moving within the rules, but also it requires the ongoing creation of the rules. And if this were not enough, it involves the ever shifting profile of objectives, beliefs and concerns of each author as the writing proceeds. Our one important thought in this book is that game theory will remain deficient until it develops an interest in games like the one we experienced over the last two years. Is it any wonder that this is *A Critical Introduction*?

Lastly, there are the people and the cats: Lucky, Margarita, Pandora, Sue, Thibeau and Tolstoy – thank you.

Shaun P. Hargreaves Heap
Yanis Varoufakis
May 1994

1

AN OVERVIEW

1.1 INTRODUCTION

1.1.1 Why study game theory?

Game theory is everywhere these days. After thrilling a whole generation of post–1970 economists, it is spreading like a bushfire through the social sciences. Two prominent game theorists, Robert Aumann and Oliver Hart, explain the attraction in the following way:

> Game Theory may be viewed as a sort of umbrella or 'unified field' theory for the rational side of social science . . . [it] does not use different, ad hoc constructs . . . it develops methodologies that apply in principle to all interactive situations.
>
> (Aumann and Hart, 1992)

Of course, you might say, two practitioners would say that, wouldn't they. But the view is widely held, even among apparently disinterested parties. Jon Elster, for instance, a well-known social theorist with very diverse interests, remarks in a similar fashion:

> if one accepts that interaction is the essence of social life, then . . . game theory provides solid microfoundations for the study of social structure and social change.
>
> (Elster, 1982)

In many respects this enthusiasm is not difficult to understand. Game theory was probably born with the publication of *The Theory of Games and Economic Behaviour* by John von Neumann and Oskar Morgenstern (first published in 1944 with second and third editions in 1947 and 1953). They defined a game as any interaction between agents that is governed by a set of rules specifying the possible moves for each participant and a set of outcomes for each possible combination of moves. One is hard put to find an example of social phenomenon that cannot be so described. Thus a theory of games promises to apply to almost any social interaction where

individuals have some understanding of how the outcome for one is affected not only by his or her own actions but also by the actions of others. This is quite extraordinary. From crossing the road in traffic, to decisions to disarm, raise prices, give to charity, join a union, produce a commodity, have children, and so on, it seems we will now be able to draw on a single mode of analysis: the theory of games.

At the outset, we should make clear that we doubt such a claim is warranted. This is a *critical* guide to game theory. Make no mistake though, we enjoy game theory and have spent many hours pondering its various twists and turns. Indeed it has helped us on many issues. However, we believe that this is predominantly how game theory makes a contribution. It is useful mainly because it helps clarify some fundamental issues and debates in social science, for instance those within and around the political theory of liberal individualism. In this sense, we believe the contribution of game theory to be largely pedagogical. Such contributions are not to be sneezed at.

If game theory does make a further substantial contribution, then we believe that it is a negative one. The contribution comes through demonstrating the limits of a particular form of individualism in social science: one based *exclusively* on the model of persons as preference satisfiers. This model is often regarded as the direct heir of David Hume's (the 18th century philosopher) conceptualisation of human reasoning and motivation. It is principally associated with what is known today as rational choice theory, or with the (neoclassical) economic approach to social life (see Downs, 1957, and Becker, 1976). Our main conclusion on this theme (which we will develop through the book) can be rephrased accordingly: we believe that game theory reveals the limits of 'rational choice' and of the (neoclassical) economic approach to life. In other words, game theory does not actually deliver Jon Elster's 'solid microfoundations' for all social science; and this tells us something about the inadequacy of its chosen 'microfoundations'.

The next section (1.2) sketches the philosophical moorings of game theory, discussing in turn its three key assumptions: **agents are instrumentally rational (section 1.2.1); they have common knowledge of this rationality (section 1.2.2); and they know the rules of the game (section 1.2.3)**. These assumptions set out where game theory stands on the big questions of the sort 'who am I, what am I doing here and how can I know about either?'. The first and third are ontological.[1] They establish what game theory takes as the material of social science: in particular, what it takes to be the essence of individuals and their relation in society. The second raises epistemological issues[2] (and in some games it is not essential for the analysis). It is concerned with what can be inferred about the beliefs which people will hold about how games will be played when they have common knowledge of their rationality.

We spend more time discussing these assumptions than is perhaps usual in texts on game theory because we believe that the assumptions are both controversial and problematic, in their own terms, when cast as general propositions concerning interactions between individuals. This is one respect in which this is a critical introduction. The discussions of instrumental rationality and common knowledge of instrumental rationality (sections 1.2.1 and 1.2.2), in particular, are indispensable for anyone interested in game theory. In comparison section 1.2.3 will appeal more to those who are concerned with where game theory fits in to the wider debates within social science. Likewise, section 1.3 develops this broader interest by focusing on the potential contribution which game theory makes to an evaluation of the political theory of liberal individualism. We hope you will read these later sections, not least because the political theory of liberal individualism is extremely influential. Nevertheless, we recognise that these sections are not central to the exposition of game theory *per se* and they presuppose some familiarity with these wider debates within social science. For this reason some readers may prefer to skip through these sections now and return to them later.

Finally, section 1.4 offers an outline of the rest of the book. It begins by introducing the reader to actual games by means of three classic examples which have fascinated game theorists and which allow us to illustrate some of the ideas from sections 1.2 and 1.3. It concludes with a chapter-by-chapter guide to the book.

1.1.2 Why read this book?

In recent years the number of texts on game theory has multiplied. For example, Rasmussen (1989) is a good 'user's manual' with many economic illustrations. Binmore (1990) comprises lengthy, technical but stimulating essays on aspects of the theory. Kreps (1990) is a delightful book and an excellent eclectic introduction to game theory's strengths and problems. More recently, Myerson (1991), Fudenberg and Tirole (1991) and Binmore (1992) have been added to the burgeoning set. Dixit and Nalebuff (1993) contribute a more informal guide while Brams (1993) is a revisionist offering. One of our favourite books, despite its age and the fact that it is not an extensive guide to game theory, is Thomas Schelling's *The Strategy of Conflict*, first published in 1960. It is highly readable and packed with insights few other books can offer. However, *none* of these books locates game theory in the wider debates within social science. This is unfortunate for two reasons.

Firstly, it is liable to encourage further the insouciance among economists with respect to what is happening elsewhere in the social sciences. This is a pity because mainstream economics is actually founded on philosophically controversial premises and game theory is potentially in

rather a good position to reveal some of these foundational difficulties. In other words, what appear as 'puzzles' or 'tricky issues' to many game theorists are actually echoes of fundamental philosophical dispute; and so it would be unfortunate to overlook this invitation to more philosophical reflection.

Secondly, there is a danger that other social sciences will greet game theory as the latest manifestation of economic imperialism, to be championed only by those who prize technique most highly. Again this would be unfortunate because game theory really does speak to some of the fundamental disputes in social science and as such it should be an aid to all social scientists. Indeed, for those who are suspicious of economic imperialism within the social sciences, game theory is, somewhat ironically, a potential ally. Thus it would be a shame for those who feel embattled by the onward march of neoclassical economics if the potential services of an apostate within the very camp of economics itself were to be denied.

This book addresses these worries. It has been written for all social scientists. It does not claim to be an authoritative textbook on game theory. There are some highways and byways in game theory which are not travelled. But it does focus on the central concepts of game theory, and it aims to discuss them critically and simply while remaining faithful to their subtleties. Thus we have trimmed the technicalities to a minimum (you will only need a bit of algebra now and then) and our aim has been to lead with the ideas. We hope thereby to have written a book which will introduce game theory to students of economics and the other social sciences. In addition, we hope that, by connecting game theory to the wider debates within social science, the book will encourage both the interest of non-economists in game theory and the interest of economists to venture beyond their traditional and narrow philosophical basis.

1.2 THE ASSUMPTIONS OF GAME THEORY

Imagine you observe people playing with some cards. The activity appears to have some structure and you want to make sense of what is going on; who is doing what and why. It seems natural to break the problem into component parts. First we need to know the rules of the game because these will tell us what actions are permitted at any time. Then we need to know how people select an action from those that are permitted. This is the approach of game theory and the first two assumptions in this section address the last part of the problem: how people select an action. One focuses on what we should assume about what motivates each person (for instance, are they playing to win or are they just mucking about?) and the other is designed to help with the tricky issue of what each thinks the other will do in any set of circumstances.

1.2.1 Individual action is instrumentally rational

Individuals who are instrumentally rational have preferences over various 'things', e.g. bread over toast, toast and honey over bread and butter, rock over classical music, etc., and they are deemed rational because they select actions which will best satisfy those preferences. One of the virtues of this model is that very little needs to be assumed about a person's preferences. Rationality is cast in a means–end framework with the task of selecting the most appropriate means for achieving certain ends (i.e. preference satisfaction); and for this purpose, preferences (or 'ends') must be coherent in only a weak sense that we must be able to talk about satisfying them more or less. Technically we must have a 'preference ordering' because it is only when preferences are ordered that we will be able to begin to make judgements about how different actions satisfy our preferences in different degrees. In fact this need entail no more than a simple consistency of the sort that when rock music is preferred to classical and classical is preferred to muzak, then rock should also be preferred to muzak (the interested reader may consult Box 1.1 on this point).[3]

Thus it appears a promisingly general model of action. For instance, it could apply to any type of player of games and not just individuals. So long as the State or the working class or the police have a consistent set of objectives/preferences, then we could assume that it (or they) too act instrumentally so as to achieve those ends. Likewise it does not matter what ends a person pursues: they can be selfish, weird, altruistic or whatever; so long as they consistently motivate then people can still act so as to satisfy them best.

