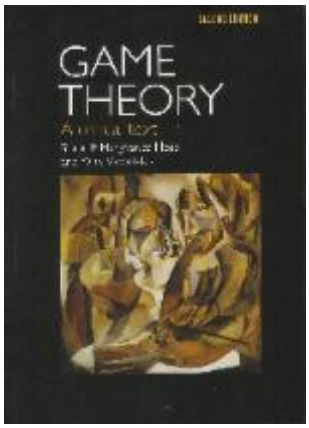


Game Theory: A critical text

Routledge 2004

This book started life as the second edition of an earlier one, also written in collaboration with Shaun Hargreaves-Heap (of the University of East Anglia), published back in 1995. The latter was entitled *Game Theory: A critical introduction.*

Around 2000 we started working on a second edition, following the unexpected success of the 1995 effort. However, as the writing continued, we realised that we were, in effect, writing a totally new book; one that shared very few passages with the 1995 edition (and was much longer and more comprehensive to boot). To make this clear, we effected a subtle but important change in the subtitle: The 2004 edition is referred to as a *Text*, in contrast to the 1995 one which was a 'mere' *Introduction*. Below you can read the Preface of the 2004 text, followed by the table of contents. Individual chapters are downloadable by pressing on the relevant chapter title below (the versions available are preproduction drafts - not the final copy as it appears in the book)

GAME THEORY: A Critical Text

PREFACE

As ever there are people and cats to thank. Some have left us, others have joined in, all are crucial players in our minds. There is also on this occasion electronic mail. Separated by oceans, endless landmasses, and the odd desert, we could not have written this book otherwise.

Its genesis goes back to a time in the eighties when we were colleagues at the University of East Anglia, Norwich, where game theory had become the object of interdisciplinary scrutiny. Both of us had been toying with game theory in an idiosyncratic way (see SHH's 1989 and YV's 1991 books) and it was a matter of time before we did it in an organised manner. The excuse for the book developed out of some joint work we did in Sydney in 1990, where YV had sought refuge (from Maggie's England) and SHH was visiting (on a year's 'release'). During this gestation period a number of colleagues played an important role in helping us see the trees from the forest and guided us around many pitfalls. The late Martin Hollis was one of them and we miss him dearly.

The first draft of this book took shape in various cafeterias in Florence during YV's visit to Europe in 1992 and matured on beaches and in restaurants during SHH's second visit to Sydney in 1993. For the next two years, the email wires between Sydney and Norwich, or wherever they are, were rarely anything other than warm to hot. When the first edition of this book came out in 1995, we blamed the Internet for all errors. On viewing the galley proofs of the first edition, the uncertainty of our feelings about the whole enterprise was such that we almost fell out.

The actual reception was far more gratifying than at least one of us had imagined. Many social theorists, from distant lands and near, wrote in appreciation. It is they, and Rob Langham (our indefatigable Routledge editor), who must be blamed entirely for this new effort. Had their reaction not been as heart warming, we doubt we would have found the energy to go back to the drawing board. For this is precisely what we did: While maintaining the original book's style and philosophy, we started from scratch, only occasionally plundering the first edition. Help came our way from two quarters: The Australian Research Council which funded the experimental work referred to in Chapters 6&7, and EUSSIRF which funded YV's research at the LSE Library in 2002.

The current book's creation coincided with YV's latest migration, this time to his native Greece and to hitherto unknown administrative chores. It also coincided with SHH going 'native' at East Anglia; namely, becoming that University's Pro-Vice-Chancellor. Evidently, unlike the first edition, this one did not come of age on golden Antipodean beaches or in Florentine cafeterias. We hope we succeeded in concealing this sad reality in the pages that follow.

It is natural to reflect on whether the writing of a book exemplifies its theme. Has the production of these two books been a game? In a sense it has. The opportunities for conflict abounded within a two-person interaction which would have not generated this book unless strategic compromise was reached and co-operation prevailed. In another sense, however, this was definitely no mere game. The point about games is that objectives and rules are known in advance. The writing of a book (let alone two in succession, and on the same subject) is a different type of game, one that game theory does not consider. It not only involves moving within the rules, but also 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 while writing this book over the last ten years or so. Is it any wonder that this is A Critical Text?

Lastly, there are the people and the cats: Empirico, Joε, Lindsay, Margarita, Pandora, Thibeau and Tolstoy – thank you.

Yanis Varoufakis and Shaun Hargreaves-Heap, July 2003

Chapter 1: OVERVIEW

1.1 INTRODUCTION
1.1.1 Why study game theory?
1.1.2 What is game theory?
1.1.3 Why this book?
1.1.4 Why a second edition?
1.2 THE ASSUMPTIONS OF GAME THEORY
1.2.1 Individual action is instrumentally rational
1.2.2 Common knowledge of rationality
1.2.3 Common priors
1.2.4 Action within the rules of the game
1.3 LIBERAL INDIVIDUALISM, THE STATE AND GAME THEORY
1.3.1 Methodological individualism
1.3.2 Game theory's contribution to liberal individualism
1.4 A GUIDE TO THE REST OF THE BOOK
1.4.1 Three classic games: Hawk-dove, Co-ordination and the Prisoner's Dilemma
1.4.2 Chapter-by-chapter guide
1.5 CONCLUSION

Chapter 2: THE ELEMENTS OF GAME THEORY

2.1 INTRODUCTION
2.2 THE REPRESENTATION OF STRATEGIES, GAMES AND INFORMATION SETS
2.2.1 Pure and mixed strategies
2.2.2 The Normal Form, the Extensive Form and the Information Set
2.3 DOMINANCE REASONING
2.3.1 Strict and weak dominance
2.3.2 Degrees of common knowledge of instrumental rationality
2.4 RATIONALISABLE BELIEFS AND ACTIONS
2.4.1 The successive elimination of strategically inferior moves
2.4.2 Rationalisable strategies and their connection with Nash's equilibrium
2.5 NASH EQUILIBRIUM
2.5.1 John Nash's beautiful idea
2.5.2 Consistently aligned beliefs, the hidden Principle of Rational Determinacy and the Harsanyi-Aumann Doctrine
2.5.3 Some Objections to Nash: Part I
2.6 NASH EQUILIBRIUM IN MIXED STRATEGIES
2.6.1 The scope and derivation of Nash equilibria in mixed strategies
2.6.2 The reliance of NEMS on CAB and the Harsanyi doctrine
2.6.3 Aumann's defence of CAB and NEMS
2.7 CONCLUSION

Chapter 3: BATTLING INDETERMINACY - Refinements of Nash's equilibrium in static and dynamic games

3.1 Introduction
3.2 THE STABILITY OF NASH EQUILIBRIA
3.2.1 Trembling hand perfect Nash equilibria
3.2.2 Harsanyi's Bayesian Nash equilibria and his defence of NEMS
3.3 DYNAMIC GAMES
3.3.1 Extensive form and backward induction
3.3.2 Subgame perfection, Nash and CKR
3.3.3 Sequential equilibria
3.3.4 Bayesian learning, sequential equilibrium and the importance of reputation
3.3.5 Signalling equilibria
3.4 FURTHER REFINEMENTS
3.4.1 Proper equilibria
3.4.2 Forward induction
3.4 SOME LOGICAL OBJECTIONS TO NASH, PART III
3.4.1 A critique of subgame perfection
3.4.2 A negative rejoinder (based on the Harsanyi-Aumann doctrine)
3.4.3 A positive rejoinder (based on sequential equilibrium)
3.3.4 Conclusion: Out-of-equilibrium beliefs, patterned trembles and consistency
3.5 CONCLUSION
3.5.1 The status of Nash and Nash refinements
3.5.2 In defence of Nash
3.5.3 Why has game theory been attracted 'so uncritically' to Nash?

Chapter 4: BARGAINING GAMES - Rational agreements, bargaining power and the Social Contract

- 4.1 INTRODUCTION
- 4.2 CREDIBLE AND INCREDIBLE TALK IN SIMPLE BARGAINING GAMES
- 4.3 JOHN NASH'S GENERIC BARGAINING PROBLEM AND HIS SOLUTION

4.3.1 The bargaining problem
4.3.2 Nash's solution – An example
4.3.3 Nash's solution as an equilibrium of fear
4.3.4 Nash's axiomatic account
4.3.5 Do the axioms apply?
4.3.6 Nash's solution – a summary

- 4.4 ARIEL RUBINSTEIN AND THE BARGAINING PROCESS: The return of Nash backward induction

4.4.1 Rubinstein's solution to the bargaining problem
4.4.2 A proof of Rubinstein's theorem
4.4.3 The (trembling hand) defence of Rubinstein's solution
4.4.4 A final word on Nash, trembling hands and Rubinstein's bargaining solution

- 4.5 JUSTICE IN POLITICAL AND MORAL PHILOSOPHY

4.5.1 The negative result and the opening to Rawls and Nozick
4.5.2 Procedures and outcomes (or 'means' and ends) and axiomatic bargaining theory

- 4.6 CONCLUSION

Chapter 5: THE PRISONERS' DILEMMA - The riddle of co-operation and its implications for collective agency

5.1 INTRODUCTION: THE STATE AND THE GAME THAT POPULARISED GAME THEORY
5.2 EXAMPLES OF HIDDEN PRIOSNERS' DILEMMAS AND FREE RIDERS IN SOCIAL LIFE
5.3 SOME EVIDENCE ON HOW PEOPLE PLAY THE GAME
5.4 EXPLAINING COOPEATION
5.4.1 Kant and morality: Is it rational to defect?
5.4.2 Altruism
5.4.3 Inequality aversion
5.4.4 Choosing a co-operative disposition instrumentally
5.5 CONDITIONAL CO-OPERATION IN REPEATED PRISONERS' DILEMMAS
5.5.1 Tit-for-tat in Axelrod's tournaments
5.5.2 Tit-for-tat as a Nash equilibrium strategy when the horizon is unknown
5.5.3 Spontaneous public good provision
5.5.4 The Folk Theorem, Indeterminacy and the State
5.5.5 Does a finite horizon wreck co-operation? The theory and the evidence
5.6 CONCLUSION: COOPERATION AND THE STATE IN LIBERAL THEORY
5.6.1 Rational cooperation?
5.6.2 Liberalism and the prisoners' dilemma
5.6.3 The limits of the prisoners' dilemma

Chapter 6: EVOLUTIONARY GAMES - Evolution, Games and Social Theory

6.1 GAME THEORY'S ENCOUNTER WITH EVOLUTIONARY BIOLOGY
6.1.1 The origins of Evolutionary Game Theory
6.1.2 Evolutionary stability and equilibrium: An introduction
6.1.3 Spontaneous order versus political rationalism
6.2 SYMMETRICAL EVOLUTION IN HOMOGENEOUS POPULATIONS
6.2.1 Static games
6.2.1 Dynamic games
6.3 EVOLUTION IN HETEROGENEOUS POPULATIONS
6.3.1 Asymmetrical (or two-dimensional) evolution and the demise of Nash equilibria in mixed strategies (NEMS)
6.3.2 Does Evolutionary Game Theory apply to humans as well as it does to birds, ants, etc.? An experiment with 2-dimensional evolution in the Hawk-Dove game
6.3.3 Multi-dimensional evolution and the conflict of conventions
6.3.4 The origin of conventions and the challenge to methodological individualism
6.3.5 The politics of mutations: Conventions, inequality and revolt
6.3.6 Discriminatory Conventions: A brief synopsis
6.4 SOCIAL EVOLUTION: POWER, MORALITY AND HISTORY
6.4.1 Social versus natural selection
6.4.2 Conventions as covert social power
6.4.3 The evolution of predictions into moral beliefs: Hume on morality
6.4.4 Gender, class and functionalism
6.4.5 The evolution of predictions into ideology: Marx against morality
6.5 CONCLUSION

Chapter 7: PSYCHOLOGICAL GAMES - Demolishing the divide between motives and beliefs

7.1 INTRODUCTION
7.2 DIFFERENT KINDS OF 'OTHER REGARDING' MOTIVES
7.2.1 The 'other'-regarding motives of Homo Economicus
7.2.2 Beliefs as predictions and as motives
7.3 THE MOTIVATING POWER OF NORMATIVE BELIEFS
7.3.1 Fairness equilibria
7.3.2 Computing Fairness equilibria
7.3.3 Assessing Rabin
7.3.4 An alternative formulation linking entitlements to intentions
7.3.5 Team thinking
7.4 PSYCHOLOGY AND EVOLUTION
7.4.1 On the origins of normative beliefs: an adaptation to experience
7.4.2 On the origins of normative beliefs: The resentment-aversion versus the subversion-proclivity hypotheses
7.5 CONCLUSION: SHARED PRAXIS, SHARED MEANINGS

POSTSCRIPT

SOLUTIONS TO PROBLEMS

NOTES

AUTHORS' INDEX

SUBJECT INDEX

REFERENCES

CHAPTER 1
OVERVIEW

1.1 INTRODUCTION
- 1.1.1 Why study game theory?
- 1.1.2 What is game theory?
- 1.1.3 Why this book?
- 1.1.4 Why this edition?

1.2 THE ASSUMPTIONS OF GAME THEORY
- 1.2.1 Individual action is instrumentally rational
- 1.2.2 Common knowledge of rationality
- 1.2.3 Common priors
- 1.2.4 Action within the rules of the game

1.3 LIBERAL INDIVIDUALISM, THE STATE AND GAME THEORY
- 1.3.1 Methodological individualism
- 1.3.2 Game theory's contribution to liberal individualism

1.4 A GUIDE TO THE REST OF THE BOOK
- 1.4.1 Three classic games: Hawk-Dove, Co-ordination and the Prisoner's Dilemma
- 1.4.2 Chapter-by-chapter guide

1.5 CONCLUSION

1
OVERVIEW

1.1 INTRODUCTION

1.1.1 Why study game theory?

This book's first edition began with the observation that game theory was everywhere; that after thrilling a whole generation of post-1970 economists, it was spreading like a bushfire through the social sciences. In addition, game theorists had begun to advance some pretty ambitious claims regarding the potential of the theory to become for the social sciences what mathematics is to the natural sciences: a unifying force able to bring together politics, economics, sociology, anthropology etc. under one roof and turn them into sub-disciplines of some broader 'science of society'.

As a glimpse of game theory's increasing confidence, we cited two prominent game theorists' explanation of the attraction:

> 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)

To overcome the reader's suspicion that such exuberance was confined to game theory's practitioners, we also cited Jon Elster, a well-known social theorist with very diverse interests, whose views on the usefulness of game theory did not differ significantly from that of the practitioners:

> [I]f 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)

Our point was that, if seemingly disinterested social theorists held game theory in such esteem, those studying social processes and institutions can ill-afford to pass game theory by. Our book intended to subject its grand claims to critical scrutiny while, at the same time, presenting a concise, simplified yet analytically advanced account of game theory's techniques. We concluded our inquiry by arguing that the study of game theory is extremely useful for social scientists, albeit not for the reasons put forward by the game theorists.

In the first few years after our first edition saw the light of day, the enthusiasm continued to grow unabated. In his impressive 1999 survey, Roger Myerson compared the discovery of game theory's main concept (the Nash Equilibrium) with that of the DNA double helix and claimed that it has transformed economics to such a remarkable degree that the latter can now pose credibly as *the* foundational 'science of society'.

It is often said that the seed of Decline begins to take root at the height of an Empire's power and optimism. As game theory was conquering in the 1990s diverse fields from economics and anthropology to philosophy and biology, doubt emerged concerning its real value for social theorists. Interestingly, this apostasy came not from some radical group opposed to game theory for self-interested reasons but, rather, from practising game theorists (see Mailath, 1998, and Samuelson, 2002).

Our book's point in 1995 was that game theory is best studied critically (hence our subtitle). Insights of great substance are to be had from understanding two things at once:

(A) Why game theory inspired such enthusiasm among intelligent social theorists spanning many disciplines, and
(B) Why the jury is still out regarding all of the theory's foundational notions.

At the time of our first edition's publication, our commitment to seeking enlightenment about social processes through immanent criticism of game theory's concepts was treated by some as eccentric (even heretical).

Naturally we feel vindicated by the game theorists' recent espousal of the method of immanent criticism. However, what is of greater importance is that, precisely because the game theorists themselves are becoming increasingly interested in the weaknesses of their foundations, social theorists stand to gain substantially from a critical engagement with the debates within game theory. To put it bluntly, *understanding why game theory does not, in the end, constitute the science of society (even though it comes close) is terribly important in understanding the nature and complexity of social processes*. This is, in our view, the primary reason why social theorists should be studying game theory.

1.1.2 What is Game Theory?

Game theory really begins 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, the claim was made that we shall now be able to draw on a single mode of analysis: the theory of games.

