identifying three puzzles.

(1) Can game-theoretic analysis account for all manner of social norms or must it presuppose at least some of them?
(2) How does its abstract analysis of an ideal-type world peopled by ideally rational agents relate to our ordinary world of unidealised persons?
(3) Where it does reveal significant features of social interaction, is it an exercise in explanation or in understanding?