Readers familiar with neoclassical *Homo economicus* will need no further introduction. This is the model found in standard introductory texts, where preferences are represented by indifference curves (or utility functions) and agents are assumed rational because they select the action which attains the highest feasible indifference curve (maximises utility). For readers who have not come across these standard texts or who have forgotten them, it is worth explaining that preferences are sometimes represented mathematically by a utility function. As a result, acting instrumentally to satisfy best one's preferences becomes the equivalent of utility maximising behaviour. In short, the assumption of instrumental rationality cashes in as an assumption of utility maximising behaviour. Since game theory standardly employs the metaphor of utility maximisation in this way, and since this metaphor is open to misunderstanding, it is sensible to expand on this way of modelling instrumentally rational behaviour before we discuss some of its difficulties.

Ordinal utilities, cardinal utilities and expected utilities

Suppose a person is confronted by a choice between driving to work or catching the train (and they both cost the same). Driving means less waiting

in queues and greater privacy while catching the train allows one to read while on the move and is quicker. Economists assume we have a preference ordering: each one of us, perhaps after spending some time thinking about the dilemma, will rank the two possibilities (in case of indifference an equal ranking is given). The metaphor of utility maximisation then works in the following way. Suppose you prefer driving to catching the train and so choose to drive. We could say equivalently that you derive X utils from driving and Y from travelling on the train and you choose driving because this maximises the utils generated, as $X > Y$.

Box 1.1

UTILITY MAXIMISATION AND CONSISTENT CHOICE

Suppose that a person is choosing between different possible alternatives which we label x_1, x_2, etc. We shall use the following notation to describe the preferences which inform these choices: $x_1 \; x_2$ means that the person 'prefers x_1 to x_2 **or** is indifferent between them'; $x_1 \geq x_2$ means that he or she 'prefers x_1 to x_2'; and $x_1 = x_2$ means that he or she is 'indifferent between the two'. A person is deemed *instrumentally rational* if he or she has preferences which satisfy the following conditions:

(1) *Reflexivity*: For any x_i, $x_i \geq x_i$
(2) *Completeness*: For any x_i, x_j, either $x_i \geq x_j$ or $x_i \leq x_j$
(3) *Transitivity*: For any x_i, x_j, x_k, if $x_i \geq x_j$ **and** $x_j \geq x_k$, then $x_i \geq x_k$
(4) *Continuity*: For any x_i, x_j, x_k, if $x_i > x_j > x_k$, then there must exist some 'composite' of x_i and x_k, say y, which gives the same amount of utility as x_j; that is, $y = x_j$ and our individual is indifferent between them.

In the definition of *continuity* above there are more than one way of interpreting the 'composite' alternative denoted by y. One is to think of y as a basket containing bits of x_i and bits of x_k. For example, if x_i is '5 croissant', x_j is '3 bagels' and x_k is '10 bread rolls', then some combination of croissant and bread rolls (e.g. 2 croissant and 4 bread rolls) must be equally valued as the 3 bagels. Another interpretation of y is probabilistic. Imagine that y is a lottery which gives the individual x_i with probability p $(0 < p < 1)$ and x_k with probability $1 - p$. Then the continuity axiom says that there exists some probability p (e.g. 0.3) such that this lottery (that is, alternative y) is valued by the individual exactly as much as x_j.

When axioms (1), (2) and (3) hold, then the individual has a well-defined preference ordering. When (4) also holds, this preference ordering can be represented by a utility function. (A utility function takes what the individual has, e.g. x_i, and translates it into a unique level of utility. Its mathematical representation in this case is $U(x_i)$.) Thus the individual who makes choices with a view to satisfying his or her preference ordering can be conceived as one who is maximising this utility function.

Box 1.2

REFLECTIONS ON INSTRUMENTAL RATIONALITY

Instrumental rationality is identified with the capacity to choose actions which best satisfy a person's objectives. Although there is a tradition of instrumental thinking which goes back to the pre-Socratic philosophers, it is David Hume's *Treatise on Human Nature* which provides the clearest philosophical source. He argued that 'passions' motivate a person to act and 'reason' is their servant.

> We speak not strictly and philosophically when we talk of the combat of passion and reason. Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them.

Thus reason does not judge or attempt to modify our 'passions', as some might think. This, of course, does not mean that our 'passions' might not be 'good', 'bad', 'wishy-washy' when judged by some light or other. The point is that it is not the role of reason to form such judgements. Reason on this account merely guides action by selecting the best way to satisfy our 'passions'.

This hypothesis has been extremely influential in the social sciences. For instance, the mainstream, neoclassical school of economics has accepted this Humean view with some modification. They have substituted preferences for passions and they have required that these preferences should be consistent. This, in turn, yields a very precise interpretation for how instrumental reason goes to work. It is as if we had various desires or passions which when satisfied yield something in common; call it 'utility'. Thus the fact that different actions are liable to satisfy our different desires in varying degrees (for instance, eating some beans will assuage our desire for nourishment while listening to music will satisfy a desire for entertainment) presents no special problem for instrumental reason. Each action yields the same currency of pleasure ('utils') and so we can decide which action best satisfies our desires by seeing which generates the most 'utility' (see Box 1.1 on consistent choice).

This maximising, calculative view of instrumental reason is common in economics, but it needs careful handling because it is liable to suggest an unwarranted connection with the social philosophy of 'utilitarianism' as presented by Jeremy Bentham and later John Stewart Mill (especially since J.S. Mill is a key figure associated with both the beginnings of neoclassical economics and the social philosophy of utilitarianism). The key difference is that Bentham's social philosophy envisioned a universal currency of happiness for all people. Everything in people's lives either adds to the sum total of utility in society (i.e. it is pleasurable) or subtracts from it (i.e. is painful) and the good society is the one that maximises the sum of those utilities, or average utility (see also Box 4.5 in Chapter 4). This was a radical view at the time because it broke with the tradition of using some external authority (God, the Church, the Monarch) to judge social

outcomes, but it is plainly controversial now because it presumes we can compare one person's utility with another's. Neither neoclassical economics nor Humean philosophy is committed to such a view as the utility indices are purely personal assessments on these accounts and cannot be compared one with another.

The influence of instrumental reasoning stretches well beyond economics. Neoclassical economists have themselves exported this model of 'rational choice' to many other parts of the social sciences through the so-called 'economic' or 'rational choice' models of politics, marriage, divorce, suicide, etc. (see Becker, 1976). There is even the 'rational choice' version of Marxism (see Elster, 1986b). In turn, these efforts join forces with those of other social theorists. For example, Max Weber famously sees purposive rational action as one of the ideal types through which we can develop a rational understanding of individual action; and he regards the way that western institutions increasingly embody the character of calculative reason as one of the hallmarks of 'modernity'.

However, while (neoclassical) economists typically work only with instrumental reason, social theorists, like Weber and Jürgen Habermas, recognise other motivations. Thus instrumental reason is to be contrasted for Weber with 'value rational' action: that is, action which is to be understood *not* as a means to an end but as valuable in its own right. Likewise for Habermas the 'life form' of the human being cannot be simply reduced to the mastery over nature which is symptomatic of purposive (instrumentally) rational action. Our life form is distinguished by the fact that we reach understanding through language and this is the source of another kind of rationality, the rationality of communicative action. This recognition of alternative types of rationality enriches the work of these social theorists in ways which are typically lost on economists. For example, it creates the possibility of tensions developing between the different types of reason and it offers a vantage point from which to assess both instrumental reasoning and 'modernity'.

It will be obvious though that this assignment of utility numbers is arbitrary in the sense that any X and Y will do provided $X > Y$. For this reason these utility numbers are known as *ordinal utility* as they convey nothing more than information on the ordering of preferences.

Two consequences of this arbitrariness in the ordinal utility numbers are worth noting. Firstly the numbers convey nothing about strength of preference. It is as if a friend were to tell you that she prefers Verdi to Mozart. Her preference may be marginal or it could be that she adores Verdi and loathes Mozart. Based on ordinal utility information you will never know. Secondly there is no way that one person's ordinal utility from Verdi can be compared with another's from Mozart. Since the ordinal utility number is meaningful only in relation to the *same* person's satisfaction from something else, it is meaningless across persons. This is why the talk of utility maximisation does not automatically connect neoclassical economics

and game theory to traditional utilitarianism (see Box 1.2 on the philosophical origins of instrumental rationality).

Ordinal utilities are sufficient in many of the simpler decision problems and games. However, there are many other cases where they are not enough. Imagine for instance that you are about to leave the house and must decide on whether to drive to your destination or to walk. You would clearly like to walk but there is a chance of rain which would make walking awfully unpleasant. Let us say that the predicted chance of rain by the weather bureau is 50–50. What does one do? The answer must depend on the strength of preference for walking in the dry over driving in the dry, driving in the wet and walking in the wet. If, for instance, you relish the idea of walking in the dry a great deal more than you fear getting drenched, then you may very well risk it and leave the car in the garage. Thus, we need information on strength of preference.

Cardinal utilities provide such information. If 'walking in the dry', 'driving in the wet', 'driving in the dry' and 'walking in the wet' correspond to 10, 6, 1 and 0 cardinal utils respectively, then not only do we have information regarding ordering, but also of how much one outcome is preferred over the next. Walking in the dry is ten times better for you than driving in the dry. Such cardinal utilities allow the calculus of desire to convert the decision problem from one of utility maximisation to one of utility maximisation *on average*; that is, to the maximisation of *expected utility*. It works as follows (see Box 1.3 on how expected utility maximisation is an extension of the idea of consistent choice to uncertain decision settings).