1.1.3 Why this book?

Our motivation for writing this book originally was an interesting contradiction. On the one hand, we doubted that the claim in Section 1.1.2 was warranted. This explains the book's subtitle. On the other hand, however, we enjoyed game theory and had spent many hours pondering its various twists and turns. Indeed it had helped us on many issues. However, we believed 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 believed the contribution of game theory to be largely paedagogical. Such contributions are not to be sneezed at.

We also felt that game theory's further substantial contribution was 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 first edition's main conclusion (which was developed through the book) was that game theory exposes the limits of these models of human agency. 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'.

Game theory books had proliferated in number even before our first edition in 1995. For example, Rasmussen (1989) was a good 'user's manual' with many economic illustrations. Binmore (1990) comprised lengthy, technical but stimulating essays on aspects of the theory. Kreps (1990) was a delightful book and an excellent eclectic introduction to game theory's strengths and problems. Myerson (1991), Fudenberg and Tirole (1991) and Binmore (1992) added worthy entrants to a burgeoning market. Dixit and Nalebuff (1993) contributed a more informal guide while Brams (1993) was a revisionist offering. One of our favourite books, despite its age and the fact that it is not an extensive guide to game theory, was Thomas Schelling's *The Strategy of Conflict,* first published in 1960. It is highly readable and packed with insights few other books can offer.

Despite the large number of textbooks available at the time, *none* of them located game theory in the wider debates within social science. We thought it important to produce an introductory book which does *not* treat game theory as a series of solved problems to be learnt by the reader. Indeed, we felt that the most fruitful way of conveying game theory was by presenting its concepts and techniques *critically*. Engineers can afford to impart their techniques assertively and demand that the uninitiated go through the motions until they acquire the requisite knowledge. Game theorists doing the same devalue their wares. Our first edition was, thus, motivated by the conviction that presentations of game theory which simply plunder the social sciences for illustrations (without however locating the theory properly within the greater debates of social science) are unfortunate for two reasons:

First, they were 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 was 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.

The first edition addressed these worries. It was written for all social scientists. It did not claim to be an authoritative textbook on game theory. There are some highways and byways in game theory which were not travelled. But it did focus on the central concepts of game theory, and discussed them critically and simply while remaining faithful to their subtleties. The technicalities were trimmed to a minimum (readers needed a bit of algebra now and then) and our aim was to lead with the ideas.

1.1.3 Why this second book?

Since our first edition, the list of game theory textbooks has grown to such an extent that it would be futile to enumerate them.[1] Most of them are competent and some of them are excellent. Of the relatively (technically) advanced introductions, we have found Osborne and Rubinstein (1994) to be the most useful and thoughtful offering. Among the many texts on the market, there have been quite a few good guides on game theory's applications to political and other social sciences (our preferred one is Dixit and Skeath, 1999).

Nevertheless, we still feel that there is still no other text undertaking the task we set our selves ten years ago: of combining an introduction to game theory with a critical attempt to locate the latter within the broader social science debates. So, why a new version of our 1995 effort? For two reasons: First, because there have been many developments in game theory which, once understood, reinforce our book's original argument but also open up windows onto some interesting new vistas. Indeed, the same developments, if misunderstood, may cause confusion and sidetrack the social theorist who cares not for the technicalities but for the *meaning* of these developments. This new edition hopes to offer readers a guide through this theoretical maze of increasing complexity.

Secondly, many readers and colleagues suggested a new edition which would cover game theory's *techniques* more accurately and comprehensively. In short, it was suggested to us that, while retaining our emphasis on 'leading with the ideas', the book should offer more on techniques so as to be *useable as a self-contained textbook*. Some even demanded solved problems at the end of each chapter (a 'demand' that has been met).

As a result of taking in more material and trying to maintain the critical aspect of the book, while at the same time turning it into an accomplished textbook, the second edition is much longer than the first. But even though most of the book is almost new (sharing only very few passages with the first edition), its spirit and its philosophy have remained intact. The first four chapters retain their original titles, barring some new subtitles. The organisational changes begin with Chapter 5. In the first edition Chapter 5 was dedicated to the *prisoner's dilemma* and Chapter 6 to a dynamic extension of it (and of other games). Here, these two chapters have been merged into the new Chapter 5.

Furthermore, there is no longer a separate chapter discussing the empirical evidence on how people actually play games (thus Chapter 8 of the first edition does not appear in this one). This is not because we believe empirical evidence from the laboratory to be less significant; indeed, quite the opposite is true. As the empirical evidence has grown, it is natural to refer to the relevant evidence when theory is being discussed. So, each chapter now is littered with experimental evidence. This adds an important, and we hope helpful, dimension to the critical aspects of the argument at each stage because the evidence adds weight to these critical theoretical observations.

Chapter 6 covers evolutionary game theory (as did Chapter 7 of the first edition) and a new Chapter 7 is introduced on psychological games. The latter received a passing mention in the first edition but has been upgraded to a fully-fledged chapter here. The combination of the last two chapters (on evolutionary and psychological games) is of central importance to this edition for reasons which will become obvious below.

As we have already stressed, the principal cause of our critical stance in the first edition was the failure of game theory to explain action in a variety of social settings. We argued that this was directly related to weakness of the 'rational choice' model on which game theory is founded. There are two aspects to the problem.

First, game theory fails to make predictions about what rational people will do in many settings. This is because often the connection between the kind of rationality agents are presumed to have and game theory's predictions (e.g. the so-called Nash equilibrium) is tenuous. Additionally, there are many social settings in which game theory predicts too many outcomes at once (the case of so-called multiple equilibria).

Secondly, when game theory does make predictions, these are often not upheld in practice. This has become even clearer since the first edition and we now include more references to and discussion of the empirical evidence on how people play games.

The two developments that we focus on in this new edition are essentially responses to these two problems. One is evolutionary game theory (see Chapter 6). The first edition's chapter on the latter has been completely revised here. In part this reflects the way that evolutionary game theory promises to provide an account of equilibrium selection and so directly addresses one aspect of the weakness with respect to predicting behaviour. It is also a consequence of the way evolutionary arguments have acquired much greater significance in the social sciences since we wrote the first

edition. For instance, the arguments from evolutionary psychology have become popular sources not just for the Sunday colour magazines but also for the debates concerning the origin of language and morality (see Pinker, 1997, and Binmore, 1998). We have some sympathy for evolutionary game theory not least because it actually supplies a useful corrective both to arguments in evolutionary psychology and to the more casual appeal to evolutionary ideas that has become commonplace.

Evolutionary game theory marks a relatively minor departure from the rational choice model. In contrast, the other area of development concerns alternative models of rational action. Our brand new Chapter 7 (on psychological games) sets out some of these theories. They share a recognition that in many social settings behaviour can only be understood with reference to the prevailing norms which give actions symbolic properties. In other words, Chapter 7 challenges even the possibility of describing a game's structure prior to understanding the social norms in which the players are entangled. In the first edition, we had made some noises about the Wittgensteinian idea of rule-governed behaviour and how it could be seen as a potential source for a necessary corrective to the simple rational choice model. Here, see Chapter 7, we utilise the improved understanding of how norms help in framing decision-making, in order to illustrate better the pertinence of the Wittgensteinian insight.

These changes reflect our experience from teaching with the first edition as does the inclusion now of problems after each chapter with answers at the back. We feel that the new subtitle does our book justice: for this is a fully-fledged *Critical Text* (unlike the first edition which was only meant as a critical introduction to game theory). Besides more technical sophistication, the current book has an air of excitement not found in the first. It narrates many new developments, partially fuelled by the growing empirical evidence on how people play games. They promise to address the problems we identified in the first edition and, by doing so, they threaten to change the foundations of game theory. In short, game theory has become a site where the dominant 'rational choice' model in the social sciences is being subverted by socially richer models of agency. Our ambition for the present book is to be a reliable and relatively complete guide not only to game theory *per se* but also to the in-tense relation between game theory and the social sciences.

1.1.4 The rest of this chapter

We begin the argument of the book, as in the first edition, by sketching (see Section 1.2) the philosophical moorings of conventional game theory, discussing in turn its four key assumptions: *Agents are instrumentally rational* (Section 1.2.1); they have *common knowledge* of this rationality (Section 1.2.2); they hold *common priors* (Section 1.2.3); and they *know the rules* of the game (Section 1.2.4). These assumptions set out where standard 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 fourth are ontological.[2] 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 and third are epistemological in nature[3] (and in some games they, particularly the third, are not essential for the analysis). They help establish what can be inferred about the beliefs which rational people will hold about how games will be played.

We spend more time discussing these assumptions than is perhaps usual in texts on game theory precisely because we believe that the assumptions are both controversial and problematic, in their own terms and when cast as general propositions concerning interactions between individuals. The discussions of instrumental rationality, common knowledge of instrumental rationality and common priors (Sections 1.2.1, 1.2.2 and 1.2.3), in particular, are indispensable for anyone interested in game theory. In comparison Section 1.2.4 will appeal more to those who are concerned with where game theory fits in to the wider debates within social science and to those who are particularly interested in the new developments with respect to normative reason and psychological games.

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 conventional 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 that 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.

1.2 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 three 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 two are 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-ends 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).[4]

Thus a promisingly general model of action seems to the heart of game theory. For instance, it could apply to any type of player 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) also 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 introductory economic 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 conveniently forgotten them, it is worth explaining that preferences are sometimes represented mathematically by a utility function. This needs careful handling.

'Utility' here should not be confused with the philosophy of *Utilitarianism*. A utility function is just a device for mathematically representing a person's preferences. The function gives numbers to outcomes such that the most preferred outcome has the highest number, the next most preferred has the second highest number, and so on until the least desirable outcome gets the lowest number (or rank). In this way, selecting the action that best satisfies one's preferences is the equivalent of choosing the action with the highest 'utility' number (i.e. maximising utility).

The designation of this function as a utility function and the associated numbers as 'utils' is a (gratuitous) gloss on what is actually a simple mathematical device for representing a preference ordering. The function could as well be called a preference function or some such. Nevertheless, it is the practice of game theory and economics to refer to these functions as utility functions so that the numerical pay-offs associated with each outcome are counted in 'utils'; and we follow this here.

However, since the resulting metaphor of utility maximisation 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 and catching the train (assume 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 2 utils from driving and 1 util from travelling on the train and you choose driving because this maximises the utils generated (as 2 >1).

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. A person is deemed *instrumentally rational* if he or she has preferences which satisfy the following conditions:

(1) *Reflexivity*: No alternative x_i is *less* desired than itself.

(2) *Completeness*: For any two alternatives x_i, x_j, either x_i is preferred to x_j, or x_j is preferred to x_i, or the agent is indifferent between the two

(3) *Transitivity*: For any x_i, x_j, x_k, if x_i is no less desired than x_j, *and* x_j is no less desired than x_k, then x_i cannot be less desired than x_k

(4) *Continuity* : For any x_i, x_j, x_k, if x_i is preferred to x_j and x_j is preferred to 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.

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 croissants', x_j is '3 bagels' and x_k is '10 bread rolls', then there must exist some combination of croissants and bread rolls (e.g. 2 croissants and 4 bread rolls) which is equally valued with 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 (i.e. the 3 bagels).

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 behaving *as if* to maximise this utility function.

It will be obvious though that this assignment of utility numbers is arbitrary in the sense that any number of utils X and Y will do respectively for driving and travelling by rail respectively provided X >Y whenever the person prefers the former. 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. First, 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 one reason 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).

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 of 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. (Hume, 1740,1888)

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 Stuart 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'.

Suppose now that the choice problem is complicated by the presence of uncertainty. 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. In such cases, we assume that people have a preference ordering over what are called 'prospects': these are the outcomes and their probabilities associated with each action. Let us say that the predicted chance of rain by the weather bureau is 50-50. The prospects here using the standard notation are: ('walking in dry', 'walking in rain'; 0.5, 0.5) and ('driving in dry', 'driving in rain'; 0.5, 0.5). That is, when you decide to 'walk' there is probability of 0.5 that it will be a 'walk in the rain' and a 0.5 chance that it will be a 'walk in the dry' and likewise for driving. If in addition we assume that people's preferences satisfy some further axioms regarding how the probability of an outcome can affect the preference for a prospect, then this ordering can be represented as acting so as to maximise *expected utility* (see Box 1.3 for details).

For example, suppose that the person prefers to walk and has a preference ordering over 'walking in the dry', driving in the wet', 'driving in the dry' and 'walking in the wet' that can be represented by a utility function which gives the following numbers (or utils) respectively to these outcomes: 10, 6, 1, and 0. Then it will be possible to reconstruct this choice in terms of expected utility maximisation. The expected utility from walking is $(0.5)\times(10) + (0.5)\times(0) = 5$ (i.e. a 50-50 chance of getting 10 or 0). The expected utility from driving is $(0.5)\times(6) + (0.5)\times(1) = 3.5$. Hence the person chooses to walk because it yields the higher expected utility.

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 $1\text{-}p$, 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 is preferred to y_j and $y_m = (y_i, y_j; p_1, 1\text{-}p_1)$, $y_n = (y_i, y_j; p_2, 1\text{-}p_2)$, then y_m is preferred to y_n only if $p_1 > p_2$.

(6) *Independence:* For three prospects y_i, y_j and y_k, if y_i is preferred to y_j, then there exists a probability λ such that a $(\lambda, 1\text{-}\lambda)$ probability mix of y_i and y_j must be at least as good as a $(\lambda, 1\text{-}\lambda)$ probability mix of y_i and y_k. In our notation for prospects, $(y_i, y_j; \lambda, 1\text{-}\lambda)$ is no less desired than $(y_i, y_k; \lambda, 1\text{-}\lambda)$.

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.

In this way, the use of the metaphor of utility maximisation to describe acting so as to best satisfy preferences can be extended to choice under uncertainty where it becomes expected utility maximisation. This will probably make immediate intuitive sense since actions in these settings are no longer uniquely associated with one outcome and so people will have to form an expectation regarding the consequences of any action. But there is one wrinkle which is worth exploring. In

choice under uncertainty, the function which represents a person's ordering of outcomes is a *cardinal* (as opposed to ordinal) utility function. This means that utility numbers assigned to outcomes do not simply capture the person's ordering, they also provide a measure of the intensity of a person's preference. Thus if 'walking in the dry' is 10 times better than 'driving in the dry', the cardinal utility function that represents this ordering would assign numbers like 10 and 1 to 'walking in the dry' and 'driving in the dry' respectively. 20 and 2 would do as well or 30 and 3 and so on, since these numbers would also convey this same intensity of preference. Thus the precise numbers in a cardinal utility function remain arbitrary in this sense (and as a result it still makes no sense to attempt any comparison across individuals). Nevertheless, they do yield more information regarding a person's preferences than do ordinal functions (where the numbers would only have to satisfy the constraint of X>Y).

Again, it makes some intuitive sense to adopt cardinal utility functions so as to extend the metaphor of utility maximisation to settings in which outcomes of choices are not known in advance. The decision between driving and walking 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 *as well as on the likelihood of rain*. 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. Alternatively, if walking in the dry is only slightly more preferred to getting wet, then you are less likely to take the risk of getting wet. Hence when we use the metaphor of utility maximisation to describe risky decisions we will have to work with utility functions that encode information regarding intensity of preference.

Cardinal utilities and the assumption of expected utility maximisation are important because uncertainty is ubiquitous in games. Consider the following variant of an earlier example. You must choose between walking to work and 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 and strongest preference is that you walk together. If you cannot have this stroll with our friend, you would rather drive (your second preference). Last in your preference ordering is that you walk only to find out that your friend has driven to work. We will represent these preferences with some cardinal utility numbers in matrix form - see Game 1.1.

Suppose that, from past experience, you believe that there is ⅔ 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 Game 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 ⅔ chance that you will meet up with your friend (yielding 2 utils for you) and a ⅓ chance of walking alone (0 utils). On average, walking will give you 4/3 utils (⅔ times 2 plus ⅓ 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. p=⅔) 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 ½.

		Friend	
		Drive	**Walk**
You	**Drive**	1	1
	Walk	0	2

Game 1.1 – Walking with a friend (a co-ordination problem)

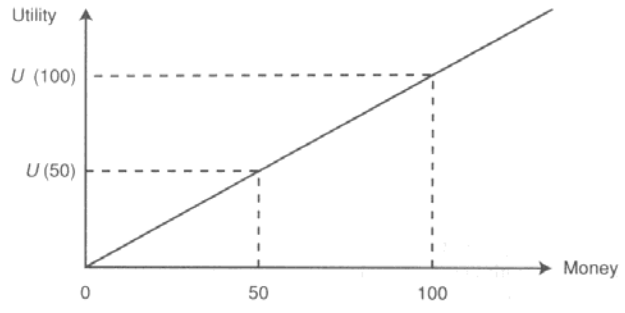
There is one final important application of cardinal utility functions. When a person makes risky decisions involving outcomes that are monetary sums (as for example in a lottery or any other gamble involving financial outcomes like investments in shares, houses, education, etc) the cardinal utility numbers associated with these monetary outcomes reflect a person's attitude towards risk. This is explained in more detail in Box 1.4. As we shall see later on in the book, this is of importance in many game situations (e.g. in bargaining – see Chapter 4).