In the previous example, we took for granted that the probability of rain is $\frac{1}{2}$. If you walk there is, therefore, a 50% chance that you will receive 10 cardinal utils and a 50% chance that you will receive 0 utils. On average your tally will be 5 utils. If, by contrast, you drive, there is a 50% chance of getting 6 utils (if it rains) and a 50% chance of ending up with only 1 cardinal util. On average driving will give you 3.5 utils. If you act as if to maximise average utility, your decision is clear: you will walk. So far we conclude that in cases where the outcome is uncertain cardinal utilities are necessary and expected utility maximisation provides the metaphor for what drives action. As a corollary, note for future reference that whenever we encounter expected utility, cardinal (and not ordinal) utilities are implied. The reason is that it would be nonsense to multiply probabilities with ordinal utility measures whose actual magnitude is inconsequential since they do not reveal strength of preference. Finally notice that, although cardinal utility takes us closer to 19th century utilitarianism, we are still a long way off because one person's cardinal utility numbers are still incomparable with another's. Thus, when we say that your cardinal utility from walking in the dry is 10, this is meaningful only in relation to the 6 utils you receive from driving in the wet. It cannot be compared with a similar

Box 1.3

CONSISTENT CHOICE UNDER RISK AND EXPECTED UTILITY MAXIMISATION

Suppose the actions which a person must choose between have uncertain outcomes in the following sense. Each action has various possible outcomes associated with it, each with some probability. For example, the purchase of a lottery ticket for \$1 where there is a probability of $\frac{1}{100}$ of winning \$50 is an action with an uncertain outcome. One could either lose \$1 or gain a net \$49 when buying the ticket and the respective probabilities of each outcome are $\frac{99}{100}$ and $\frac{1}{100}$. Notationally we call this action a *prospect* and we represent it as a pairing of the possible outcomes with their respective probabilities: $(-\$1, \$49; \frac{99}{100}, \frac{1}{100})$. Then the question is: how do people choose between (risky) prospects?

As we saw in Box 1.1, the theory of instrumentally rational choices specifies certain conditions (or axioms) which the preferences of an individual must satisfy if they are to be consistent. The following axioms need to be added to the list in Box 1.1 in order to make preferences over prospects consistent also.

(1), (2) and (3) remain as in Box 1.1.

(4) *Continuity* also remains as in Box 1.1 but with a minor alteration to extend its relevance to preferences over prospects. Consider three prospects y_i, y_j and y_k and imagine that the individual prefers the first to the second and the second to the third. Then there exists *some* probability p such that if we were to let the individual have prospect y_i with probability p and prospect y_k with probability $1 - p$, then our individual would be equally happy with this situation as he or she would be with prospect y_j. (Notice the similarity with the second interpretation of the continuity axiom in Box 1.1.)

(5) *Preference increasing with probability*: If $y_i > y_j$ and $y_m = (y_i, y_j; p_1, 1 - p_1)$, $y_n = (y_i, y_j; p_2, 1 - p_2)$, then $y_m > y_n$ only if $p_1 > p_2$.

(6) *Independence*: For three prospects y_i, y_j, y_k, if $y_i > y_j$, then there exists a probability λ such that a $(\lambda, 1 - \lambda)$ probability mix of y_i and y_j must be at least as good as a $(\lambda, 1 - \lambda)$ probability mix of y_i and y_j. In our notation for prospects, $(y_i, y_j; \lambda, 1 - \lambda) \geq (y_i, y_k; \lambda, 1 - \lambda)$.

The theory of instrumentally rational choice shows that if an individual's preferences satisfy conditions (1) to (6) then an individual who acts on his or her preference ordering acts *as if* in order to maximise his or her *expected utility function.*

number relating somebody else's cardinal utility from driving in the wet, walking in the dry and so on.

Cardinal utilities and the assumption of expected utility maximisation to game theory are important because uncertainty is ubiquitous in games.

Consider the following variant of an earlier example. You must choose between walking to work or driving. Only this time your concern is not the weather but a friend of yours who also faces the same decision in the morning. Assume your friend is not on the phone (and that you have made no prior arrangements) and you look forward to meeting up with him or her while strolling to work (and if both of you choose to walk, your paths are bound to converge early on in the walk). In particular your first preference is that you walk together. Last in your preference ordering is that you walk only to find out that your friend has driven to work. Of equal second best ranking is that you drive when your friend walks and when your friend drives. We will capture these preferences in matrix form – see Figure 1.1.

If the numbers in the matrix were ordinal utilities, it would be impossible to know what you will do. If you expect your friend to drive then you will also drive as this would give you 1 util as opposed to 0 utils from walking alone. If on the other hand you expect your friend to walk then you will also walk (this would give you 2 utils as opposed to only 1 from driving). Thus your decision will depend on what you expect your friend to do and we need some way of incorporating these expectations (that is, the uncertainty surrounding your friend's behaviour) into your decision making process.

Suppose that, from past experience, you believe that there is $\frac{2}{3}$ chance that your friend will walk. This information is useless unless we know how much you prefer the accompanied walk *over* the solitary drive; that is, unless your utilities are of the cardinal variety. So, imagine that the utils in the matrix of Figure 1.1 are cardinal and you decide to choose an action on the basis of expected utility maximisation. You know that if you drive, you will certainly receive 1 util, regardless of your friend's choice (notice that the first row is full of ones). But if you walk, there is a $\frac{2}{3}$ chance that you will meet up with your friend (yielding 2 utils for you) and a $\frac{1}{3}$ chance of walking alone (0 utils). On average, walking will give you $\frac{4}{3}$ utils ($\frac{2}{3}$ times 2 plus $\frac{1}{3}$ times 0). More generally, if your belief about the probability of your friend walking is p (p having some value between 0 and 1, e.g. $\frac{2}{3}$) then your expected utility from walking is $2p$ and that from driving is 1. Hence an expected utility maximiser will always walk as long as p exceeds $\frac{1}{2}$.

Game theory follows precisely such a strategy. It assumes that it is 'as if'

		Friend	
		Drive	Walk
You	Drive	1	1
	Walk	0	2

Figure 1.1

you had a cardinal utility function and you act so as to maximise expected utility. There are a number of reasons why many theorists are unhappy with this assumption.

The critics of expected utility theory (instrumental rationality)

(a) Internal critique and the empirical evidence

The first type of worry is found within mainstream economics (and psychology) and stems from empirical challenges to some of the assumptions about choice (the axioms in Box 1.3) on which the theory rests. For instance, there is a growing literature that has tested the predictions of expected utility theory in experiments and which is providing a long list of failures. Some care is required with these results because when people play games the uncertainty attached to decision making is bound up with anticipating what others will do and as we shall see in a moment this introduces a number of complications which in turn can make it difficult to interpret the experimental results. So perhaps the most telling tests are not actually those conducted on people playing games. Uncertainty in other settings is simpler when it takes the form of a lottery which is well understood and apparently there are still major violations of expected utility theory. Box 1.4 gives a flavour of these experimental results.

Of course, any piece of empirical evidence requires careful interpretation and even if these adverse results were taken at their face value then it would still be possible to claim that expected utility theory was a prescriptive theory with respect to rational action. Thus it is not undermined by evidence which suggests that we fail in practice to live up to this ideal. Of course, in so far as this defence is adopted by game theorists when they use the expected utility model, then it would also turn game theory into a prescriptive rather than explanatory theory. This in turn would greatly undermine the attraction of game theory since the arresting claim of the theory is precisely that it can be used to explain social interactions.

In addition, there are more general empirical worries over whether all human projects can be represented instrumentally as action on a preference ordering (see Sen, 1977). For example, there are worries that something like 'being spontaneous', which some people value highly, cannot be fitted into the means–ends model of instrumentally rational action (see Elster, 1983). The point is: how can you decide to 'be spontaneous' without undermining the objective of spontaneity? Likewise, can all motives be reduced to a utility representation? Is honour no different to human thirst and hunger (see Hollis, 1987, 1991)? Such questions quickly become philosophical and so we turn explicitly in this direction.

Box 1.4

THE ALLAIS PARADOX

Kahneman and Tversky (1979) offer the following reworking of the famous study in Allais (1953) (see also Sugden (1991) for an up to date survey of the literature).

You are asked to choose between two lotteries, lottery 1 and lottery 2.

Lottery 1	\$2500 with probability 33%
	\$2400 with probability 66%
	0 with probability 1%
Lottery 2	\$2400 with certainty

(Notice that lottery 2 is a lottery only in name since it offers a certain pay-off.)

Which do you choose? Once you have made a choice consider two other lotteries:

Lottery 3	\$2500 with probability 33%
	0 with probability 67%
Lottery 4	\$2400 with probability 34%
	0 with probability 66%

Which do you choose now? Many people choose lotteries 2 and 3. It seems that in choosing between lotteries 1 and 2 they are not prepared to take the small risk of receiving nothing in order to have a small chance of getting an extra \$100. They prefer the safety of the second lottery instead. However, when it comes to a choice between lotteries 3 and 4, lottery 3 seems only slightly riskier than lottery 4 and people are more willing to take that extra risk in order to boost their pay-offs.