Box 1.4

UTILITY FUNCTIONS AND RISK AVERSION

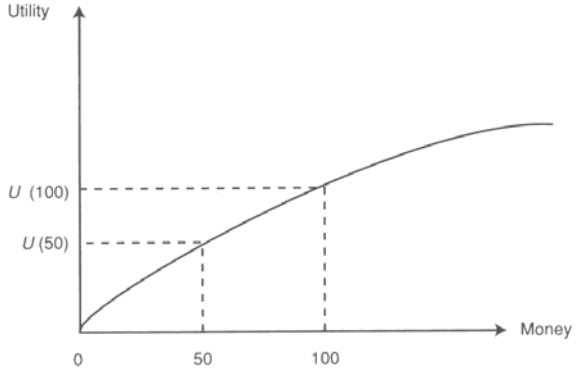
Suppose an individual is offered a 50-50 chance of winning \$100 in a lottery and a lottery ticket costs \$50. We say such persons are *risk neutral* when they are indifferent between buying the lottery ticket and forsaking the opportunity. If they buy the lottery ticket, then we call them *risk lovers*; and when they will not buy the ticket, we call them *risk averse*. The intuition behind these descriptions will be obvious as the expected return from buying the lottery ticket is \$50 and so if you positively want to buy this prospect for \$50 you must love a gamble. Conversely, if you are indifferent between them, you are neutral to the risk; whereas the *risk averse* person obviously will not buy the ticket as there is nothing in it for him or her except the risk which they do not like.

When we plot utility as a function of dollars and we assume that the individual is an expected utility maximiser, the curvature of the utility function can be directly linked to these varying attitudes to risk. To see the point consider someone who has a linear utility function in money, as in the figure below.



For this person, the utility of \$50, $U(50)$ is plainly equal to the expected utility of the lottery ticket (= $0.5U(0) + 0.5U(100)$). Thus this is the kind of person we have referred to as *risk neutral*. Now consider another person with a utility function which is convex in money, as in the figure below.

For this person, the utility of \$50, $\underline{U}(50)$ is plainly greater than the expected utility of the lottery ticket (= $0.5U(0) + 0.5U(100)$) because of the curvature of the utility function. Thus this is the kind of person we have referred to as *risk averse*. Had the utility curved upwards in the opposite direction, then the result would have been the exact opposite and we would have a person who was a *risk lover*.



The ability to represent the idea that people are instrumentally rational in the sense that they act so as best to satisfy their preferences through the metaphor of utility functions and the assumption that people maximise expected utility is an analytic convenience. It greatly simplifies the way that choice problems are represented and solved. This will be evident in what follows, but there is already a hint of this in the example above where the choice between walking and driving depends on what your friend decides to do. This dependence was set out quite simply in the matrix form of Game 1.1 using utility numbers and the decision was then easily analysed and found to depend on your probability assessment regarding your friend's action. In this way, the assumption of instrumental rationality is attractively tractable and it cues the next set of assumptions which are concerned with what you should rationally expect your friend to do. Nevertheless, there are a number of reasons why many theorists are unhappy with the assumption of instrumental rationality and we turn to the source of these doubts before we consider the particular assumptions made in game theory regarding rational beliefs.

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 1.5 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 still possible to maintain allegiance to expected utility theory even after acknowledging the evidence which suggests that people fail, in practice, to live up to its recommendations. However, in so far as game theorists adopt this defence of expected utility theory, game theory runs the risk of appearing as a prescriptive theory, as opposed to a 'positive' account of how people actually play games. 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.

Box 1.5

THE ALLAIS PARADOX

Kahneman and Tversky (1979) offer the following reworking of the famous study in Allais (1953) (see also Sugden, 1991, for a comprehensive 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 over 1 and 3 over 4. 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% *and* \$2500 with probability 33%
Lottery 2	**\$2400 with probability 66%** \$2400 with probability 34%
Lottery 3	**0 with probability 66%** 0 with probability 1% *and* \$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 (heavily bolded) 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 (i.e. \$2400 with probability 66%) 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 vice versa. Yet, the majority of people participating in such experiments seem to violate the independence axiom and choose lotteries 2 (over 1) and 3 (over 4). The fact that expected utility theory receives little empirical support is potentially worrying for game theory because it relies so heavily on it.

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)? On many accounts honour comes from acting on a code of behaviour that will often run counter to the dictates of individual interest. Acting morally seems frequently to do the same. So making 'honour' or 'morality' just another preference is likely to defeat the object of acting 'honourably' or 'morally' by subsuming them under an account of individual preference satisfaction.

At best they can only survive as distinct motives if we consider the character of preferences in more detail and introduce a two-tier structure such that the extent of ethical or honourable preference satisfaction (the second tier) can be gauged by comparing actual action with what would be dictated by the pursuit of one's self interested preferences (the first tier). This is the approach of the new theories of rational action that we consider in Chapter 7. Two observations are worth making at this stage about this strategy because they mark the ways in which this move is a departure from the model of instrumental rationality used in conventional game theory.

The judgements that individuals make about their actions are likely to depend on some external standard of 'honour' or 'morality', otherwise they will be prone to the suspicion of being self serving and so fall into the same trap of being indistinguishable from self interested preference satisfaction. The shared judgement of others that is encoded in the norms of a group supplies one obvious such external standard, but it begs a question as to their origin. Can the norms be derived in some way from the underlying self-interested preferences of the individuals of the group? Or are the group and the individual mutually constituted through the presence of norms (see the discussion of Wittgenstein in the next section)?

Secondly, these judgements about the value of an action with respect to any code of conduct are likely to be highly context dependent. Sen (1994) supplies a famous illustration when considering how a person might choose between an apple and an orange. Suppose they choose the apple. Now add a second, smaller apple to the menu and it is not implausible to imagine that a person would now choose the orange because taking the apple that is now 'big' in the context of the addition of the smaller apple could make the person look greedy. This involves an obvious inconsistency of choice if the objects of choice are defined purely by their physical attributes: the person first prefers the apple to the orange and then the orange to the apple.

To avoid the inconsistency one can distinguish objects by their context, so that the first apple is not the same on the second menu even though it shares the same physical characteristics because of the presence of the smaller apple on this menu. This creates problems for the axiomatic representation of instrumentally rational choice (see Hargreaves Heap, 2000) and also takes us directly back to the need to understand the norms which underpin the symbolic evaluations of action. What are the norms and where do they come from? Such questions quickly become philosophical and so we turn explicitly in this direction.

(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, just 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 restraint not only on the actions *but also on the objectives* which one deems permissible. The constraint is akin to an extended biblical order that you should do unto others (and wish for them) as you would have done to (and wished for) 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).

In this vein, reason is instead supposed to guide the ends we pursue. In other words, to return to the case of the stolen toy, reason might help the child to see that it *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, 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.

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.

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 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. In fact, otherwise civilised Europeans did not automatically accepted the humanity of slaves (or women for that matter) until well into the 19th century.

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 smokers' systematically biased views of the dangers of smoking, or workers in risky occupations who similarly underestimate the risks of their jobs. Indeed, in a modified form, we are all familiar with a problem of consumer choice when it seems impossible to decide between different brands. We consult consumer reports, specialist magazines and the like and it does not help because all this extra information only reveals how uncertain we are about what we want.

The problem is we do not know whether safety features of a car, for instance, matter to us more than looks, or speed, or cost. And when we choose one rather than another we are in part choosing to make, say, 'safety' one of our 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 so-called Austrian School's critiques of the socialist central planning system: according to the Austrians (but also other critics) planning can never substitute for the market because it *presupposes* information regarding preferences which is in part created in markets when consumers choose.[5]

Of course, having used the fluidity of preference against socialist alternatives to the market, the Austrian School and other neo-liberals thinkers have no alternative but to revert to the model of persons who, ultimately, like what they do and do what they like; the instrumentally rational *Homo Economicus* in other words. Just as Kantian logic is hard to escape once one thinks in terms of categorical imperatives (i.e. it is *right* for all of us, and thus for myself, to do X_i), the Humean model is the natural destination of hypothetical imperatives (i.e. *If* I expect others do Y_i, my best response is to do X_j). While the Humean agent's reasons for action are utterly 'internal', the Kantian acts on purely external reasons.

The beauty of these extreme formulations of human agency (Humean and Kantian) is that, in their respective contexts, rationality can be defined *a priori*. Their downfall, however, is that it may be a bad idea to try to define rationality (or, indeed, freedom) *a priori*. For such definitions, despite satisfying a primal urge to know what it means to be rational (and free), may cause us to lose sight of other important dimensions of our social location and individual character; for instance once rationality is defined in either the Humean or Kantian manner, it becomes impossible even to conceive of the notion of *solidarity* (see Varoufakis, 2002 and Arnsperger and Varoufakis, 2003).

In reading the following chapters on game theory, the reader should remain fully aware of the alacrity with which game theory sweeps the above issues under the carpet, opting with little concern for the neoclassical variant of Hume: the person as a utility function maximiser. This is important not least because we shall soon find that game theory lands in explanatory trouble. At that point it will be important to consider the possibility that the root cause of its problems may not be unconnected to the model of persons at its foundations.

(c) The source of beliefs

You will recall in the example contained in Game 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 she will want to walk; and you will walk if you think she will walk. So what she thinks you think will in fact influence what 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 1.7 below.

Box 1.7a

BAYES'S RULE: 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, however, 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 the test proves positive when you have X, Pr(test is positive|X) = 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

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \times \Pr(A)}{\Pr(B \mid A) \times \Pr(A) + \Pr(B \mid not\ A) \times \Pr(not\ A)}$$ where 'not A' means that event A did not occur

To see how it applies in our example, think of A as event: 'You have X' and 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)×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 discursively; a touch above 9%.

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 (see also the cautionary note in Box 1.7b). We focus here on a slightly different problem. Thomas Bayes provided a rule for *updating* our expectations. But where do our *original* (or *prior*) expectations come from? Or to put the question in a different way: In the absence of evidence, how do agents form initial probability assessments governing events like the behaviour of others?

Box 1.7b

BAYES RULE: THE DECISION TO 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, 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

$$\Pr(G \mid C) = \frac{\Pr(C \mid G) \times \Pr(G)}{\Pr(C \mid G) \times \Pr(G) + \Pr(C \mid NG) \times \Pr(NG)}$$

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.

Caution: 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 what mathematicians refer to as stationary probability distributions. This means that, in this example, 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 cannot 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). Therefore, 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.

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 in utility terms the same amount as the amount of utility 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 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 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 she can expect on average from a little more search (i.e. the expected marginal benefits) because 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 1.8 for some experimental evidence which casts doubt on the consistency of such subjective assessments and more generally on the probabilistic representations of uncertainty).

Box 1.8

THE ELLSBERG PARADOX, UNCERTAINTY, PROBABILITY ASSESSMENTS, AND CONFIDENCE

Suppose an urn contains 90 balls and you are told that 30 are red and that the remaining 60 balls are either black or yellow. However, you are *not* told how many of the 60 black or yellow balls are actually black or yellow. Indeed, they may all be yellow, all black or any combination of black and 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*, she reveals that 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* she knows the exact probability of winning $100: it is 30%. By contrast, were she to choose *Option II*, the probability of winning would have been unknown (since the proportion of black balls is unknown). Now look again at *Options III* and *IV*. By choosing *Option IV* one knows the *exact* probability of winning: 60%. On the other hand, the probability of winning $100 when choosing *Option III* is ambiguous (as the proportion of red *and* yellow balls is unknown). 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, 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).

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 and we turn to it now.

1.2.2 Common knowledge of rationality (CKR)

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.

The common knowledge assumption 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 as follows:

(a) each person is instrumentally rational
(b) each person knows (a)
(c) each person knows (b)
(d) 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 seemed like a promising assumption suddenly looks 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 too (a kind of common knowledge, like 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. It is to counter this type of problem that game theorists often (but not always, see Bernheim, 1984 and Pierce, 1984) make a further assumption concerning how rational players will form beliefs.

1.2.3 Common priors

This assumption holds that rational agents will draw the same inferences on how a game is to be played. The actual term 'common priors' is a reference back to the question regarding where the 'prior' probability estimates come from in Bayes's rule and this assumption holds that, whatever these prior estimates are, rational agents will share the same view of what they are. The assumption entails what we refer to in this book as the *consistent alignment* of people's *beliefs*. This alignment is the hallmark of the most influential solution concept in game theory, the Nash equilibrium (see Chapter 2). 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 connection with the 'common priors' assumption is this: If you are rational, *and* know the other person is rational *and*, courtesy of this assumption, you know your thoughts about what your rational opponent might be doing have the same origin (and will guide you to thoughts on the same lines as her own thoughts), then their action should never surprise you. So your beliefs about what your opponent will do are consistently aligned in the sense that, if you actually knew what her plans were, you would not want to change your beliefs about those plans. And if she knew your plans she would not want to change the beliefs she holds about you and which support her own planned actions. (Note that this does not mean that everything can be accurately predicted. For example, if you observe rain when sunshine was expected with probability 3/4, you should not be surprised since there was always a good chance (1/4) of foul weather. You may be disappointed, but you are not surprised!)

This assumption is usually justified by an appeal to the so-called *Harsanyi-Aumann 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. Robert Aumann defended this position staunchly and thus the naming of the said doctrine. 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 how rational people will dress.

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 1.9 for the complete argument.) One can almost discern a dialectical argument here where, 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 plus the Harsanyi-Aumann doctrine spawns common priors and CAB.

Box 1.9

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

Suppose you believe that the probability of rain tomorrow is ¾. And suppose that I believe it to be ¼. 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 will always sum to zero), this is not so with our expected *ex ante* pay-offs. Each one of us expects payoffs: $1 with probability ¾ and -$1 with probability ¼. On average, each expects to make 50 cents [$1×¾ - $1×¼ = 50 cents]. Thus our expectations are inconsistent with each other. For 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. At that point of convergence, we shall share the same beliefs and neither will be prepared to bet against the other. Thus, according to Aumann, rational and identically informed agents cannot agree to disagree.

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 part of the argument (due to John Harsanyi). 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 invariably to co-ordinate their look 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 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 Hegel, outside of human history there are *no* answers to the question of what one's reason demands of other people's reason. 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,1989, 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. In fact, Hume was pessimistic about the powers of reason to acquire knowledge and this pessimism can be seen clearly in his (empiricist) view that the best we can do in a our effort to understand the world is to observe some empirical regularities using our sensory devices.

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.

1.2.4 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 3 gives a full account of it, 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 to convert the game into one of *complete information* is that you hold common prior expectations with your opponent (the Harsanyi/Aumann doctrine) about the likelihood that your opponent will turn out to be one type of player or another.

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. Each player thinks of the game as one played against some opponent who has been drawn *as if* by some lottery from a perfectly known variety (or distribution) of players. 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, according to game theory, 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 your pay-offs. 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 *is* of type A in the first place.

The difficulty here, 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, another 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 just ways of increasing the team's score. 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 as well as the material in this book's final chapter. 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. As we shall see in Chapter 7, the more ambitious game theory becomes, the less able it is to avoid these philosophical 'complications'.

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 aboriginal *asocial* 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 *asocial* 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 Thomas Hobbes, the early English philosopher, who suggested, in *Leviathan*, that individuals would contract with each other to form a State *out of fear of each other*. 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 'nasty, brutish and short'.

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 Jean-Jacques Rousseau.

Hobbes has the State emerging from a contract between individuals because it serves the pre-existing 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.[6]

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 bring about institutions or, at the very least, 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 (i.e. *multiple equilibria*, in game theoretical language). 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. The methodological individualist's sole commitment is to 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'.

First, 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 an affirmative answer to '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. Although tempted, 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, according to the New Right wing of liberalism, we do not need to create a collective action agency like the State to escape from Hobbes's nightmare; and again game theory is well placed to examine this claim through the study of dynamic (or repeated) games.

1.4 A GUIDE TO THE REST OF THE BOOK

1.4.1 Three classic games: Hawk-Dove, Co-ordination and the Prisoners' Dilemma

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 co-operation: it is called *Hawk-Dove* or *Chicken*. For instance, two people, Jack and Jill, come across a $100 note on the pavement and each has a basic choice between demanding the lion's share (playing *hawkishly* or '*h*') or acquiescing in the other person taking the lion's share ('playing *dove*' or '*d*'). 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 Game 1.2; the pay-off to the row player, Jill, is the first sum and the pay-off to the column player, Jack , is the second sum.

	h	*d*
h	0,0	90,10
d	10,90	50,50

Game 1.2 – *Hawk-Dove* or *Chicken*

Plainly both parties will benefit if they can avoid simultaneous hawk-like behaviour, so there are gains from some sort of co-operation. On the other hand, the motive to act aggressively is powerful and this gives rise to conflict. To see this, note that, as long as each wants to maximise their dollar take, if Jill expects that Jack will play *d* she has a strong incentive to play *h* as doing so will net her pay-out $90 pay-out. But this also applies in reverse: if Jack expects Jill to be acquiescent (play *d*), he has every reason to be aggressive (play *h*). The interesting questions are: Do the players avoid the fight, and if they do how is the $100 divided? Equally or asymmetrically? (We shall return to this question when analysing Game 2.16 in Chapter 2.)

To illustrate a *co-ordination game*, suppose in our earlier example of your attempt to walk to work along with a friend (Game 1.1) that your friend has similar preferences and is trying to make a similar decision. Thus Game 1.3 represents the joint decision problem.

	Drive	*Walk*
Drive	1,1	1,0
Walk	0,1	2,2

Game 1.3 – *Co-ordination between friends going to work*

Will the two of you succeed in co-ordinating your decisions and, if you do, will you walk together or drive separately? (In Chapter 2 we delve into the co-ordination problem in the context of Games 2.17 and 2.18).

Finally, there is the *prisoners' dilemma* game (to which we have dedicated the whole of Chapter 5). Recall the time when there were still two superpowers each of which would like to dominate the other, if possible. They each faced a choice between arming and disarming. When both arm or both disarm, neither is able to dominate the other. Since arming is costly, when both decide to arm this is plainly worse than when both decide to disarm. However, since we have assumed each would like to dominate the other, it is possible that the best outcome for each party is when that party arms and the other disarms since, although this is costly for the arming side, it enables it to dominate the other. These preferences are reflected in the utility pay-offs depicted in Game 1.4. E.g. the both players highest utility corresponds to outcome 'I arm when the other disarms', the lowest to the outcome 'I disarm when the other arms', the second lowest to outcome 'we both arm' and the second highest to outcome 'we both disarm'.

	disarm	*arm*
disarm	2,2	0,3
arm	3,0	1,1

Game 1.4 – *The Prisoner's Dilemma*

Game theory makes a rather stark prediction in this game: both players will arm (the reasons will be given later). It is a paradoxical result because the two antagonists are facing the same dilemma and, given their identical preferences and rationality, will come to the same decision. In terms of the above matrix, they will end up on one of the two outcomes on the diagonal: (2,2) or (1,1). Moreover, they know all this (as they are presumed fully rational). Should they not choose the former over the latter? They should. But (according to game theory) they won't!

The prisoner's dilemma turned out to be one of game theory's great advertisements. The elucidation of this paradox, and the demonstration of how each player brings about a collectively self-defeating outcome, because she is *rationally* pursuing her own interest, was one of game theory's early achievements which established its reputation among the social scientists. The prevalence of this type of interaction between, not only states but, also, among individuals in every day circumstances, together with the inference that rational players will fail to serve their interests when left to their own devices (e.g. the two states in Game 1.4 will fail to bring about mutually advantageous disarmament), has provided one of the strongest arguments for the creation of a State. This is, in effect, Thomas Hobbes's argument in *Leviathan*. And since our players here are themselves States, both countries should agree to submit to the authority of a higher State which will enforce an agreement to disarm. Does this translate into an argument for a strong, independent, United Nations? It does. However, the democratically minded reader should beware that it translates *equally* into an argument for a global tyranny!

1.4.2 Chapter-by-chapter guide

The next two chapters set out the key elements of game theory. For the most part the discussion here relates to games in the abstract. There are few concrete examples of the sort that 'Jack and Jill must decide how to fill a pail of water'. Our aim is to introduce the central organising ideas as simply and as clearly as possible so that we can draw out the sometimes controversial way in which game theory applies abstract reasoning.

Chapter 2 introduces the basics: the different kinds of strategy available to players (including ones that appear to involve choosing actions randomly, so-called 'mixed strategies'); the logical weeding out of strategies because they are not compatible with instrumental rationality and common knowledge of this rationality (this is called using 'dominance reasoning'); and the selection of strategies using the most famous solution concept for games, the *Nash equilibrium*. John Nash developed the latter in the early 1950s and he has become more familiar since the first edition of this book as a result of the Oscar winning film based on his life, *A Beautiful Mind*. This solution concept has proved central to game theory and, thus, we discuss its meaning and uses extensively. Much attention is also paid to the critical aspects of its use. In particular, we take up some of the issues foreshadowed in Sections 1.2.1 and 1.2.2 above.

Chapter 3 focuses on the *refinements of the Nash equilibrium concept* that have been proposed in order to combat what otherwise seems like indeterminacy (i.e. when there are multiple Nash equilibria). Many of these refinements apply to games with an explicit dynamic structure where players take decisions sequentially and we include a discussion of various concepts with puzzling names such as *subgame perfection*, *sequential equilibria*, *proper equilibria* and the ideas of *backward* and *forward induction*. The idea of 'trembles' (small errors in the execution of the chosen strategy) is crucial to the coherent construction of most of these Nash refinements and we spend much space discussing whether, once the idea of such 'trembles' is admitted, people might 'rationally' decide to 'tremble' in particular ways. The chapter concludes with an assessment of the Nash equilibrium and its refinements.

Chapter 4 is devoted to the analysis of bargaining games. These are games which have a structure which is similar to the *Hawk-Dove* (or chicken) game above. Somewhat confusingly, John Nash proposed a particular solution for this type of game which has nothing to do with his earlier equilibrium concept (although this solution does emerge as a Nash equilibrium in the bargaining game). So be warned: the *Nash solution* to a bargaining game in Chapter 4 is *not* the same thing as the *Nash equilibrium* concept in Chapters 2 and 3. Much of the most recent work on this type of game has taken place in the context of an explicit dynamic version of the interaction and so Chapter 4 also provides some immediate concrete illustrations of the ideas from Chapter 3.

Chapter 4 also introduces the distinction between *co-operative* and *non-cooperative* game theory. The distinction relates to whether agreements made between players are binding. *Co-operative game theory* assumes that such agreements are binding, whereas *non-cooperative game theory* does not. For the most part the distinction is waning because most sophisticated *co-operative game theory* is now based on a series of *non-cooperative* games for the simple reason that, if we want to assume binding agreements, we shall want to know what makes such agreements binding and this will require a *non-cooperative* approach. In line with this trend, and apart from the discussion contained in Chapter 4, this book is concerned only with *non-cooperative game theory*.

Chapter 5 focuses exclusively on the *Prisoners' Dilemma*. This game has fascinated social scientists both because it seems to be ubiquitous and because of its paradoxical conclusion that rational people, when acting apparently in their own best interest, actually produce a collectively inferior outcome to one which is available. We consider a variety of attempts to explain how people might rationally overcome this dilemma, ranging from Kant's conception of rationality through weaker forms of moral agency, to the idea that we can overcome the dilemma by choosing *dispositions*. We also include a discussion of how the dilemma might be overcome once the game is repeated. This is where this book deals formally with the topic of repeated games.

Repetition makes a difference because it enables people to develop much more complicated strategies. For example, there is the scope for punishing players for what they have done in the past and there are opportunities for developing reputations for playing the game in a particular way. These are much richer types of behaviour than are possible in one-shot games and it is tempting to think that these repeated games provide a model for historical explanation. In fact, the richness of play comes at a price: almost anything can happen in these repeated games! In other words, repeated games pose even more sharply the earlier problem of indeterminacy due to multiple Nash equilibria; that is, of knowing what is (rationally) possible when almost anything goes. Finally, as suggested earlier, the analysis of this game has been central to discussions in liberal political theory of the State and so we draw out the implications of this game for political philosophy.

Chapter 6 is concerned with the evolutionary approach to repeated games. This approach potentially both provides an answer to the question of how actual historical outcomes come into being (when all sorts of outcomes could have occurred) and it circumvents some of the earlier doubts expressed in Chapters 2, 3 and 4. It does this by moving away from instrumental reason, concentrating instead on the adaptations, or evolution, of behaviour in various social contexts. The analysis of evolutionary games is particularly useful in assessing the claims in liberal theory regarding *spontaneous order*. It also allows a more general assessment of the kind of evolutionary arguments that have become commonplace. The chapter ends with allusions to the origin of morality and the discriminatory norms along class, race and gender lines.

Chapter 7 is devoted to psychological games. These are the games analysed once we abandon the strict separation of preference and belief (which neoclassical economists often, mistakenly, assume to be part of one's rationality). We start the chapter by looking at what happens when our view of what others expect of us feeds directly into what we wish for. Later, we extend this analysis to situations in which norms of behaviour affect beliefs which, in turn, affect preferences.

In this way the agents are importantly socially located before they interact in games and this supplies the link to some of Wittgenstein's arguments already mentioned above. There is a history to such models of action in economics that goes back, at least, to some of Adam Smith's ideas on the role played by sympathy (or feelings of fellowship) in motivating action (see Sugden, 2002, and Arnsperger and Varoufakis, 2003), but we focus here on two modern variants that have been explicitly used in the analysis of games. One is associated with Rabin (1993), who has built on the model of Geanakoplos et al (1989) which was also discussed, albeit briefly, in our first edition. The other is primarily associated in game theory with Bacaharach (1999) and Sugden (2000a) and turns on the idea of norm-driven preferences.

Throughout, we have tried to provide the reader with a helpful mix of pure, simple game theory and of a commentary which would appeal to the social scientist. In some chapters the mix is more heavily biased towards the technical exposition (e.g. Chapters 2 and 3). In others we have emphasised those matters which will appeal mostly to those who are keen to investigate the implications of game theory for social theory (e.g. Chapters 4 -7).

1.5 CONCLUSION

We first noticed the spreading of game theory into popular culture while watching a scene in a BBC prime time drama series. A smiling police inspector told a sergeant: "That puts them in a prisoner's dilemma." The sergeant asked what "this dilemma" was and the inspector replied as he walked off: "Oh, it's something from this new [sic] theory of games." The inspector may not have thought it worth his while, or the sergeant's, to explain this "theory of games", but it is surely significant that game theory was featured as part of the vocabulary of a popular television drama. Since then the huge commercial success of *A Beautiful Mind*, and the touching efforts of screen-writers and director to convey to mass audiences the gist of Nash's equilibrium, has brought game theory even deeper into popular culture.

In an assessment of game theory, Tullock (1992) has remarked somewhat similarly that "…game theory has been important in that it has affected our vocabulary and our methods of thinking about certain problems." Of course, he was thinking of the vocabulary of the social scientist. However, the observation is even more telling when the same theory also enters into a popular vocabulary, as it seems to have done. As a result, the need to understand what that theory tells us "about certain problems" becomes all the more pressing. In short, we need to understand what game theory says, if for no other reason than that many people are thinking about the world in that way and using it to shape their actions.

We hope to have written a book that does this, and more. Thus we hope to persuade the reader that game theory is useful not just for thinking about a range of interactions and for assessing debates in liberal political theory, but also for revealing the problems with the popular model of instrumental rationality. This task was accomplished by our first edition, but game theory is good for more than critique now: it has become an important site (perhaps the only one) where economic orthodoxy is subverted and richer models of rational agency are being developed (see Chapters 6 and 7). In other words, it has begun to exhibit what Hegel referred to as the *cunning of reason*; and this is why we wrote a second edition.

NOTES

[1] A good source of constantly updated information on game theory texts is the website of the *Game Theory Society*. See www.gametheory.net

[2] The ontological question addresses the essence of *what is* (its etymology comes from the Greek *onta* which is plural for being).

[3] An epistemological question (*episteme* meaning the knowledge acquired through engagement) asks about *what is known* or about *what can be known*.

[4] In fact, some economists prefer to talk solely about 'consistent' choice rather than acting to satisfy best one's preferences. The difficulty with such an approach is to know what sense of rational motivation, if it not instrumental, leads agents to behave in this 'consistent' manner. In other words, the obvious motivating reason for acting consistently is that one has objectives/preferences which one would like to see realized/satisfied. In which case, the gloss of 'consistent' choice still rests on an instrumentally rational motivating psychology.

[5] Note, however, that the same criticism applies for neoclassical economic theory which, too, takes preferences as data and seeks out equilibrium prices and quantities that correspond to that data.

[6] You will notice how the Rousseau version not only blurs the contribution of the individual by making the process of institution building transformative, it also breaches the strict separation between action and structure. In fact, this difference also lies at the heart of one of the great cleavages in Enlightenment thinking regarding liberty (see Berlin,

1958). The strict separation of action and structure sits comfortably with the negative sense of freedom (which focuses on the absence of restraint in pursuit of individual objectives) while the fusion is the natural companion for the positive sense of freedom (which is concerned with the ability of individuals to choose their objectives autonomously).

Chapter 2: THE ELEMENTS OF GAME THEORY

CONTENTS

2.1 INTRODUCTION
2.2 THE REPRESENTATION OF STRATEGIES, GAMES AND INFORMATION SETS
    2.2.1 Pure and mixed strategies
    2.2.2 The Normal Form, the Extensive Form and the Information Set
2.3 DOMINANCE REASONING
    2.3.1 Strict and weak dominance
    2.3.2 Degrees of common knowledge of instrumental rationality
2.4 RATIONALISABLE BELIEFS AND ACTIONS
    2.4.1 The successive elimination of strategically inferior moves
    2.4.2 Rationalisable strategies and their connection with Nash's equilibrium
2.5 NASH EQUILIBRIUM
    2.5.1 John Nash's beautiful idea
    2.5.2 Consistently aligned beliefs, the hidden Principle of Rational Determinacy and the Harsanyi-Aumann Doctrine
    2.5.3 Some Objections to Nash: Part I
2.6 NASH EQUILIBRIUM IN MIXED STRATEGIES
    2.6.1 The scope and derivation of Nash equilibria in mixed strategies
    2.6.2 The reliance of NEMS on CAB and the *Harsanyi doctrine*
    2.6.3 Aumann's defence of CAB and NEMS
2.7 CONCLUSION

2
THE ELEMENTS OF GAME THEORY

2.1 INTRODUCTION

When game theorists refer to a game's 'solution', or a 'solution concept', they have in mind how rational people might play the game. A 'rational solution' corresponds to a set of strategies (one for each player) that the players of the game have no (rational) reason to regret choosing. It is in this sense that a game's 'solution' is often used interchangeably with the term 'equilibrium'. The most important solution concept in game theory is the *Nash equilibrium*. We shall see later how Nash's most influential 'solution' is based on the idea that, in equilibrium, players' strategies must be best replies to one another.

Best Reply Strategies (Definition)
A strategy for player R, R_i, is *best reply* (or *best response*) to C's strategy C_j if it gives R the largest pay-off given that C has played C_j. (Similarly for player C.)

In Game 2.1, we have indicated R's best reply to each of C's strategies with a $(^{+})$ sign and likewise C's best response to each of R's strategies with a $(^{-})$ sign. The virtue of the $(^{+})$ and $(^{-})$ markings is that they help us spot immediately the player's best replies. So, in Game 2.1 the presence of two $(^{+})$ markings in row R1 help us conclude, without much thought, that strategy R1 is a best reply to either of C's possible strategic choices (C1 *and* C2). By contrast, the presence of a $(^{-})$ sign in column C1 and another one in column C2 tells us that player C does *not* possess a single best reply. In fact, C1 is his best reply to R2. And C2 is his best reply to R1. [Note how the $(^{-})$ corresponding to R1 is in column C2 and that corresponding to R2 lies in column C1).]

Once the reader becomes familiar with the identification of the players' best replies to each strategy of their opponent, the next step is to notice an interesting coincidence: There is a cell in which the $(^{+})$ and $(^{-})$ markings actually coincide. It is the utility outcome (1,5) which comes about when R plays R1 and C plays C2. Whenever such a coincidence occurs, we have a *Nash equilibrium*.

	C1	C2
R1	$^{+}10,4$	$^{+}1,5^{-}$
R2	$9,9^{-}$	0,3

The $(^{+},^{-})$ marks next to pay-offs indicate 'best reply' strategies

Game 2.1

The Nash equilibrium (preliminary definition)
The outcome of strategies R_i for R and C_j for C is a *Nash equilibrium* of the game, and thus a potential 'solution', if R_i is the best reply to C_j and, at once, C_j is the best reply to R_i. If R_i and C_j are, indeed, best replies to one another, then their adoption fully justifies the beliefs of each player which led to that adoption. That is, were R to be interrogated along the lines: "Why did you choose R_i?" her credible response would be: "Because I was expecting C to choose C_j". Thus C's choice of C_j would validate R's thinking. Moreover, since C_j is also the best reply to R_i, C's own reasons for playing C_j will be automatically vindicated by the observation that R chose R_i.