However, expected utility theory is categorical here. If you have chosen lottery 1 you must also choose lottery 3. And if you have chosen lottery 2, you must choose lottery 4. To see why expected utility theory says this, let us rewrite the above lotteries as follows:

Lottery 1	\$2400 with probability 66%
	0 with probability 1%
	\$2500 with probability 33%
Lottery 2	\$2400 with probability 66%
	\$2400 with probability 34%
Lottery 3	0 with probability 66%
	0 with probability 1%
	\$2500 with probability 33%
Lottery 4	0 with probability 66%
	\$2400 with probability 34%

Notice that lotteries 1 and 2 contain a common 'element' in the first line: \$2400 with probability 66%. Expected utility theory insists that if you have a preference between lotteries 1 and 2 then this must be so because of the other 'elements' in these lotteries. And if you were to substitute that common element with some other common element, then your original preference should be preserved. For example, suppose that you amended the first line of lotteries 1 and 2 so that instead of '\$2400 with

probability 66%' it read '$200 with probability 66%'. If you preferred lottery 2 to lottery 1 (say) before the amendment, expected utility theory argues that you must preserve this preference after the amendment since only the common element has been changed. This is the so-called independence axiom of expected utility theory (see Box 1.3). Now consider lotteries 3 and 4. The way we have rewritten them above, they are identical to lotteries 1 and 2 excepting the common element which has been changed from '$2400 with probability 66%' to '0 with probability 66%'. Thus, according to expected utility theory, if you prefer lottery 2 to lottery 1, you must also prefer lottery 4 to lottery 3. And yet, the majority of people participating in such experiments seem to violate the independence axiom and choose lotteries 2 and 3. The fact that expected utility theory receives little empirical support is potentially worrying for game theory because it relies so heavily on it.

(b) Philosophical and psychological discontents

This is not the place for a philosophy lesson (even if we were competent to give it!). But there are some relatively simple observations concerning rationality that can be made on the basis of common experiences and reflections which in turn connect with wider philosophical debate. We make some of those points and suggest those connections here. They are not therefore designed as decisive philosophical points against the instrumental hypothesis. Rather their purpose is to remind us that there are puzzles with respect to instrumental rationality which are openings to vibrant philosophical debate. Why bother to make such reminders? Partially, as we have indicated, because economists seem almost unaware that their foundations are philosophically contentious and partially because it seems to us and others that the only way to render some aspects of game theory coherent is actually by building in a richer notion of rationality than can be provided by instrumental rationality alone. For this reason, it is helpful to be aware of some alternative notions of rational agency.

Consider first a familiar scene where a parent is trying to 'reason' with a child to behave in some different manner. The child has perhaps just hit another child and taken one of his or her toys. It is interesting to reflect on what parents usually mean here when they say 'I'm going to reason with the blighter.'

'Reason' here is usually employed to distinguish the activity from something like a clip around the ear and its intent is to persuade the 'blighter' to behave differently in future. The question worth reflecting upon is: what is it about the capacity to reason that the parent hopes to be able to invoke in the child to persuade him or her to behave differently?

The contrast with the clip around the ear is quite instructive because this action would be readily intelligible if we thought that the child was only

instrumentally rational. If a clip around the ear is what you get when you do such things then the instrumentally rational agent will factor that into the evaluation of the action, and this should result in it being taken less often. Of course, 'reasoning' could be operating in the same way in so far as listening to parents waffling on in the name of reason is something to be avoided like a clip around the ear. Equally it could be working with the grain of instrumental rationality if the adult's intervention was an attempt to rectify some kind of faulty 'means–ends' calculation which lay behind the child's action. However, there is a line of argument sometimes used by adults which asks the child to consider how they would like it if the same thing was to happen to them; and it is not clear how a parent could think that such an argument has a purchase on the conduct of the instrumentally rational child. Why should an instrumentally rational child's reflection on their dislike of being hit discourage them from hitting others unless hitting others makes it more likely that someone will hit them in turn? Instead, it seems that the parents when they appeal to reason and use such arguments are imagining that reason works in some other way. Most plausibly, they probably hope that reason supplies some kind of internal constraint on the actions and objectives which one deems permissible, where the constraint is akin to the biblical order that you should do unto others as you would have done to yourself.

Of course, reason may not be the right word to use here. Although Weber (1947) refers to *wertrational* to describe this sort of rationality, it has to be something which the parent believes affects individual actions in a way not obviously captured by the instrumental model. Furthermore there is a philosophical tradition which has associated reason with supplying just such additional constraints. It is the tradition initiated by Immanuel Kant which famously holds that reason is ill equipped to do the Humean thing of making us happy by serving our passions.

> Now in a being which has reason and will, if the proper object of nature were its conservation, its welfare, in a word, its happiness, then nature would have hit upon a very bad arrangement in selecting reason to carry out this purpose. . . . For reason is not competent to guide the will with certainty in regard to its objects and the satisfaction of all our wants (which to some extent it even multiplies) . . . . its true destination must be to produce a will, not merely good as a means to something else, but good in itself, for which reason was absolutely necessary.
>
> (Kant, 1788, pp. 11–12).

Thus reason is instead supposed to guide the ends we pursue. In other words, to return to the case of the child taking the toy, reason might help us to see that we should not want to take another child's toy. How might it specifically do this? By supplying a negative constraint is Kant's answer. For Kant it is never going to be clear what reason specifically instructs, but

Box 1.5

KANT'S CATEGORICAL IMPERATIVE

Kant summarises the categorical imperative thus: 'Act only on that maxim whereby thou canst at the time will that it should become a universal law.'

As an example of how the categorical imperative might be applied and how it differs from instrumental reasoning, consider a person wondering whether to pay his or her taxes. Non-payment could be instrumentally rational in so far as the person is concerned only with his or her welfare and the chances of being fined for non-payment are slight. However, such an action would not pass the test of the categorical imperative. If the person were (hypothetically) to consider not paying his or her taxes while at the same time accepting the premise that others are similarly rational, then he or she would be committed to the predictable result that society would break down and life would become nasty, brutish and probably short as government support for law and order, health care, road building, etc., collapsed without the necessary funding from taxes. Thus for Kant the rational person should not allow reason to be a slave to the passions (which might lead to non-payment); instead our rationality, and the fact that we share it, should lead us to the categorical imperative and the payment of taxes.

since we are all equipped with reason, we can see that reason could only ever tell us to do something which it would be possible for everyone to do. This is the test provided by the categorical imperative (see Box 1.5) and reason guides us by telling us to exclude those objectives which do not pass the test. Thus we should not want to do something which we could not wish would be done by everyone; and this might plausibly explain why reason could be invoked to persuade the child not to steal another child's toy.

Even when we accept the Kantian argument, it is plain that reason's guidance is liable to depend on characteristics of time and place. For example, consider the objective of 'owning another person'. This obviously does not pass the test of the categorical imperative since all persons could not all own a person. Does this mean then we should reject slave-holding? At first glance, the answer seems to be obvious: of course, it does! But notice it will only do this if slaves are considered people. Of course we consider slaves people and this is in part why we abhor slavery, but ancient Greece did not consider slaves as people and so ancient Greeks would not have been disturbed in their practice of slavery by an application of the categorical imperative.

This type of dependence of what is rational on time and place is a feature

of many philosophical traditions. For instance, Hegel has reason evolving historically and Marx tied reason to the expediency of particular modes of production. It is also a feature of the later Wittgenstein who proposes a rather different assault on the conventional model of instrumental reason. As we shall say more about this in section 1.2.3, it suffices for now to note that Wittgenstein suggests that if you want to know why people act in the way that they do, then ultimately you are often forced in a somewhat circular fashion to say that such actions are part of the practices of the society in which those persons find themselves. In other words, it is the fact that people behave in a particular way in society which supplies the reason for the individual person to act: or, if you like, actions often supply their own reasons. This is shorthand description rather than explanation of Wittgenstein's argument, but it serves to make the connection to an influential body of psychological theory which makes a rather similar point.

Festinger's (1957) cognitive dissonance theory proposes a model where reason works to 'rationalise' action rather than guide it. The point is that we often seem to have no reason for acting the way that we do. For instance, we may recognise one reason for acting in a particular way, but we can equally recognise the pull of a reason for acting in a contrary fashion. Alternatively, we may simply see no reason for acting one way rather than another. In such circumstances, Festinger suggests that we experience psychological distress. It comes from the dissonance between our self-image as individuals who are authors of our own action and our manifest lack of reason for acting. It is like a crisis of self-respect and we seek to remove it by creating reasons. In short we often rationalise our actions *ex post* rather than reason *ex ante* to take them as the instrumental model suggests.

This type of dissonance has probably been experienced by all of us at one time or another and there is much evidence that we both change our preferences and change our beliefs about how actions contribute to preference satisfaction so as to rationalise the actions we have taken (see Aronson, 1988). Some of the classic examples of this are where smokers have systematically biased views of the dangers of smoking or workers in risky occupations similarly underestimate the risks of their jobs. Indeed in a modified form, we will all be familiar with a problem of consumer choice when it seems impossible to decide between different brands. You consult consumer reports, specialist magazines and the like and it does not help because all this extra information only reveals how uncertain you are about what you want. The problem is you do not know whether safety features of a car, for instance, matter to you more than looks or speed or cost. And when you choose one rather than another you are in part choosing to make, say, 'safety' one of your motives. Research has shown that people seek out and read advertisements for the brand of car they have just bought. Indeed, to return us to economics, it is precisely this insight which has been at the heart of one of the Austrian and other critiques of the central planning system when it is argued that planning can never substitute for the market

because it presupposes information regarding preferences which is in part created in markets when consumers choose.

(c) The source of beliefs

You will recall in the example contained in Figure 1.1 that in deciding what to do you had to form an expectation regarding the chances that your friend would walk to work. Likewise in an earlier example your decision over whether to walk or drive depended on an expectation: the probability of rain. The question we wish to explore here is where these beliefs come from; and for this purpose, the contrast between the two decision problems is instructive.