In this chapter we explore the assumptions that underpin Nash's equilibrium concept. These assumptions are controversial and, while the Nash equilibrium is absolutely central to game theory, it is not without its critics. There are other approaches to solving games and we begin with a less demanding one that turns on 'dominance reasoning' in Section 2.3 (also see Box 2.1 for another). In some games, when players are rational and there is *common knowledge of rationality*, this type of reasoning actually predicts a Nash equilibrium. But as we see in Section 2.4, this is not always the case. In general, a further assumption is required, the *consistent alignment of beliefs*, and this is discussed in Section 2.5.

One of the attractions of the Nash equilibrium is that every (finite) game has (at least) one Nash equilibrium. So, it can be applied as a solution concept to all games. Since some of these equilibria involve a particular kind of strategy, a so-called *mixed strategy*, that some people find rather strange, we spend some time on these equilibria in Section 2.6. We offer a preliminary assessment of the Nash equilibrium concept in Section 2.5 and conclude with some further comments on the controversy which surrounds the Nash equilibrium solution concept in Section 2.7. Some groundwork is necessary, however, before we can begin all this and the next section sets out some of the technicalities of game theory with respect to how actions and interactions are represented.

Box 2.1.

JOHN VON NEUMANN'S MINIMAX THEOREM

John von Neumann, in a famous paper published (in German) in 1928, presented a solution applicable to a *wide class of games*. Consider the game below involving two players. Such interactions are known as *zero-sum games* because the sum of the two opponents' payoffs for each outcome equals zero. John von Neumann studied these games and proved that they can all be 'solved' in the same manner.

	C1	C2	C3
R1	-2,2	1,-1	10,-10
R2	-1,1	2,-2	0,0
R3	-8,8	0,0	-15,15

Von Neumann's proposed solution was based on the idea that each player should maximise the value of the worst possible outcome. So if R chooses R1, the worst outcome is payoff of –2, which occurs when C plays C1. Were she to choose R2, the worst outcome is –1, when C selects C1. Lastly, the worst outcome for R if she chooses R3 is -15 when C opts for C3. Of these 'bad' outcomes R's best one is the –1 which obtains when she chooses R1. This he called R's *maximin* strategy, i.e. the one which *maximises* her *minimum*. Thus R's *maximin* strategy is given by R1 with a payoff of –1. Investigating C's choices from the same perspective, we obtain strategy C1 as his *maximin* with a pay-off of +1.

Notice the sum of the two players' *maximin* values equals zero, as R's *maximin* equalled –1 and C's +1. John von Neumann's remarkable theorem is that this is so *in all zero-sum games:* that is, in every two-person, zero-sum interaction, the sum of the players' *maximin* payoffs equals zero. Another way of putting this theorem is that when each player selects their *maximin* strategies, the *maximin* pay-offs result. Why is this significant? It clearly is if we assume that both players are deeply pessimistic because they will then both select their *maximin* strategy.

The reason why von Neumann thought that players would rationally choose to be so pessimistic is related to his theorem above. Recall that in a zero-sum game your gain is your opponent's loss, and vice versa. So, being a pessimist does not mean necessarily that you are the sort of person fearing to step out of the house in fear of being hit by lighting; in the context of zero-sum games it may simply mean that you are a realist who is aware of the fact that your opponent will always try to inflict maximum damage on you. Why? Because your loss is her gain! Or to put this slightly differently, John von Neumann was convinced that no rational player would settle for anything less than her or his attainable *maximin* payoff; and if both were determined to avoid a payoff below their *maximin*, neither would *expect to get*, *try to get* or actually *get* more than their *maximin* payoff.

Notice also that the *maximin* solution is the same as the Nash equilibrium in this game (note the coincidence of best reply ($^{+}$) and ($^{-}$) sings in the same cell). This is always the case for zero-sum games, but not for non-zero-sum ones.

2.2 THE REPRESENTATION OF STRATEGIES, GAMES AND INFORMATION SETS

2.2.1 Pure and mixed strategies

'Games' are interactions in which the outcome is determined by the *combination* of each person's action. 'Strategies' are the plans for action.

The simplest kind of strategy selects unambiguously some specific course of action (also referred to as a 'move'); e.g. 'help an old person cross the road', or 'shoot an opponent'. This is called a *pure strategy*. However, there are times when you are uncertain about what is the best *pure strategy*. In these cases, you may choose *as if* at random between two or more pure strategies: e.g. in the absence of reliable meteorological information, you may decide on whether to carry an umbrella by tossing a coin. This type of strategy is called a *mixed strategy*, in the sense that you choose a specific 'probabilistic mix' of a set of pure strategies. In our trivial umbrella example, your *mixed* strategic choice can be thought of as 'Take umbrella with probability p= ½ and leave it behind with probability 1-p= ½.'

The idea of mixed strategies can seem a bit strange, so here are a couple of further examples where the decision maker is engaged in a genuine strategic interaction (as opposed to a 'game' against the weather). Suppose that Anne is about to go to a party and must choose between two pure strategies: *'Wear black shoes'* (***B***) or *'Wear red shoes'* (***R***). Let us also assume that she would have chosen ***B*** if she had reliable information that most of her friends would wear red shoes. If on the other hand her 'informers' warn her that her friends will be wearing mostly black shoes, she will opt for red shoes (in a bid to be different). What if, however, she lacks reliable information regarding the proportion of red-shoe-wearers that she will encounter at the party? In that case, she might as well randomise between the two by adopting the following *mixed strategy*: select pure strategy ***B*** with probability p and pure strategy ***R*** with probability 1-p.

Mixed and Pure Strategies (Definition)

If a player has N available pure strategies (S_1,S_2,...,S_N), a *mixed strategy* M is defined by the probabilities (p_1, p_2,..., p_N) with which each of her pure strategies will be selected. Note that for M to be well defined, each of the probabilities (p_1, p_2,..., p_N) must lie between 0 and 1 and they must sum to unity. Note also that to choose a *mixed strategy* (p_1=0, p_2=0...p_j=1..., p_N=0) is equivalent to choosing pure strategy S_j.

Anne opted for a *mixed strategy* (regarding her fashion statement) because of the *uncertainty about what others will do*. Another reason for opting for a *mixed strategy* is that *one may want to keep one's opponents guessing*. And if one wants *others* to be uncertain as to what one is about to do, perhaps the best way is to keep *oneself* equally uncertain as to what one will do. This is tantamount to acting *as if* by following a randomisation between pure strategies, i.e. a *mixed strategy*. Suppose, for instance, you are a striker about to take a penalty. The opposing goalkeeper would love to know whether you will shoot to his left or to his right. To keep him guessing, you may choose the *mixed strategy*: select pure strategy 'Shoot left' with probability 40%, pure strategy 'Shoot right' with probability 40%, and pure strategy 'Shoot straight on' with probability 20%.

2.2.2 The Normal Form, the Extensive Form and the Information Set

The two main representations of the way in which players' strategies interact are the *normal form* of a game and the *extensive form*.

The *normal form* of a game usually looks like a matrix (and is thus also known as the *matrix form* or the *strategic form*). It associates combinations of pure strategies with outcomes by means of a matrix showing each player's pay-offs (or preferences) for each combination of pure strategies – e.g. see Game 2.1, which is reproduced below. Since the matrix columns and rows are pure strategies, we shall, for the time being, concentrate on pure strategies.[1]

In this chapter the player choosing between the rows (or columns) will be labelled the row (or column) player, henceforth abbreviated as R (or C). R will be thought of as female and C as male. R's first strategy option is the first row denoted by R1, and so on. When R chooses R2 and C chooses C1, designated by (R2, C1), the outcome is given by the numbers in the cell defined by R2 and C1. In this example, R receives 9 utils and so does C. The first entry in any element of the pay-off matrix is R's utility pay-off while the second belongs to C. For instance, outcome (R2, C2) gives 3 utils to C and nothing to R.

	C1	C2
R1	10,4	1,5
R2	9,9	0,3

Game 2.1 – The *normal (or matrix) form* representation of a game

Note that the *normal form* says nothing about the process, or sequence, of the game and the implication is that players move simultaneously. When their choices are indeed simultaneous, the *normal form* is adequate. However, when one player acts before another, the *normal form* is incapable of conveying this strategically crucial piece of information and we need a different representation: the *extensive form*, also known as *dynamic form* or *tree-diagram form.*

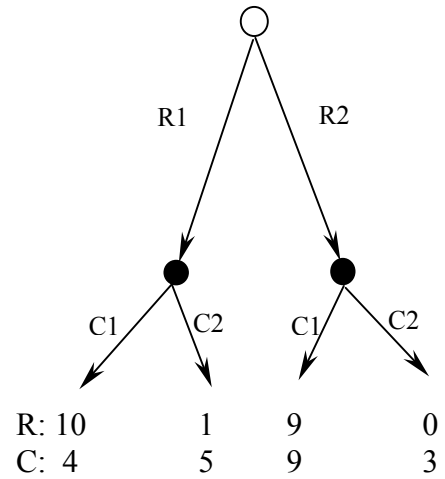
In Game 2.2 we represent two versions of the above game in an *extensive form*, one in which R moves first [see Game 2.2(a)], and one in which C takes the first step [see Game 2.2(b)]. Depending on the chosen path from one node to the other (nodes are represented with circles, with the initial node being the only one which is not full), we travel down a sequence of branches towards a final outcome at the bottom of the tree diagram. We follow the convention of representing the pay-offs of each player in the final outcome in the order in which they played. Thus in 2.2(a), R's pay-offs appear first and in 2.2(b) C's come first.

There is one marking on these diagrams to note well: the broken line in Game 2.2(b) that appears when R is called upon to choose. We have added it to Game 2.2(b) - and not in Game 2.2.(a) - in a bid to introduce what game theorists refer to as a player's *information set*. Just before a player makes a choice, all the information at his or her disposal is contained in that player's *information set*. Before we predict what players will do, it is important to know what they know; to have specified their *information set* fully. Had Oedipus known that the King he chanced upon at that ill-fated junction was his father, there would have been little drama in Thebes. Thus the contents of a player's information set matters a great deal.

In *normal form* games, e.g. Game 2.1, players are fully informed about the structure of the game as long as they know the matrix. However, when they move in sequence there is more to know; e.g. if R moves first, C may or may not have observed R's choice before being called upon to make his own choice. Note that if C knows whether R selected strategy R1 or R2, C knows which branch of Game 2.2 he is at before moving. If not, he could be at either branch.

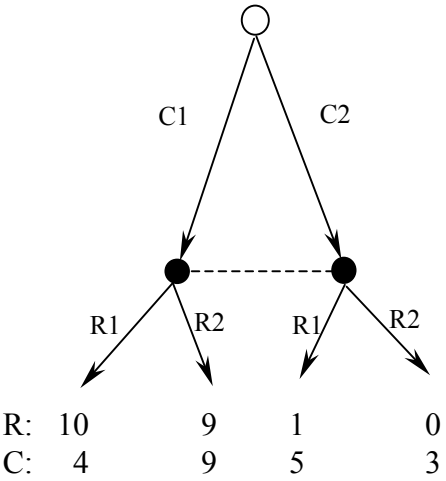
Returning to the broken line in Game 2.2(b), it means that C has moved first and has determined on which of the two branches the game will progress but R does not know, prior to choosing between R1 and R2, which it is. As far as C is concerned, he could be at either the right or the left decision node and this fact is captured by the broken line.

In summary, when in an *extensive form* game we see that a broken line joins two or more of a player's nodes, this means that the said player must choose without knowing which of those nodes represents her or his *current* position in the game tree.



Game 2.2a

The game in extensive form with R moving first (choosing between R1 and R2), C observing R's choice and then selecting either C1 or C2.



Game 2.2b

The game in extensive form with C moving first (choosing between C1 and C2), and then R selecting either C1 or C2 in ignorance regarding C's earlier choice (recall the meaning of the broken line connecting C's nodes).

Finally, it is worth noting (for more complicated games) that there is a convention when drawing a game in *extensive form* never to allow branches of the tree diagram to loop back into one another. In other words, the sequence of decisions is always drawn in an 'arboresque' or tree-like manner with branches growing out (and *not* into one another). So even when an individual faces the *same* choice (between, say, being 'nice' to R or not) after several possible sequences of previous choices by the players, the choice must be separately identified for each of the possible sequences leading up to it. The point is that, even when people face the same choice more than once within a sequence of choices, we wish to distinguish them when they have different histories and this is what the prohibition on looping back ensures.

***Logical Time* versus *Historical Time*, and *Static* versus *Dynamic games* (definitions)**
When games are discussed in their *normal form* (or equivalently, their *matrix* or *strategic form*) there is a presumption that players choose among their strategies *once*, *simultaneously* and *without communicating* with one another. Since they play once and observe their opponent's choice only after the game is well and truly over, there is no sense in which the game can be thought of as dynamic (that is, as one unfolding over time and involving trial-and-error or any other type of learning). In this context, time is ticking away only while players are thinking about their one and only choice; e.g. a player ponders "If my opponent does *X* I should do *Y*; but then again, will he not know that this is what I am about to do? Probably. So, perhaps it is best to do *Z*..." This type of pre-play thinking is not part of a dynamic game; of a feedback loop from

observations to beliefs to actions and back to observations. In the absence of such a dynamic process, we say that the (static) game unfolds in *logical time*. Juxtaposed against *logical time* we have *historical time* (or *real time*) in which actions, as well as thoughts, occur and feed off each other.

In the remainder of this chapter we concentrate on *one-shot* games (that is, static games in which players choose only once and simultaneously. The interactions are represented in *normal form* and all thinking unfolds in *logical time*.

2.3 DOMINANCE REASONING

2.3.1 Strict and weak dominance

We call moves which are best replies to anything the opposition might do *strictly dominant strategies*. Conversely, strategies resulting in reliably inferior outcomes compared to some other strategy (e.g. R2 in Game 2.1) are known as *strictly dominated strategies*. In Game 2.1, player R will do better playing R1 *regardless of C's choice*: it is strictly dominant and R2 is strictly dominated. Player C, however, has no *dominant strategy* in Game 2.1. C1 is his best response to R2 and C2 his best response to R1.

Strictly dominant/dominated strategies (definition)
A strategy is *strictly dominant* if it guarantees a player *higher* payoffs than she or he would receive playing any other strategy, against all possible strategies of the opposition. Reversing this definition, a strategy is *strictly dominated* if it guarantees a player *lower* payoffs than she or he would receive playing some other strategy, against all possible strategies of the opposition.

The presence of *strictly dominant* strategies can contribute to a straightforward 'solution' in some games. For example, in Game 2.1 we know that R1 will be played by an instrumentally rational person because it is strictly dominant and, although there is no analogous strictly dominant strategy for C, once C recognises that R will play R1 then his best response is C2. Thus if R is rational, C knows this, and C is rational himself, then on the basis of dominance reasoning we predict the solution for this game will be (R1, C2). In other words, if R and C are instrumentally rational and there is *common knowledge* of this rationality, then the solution to this game is (R1, C2).

The rationalisation in this way of a game's equilibrium (when that is possible) can require different depths of mutual knowledge depending on the game's structure. In a game of the *Prisoner's Dilemma* type (see Game 2.3 but also 2.18), no common knowledge of (instrumental) rationality is necessary because each player has a strictly dominant strategy.[2] As long as players *are* instrumentally rational, they will recognise their strictly dominant strategies and play them *whatever their thoughts concerning their opponent's rationality*.

	C1	C2
R1	$10,10^{-}$	$^{-}5,20$
R2	$^{+}20,-5^{-}$	$^{+}0,0^{-}$

Game 2.3 – *A game with the structure of the prisoner's dilemma* (see Chapters 1 and 5)

We define common knowledge of rationality as follows:

N^{th}-order Common Knowledge of (instrumental) Rationality or CKR (definition)

Zero-order CKR describes a situation in which players are instrumentally rational but they know nothing about each other's rationality. By contrast, *1^{st}-order CKR* means that not only are they instrumentally rational but also each believes that the other is rational. It follows that, if n is an even number, N^{th}-order CKR conveys the following sentence: 'R believes that C believes that R that . . . C believes that R is instrumentally rational'; a sentence containing the verb 'believes' N times. When n is odd, then n^{th}-order CKR means: 'R believes that C believes that R that . . . that C is instrumentally rational'; also a sentence containing the verb 'believes' N times. When game theorists refer to CKR without mentioning any specific order, the implication is that N tends to infinity and that 'one player believes that the others believe that one believes... that all believe that each is instrumentally rational *ad infinitum*'; that is, we are asked to imagine a sentence of infinite length which, naturally, contains the verb 'believe(s)' an infinite number of times.