At first sight it seems plausible to think of the two problems as similar. In both instances we can use previous experience to generate expectations. Previous experience with the weather provides probabilistic beliefs in the one case, and experience with other people provides it in the other. However, we wish to sound a caution. There is an important difference because the weather is not concerned at all about what you think of it whereas other people often are. This is important because while your beliefs about the weather do not affect the weather, your beliefs about others can affect their behaviour when those beliefs lead them to expect that you will act in particular ways. For instance, if your friend is similarly motivated and thinks that you will walk then he or she will want to walk; and you will walk if you think he or she will walk. So what he or she thinks you think will in fact influence what he or she does!

To give an illustration of how this can complicate matters from a slightly different angle, consider what makes a good meteorological model. A good model will be proved to be good in practice: if it predicts the weather well it will be proclaimed a success, otherwise it will be dumped. On the other hand in the social world, even a great model of traffic congestion, for instance, may be contradicted by reality simply because it has a good reputation. If it predicts a terrible jam on a particular stretch of road and this prediction is broadcast on radio and television, drivers are likely to avoid that spot and thus render the prediction false. This suggests that proving or disproving beliefs about the social world is liable to be trickier than those about the natural world and this in turn could make it unclear how to acquire beliefs rationally.

Actually most game theorists seem to agree on one aspect of the problem of belief formation in the social world: how to update beliefs in the presence of new information. They assume agents will use *Bayes's rule.* This is explained in Box 1.6. We note there some difficulties with transplanting a technique from the natural sciences to the social world which are related to the observation we have just made. We focus here on a slightly

Box 1.6

BAYES'S RULE

Two examples:

(a) How seriously do you take a medical diagnosis?

Imagine you have just taken a test for a dreaded disease X and your doctor has just gloomily informed you that you have tested positive. Suppose that it is known beyond doubt that 0.1% of the population are affected by X and that 100,000 tests have been administered so far. Also it is known that the test is correct 99% of the time (that is, the test is positive 99% of the time for someone who has X and negative 99% of the time for someone who does not have it). How depressed should you be? What are the chances that you really have X?

At first sight, it seems that there is a 99% chance that you have X since you tested positive and the test is 99% accurate. Bayes's rule gives you (a scientific) cause to rejoice; at least to postpone despair. Let us reconsider the data. Of the 100,000 people tested, 0.1% will have X; that is, 100 people on average. Of those 100 X-affected people who have taken the test, 99 will prove positive (recall the test is 99% accurate). However, of the 99,900 healthy people 1% will also test positive owing to the 1% error margin of the test, i.e. 999 healthy people will have tested positive. Thus, of a total of 1098 positive tests (999 healthy plus the 99 affected people) only 99 have X. Thus the probability that you have X given (or conditional on the fact) that you have tested positive is 99/1098 which is only about 9%!

The above captures the logic of Bayes's rule for amending initial probabilistic beliefs in the light of new evidence. The initial beliefs were that (a) the probability that you have X is 0.1%; (b) the probability that you have X if the test proves positive $\Pr(X|\text{test is positive}) = 99\%$ – notice that | stands for 'given that'. The new bit of information is that you tested positive. How do you amend the probability that you have X in the light of this information?

In general, Thomas Bayes suggested the following rule which codifies our earlier calculations: the probability that event A has occurred given that event B has just been observed is written as Pr(A|B) (this is known as a conditional probability) and equals

$$\Pr(A|B) = [\Pr(B|A)\Pr(A)]/[\Pr(B|A)\Pr(A) + \Pr(B|\text{not A})\Pr(\text{not A})]$$

(where 'not A' means that event A did not occur).

To see how it applies in our example, think of event B as the new information, namely B: 'You tested positive for disease X.' Then the question is, what is Pr(A|B)? That is, what is the probability that you have X given that the test was positive? Let us put together the right hand side of Bayes's rule. Pr(B|A) is the probability that you will test positive given that you have X. It equals 99% (from (b) above). Pr(A) is the probability that you have X as assessed before the test (i.e. the new information): it equals 0.1% (from (a) above). Thus the numerator equals 99% times 0.1%, i.e. 9.9%. The denominator equals 9.9% plus Pr(B|not A) times

Pr(not A). The probability of 'not A', i.e. that you do not have X, is 99.9% while the probability of testing positive if you do not have it (i.e. Pr(B|not A)) equals 1%. Therefore the whole denominator equals 109.8%. It turns out that the probability that you have X given that you tested positive equals 9.9/109.8, which is exactly what we found earlier; a touch above 9%.

(b) Should you prosecute?

Let us suppose that you are the district attorney who must decide whether to prosecute the person who the police say has committed the crime. You adopt a simple rule of thumb: if it seems that there is more than a 50% chance, based on the evidence presented by the police, that the person did commit the crime then you prosecute. Here are the details of the case. It is known almost beyond doubt that the crime was committed by one person in a group of six people. So before any police evidence is presented, you believe that there is something fractionally less than a one-in-six chance that the person identified by the police actually did commit the crime (to allow for just some doubt that the crime could have been committed by someone outside the group), say 0.15. The police offer one piece of evidence to support their claim that their candidate committed the crime: this person's confession. It is also 'well known' that what people say to the police is only 80% reliable. Should you prosecute?

Bayes's rule tells us that the probability that the person is G (guilty) conditional on the information C (the evidence of a confession) is given by

$$\Pr(G|C) = \Pr(G \text{ and } C)/\Pr(C) = \Pr(C|G)\Pr(G)/[\Pr(C|G)\Pr(G)+\Pr(C|NG)\Pr(NG)]$$

where Pr(C|G) is the probability of confessing when guilty (which is the 80% reliability rate), Pr(C|NG) is the probability of the person confessing when not guilty (that is, the unreliability rate of 20%) and the Pr(G) and Pr(NG) are the prior probability assessments of guilty and not guilty (respectively 15 and 85%).

When the substitutions are performed, Bayes's rule yields the inference that the probability of guilt is revised to 0.41, which is less than the 50% and the DA tells the police to get more evidence if they want a prosecution! The result is perhaps somewhat surprising but you can see how it is derived by imagining a population of 100 people with 15 guilty people in it. You ask each to confess and given the 80% reliability rate, 12 of the guilty will and 3 will not, and 68 of the 85 innocents will not confess (= 80% reliable) and 17 innocents will confess. Thus there are 29 confessions altogether, but only 12 (that is, a proportion equal to 0.41) come from people who are genuinely guilty.

There are a couple of points to notice about Bayes's rule. The first is that it is a rule of statistical inference and it will only apply to stationary probability distributions. So in this instance you cannot apply it if the chance of the guilty person coming from the group of six suspects, rather than some larger group, kept changing. Secondly, the rule can only be applied when the new information, the event, has a prior probability assessment of zero (this can be seen from the expression above because it is not defined when the probability of a confession is zero). In other

words, if something happens which you had never anticipated, but which is actually relevant, then you cannot use Bayes's rule to take it into account.

different problem. Bayes provides a rule for updating, but where do the original (prior) expectations come from? Or to put the question in a different way: in the absence of evidence, how do agents form probability assessments governing events like the behaviour of others?

There are two approaches in the economics literature. One responds by suggesting that people do not just passively have expectations. They do not just wait for information to fall from trees. Instead they make a conscious decision over how much information to look for. Of course, one must have started from somewhere, but this is less important than the fact that the acquisition of information will have transformed these original 'prejudices'. The crucial question, on this account, then becomes: what determines the amount of effort agents put into looking for information? This is deceptively easy to answer in a manner consistent with instrumental rationality. The instrumentally rational agent will keep on acquiring information to the point where the last bit of search effort costs her or him in utility terms the same amount as the amount of utility he or she expects to get from the information gained by this last bit of effort. The reason is simple. As long as a little bit more effort is likely to give the agent more utility than it costs, then it will be adding to the sum of utilities which the agent is seeking to maximise.

This looks promising and entirely consistent with the definition of instrumentally rational behaviour. But it begs the question of how the agent knows how to evaluate the potential utility gains from a bit more information *prior to gaining that information*. Perhaps he or she has formulated expectations of the value of a little bit more information and can act on that. But then the problem has been elevated to a higher level rather than solved. How did he or she acquire that expectation about the value of information? 'By acquiring information about the value of information up to the point where the marginal benefits of this (second-order) information were equal to the costs', is the obvious answer. But the moment it is offered, we have the beginnings of an infinite regress as we ask the same question of how the agent knows the value of this second-order information. To prevent this infinite regress, we must be guided by something *in addition* to instrumental calculation. But this means that the paradigm of instrumentally rational choices is incomplete. The only alternative would be to assume that the individual *knows* the benefits that he or she can expect on average from a little more search (i.e. the expected marginal benefits)

Box 1.7

THE ELLSBERG PARADOX, UNCERTAINTY, PROBABILITY ASSESSMENTS, AND CONFIDENCE

Suppose you are shown an urn with 90 balls in it and you are told that 30 are red and that the remaining 60 balls are either black or yellow. One ball is going to be selected at random and you are given the following choice. Option I will give you \$100 if a red ball is drawn and nothing if either a black or a yellow ball is drawn; option II will give you \$100 if a black ball and nothing if a red or a yellow ball is drawn. Here is a summary of the options:

	Red	Black	Yellow
Option I	\$100	0	0
Option II	0	\$100	0

Make a note of your choice and then consider another two options based on the same random draw from this urn:

	Red	Black	Yellow
Option III	\$100	0	\$100
Option IV	0	\$100	\$100

Which of these would you choose?

Ellsberg (1961) reports that, when presented with this pair of choices, most people select options I and IV. Adopting the approach of expected utility theory (see Box 1.3), this reveals a clear inconsistency in probability assessments. On this interpretation, when a person chooses option I over option II, he or she is revealing a higher subjective probability assessment of a 'red' than a 'black'. However, when the same person prefers option IV to III, he or she reveals that his or her subjective probability assessment of 'black' or 'yellow' is higher than a 'red' or 'yellow', and this implies that a 'black' has a higher probability assessment than a 'red'!

Perhaps the simplest explanation of this pair of choices turns on the confidence which a person attaches to probability assessments. For example, when choosing between options I and II, if the person opts for I he or she knows the *exact* probability of winning \$100: it is $\frac{1}{3}$. By contrast, were he or she to choose option II, the probability of winning would have been unknown. Now look again at options III and IV. By choosing option IV one knows the exact probability of winning: $\frac{2}{3}$. On the other hand, the probability of winning \$100 when choosing option III is ambiguous. In other words, the choices of I and IV can be explained by an aversion to ambiguity and a preference for prospects which come with precise, objective, information about the probability of winning or losing. This kind of preference violates expected utility theory but can by no means be dismissed as irrational.

In so far as this explanation seems plausible, then the Ellsberg paradox points to a deeper problem with respect to the conventional expected utility maximising model because it suggests that probability assessments inadequately capture the way that uncertainty enters into decision making. In fact, it is precisely this observation which lies at the

famous distinction between risk (i.e. as in lotteries where you do not know what will happen but you know all the possible outcomes and the probability for each) and uncertainty (i.e. cases in which you are in the dark) in economics (see Knight, 1921, and Keynes, 1936).

because he or she knows the full information set. But then there is no problem of how much information to acquire because the person knows everything!

The second response by neoclassical economists to the question 'Where do beliefs come from?' is to treat them as purely subjective assessments (following Savage, 1954). This has the virtue of avoiding the problem of rational information acquisition by turning subjective assessments into data which is given from outside the model along with the agents' preferences. They are what they are; and they are only revealed *ex post* by the choices people make (see Box 1.7 for some experimental evidence which casts doubt on the consistency of such subjective assessments and more generally on the probabilistic representations of uncertainty). The distinct disadvantage of this is that it might license almost any kind of action and so could render the instrumental model of action close to vacuous. To see the point, if expectations are purely subjective, perhaps any action could result in the analysis of games, since any subjective assessment is as good as another. Actually game theory has increasingly followed Savage (1954), by regarding the probability assessments as purely subjective, but it has hoped to prevent this turning itself into a vacuous statement (to the effect that 'anything goes') by supplementing the assumption of *instrumental rationality* with the assumption of *common knowledge of rationality* (CKR). The purpose of the latter is to place some constraints on people's subjective expectations regarding the actions of others.

1.2.2 Common knowledge of rationality (CKR) and consistent alignment of beliefs (CAB)

We have seen how expectations regarding what others will do are likely to influence what it is (instrumentally) rational for you to do. Thus fixing the beliefs that rational agents hold about each other is likely to provide the key to the analysis of rational action in games. The contribution of CKR in this respect comes in the following way.

If you want to form an expectation about what somebody does, what could be more natural than to model what determines their behaviour and then use the model to predict what they will do in the circumstances that interest you? You could assume the person is an idiot or a robot or whatever, but most of the time you will be playing games with people

who are instrumentally rational like yourself and so it will make sense to model your opponent as instrumentally rational. This is the idea that is built into the analysis of games to cover how players form expectations. We assume that there is common knowledge of rationality held by the players. It is at once both a simple and complex approach to the problem of expectation formation. The complication arises because with common knowledge of rationality I know that you are instrumentally rational and since you are rational and know that I am rational you will also know that I know that you are rational and since I know that you are rational and that you know that I am rational I will also know that you know that I know that you are rational and so on . . . . This is what common knowledge of rationality means. Formally it is an infinite chain given by

(a) that each person is instrumentally rational
(b) that each person knows (a)
(c) that each person knows (b)
(d) that each person knows (c) And so on *ad infinitum.*

This is what makes the term *common knowledge* one of the most demanding in game theory. It is difficult to pin down because common knowledge of X (whatever X may be) cannot be converted into a finite phrase beginning with 'I know . . . '. The best one can do is to say that if Jack and Jill have common knowledge of X then 'Jack knows that Jill knows that Jack knows . . . that Jill knows that Jack knows . . . X' – an infinite sentence. The idea reminds one of what happens when a camera is pointing to a television screen that conveys the image recorded by the very same camera: an infinite self-reflection. Put in this way, what looked a promising assumption suddenly actually seems capable of leading you anywhere.

To see how an assumption that we are similarly motivated might not be so helpful in more detail, take an extreme case where you have a desire to be fashionable (or even unfashionable). So long as you treat other people as things, parameters like the weather, you can plausibly collect information on how they behave and update your beliefs using the rules of statistical inference, like Bayes's rule (or plain observation). But the moment you have to take account of other people as like-minded agents concerned with being fashionable, which seems to be the strategy of CKR, the difficulties multiply. You need to take account of what others will wear and, with a group of like-minded fashion hounds, what each of them wears will depend on what they expect others (including you) to wear, and what each expects others to wear depends on what each expects each other will expect others to wear, and so on . . . . The problem of expectation formation spins hopelessly out of control.

Nevertheless game theorists typically assume CKR and many of them, and certainly most people who apply game theory in economics and other disciplines, take it further: in order to come up with precise predictions on

rational behaviour they assume not only CKR, but also they make (what we call) the assumption of consistently aligned beliefs (CAB). In other words they assume that everybody's beliefs are consistent with everybody else's. CAB gives great analytical power to the theorist, as we will see in later chapters. Nevertheless, the jump from CKR to CAB is controversial, even among game theorists (see Kreps, 1990, Bernheim, 1984, and Pearce, 1984).

Put informally, the notion of *consistent alignment of beliefs* (CAB) means that no instrumentally rational person can expect another similarly rational person who has the same information to develop different thought processes. Or, alternatively, that no rational person expects to be surprised by another rational person. The point is that if the other person's thought is genuinely moving along rational lines, then since you know the person is rational and you are also rational then your thoughts about what your rational opponent might be doing will take you on the same lines as his or her own thoughts. The same thing applies to others provided they respect *your* thoughts. So your beliefs about what your opponents will do are consistently aligned in the sense that if you actually knew their plans, you would not want to change your beliefs; and if they knew your plans they would not want to change the beliefs they hold about you and which support their own planned actions.

Note that this does not mean that everything can be deterministically predicted. For example, both you and others may be expecting good weather with probability $\frac{3}{4}$. In that sense your beliefs are consistently aligned. Yet it rains. You may be disappointed but you are not surprised, since there was always a $\frac{1}{4}$ chance of rain. What partially underpins the jump from CKR to CAB is the so-called Harsanyi doctrine. This follows from John Harsanyi's famous declaration that when two rational individuals have the same information, they **must** draw the same inferences and come, independently, to the same conclusion. So, to return to the fashion game, this means that when two rational fashion hounds confront the same information regarding the fashion game played among fashion hounds, they should come to the same conclusion about what rational players will wear.

As stated this would still seem to leave it open for different agents to entertain different expectations (and so genuinely surprise one another) since it only requires that rational agents draw the same inferences from the same information but they need not enjoy the same information. To make the transition from CKR to CAB complete, Robert Aumann takes the argument a stage further by suggesting that rational players will come to hold the same information so that in the example involving the expectations on whether it will rain or not, rational agents could not 'agree to disagree' about the probability of rain. (See Box 1.8 for the complete argument.) One can almost discern a dialectical argument here; where

Box 1.8

ROBERT AUMANN'S DEFENCE OF THE ASSUMPTION OF A CONSISTENT ALIGNMENT OF BELIEFS

Suppose you believe that the probability of rain tomorrow is $\frac{3}{4}$. And suppose that I believe it to be $\frac{1}{4}$. On this basis, you could agree to pay me \$1 if it does not rain and I could agree to pay you \$1 if it does. Sounds reasonable? Not to game theorists in this tradition. Notice that although the final payoff tomorrow will sum to zero (that is, what I will win/lose and what you will lose/win sum to zero), this is not so with the pay-offs we expect today. Each one of us expects payoffs: \$1 with probability $\frac{3}{4}$ and $-\$1$ with probability $\frac{1}{4}$. On average, each expects to make 50 cents $[\$1 \times (\frac{3}{4}) - \$1 \times (\frac{1}{4}) = \$(\frac{1}{2})]$. Thus our expectations are inconsistent with each other. If we are both rational we can only disagree because we have different evidence or information sets. In offering to make the bet, each one of us reveals to the other some of what was previously 'privately' held information. You reveal that you have evidence which ought to temper my confidence that it will be dry tomorrow and similarly I reveal to you some of my evidence which ought to temper your confidence in rain. Consequently each will want to revise their expectation of rain tomorrow. This exchange of information will continue so long as we disagree and with each exchange the disagreement narrows until finally it disappears. Thus according to Aumann, rational agents cannot agree to disagree.

following Socrates, who thought unique truths can be arrived at through dialogue, we assume that an opposition of incompatible positions will give way to a uniform position acceptable to both sides once time and communication have worked their elixir. Thus, CKR spawns CAB.

Such a defence of CAB is not implausible, but it does turn on the idea of an explicit dialogue in real (i.e. historical) time. Aumann does not specify how and where this dialogue will take place, and without such a process there need be no agreement (Socrates' own ending confirms this). This would seem to create a problem for Aumann's argument at least as far as one-shot games are concerned (that is, interactions which occur between the same players only once and in the absence of communication). You play the game once and then you might discover *ex post* that you must have been holding some divergent expectations. But this will only be helpful if you play the same game again because you cannot go back and play the original game afresh.