Thus a solution is possible in Game 2.1 through dominance reasoning provided there is 1^{st} order CKR; that is, once C knows that R is rational, he is convinced that she will choose R1, in which case his best reply is C2. In contrast, zero-order CKR suffices for Game 2.3 since neither player needs to know anything about the other in order to work out their best strategies (this is, of course so, courtesy of the presence of dominant strategies). Boxes 2.2 and 2.3 supply further illustrations of how dominance reasoning leads players to a unique course of action. You will notice that the actions identified through dominance reasoning in Games 2.1 and 2.3 corresponded to the Nash equilibria of those games. In the next section we explore whether this is always the case.

Box 2.2

TRUTHFUL BIDDING IN SEALED-BID 2^{ND} PRICE AUCTIONS IS A DOMINANT STRATEGY

Suppose there are N bidders for a house and each of them submits a sealed bid to an auctioneer. The highest bidder wins but pays a price equal to the 2^{nd} highest bid. It is easy to prove that *bidding one's true valuation is a dominant strategy*: Suppose you value a house at V. If, on winning, you had to pay a price equal to your (sealed) bid, you might rightly worry that the 2^{nd} highest bid might be considerably less than V, in which case it would make sense to bid less than V. However, if you know that you will not have to pay more than the 2^{nd} highest bid, you have no reason to bid less than *V regardless of your expectations regarding your opponents' bids*. Thus, truthful bidding is a *dominant strategy*.

Note that this idea has applications in other situations, e.g. when a group of people are trying to decide how much each will contribute to some public good but all have an incentive to understate their private valuation of it. Truthfulness is then encouraged when individuals make sealed offers but are only obliged to contribute less than that (e.g. donate the second largest offer).

Box 2.3

DOMINANT STRATEGIES AND THE TRAGEDY OF THE COMMONS

Consider a game involving $N(>3)$ players and a common asset. Suppose further that players, independently of one another and only once, attempt to appropriate chunks of that public good for private use. Let us normalise the individual's 'greed' restricting X to the range [0,1]; where X=0 means that she has abstained altogether from grabbing parts of the common asset, and X=1 that she has grabbed the most that is possible for a single individual to grab. Lastly, let the payoffs of player i be given, for all i=1,...,N, by : $P_i = 1-3\mu+2X_i$, where μ is the average choice of X in the population of N players.

The idea here is that the greedier the players (i.e. the closer μ is to 1) the greater the depletion of the public good and the less is left for all, and each, to enjoy. Note that the payoffs are normalised so that when the public good is intact each person enjoys 1 unit of it (that is, if everyone chooses X=0, each receives P_i=1). Their tragedy (often referred to as the *tragedy of the commons*) is that each has a pressing private reason to set X_i=1! What is this reason? It is that X_i=1 is a best reply to all values of μ; i.e. X_i=1 is a *strictly dominant* strategy.

Proof: Suppose i were to predict that each of the others would choose X=0 (that is, be abstemious). Then i's payoff will equal 1 if she also sets her X_i=0 and it will equal $1-3(1/N)+2$ if she sets X_i=1. But (since $N>3$) $1-3(1/N)+2 > 1$ and X_i=1 pays more than X_i=0. Is this a *strictly dominant* strategy? To see that it is, suppose i were to

predict that each of the others would be maximally greedy (as opposed to abstemious) and choose X=1. Then, were she to choose X_i=0, her own payoff would be $P_i = 1 - 3\left(\frac{N-1+0}{N}\right) + 2 \times 0 = 1 - 3\left(\frac{N-1}{N}\right)$. And if she were to choose X_i=1, she would collect $P_i = 1 - 3\left(\frac{N-1+1}{N}\right) + 2 \times 1 = 1 - 3 + 2 = 0$. Again (since N>3) the former P_i value is always negative and thus her best rely is X_i=1. Indeed, whatever beliefs she may entertain regarding the choices of others, i.e. whatever her estimate of μ, X_i=1 is her best reply; namely, her *strictly dominant strategy*. But since this is a dominant strategy for all players, each will select X=1, the average μ will become unity, the common asset will be depleted, and each player's payoffs will be zero.

We now turn to a weaker form of dominance reasoning. In Game 2.4 we find three cells in which our best response markings ($^+$) and ($^-$) coincide: (R1,C1), (R1,C2) and (R2,C1). Thus, all three constitute Nash equilibria in pure strategies. In fact, the only outcome *not* corresponding to a Nash equilibrium is (R2,C2). Can *dominance reasoning* shed some light on this situation? On inspection, we note that R1 and R2 are equally good replies to C1 [thus the presence of markings ($^+$) in both cells corresponding to C1]. This means that R will be *indifferent* between playing R1 or R2 when she expects C to play C1. However, R has a clear preference when she expects C to choose C2: her best reply is R1. In this light, R1 is a *weakly dominant* strategy: that is, a strategy which does better against one of the opposition's strategies (C2) and no worse than the player's alternative (R2) against the other strategy of the opponent (C1). Evidently, since there are only two strategies in this game, the weak dominance of R1 means that R2 is *weakly dominated*. The same holds viz. player C whose strategy C1 is *weakly dominant* and his C2 is *weakly dominated*.

	C1	C2
R1	$^+10,10^-$	$^+5,10^-$
R2	$^+10,5^-$	0,0

Game 2.4 – *Weakly dominated strategies*

Weakly dominant/dominated strategies (definition)

A strategy is *weakly dominant* if it guarantees a player, for each choice of the opposition, at least as good a payoff as does any other of his or her strategies, as well as higher payoffs for at least one choice by the opposition. As for *weakly dominated* strategies, they guarantee a player lower payoffs for at least one of the opposition's moves and, as far the remainder are concerned, they guarantees payoffs which are no better but also no worse than the ones she or he would have received from some other strategy.

Weak dominance explains why some outcomes do not qualify as Nash equilibria of a game [e.g. outcome (R2,C2), since it will occur only if players choose their weakly dominated strategies] but this is more of a theoretical nicety than a result with practical value. For even though it is true that (R2,C2) will not occur when R expects C to choose C2 and C expects R to choose R2, it may still occur if R expects C to choose his weakly dominant strategy (C1) and C expects R to choose her weakly dominant strategy (R1). Indeed, given these expectations, R is indifferent between playing R1 or R2 and, similarly, C is indifferent between playing C1 and C2. Consequently, they may well end up playing (R2,C2). The conclusion here is as follows:

Strict dominance, weak dominance and the stability of Nash equilibria
Whereas a Nash equilibrium supported by strict dominance reasoning is stable [e.g. (R1,C2) in Game 2.1], Nash equilibria supported solely by weak dominance reasoning are unstable [e.g. the Nash equilibria in Game 2.4]

2.4 RATIONALISABLE BELIEFS AND ACTIONS

2.4.1 The successive elimination of strategically inferior moves

Does dominance reasoning always identify a game's Nash equilibria? In other words, do the assumptions of rationality and CKR always point to a Nash equilibrium? Or is it possible that dominance reasoning may lead us to rationalise strategies that *do not* correspond to a Nash equilibrium?

Consider another interaction, Game 2.5. It is a version of Game 2.1 extended by the addition of strategies R3 and C3 which offer rich rewards to both parties. Nevertheless, players with some degree of confidence in each other's instrumental rationality will choose neither strategy. To see why this is so consider whether C would ever choose C3 rationally. The answer is *never*, because C3 is *strictly dominated* by both C1 and C2. Thus, C has absolutely no reason ever to play C3. The same applies to R regarding R2 (which is still strictly dominated by R1).

Would R ever play R3? Yes, provided that she expects C to play C3 – you can see this because of the ($^{+}$) marking that can be found next to 100 on the bottom right of the matrix. ***But***, have we not just concluded that a rational C will never play C3? We have indeed. Thus, as long as R knows that C is instrumentally rational, she will never play R3 (since R3 only makes sense as a best response to C3, a strictly dominated strategy which C is bound to avoid if rational). Therefore 1st-order CKR eliminates the possibility that R will choose R3 and leaves R1 as the only strategic option open to R. At the same time, 1st-order CKR ensures that C will never expect R to choose R1 and therefore convinces him to play C2. In conclusion, in Game 2.5, 1st-order CKR leads our players inexorably to the unique Nash equilibrium (just as it did in Game 2.1).[3]

Are Nash's equilibria synonymous with outcomes of rational play? Seven examples

	C1	C2	C3
R1	$^{+}$10,4	$^{+}$1,5^{-}	99,3
R2	9,9^{-}	0,3	98,2
R3	1,99	0,100^{-}	$^{+}$100,98

Game 2.5

	C1	C2	C3	C4
R1	$^{+}$5,12	-1,11	1,20^{-}	10,10
R2	4, -1	$^{+}$1,1^{-}	2,0	20, -1
R3	3,2	0,4^{-}	$^{+}$4,3	50,1
R4	2,93^{-}	-1,92	0,91	$^{+}$100,90

Game 2.6

	C1	C2	C3
R1	$^{+}$10,4	$^{+}$1,5^{-}	99,5^{-}
R2	9,9^{-}	0,3	98,2
R3	1,99	0,100^{-}	$^{+}$100,99

Game 2.7

	C1	C2
R1	$^{+}$5,5^{-}	0,0
R2	$^{+}$5,5^{-}	$^{+}$1000,5^{-}
R3	0,0	$^{+}$1000,5^{-}

Game 2.8

	C1	C2	C3
R1	$^{+}$100,99	0,0	99,100^{-}
R2	0,0	$^{+}$1,1^{-}	0,0
R3	99,100^{-}	0,0	$^{+}$100,99

Game 2.9

	C1	C2	C3
R1	$^{+}$2,1	0,0	1,2^{-}
R2	0,0	$^{+}$1000,1000^{-}	0,0
R3	1,2^{-}	0,0	$^{+}$2,1

Game 2.10

Nb. Nash equilibria are highlighted and can be easily spotted due to the coincidence of the ($^{+}$),($^{-}$) markings in the same cell.

A summary of the success of successive elimination at leading to a Nash equilibrium

Games 2.5 and 2.6: Successive elimination of *strictly* dominated strategies leads, uncontroversially, to the unique Nash equilibrium
Game 2.7: A similar procedural logic based on the elimination of *weakly* dominated strategies also leading, convincingly, to a unique Nash equilibrium
Game 2.8: A game featuring more than one Nash equilibrium and only weakly dominated strategies which can be eliminated. The process of elimination leads to one of the games (four) Nash equilibria. However we observe path-dependence; that is, we arrive at significantly different Nash equilibria depending on which weakly dominated strategy is eliminated first
Games 2.9 and 2.10: A comparison of two games with identical strategic structure. Is strategic structure, however, *all* the matters (as Nash suggests)? Or are non-Nash outcomes more likely outcomes of rational play in the presence of sufficiently strong incentives?

A piece of shorthand will be helpful in what follows. Let the verb 'chooses' or 'will choose' be denoted by a colon ':' and the verb 'believes' by the letter 'b'. Then, 1st-order CKR in Game 2.5 means that 'RbC is rational & CbR is rational' and therefore (as we concluded above on the basis of 1st-order CKR) RbC:C2 & CbR:R1. Hence, 1st-order CKR leads to Nash equilibrium (R1,C2).

We turn now to a 'larger' game, Game 2.6. Our first step is to add our best response markings to the payoff matrix and observe that a ($^{+}$) and a ($^{-}$) coincide only in cell (R2,C2). This must be the game's Nash equilibrium, in the sense that R2 is a best response to C2 and C2 a best response to R2. But do players R and C have good reason to play these strategies, notwithstanding the rather unattractive payoffs of (R2,C2) [when compared to, say, those of (R4,C4)]? We begin our investigation with an observation: C4 is strictly dominated. In other words, only an instrumentally irrational C would play it. Turning to R, she has no strictly dominated strategies (notice that each of her strategies is a best response to some of C's strategies).

However, we also note that R4 is only rationally playable as a best response to C4 and we know that C4 is *not* rationally playable by C. Hence, if R believes that C is rational, she will not play R4. It seems that zero-order CKR rules out C4 and 1st- order CKR rules out R4. But then, if C knew that R knows that C is rational (2nd-order CKR), C would not expect R to play R4. This is important because C1 makes sense to C only as a best response to R4 and, therefore, 2nd-order CKR eliminates C1 from C's list of rationally playable strategies.

So far, 2nd-order CKR has 'blacklisted' C4, R4 and C1 – in precisely that order. What if R believes that C has worked all these thoughts out for himself and has, thus, eliminated from his list not only C4 but C1 also? R will add R1 to her blacklist, since once C1 and C4 have been eliminated, R1 is strictly dominated by R2. In short 3rd-order CKR erases R1 from the list of rationally playable strategies. Which strategies are the players left with? The surviving ones are: R2,R3,C2 and C3 (the central part of the payoff matrix).

It takes no more than a cursory look to spot that if C had worked out all of the above (i.e. under the assumption of 4th-order CKR), C would never play C3. Why? Because, now that R1,C1,R4 and C4 are 'out of play', the surviving bits of C3 are strictly dominated by the surviving bits of C2. And if R knows this (5th-order CKR), she loses all interest in R3 because she can now expect with certainty that C will choose C2 (his only surviving strategy) and, therefore, she has a choice between payoff 1 (if she chooses R2) and payoff 0 (if she selects R3). Being a pay-off maximiser, she cannot resist R2.

The above process is known as the *Successive Elimination of Strictly Dominated Strategies* (SESDS). [Note, however, that some other texts refer to it as *Iterated Dominance*, reflecting the fact that different strategies are eliminated at different *iterations* (depending on the order of CKR).] At the outset, players 'blacklist' their strictly dominated strategies (what we have called zero-order CKR) and then, as the order of CKR rises, more strategies appear to be strictly dominated and are similarly blacklisted. The SESDS process, as applied to Game 2.6, is illustrated below.

Step	Order of CKR	Process of successive elimination in Game 2.8
1	*0-order*	C4 is eliminated as strictly dominated by C2 and C3)
2	*1st -order:*	R4, a best response to C4 only, is eliminated; note that once C4 is 'out', R4 is strictly dominated by R3
3	*2nd -order*	C1, a best response to R4 only, is eliminated; note that once R4 is 'out', C1 is strictly dominated by C3
4	*3rd -order*	R1, a best response to C1 only, is eliminated; note that once C1&C4 are 'out', R1 is strictly dominated by R2&R3
5	*4th -order*	C3, a best response to R1 only, is eliminated; note that once R1&R4 are 'out', C3 is strictly dominated by C2
6	*5th -order*	R3, a best response to C3 only, is eliminated; note that once C1,C3&C4 are 'out', R3 is strictly dominated by R2 **At this stage, each player is only left with a single rationally playable, or rationalisable, strategy**: *The Nash equilibrium strategies R2 and C2.*

Successive Elimination of Strictly Dominated Strategies (definition)
At the beginning, each player identifies her strictly dominated strategies and those of his opponent. These are eliminated (0-order CKR). Then each player eliminates those other strategies which have become strictly dominated once the cells of the strategies discarded in the previous iteration are ignored (1st-order CKR). In the next round (or iteration), more strategies are eliminated if they are rendered strictly dominated by the earlier elimination (2nd-order CKR). And so on until no strategies can be further eliminated.

In some games, as demonstrated above (e.g. Games 2.5 and 2.6), SESDS leads to the unique pure strategy Nash equilibrium. The presence of a strictly dominated strategy for at least one of the two players is a necessary (but insufficient) condition for this to happen. However, when no player faces strictly dominated strategies, SESDS cannot even get off the ground (e.g. Game 2.10). But even when it does, the process of elimination may grind to an abrupt halt before it can pinpoint a unique equilibrium.[4]

Perusing Games 2.5 to 2.10 above, it is easy to see that SESDS leads players along the path to a Nash equilibrium (that is, rationalises the Nash equilibrium concept as the result of actions reflecting rational beliefs) in only two cases: Games 2.5 and 2.6. In Game 2.7 SESDS also gets off to a promising start with the dismissal of R2 (since R2 is strictly dominated by R1) and the subsequent rejection of C1 (since C1 is best response to the discarded R2 and, once R2 is 'out' C1 is dominated strictly by C2). However, this is where it stops. Unlike the process of elimination in the very similar Game 2.5, here (in Game 2.7) C3 is *weakly* (but not strictly) dominated by C2. For even though C3 is a bad response to both R2 and R3, it is not worse than C2 when R chooses R1.[5] The problem however is that, unless we eliminate C3, we cannot dismiss R3. In conclusion, if we are determined to delete only strictly dominated strategies, and allow weakly dominated ones to survive, the game might converge on any of the following cells: (R1,C2), (R1,C3), (R3,C2) and/or (R3,C3).

Game 2.7 demonstrates that, in some contexts, an unwillingness to discard weakly dominated strategies, on the strength that they are *not* irrational responses to *some* belief of the opposition, prevents the process of elimination to proceed to the Nash equilibrium (even when there exists a unique one). If we weaken our criterion for dismissing strategies so as to exclude not only strictly dominated choices but also weakly dominated ones, then the process can continue.

In Game 2.7 it seems quite plausible to amend SESDS and turn it, more generally, into the *Successive Elimination of Dominated Strategies* (of both the strictly and the weakly dominated variety): As before, we would eliminate R2 (*strictly dominated* by R1) and then C1 (*strictly dominated* by C2, once R2 is 'out'). Looking at C's motives, we should quickly dispense with C3 (*weakly dominated* by C2) and thus erase R3 (*strictly dominated* by R1, once the *weakly dominated* C3 has been thrown out). With this last iteration[6] SEDS has left a single possible outcome standing: the Nash equilibrium.

The plausibility of the weakening of SEDS in Game 2.7 notwithstanding, it is important to be cautious when eliminating *weakly dominated* strategies. In Game 2.7 only one *weakly dominated* strategy was felled during the SEDS process (which eventually led us to the Nash equilibrium). But when there is more than one *weakly dominated* strategy, the logic of SEDS can generate a headache known in the literature as *path dependency*. To see this, we consider Game 2.8 featuring two *weakly dominated* strategies for player R: R3 (which is *weakly dominated* by R2); and R1 (also *weakly dominated* by R2).

Suppose we eliminate R3 at the outset. Suddenly C2 becomes a *weakly dominated* strategy for C. (NB. with R3 out of the game, C2 can do no better than C1 when R plays R2 and in fact does worse when R selects R1.) If we are to dismiss C2 on these grounds (namely, its *weakly dominated* status), we will have reached one of the two identical Nash equilibria distributing payoff 5 to each player [that is, they will end up playing either combination (R1,C1) or (R2,C2), both yielding payoff combination (5,5)]. On the other hand, if we start by eliminating R1 rather than R3, R1's demise will render C1 *weakly dominated*. With R1 and C1 out of the picture, players will converge on one of the two highly asymmetrical Nash equilibria yielding payoff 1000 for R and 5 for C. This is what is called *path-dependence*: the destination (i.e. the resulting Nash equilibrium) will depend on the first move at the point of departure.

Path-dependence is a problem because we would like to think that our logic should lead to the same conclusion (regarding rational play) independently of the logical step we take initially. If wildly different conclusions are to be had depending on what rational thought happened to cross our mind first, then SEDS fails reliably to rationalise the convergence of rational play to Nash equilibrium.

One way of summarising our results so far in this section is to say that, when the successive elimination of strictly dominated strategies (SESDS) converges on some Nash equilibrium, our confidence in the latter emerging as the outcome of rational play reduces to how confident we are that the necessary order of common knowledge of rationality (CKR) characterises our players' thoughts. E.g. in Game 2.6 we can be certain that, as long as there exists at least 5^{th}-order CKR, Nash strategies R1 and C2 are uniquely rationally playable or, less awkwardly, *rationalisable*.

2.4.2 Rationalisable strategies and their connection with Nash's equilibrium

Rationalisable strategies were first defined by Bernheim (1984) and Pierce (1984), even though Spöhn (1982) had foreshadowed their meaning, as follows.

Rationalisable strategies and Rationalisability (definition)
In two-person games, the strategies which survive the *successive elimination of strictly dominated strategies* (SESDS) are called *rationalisable*. All pure strategy Nash equilibrium strategies are *rationalisable* though the opposite is not true. If there exists only one *rationalisable* strategy per player, then that strategy is bound to correspond to some Nash equilibrium in pure strategies. However, when there are multiple *rationalisable* strategies per player, there is no guarantee that the rationalisable strategies will also be Nash strategies.

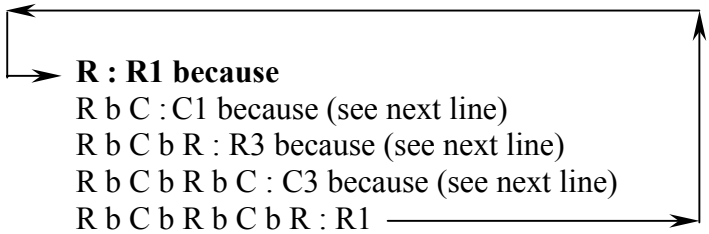
The underlying principle behind a *rationalisable strategy* is that players who choose it are invariably in a position to justify their choice, without invoking either an irrational argument or a belief that their opponent will respond irrationally to some of his or her beliefs. Their reasons for playing *rationalisably* must be expressible in terms of fully rational arguments and without resorting to any patronising views regarding their opponent's capacity to base their choices on equally rational arguments. But is this not also what Nash required as part of his equilibrium solution? Not quite.

The difference between *Nash strategies* and *rationalisable strategies* is subtle but significant. SESDS eliminates all the strategies that people will *never* play given that one knows that the other knows that one knows that the other knows... *no one will choose strictly dominated strategies*. By definition, therefore, *rationalisable* strategies are the ones left standing (after this process of elimination) and, therefore, such choices reflect the common knowledge (of sufficient order) that no player will make moves ill-supported by her beliefs. *The Nash equilibrium concept is far more demanding.* It requires not only that that the players' acts are best replies to their predictions about others' behaviour (as well as common knowledge that this will be so), but also that their predictions will be *correct* and that, additionally, it is common knowledge that they are correct! An examination of Game 2.9 helps clarify matters.

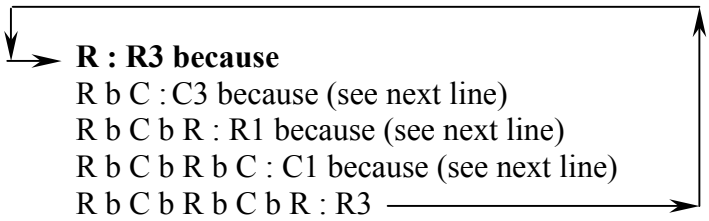
In Game 2.9 each strategy of either player is a best reply to one of the opponent's strategies;[7] thus, no strategy is dominated and all survive the successive elimination process. Therefore, according to the definition above, all of the game's strategies are *rationalisable*. What does this mean in practice? It means that players can choose any one of them and have a sound explanation to offer as to why they did so. For example, suppose R whispers in our ear that she intends to play R1. If we ask her "Why?" she could answer thus:

> I will play R1 because I expect C to play C1. Why do I expect this? Because I do not think C expects me to play R1; indeed I think he expects that I will be playing R3 (rather than the R1 which I intend to play). You can ask me why I think that he will think that. Well, perhaps because he expects that I will mistakenly think that he is about to play C3, when in reality I expect him to play C1. Of course, if he knew that I was planning to play R1, he ought to play C3. But he does not know this and, for this reason, and given my expectations, R1 is the right choice for me. Of course, had he known I will play R1, I should not do so. It is my conjecture, however, that he expects me to play R3 thinking I expect him to play C3. The reality is that I expect him to play C1 and I plan to play R1.

The above thought process can be summarised, using the shorthand introduced earlier as follows:



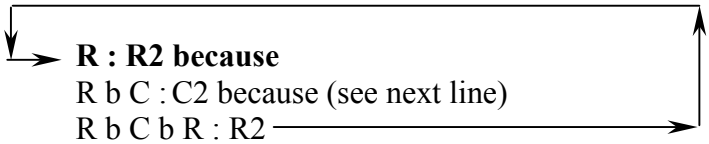
We see that 4th-order CKR is sufficient for a belief to be developed which is consistent with strategy R1. Increasing the order of CKR does not change things as the above loop will be repeated every four iterations. Thus strategy R1 can be based on expectations which are sustainable even under infinite-order CKR. Different, but equally internally consistent, trains of thought exist to support R2 and R3. For example, R3 is supported by a story very similar to the above. We offer only its shorthand exposition.



Now consider the equivalent rationalisation of R2.

> I will play R2 because I believe that C will play C2. And why do I believe that C will play C2? Because he thinks that I will play R2, thinking that I expect him to play C2. And so on.

Or, in our shorthand,



As we just saw, all three strategies (R1,R2 and R3) are *rationalisable*, because there are plausible beliefs to which each strategy is a best reaction. Consequently, the notion of *rationalisability* does not help game theorists 'solve' this game since anything goes: that is, every available strategy is rationally playable or *rationalisable*. In short, the game is *indeterminate*.

Not so, insists John Nash. For among this plethora of rationalisable strategies, there is a single one per player that stands out: R2 and C2. Why does Nash think that R2 and C2 are particularly salient for R and C? Its unique appeal springs from a remarkable feature that it alone has: R2 and C2 are strategies supported by beliefs which will *not* be frustrated by the actual choice of R2 and C2.

To see this clearly, recall that strategy R1 is rationalised by R thinking (a) that C expects her to choose R3 and (b) that C will choose C1 in reply. There are two possibilities. Either R's predictions are confirmed, or they are not. Suppose they are (and that R has indeed played R1). In that case, C's beliefs will be frustrated. We know this because the only way R's beliefs could be confirmed is if C has played C1. Why would he do this? The only rational belief that would make C play C1 is if C expected R to play R3. But R has frustrated that belief of C with his actual choice of R1. Alternatively, R's own beliefs will have been frustrated (when C chooses

something other than C1). In either case, *the play of R1 will frustrate someone's beliefs and will only be played rationally by an R who is confident that the outcome will frustrate her opponent's beliefs, rather than her own.*

Of course the same applies to R3 since, as we saw above, it also relies on a logical loop according to which R selects R3 on the strength of her belief that C will *not* expect her to play R3. But the same does *not* apply to R2. In fact by choosing R2, R is telling the world that she is expecting C to make no mistake in predicting her strategy. In other words, R2 is chosen rationally only when R has no reason to think that C will base his decision on a mistaken prediction of her choice. Similarly, C2 will be played when C has no reason to predict that R will be fooled. And when R and C choose R2 and C2, their actions will confirm their trust in one another's capacity to avoid erring.

2.5 NASH EQUILIBRIUM

2.5.1 John Nash's beautiful idea

John Nash left an indelible mark on game theory by proposing a solution to games where none seemed to exist. Game 2.9 is an excellent example. With all its strategies fully rationalisable, one is tempted to conclude that Game 2.9's outcome is indeterminate, that rationality alone cannot guide players (and thus theorists) to a single outcome. Nash, however, did not agree. He singled out the strategy combination of (R2,C2) as the only pair where the expectations upon which they were based were confirmed by the subsequent actions and proposed this as the uniquely rational set of actions because no person would have reason to regret either their belief or their action. We call this attribute of people's beliefs their *consistent alignment*.

Consistently Aligned Beliefs (definition)

Beliefs are *inconsistently aligned* when action emanating as best replies to these beliefs can potentially 'upset' them. A belief of one player (say R) is 'upset' when another player (say C) takes an action with a probability which player R has miscalculated. By contrast, beliefs are *consistently aligned* **(CAB)** when the actions taken by each player (based on the beliefs they hold about the other players) are constrained so that they do not upset those beliefs.

When people select Nash equilibrium strategies, there is no guarantee that they shall be 'happy'. There is, however, a guarantee that have no reason to change their beliefs about each other or to regret their actions (given their beliefs) after the event. This is what makes the Nash equilibrium a plausible solution to any game.

If Nash is right, a seemingly insoluble game has been solved. And, as if this remarkable result were not enough, Nash proceeded to prove that a Nash Equilibrium exists *for all (finite) games*! With this proof, game theory came of age: Nash had proposed an apparently plausible solution that could be applied to all interactions involving rational players.[8]

Before scrutinising this claim more carefully, there is another angle to the Nash Equilibrium that is hardly ever stated: the aesthetic one. Let us revisit **Loops 1** and **2**; the schematic (shorthand) representation of the rationales for playing R1 and R3 in Game 2.9. Then let us compare them with **Loop 3**; the rationale behind R2 in shorthand. There is an *aesthetic* difference between them: the disarming simplicity of the logical **loop** supporting R2. There is a conspicuous harmony in the mutual reflection of R2 into C2, back to R2, then again to C2 *ad infinitum.* The ancient Greeks were convinced that simplicity and harmony were the two essential elements of beauty and that beauty turned on some nice geometrical property. If they were right, Nash's equilibrium is a 'beautiful' concept.

Turning from the aesthetic to the philosophical, Nash's *mutually-confirming strategies almost* invokes the Socratic notion that one's views are confirmed only through reflection against another's; or perhaps the Hegelian take on the dialectic where a 'self' is well-defined only after an infinite self-reflection in the eyes of an 'other'.[9] To the extent that one has faith in the capacity of human reason to home in on self-reflective 'states', one is tempted to celebrate Nash's discovery. Before we ask, however, whether some caution is necessary, let us extend and formalise the definition of Nash's Equilibrium (first encountered in Section 2.1):

N-person Nash Equilibrium in pure strategies (comprehensive definition)
Let G be a *normal form* game, involving N players. Each player chooses among a finite set of strategies S_i: That is, player i (i=1,...,N) has access to strategy set S_i from which she/he chooses strategy σ_i (belonging to S_i). Player i's payoff Π_i then depends not only on her/his choice of strategy σ_i but on the whole set of strategic choices (σ_1, σ_2,..., σ_i,...σ_N) by all players (including her own). Thus $\Pi_i = \Pi_i (\sigma_1, \sigma_2,...,\sigma_i,...\sigma_N)$. A set of pure strategies $S^* = (\sigma_1^*, \sigma_2^*,...,\sigma_i^*,...\sigma_N^*)$ constitutes a *Nash equilibrium* if and only if pure strategy σ_i^* is a *best reply* to the combination of the strategies of all other players in S^* for all i=1,...,N.
Corollary: If *for all* i=1,...,N, i chooses his/her pure strategy in S^*, then each player's predictions of her/his opponents' behaviour will be confirmed.

2.5.2 Consistently aligned beliefs, the hidden Principle of Rational Determinacy and the Harsanyi-Aumann Doctrine

The question, then, is: Should we expect players to hold consistently aligned beliefs, as this consistency underpins the Nash equilibrium concept? Notice that when players are rational and hold common knowledge of rationality, there is one circumstance when CAB will hold: When there is a uniquely rational way to play the game for each player. If there *was* such a unique way, and each player knew that all players involved are rational (and that they all knew this to be so etc.), then no player could rationally expect others to do something other than play in this uniquely rational way. The outcome would then *have* to be a Nash equilibrium since, for non-Nash play, at least one player would have to fail either to be rational or to share in the common knowledge that everyone is rational.

However, it is one thing to agree that a uniquely rational way of playing will be adopted by rational people (under CKR) but it is quite another to assume that there always exists a uniquely rational behavioural code. Indeed, the idea that there always exists a unique way to play each and every game imaginable is equivalent to a rather strong version of Rationalism. We call it the *Principle of Rational Determinacy.*[10] So, one way of approaching the question of whether CAB is plausible is through an assessment of the *Principle of Rational Determinacy* in this context.

In Section 1.2.2 we discussed a proposition put forward by John Harsanyi (1967/8) which has come to be known as the *Harsanyi doctrine*. It suggested that, given the *same information set* regarding some event, rational agents must always draw the *same prediction* (that is, expectations of what will happen). In computer terms, it is like saying that the same input (i.e. information) must yield the same output (i.e. prediction) as long as the computer used is identical in terms of its computational capabilities (i.e. the players are equally rational).

Of course, it is possible that agents have different information sets and so draw different conclusions. But it will be recalled from Box 1.9 in Section 1.2.2 that Robert Aumann (1976), in defence of Harsanyi, discounts this possibility because two agents could not agree to disagree in such a manner since, the moment rational agents (under CKR) discover that they are holding inconsistent expectations, each has a reason to revise their beliefs until they converge and become consistent. Thus the combined *Harsanyi-Aumann doctrines* imply that rational agents,

when faced by the same information with respect to the game (the 'event' in this case), should hold the same beliefs about how the game will be played by rational agents. In short there must be a unique set of prior beliefs which rational players will hold about how a game is played rationally.

The *Harsanyi-Aumann Doctrine* (definition)

The assumption that, in every finite game, the prior beliefs (of every rational player who knows the rules of the game are the same) is known as the *Harsanyi-Aumann doctrine*. This is because all differences in the players' belief arise uniquely from differences in information since each player has the same information given by common knowledge of the rules of the game and each player's rationality. Thus, rational players with common knowledge of rationality will not be able to agree to disagree on the likelihood of any action in the game.