Furthermore, there is something distinctly optimistic about the first (Harsanyi) part of the argument. Why should we expect rational agents faced with the same information to draw the same conclusions? After all, we do not seem to expect the same fixtures will be draws when we

complete the football pools; nor do we enjoy the same subjective expectations about the prospects of different horses when some bet on the favourite and others on the outsider. Of course, some of these differences might stem from differences in information, but it is difficult to believe that this accounts for all of them. What is more, on reflection, would you really expect our fashion hounds to select the same clothing when each only knows that the other is a fashion hound playing the fashion game?

These observations are only designed to signal possible trouble ahead and we shall examine this issue in greater detail in Chapters 2 and 3. We conclude the discussion now with a pointer to wider philosophical currents. Many decades before the appearance of game theory, the German philosophers G.F.W. Hegel and Immanuel Kant had already considered the notion of the self-conscious reflection of human reasoning on itself. Their main question was: can our reasoning faculty turn on itself and, if it can, what can it infer? Reason can certainly help persons develop ways of cultivating the land and, therefore, escape the tyranny of hunger. But can it understand how it, itself, works? In game theory we are not exactly concerned with this issue but the question of what follows from common knowledge of rationality has a similar sort of reflexive structure. When reason knowingly encounters itself in a game, does this tell us anything about what reason should expect of itself?

What is revealing about the comparison between game theory and thinkers like Kant and Hegel is that, unlike them, game theory offers something settled in the form of CAB. What is a source of delight, puzzlement and uncertainty for the German philosophers is treated as a problem solved by game theory. For instance, Hegel sees reason reflecting on reason as it reflects on itself as part of the restlessness which drives human history. This means that for him there are no answers to the question of what reason demands of reason in other people outside of human history. Instead history offers a changing set of answers. Likewise Kant supplies a weak answer to the question. Rather than giving substantial advice, reason supplies a negative constraint which any principle of knowledge must satisfy if it is to be shared by a community of rational people: any rational principle of thought must be capable of being followed by all. O'Neill (1989) puts the point in the following way:

> [Kant] denies not only that we have access to transcendent metaphysical truths, such as the claims of rational theology, but also that reason has intrinsic or transcendent vindication, or is given in consciousness. He does not deify reason. The only route by which we can vindicate certain ways of thinking and acting, and claim that those ways have authority, is by considering how we must discipline our thinking if we are to think or act at all. This disciplining leads us not to algorithms of reason, but to certain constraints on all thinking, communication and

interaction among any plurality. In particular we are led to the principle of rejecting thought, act or communication that is guided by principles that others cannot adopt.

(O'Neill p. 27)

To summarise, game theory is avowedly Humean in orientation. Nevertheless a disciple of Hume will protest two aspects of game theory rather strongly. The first we have already mentioned in Box 1.2: by substituting desire and preference for the passions, game theory takes a narrower view of human nature than Hume. The second is that game theorists seem to assume *too much* on behalf of reason. Hume saw reason acting like a pair of scales to weigh the pros and cons of a certain action so as to enable the selection of the one that serves a person's passions best. Game theory demands rather more from reason when starting from CKR it moves to CAB and the inference that rational players will always draw the same conclusions from the same information. Thus when the information comprises a particular game, rational players will draw the same inference regarding how rational players will play the game. Would Hume have sanctioned such a conclusion? It seems doubtful (see Sugden, 1991). After all, even Kant and Hegel, who attach much greater significance than Hume to the part played by reason, were not convinced that reason would ever give either a settled or a unique answer to the question of what reflection of reason on itself would come up with.

1.2.3 Action within the rules of games

There are two further aspects of the way that game theorists model social interaction which strike many social scientists as peculiar. The first is the assumption that individuals know the rules of the game – that is, they know all the possible actions and how the actions combine to yield particular pay-offs for each player. The second, and slightly less visible one, is that a person's motive for choosing a particular action is strictly independent of the rules of the game which structure the opportunities for action.

Consider the first peculiarity: how realistic is the assumption that each player knows all the possible moves which might be made in some game? Surely, in loosely structured interactions (games) players often invent moves. And even when they do not, perhaps it is asking too much to assume that a person knows both how the moves combine to affect their own utility pay-offs and the pay-offs of other players. After all, our motives are not always transparent to ourselves, so how can they be transparent to others?

There are several issues here. Game theory must concede that it is concerned with analysing interactions where the menu of possible actions for each player is known by everyone. It would be unfair of us to expect

game theory to do more. Indeed this may not be so hard to swallow since each person must know that 'such and such' is a possible action before they can *decide* to take it. Of course people often blunder into things and they often discover completely new ways of action, but neither of these types of acts could have been decided upon. Blundering is blundering and game theory is concerned with conscious decision making. Likewise, you can only decide to do something when that something is known to be an option, and genuinely creative acts create something which was not known about before the action. The more worrying complaint appears to be the one regarding knowledge of other people's utility pay-offs (in other words, their preferences).

Fortunately though, game theory is not committed to assuming that agents know the rules of the game in this sense with certainty. It is true that the assumption is frequently made (it distinguishes games where information is complete from those in which it is incomplete) but, according to game theorists, it is not essential. The assumption is only made because it is 'relatively easy' to transform any game of incomplete information into one of complete information. Harsanyi (1967/1968) is again responsible for the argument. Chapter 2 gives a full account of the argument, but in outline it works like this. Suppose there are a number of different 'types' of player in the world where each type of player has different preferences and so will value the outcomes of a game in different ways. In this way we can view your uncertainty about your opponent's utility pay-offs as deriving from your uncertainty about your opponent's 'type'. Now all that is needed is that you hold common prior expectations with your opponent (the Harsanyi/Aumann doctrine) about the likelihood of your opponent turning out to be one type of player or another and the game has become one of complete information.

The information is complete because you know exactly how likely it is that your opponent will be a player of one type or another and your opponent also knows what you believe this likelihood to be. Again it is easy to see how once this assumption has been made, the analysis of play in this game will be essentially the same as the case where there is no uncertainty about your opponent's identity. We have argued before that you will choose the action which yields the highest expected utility. This requires that you work out the probability of your opponent taking various actions because their action affects the pay-offs to you from each of your actions. When you know the identity of your opponent, this means you have to work out the probability of that kind of an opponent taking any particular action. The only difference now is that the probability of your opponent taking any particular action depends not only on the probability that a rational opponent of some type, say A, takes this action but also on the probability of your opponent being type A in the first place.

The difficult thing in all likelihood, as we have argued above, is to know

always what a rational opponent of known preferences will do. But so long as we have sorted this out for each type of player and we know the chances of encountering each type, then the fact that we do not know the identity of the opponent is a complication, but not a serious one. To see the point, suppose we know left-footed people are slower moving to the right than the left and vice versa. Then we know the best thing to do in soccer is to try and dribble past a left-footed opponent on their right and vice versa. If you do not know whether your opponent is left or right footed, then this is, of course, a complication. But you can still decide what to do for the best in the sense of being most likely to get past your opponent. All you have to know are the relative chances of your opponent being left or right footed and you can decide which way to swerve for the best.

Moving on, game theory is not unusual in distinguishing between actions and rules of the game. The distinction reflects the thought that we are often constrained in the actions that we take. For instance, nobody would doubt the everyday experience that common law and the laws of Parliament, the rules of clubs or institutions that we belong to and countless informal rules of conduct provide a structure to what we can and cannot do. Likewise social theory commonly recognises that these so-called 'structures' constrain our actions. However, the way that action is separated from the rules of the game (or 'structures') positions game theory in a very particular way in discussions in social theory regarding the relation between 'action' and 'structure'.

To be specific, game theory accepts the strict separation of action from structure. The structure is provided by the rules of the game and action is analysed under the constraints provided by the structure. This may be a common way of conceiving the relation between the two, but it is not the only one. It is as if structures provide architectural constraints on action. They are like brick walls which you bump into every now and then as you walk about the social landscape. The alternative metaphor comes from language. For example Giddens (1979) suggests that action involves some shared rules just as speaking requires shared language rules. These rules constrain what can be done (or said), but it makes no sense to think of them as separate from action since they are also enabling. Action cannot be taken without background rules, just as sentences cannot be uttered without the rules of language. Equally rules cannot be understood independently of the actions which exemplify them. In other words, there is an organic or holistic view of the relation between action and structure.

The idea behind Giddens' argument can be traced to an important theme in the philosophy of Wittgenstein: the idea that action and structure are mutually constituted in the practices of a society. This returns us to a point which was made earlier with respect to how actions can supply their own reasons. To bring this out, consider a person hitting a home run in baseball

with the bases loaded or scoring a four with a reverse sweep in cricket. Part of the satisfaction of both actions comes, of course, from their potential contribution to winning the game. In this sense, part of the reason for both actions is strictly external to the game. You want to win and the game simply constrains how you go about it.

However, a part of the satisfaction actually comes from what it means in baseball to 'hit a home run with the bases loaded' or what it means in cricket to 'score a four with a reverse sweep'. Neither actions are simply ways of increasing the team's score by four. The one is an achievement which marks a unique conjunction between team effort (in getting the bases loaded) and individual prowess (in hitting the home run); while the other is a particularly audacious and cheeky way of scoring runs. What makes both actions special in this respect are the rules and traditions of the respective games; and here is the rub because the rules begin to help supply the reasons for the action. In other words, the rules of these games both help to constitute and regulate actions. Game theory deals in only one aspect of this, the regulative aspect, and this is well captured by the metaphor of brick walls. Wittgenstein's language games, by contrast, deal with the constitutive aspect of rules and who is to say which best captures the rules of social interaction.

The question is ontological and it connects directly with the earlier discussion of instrumental rationality. Just as instrumental rationality is not the only ontological view of what is the essence of human rationality, there is more than one ontological view regarding the essence of social interaction. Game theory works with one view of social interaction, which meshes well with the instrumental account of human rationality; but equally there are other views (inspired by Kant, Hegel, Marx, Wittgenstein) which in turn require different models of (rational) action.

1.3 LIBERAL INDIVIDUALISM, THE STATE AND GAME THEORY

1.3.1 Methodological individualism

Some social scientists, particularly those who are committed to individualism, like the strict separation of choice and structure found in game theory because it gives an active edge to choice. Individuals *qua* individuals are plainly doing something on this account, although how much will depend on what can be said about what is likely to happen in such interactions. Game theory promises to tell a great deal on this. By comparison other traditions of political philosophy (ranging from Marx's dialectical feedback between structure and action to Wittgenstein's shared rules) work with models of

human agents who seem more passive and whose contribution merges seamlessly with that of other social factors. Nevertheless the strict separation raises a difficulty regarding the origin of structures (which, at least, on other accounts are no more mysterious than action and choice).

Where do structures come from when they are separate from actions? An ambitious response which distinguishes methodological individualists of all types is that the structures are merely the deposits of previous interactions (potentially understood, of course, as games). This answer may seem to threaten an infinite regress in the sense that the structures of the previous interaction must also be explained and so on. But, the individualist will want to claim that ultimately all social structures spring from interactions between some set of *a*social individuals; this is why it is 'individualist'. These claims are usually grounded in a 'state of nature' argument, where the point is to show how particular structures (institutional constraints on action) could have arisen from the interaction between *a*social individuals. Some of these 'institutions' are generated *spontaneously* through conventions which emerge and govern behaviour in repeated social interactions. For example, one thinks of the customs and habits which inform the tradition of common law. Others may arise through individuals consciously entering into contracts with each other to create the institutions of collective decision making (which enact, for example statute law). Perhaps the most famous example of this type of institutional creation comes from the early English philosopher Thomas Hobbes who suggested in *Leviathan* that, out of fear of each other, individuals would contract with each other to form a State. In short, they would accept the absolute power of a sovereign because the sovereign's ability to enforce contracts enables each individual to transcend the dog-eat-dog world of the state of nature, where no one could trust anyone and life was 'short, nasty and brutish'.

Thus, the key individualist move is to draw attention to the way that structures not only constrain; they also enable (at least those who are in a position to create them). It is the fact that they enable which persuades individuals consciously (as in State formation) or unconsciously (in the case of those which are generated spontaneously) to build them. To bring out this point and see how it connects with the earlier discussion of the relation between action and structure it may be helpful to contrast Hobbes with Rousseau. Hobbes has the State emerging from a contract between individuals because it serves the interests of those individuals. Rousseau also talked of a social contract between individuals, but he did not speak this individualist language. For him, the political (democratic) process was not a mere means of serving persons' interests by satisfying their preferences. It was also a process which *changed* people's preferences. People were socialised, if you like, and democracy helped to create a new

human being, more tolerant, less selfish, better educated and capable of cherishing the new values of the era of Enlightenment. By contrast, Hobbes' men and women were the same people before and after the contract which created the State.[4]

Returning to game theory's potential contribution, we can see that, in so far as individuals are modelled as Humean agents, game theory is well placed to help assess the claims of methodological individualists. After all, game theory purports to analyse social interaction between individuals who, as Hume argued, have passions and a reason to serve them. Thus game theory should enable us to examine the claim that, beginning from a situation with no institutions (or structures), the self-interested behaviour of these instrumentally rational agents will either bring about institutions or fuel their evolution. An examination of the explanatory power of game theory in such settings is one way of testing the individualist claims.

In fact, as we shall see in subsequent chapters, the recurring difficulty with the analysis of many games is that there are too many potential plausible outcomes. There are a variety of disparate outcomes which are consistent with (Humean) individuals *qua* individuals interacting. Which one of a set of potential outcomes should we expect to materialise? We simply do not know. Such pluralism might seem a strength. On the other hand, however, it may be taken to signify that the selection of one historical outcome is not simply a matter of instrumentally rational individuals interacting. There must be something more to it outside the individuals' preferences, their constraints and their capacity to maximise utility. The question is: what? It seems to us that either the conception of the 'individual' will have to be amended to take account of this extra source of influence (whatever it is) or it will have to be admitted that there are non-individualistic (that is, holistic) elements which are part of the explanation of what happens when people interact. In short, game theory offers the lesson that methodological individualism can only survive by expanding the notion of rational agency. The challenge is whether there are changes of this sort which will preserve the individualist premise.

1.3.2 Game theory's contribution to liberal individualism

Suppose we take the methodological individualist route and see institutions as the deposits of previous interactions between individuals. Individualists are not bound to find that the institutions which emerge in this way are fair or just. Indeed, in practice, many institutions reflect the fact that they were created by one group of people and then imposed on other groups. All that any methodological individualist is committed to is being able to find the origin of institutions in the acts of individuals *qua* individuals. The political theory of liberal individualism goes a stage further and tries to pass judgement on the legitimacy of particular institutions. Institutions in this

view are to be regarded as legitimate in so far as all individuals who are governed by them would have broadly 'agreed' to their creation.

Naturally, much will turn on how 'agreement' is to be judged because people in desperate situations will often 'agree' to the most desperate of outcomes. Thus there are disputes over what constitutes the appropriate reference point (the equivalent to Hobbes's state of nature) for judging whether people would have agreed to such and such an arrangement. We set aside a host of further problems which emerge the moment one steps outside liberal individualist premises and casts doubt over whether people's preferences have been autonomously chosen. Game theory has little to contribute to this aspect of the dispute. However, it does make two significant contributions to the discussions in liberal individualism with respect to how we might judge 'agreement'.

Firstly, there is the general problem that game theory reveals with respect to all (Humean) individualist explanations: the failure to predict unique outcomes in some games (a failure which was the source of doubt, expressed at the end of section 1.3.1, about methodological individualism). This is an insight which has a special relevance for the discussion in the political theory of liberal individualism concerning the conscious creation of institutions through 'agreement'. If the test of legitimacy is 'would individuals agree to such and such?' then we need a model which tells us what individuals will agree to when they interact. In principle there are probably many models which might be used for this purpose. But, if one accepts a basic Humean model of individual action, then it seems natural to model the 'negotiation' as a game and interpret the outcome of the game as the terms of the 'agreement'. Hence we need to know the likely outcome of such games in order to have a standard for judging whether the institutions in question might have been agreed to. Thus when game theory fails to yield a prediction of what will happen in such games, it will make it very difficult for a liberal political theory premised on Humean underpinnings to come to any judgement with respect to the legitimacy of particular institutions.

Secondly game theory casts light on a contemporary debate central to liberal theory: the appropriate role for the State, or more generally any collective action agency, such as public health care systems, educational institutions, industrial relations regulations, etc. From our earlier remarks you will recall that individualists can explain institutions either as acts of conscious construction (e.g. the establishment of a tax system) or as a form of 'spontaneous order' which has been generated through repeated interaction (as in the tradition which interprets common law as the reflection of conventions which have emerged in society). The difference is important. In the past two decades the New Right has argued against the conscious construction of institutions through the actions of the State, preferring instead to rely on spontaneous order.

One of the arguments of the New Right draws on Robert Nozick's (1974) view that the condition of 'agreement', in effect, is satisfied when outcomes result from a voluntary exchange between individuals. There is no need for grand negotiations involving all of society on this view: anything goes so long as it emerges from a process of voluntary exchange. We shall say nothing on this here. But this line of argument draws further support from the Austrian school of economics, especially Friedrich von Hayek, when they argue that the benefits of institution creation (for instance the avoidance of Hobbes's dog-eat-dog world) can be achieved 'spontaneously' through the conventions which emerge when individuals repeatedly interact with one another. In other words, to escape from Hobbes's nightmare, we do not need to create a collective action agency like the State according to the New Right wing of liberalism; and again game theory is well placed through the study of repeated games to examine this claim.

1.4 A GUIDE TO THE REST OF THE BOOK

1.4.1 Three classic games: chicken, coordination and the prisoners' dilemma games

There are three particular games that have been extensively discussed in game theory and which have fascinated social scientists. The reason is simple: they appear to capture some of the elemental features of all social interactions. They can be found both within existing familiar 'structures' and plausibly in 'states of nature'. Thus the analysis of these games promises to test the claims of individualists. In other words, how much can be said about the outcome of these games will tell us much about how much of the social world can be explained in instrumentally rational, individualist terms.

The first contains a mixture of conflict and cooperation: it is called *chicken* or *hawk–dove*. For instance, two people, Bill and Jill, come across a $100 note on the pavement and each has a basic choice between demanding the lion's share (playing hawk) or acquiescing in the other person taking the lion's share (playing dove). Suppose in this instance a lion's share is $90 and when both play dove, they share the $100 equally, while when they both act hawkishly a fight ensues and the $100 gets destroyed. The options can be represented as we did before along with the consequences for each. This is done in Figure 1.2; the pay-off to the row player, Jill, is the first sum and the pay-off to the column player, Bill, is the second sum.

Plainly both parties will benefit if they can avoid simultaneous hawk-like behaviour, so there are gains from some sort of cooperation. On the other hand, there is also conflict because depending on how the fight is avoided the benefits of cooperation will be differently distributed between the two