Von Neumann and Morgenstern suggested an objective for any analysis of games in terms of writing a book on how to play games which would remain good advice once it had been read widely. In other words, it could not really be good advice if people, having read it, and having taken into account that others have read it too, do not want to follow the advice. The Nash equilibrium concept passes this test precisely because it recommends strategies which are best replies to each other. A merely rationalisable pair of strategies would not. In Game 2.9, for instance, (R1,C1) is a rationalisable outcome. However, if the ultimate game theory book were to recommend strategies R1 and C1 to its readers, it would not go down very well.

The reason is that, in the aftermath, player C would be very upset with the book's advice since his choice of C1 is not a best reply to R1 (although R would be quite happy with it). Indeed, the book could only convince him to play C1 by *lying* to him that R would play R2 (since C1 is a best reply only to R2) while, at the same time, advising R to choose R1. Of course, this is not possible unless R and C are reading different books!

The idea sounds quite plausible: A theory cannot be good unless its advice to players is available to all of them at once. Or, equivalently, a theory cannot be admitted if it predicts well only when the players are oblivious to it (or misunderstand it). But it is an idea in need of careful handling. The good book metaphor also hinges on a hidden assumption: that there exists a *uniquely good piece of advice* regarding rational play to be printed in the 'good' book. Suppose the converse: that the book gave several bits of advice. No one could be sure that following any single bit was rational or irrational once others had read the book. The reason is that they would not know which bit of the advice the others would be following.

Thus suppose rationalisability was the source of the advice given by a game theory book. In Game 2.9, the advice becomes 'play any of the strategies', as all are rationalisable. Consequently, the common knowledge that others have read the same book would not give any reason not to follow one of the rationalisable strategies. Of course, the book would not be giving particularly helpful advice here in the sense that it endorses all of the available strategies. But maybe that is all that can be said.

In summary, to support the Nash equilibrium concept through the above test, we have to assume that a unique piece of advice is possible. If there is such a thing, and it satisfies the 'good' book condition, then the advice will have to be a Nash equilibrium. But that is a big 'if' which amounts to the same as the common priors, or CAB, defence of Nash. In short, it all turns on a hidden (or implicit) assumption that there exists a uniquely rational way to play all games.

2.5.3 Some Objections to Nash: Part I

Compare Game 2.9 with Game 2.10. They are identical in terms of their strategic structure. By this we mean that *in both games* R1 is the best reply to C1, R2 the best reply to C2, R3 the best

reply to C3, C1 the best reply to R3, C2 the best reply to R2 and, finally, C3 the best reply to R1.[11] In Nash's terms, behaviour which is rational in one ought to be *identically* rational in the other. And yet, many people who come face to face with these two games would treat them as quite different interactions. Is this so because most people are not fully rational? Or is there something amiss with Nash's concept.

In Game 2.10 the Nash equilibrium (R2,C2) sticks out from miles as the 'attractor' of rational behaviour. However, one wonders whether this is so because R2 and C2 is the only pair of strategies that reinforce one another (in Nash's sense) or because outcome (R2,C2) catches our attention courtesy of its generous and symmetrical payoffs. By comparison, (R2,C2) is far less attractive in Game 2.9, even though the two games are identical from Nash's viewpoint. Our hunch here is that players engaged in Game 2.9 have greater incentives (due to the payoffs being much larger in cells outside the Nash equilibrium) to drift away from the Nash equilibrium by considering, quite rationally, the possibility of outwitting their opponent. This hunch has been recently put to the test by Goeree and Holt (2001). They report widely different behaviour from the laboratory in games with identical strategic structures (e.g. games like 2.9 and 2.10) but different actual pay-offs.

Let us now turn to two fresh games (Games 2.11 and 2.12) which bring out some further wrinkles regarding Nash's equilibrium.

	C1	C2	C3
R1	$^{+}1,1^{-}$	$^{+}100,0$	$-100,1^{-}$
R2	$0,100^{-}$	1,1	$^{+}100,1$
R3	$^{+}1,-100$	$1,100^{-}$	1,1

Game 2.11

	C1	C2	C3
R1	$^{+}1,1^{-}$	-10000,-10000	$^{+}3,0$
R2	-10000,-10000	$^{+}1,1^{-}$	0,0
R3	$0,3^{-}$	0,0	2,2

Game 2.12

Game 2.11: An example of a game featuring an unreasonable (albeit unique) Nash equilibrium (R1,C1)
Game 2.12: An example of a reasonable non-equilibrium outcome (R3,C3)

In Game 2.11 *the unique Nash equilibrium in pure strategies seems unreasonable* to us. The unique pure strategy Nash equilibrium is (R1,C1), but note that player R can secure, with absolute certainty, exactly the same payoff (1) as that associated with (R1,C1) by playing R3. In sharp contrast, R1 comes with the risk of -100 if C plays C3. Does the fact that R1 may also yield +100 utils (if C responds to her R1 with C2) cancel out the risk of playing R1?

Under CKR the answer is negative. If R plans to play R1, and thinks that C will work this out, R has no reason to expect C to choose C2 (since C2 is a bad reply to R1). On the other hand, if R thinks that C expects her to choose her Nash strategy (R1), she has every reason to fear that C might select C3 since, from C's perspective, C3 is as good a response to R1 as C1. Indeed, C is guaranteed his Nash payoff (1) if he plays C3 (just as R was guaranteed her Nash payoff if she opted for R3.) So, R is facing a very real danger of losing 100 utils if she heads for the Nash equilibrium. Why on earth would R risk it, by playing her Nash equilibrium strategy, when R3 *guarantees* her the Nash equilibrium payoff? Similarly for C. Why risk playing his Nash strategy C1 when C3 guarantees the payoff he can expect to receive at that Nash equilibrium but without the risk?

There is therefore some real doubt over whether the unique Nash equilibrium in pure strategies is reasonable. Would Nash's followers (that is, most game theorists) recognise that playing R1 could be unreasonable? No, they would not. Their argument would be as follows: R will only play R3 if somewhere along the line there is a *breakdown of CKR*. To see this, consider what would happen if R had, as it is alleged above, good cause to shift her attention to R3. Were

this a uniquely rational deduction, every one (including C) would know that R is about to play R3 *with certainty*. In that case, a C expecting R3 would rationally shift his attention to C2, his best response to R3. If C can reach this conclusion, so can R who now thinks to herself: "If C, expecting me to play R3, selects C2, should I not play R1, the best reply to C2, and collect 100 utils?" By this logic, we are back to R choosing her Nash equilibrium strategy R1.

While the logic in the previous paragraph is perfectly reasonable, defenders of Nash have to believe that no doubts ever creep in over the ability of each *exactly* to replicate each other's thoughts. With just a smidgen of apprehension here, wouldn't it be safer to opt for R3? What does she have to lose from such a shift? Nothing at all. Does she have anything to gain? Much, since R3 offers her the chance to eradicate the risk of losing 100 utils if the synchronicity of beliefs (demanded by Nash) fails to materialise.

Let us now turn to Game 2.12. We suggest that many rational people would pick strategies R3 and C3. We are not arguing this simply because R3 and C3 yield a distribution of payoffs which is mutually preferred to either of the game's two Nash equilibria in pure strategies (R1,C1) and (R2,C2). The fact that (R3,C3) is mutually advantageous (compared to the Nash equilibria of the game) is insufficient reason to suppose that rational people would select strategies R3 and C3. But, a rational R might plausibly think twice before eliminating R3. For if she did, she would face the very real prospect of a catastrophic payoff (-10000). Can we confidently argue that any R (C) player who opts for the safer option of R3 (C3) is acting irrationally?

If it were commonly known that R3 was R's best choice, then C would naturally find it hard, under the circumstances, to resist the temptation of playing C1 [in order to collect payoff 3 rather than the 2 corresponding to (R3,C3)]. And if this were common knowledge, R would respond with R1; her best reply to C1. Clearly, R3 (and C3) are *not* part of any Nash equilibrium scenario. But, both players have real risks associated with selecting a strategy here that forms part of a Nash equilibrium in pure strategies because there are two such equilibria. If they fail to co-ordinate their choices, a disastrous pay-off results (-1000). The risks for each are not just over whether the other will opt for a Nash equilibrium (as in the earlier examples). They arise instead because the players also need to co-ordinate on one of these equilibria in this game. This aspect is new and assumes further significance later in this chapter and the rest of the book.

We turn away now from specific games where the Nash equilibrium can seem implausible to a consideration of the key assumption which underpins this solution concept: the *Principle of Rational Determinacy* and its associated presumption of CAB.

Certainly there are other game theorists who have their doubts. Kreps (1990) puts it this way.

> We may believe that each player has his own conception of how his opponents will act, and we may believe that each plays optimally with respect to his conception, but it is much more dubious to expect that in all cases those various conceptions and responses will be 'aligned' or nearly aligned in the sense of an equilibrium, each player anticipating that others will do what those others indeed plan to do.

Likewise anyone who has talked to good chess players (perhaps *the* masters of strategic thinking) will testify that rational persons pitted against equally rational opponents (whose rationality they respect) do not immediately assume that their opposition will never err. On the contrary, the point is to *engender* such errors! Are chess players irrational then?

To see how such doubts might arise, we will examine the support given to the *Principle of Rational Determinacy* by the *Harsanyi-Aumann doctrine.* This has two elements. The first is that rational people who have the same information will draw the same inferences. The second is that people playing games have the same information with respect to the rules of the game.

With respect to the first part, it is difficult to see what model of reason delivers this as a general conclusion. As we have already suggested, the model that is commonly used in this context is an algorithmic one. Reason is associated with a set of rules for processing data and so when the data is processed using the same rules, one expects the same result. This is fair enough and the algorithmic model of reason plausibly covers some settings, but it is difficult to see how it can claim to be a *general* model. This is for the simple reason that no set of rules can be exhaustive: *No set of rules can contain rules for their own application*. Any new setting where the rules might be applied can always be individuated in such a way that it is not covered by the existing rules. In these circumstances, either new rules have to be created to cover every new application (and this process threatens an infinite regress), or people have to interpret creatively how the existing rules apply in new settings. But once interpretation of this sort is recognised, there is no reason to suppose that all individuals will be creative, in this sense, in the same way.

There is another way of seeing this same point. Some questions faced by rational agents come in the form 'What is 1+1?' while others ask 'Does God exist?' or 'Will there be another major European war in the next 30 years?' The first has a determinate answer and one would expect that all rational agents will converge on it (because the algorithm for solving problems of addition is relatively simple to acquire). But the last two do not seem to admit a uniquely rational answer. Rational agents could have exactly the same evidence on both these questions and yet it seems they could quite legitimately draw different conclusions.

To put this slightly differently, when two people argue over the prospect of another major war in Europe, no one is likely to cast one or other person as irrational just because they have taken a contrary position. This is because our rules of inference cannot be applied simply to such complex questions: *They lie outside the domain of their determinate operation and so require a form of creative judgement*.

This is similar to a distinction that is sometimes made in the economics literature between risk and uncertainty. In situations where the likelihood of any outcome is given by a probability distribution, one might plausibly suppose that rational agents will come to hold the same expectations. But when there is genuine uncertainty, people can quite reasonably hold different expectations. Thus Keynes's analysis of the determination of interest rates in financial markets turns on people holding different expectations with respect to the future direction of change in interest rates. Likewise, the Austrian tradition in economics highlights the role played by entrepreneurs in being better able than others to form expectations about the future when there is uncertainty.

Of course, this leaves open whether the question of rational agency in games is more like 'What is 1+1?' or 'Will there be a major war in Europe in the next 30 years?'. We suspect that games divide on this matter. Some are like one and some are like the other with the result that a general presumption towards a unique solution is wrong. At the very least, there is a burden on those who believe in uniqueness to explain why this is always the case in games since this cannot be a general proposition about rationality in all settings.

Box 2.6

AGREEING TO DISAGREE EVEN WHEN IT IS COSTLY

During the late summer of 1992, there was exceptional activity in European currency markets. First the lira was forced to devalue within the European Exchange Rate Mechanism (ERM) and then speculators turned their attentions to the pound and the Spanish peseta, selling both in the expectation that they would follow the lira.

This selling took place against a background of a deepening recession in the UK. Indeed throughout the spring and early summer as the recession in the UK economy worsened there had been renewed calls for a reduction of interest rates. But, the British government was committed to membership of the ERM and was determined to hold the exchange rate. So it kept interest rates high and it sold foreign currency from its reserves to boost the demand for pounds. There would, of course, have been no point in these circumstances in taking such actions had the government thought that the pound would eventually have to leave the ERM; since if this had been the case, then the government would have known that it was only delaying a depreciation and such a delay was obviously costly in

terms of delaying interest rate cuts, and the recovery, and through the expenditure of reserves. Thus it seems we should assume that the government thought that the pound could be maintained at the old ERM rate through these actions. Equally, there would be no point in speculators selling pounds and forsaking the relatively high interest rates to be enjoyed by holding sterling rather than DMs, unless they expected that the pound would eventually depreciate.

Hence, the speculators and the British government appeared to be in fundamental disagreement through the summer of 1992 over the likely direction of sterling. This disagreement came to a spectacular head in the week beginning on 14 September. Sterling bumped along the bottom of the ERM band on Monday and Tuesday and then on Wednesday, reportedly triggered by a newspaper interview with the President of the Bundesbank, the selling of sterling reached new peaks. Indeed, as one Bank of England Official said `I can't stress enough the sheer scale of the selling. We have never seen anything like it . . . . It was as if an avalanche was coming at us.' The Chancellor of the Exchequer raised interest rates by 2 points to 12% at 11 am; at 2.15 pm he raised them again, this time to 15%. By 4 pm, a third (about œ15 billion) of the official reserves had been used in support of sterling that day alone and the defence of sterling was over. The Chancellor of the Exchequer took sterling out of the ERM, and by 5 pm on the New York market the pound had fallen by about 10% against the DM.

It is difficult to gauge the precise costs to the Bank of England, and ultimately to the British tax payer, of this disagreement over the summer of 1992. The net loss to the reserves was the (roughly) 10% depreciation on whatever reserves had been sold (or committed) in support of sterling. The precise figures here are never clear because the Bank of England takes up positions in futures markets for sterling where margins are low, but it is plain from the use of reserves on `Black Wednesday' alone that the cost to the reserves ran into several billion pounds. In addition, there were the costs from delaying the economic recovery by pursuing high interest rates.

Of course, the other side to these costs from `agreeing to disagree' over the summer were some spectacular speculative gains. It has been suggested, for instance, that George Soros, one of the gurus of financial markets, made a billion dollars that summer from speculating against the lira and the pound. This has not been confirmed, but as a currency dealer from the Bank of America told the *News at Ten* reporter on Wednesday evening, "We've had an excellent day. We've made a lot of money . . . about £10 million" – not bad for 1 day's work for a few people operating in front of a computer screen!

There is an alternative view of rationality to the algorithmic one that paradoxically both emphasises ineliminable uncertainty and can, potentially, provide support for the Nash equilibrium concept. For Kantians reason supplies a critique of itself that is the source of negative restraints on *what* we can believe, rather than positive instructions as to what we *should* believe. Thus the categorical imperative (see Section 1.2.1) which, according to Kant, ought to determine many of our significant choices, is a sieve for beliefs and it rarely singles out one belief. Instead, there are often many which pass the test and so there is plenty of room for disagreement over what beliefs to hold. However if, as Kant suggests, rational agents should only hold beliefs which are capable of being universalised, then, taken by itself, this could prove a powerful ally to Nash. The beliefs which support R1 and R3 in Game 2.11 do not pass this test since if C were to hold those beliefs as well, C would knowingly hold contradictory beliefs regarding what R would do. In comparison, the beliefs which support R2 and C2 are mutually consistent and so can be held by both players without contradiction.

This argument needs careful handling because a full Kantian perspective is likely to demand rather more than this which game theorists would reject out of hand. Indeed such a defence of Nash would undo much of the foundations of game theory: for the *categorical imperative* would even recommend choosing dominated strategies if this is the type of behaviour that each wished everyone adopted. Such thoughts sit uncomfortably with the Humean foundations of game theory and we will not dwell on them for now. However, we shall return to this issue in Chapter 7.

We now move on to the second part of the *Harsanyi-Aumann doctrine*: The presumption that *rational agents must have the same information*. It will be recalled that Aumann argued that disagreements yield new information for both parties which causes revisions in expectations until convergence is achieved. This sounds plausible at first, but when beliefs pertain as to how to play the game, and divergent beliefs are only revealed in the playing of the game, it is more than a little difficult to see how the argument is to be applied to the beliefs which agents hold

prior to playing the game. Naturally when the game is repeated, the idea makes much more sense. But for the time being we are only examining one-shot, timeless games.

A logical difficulty also arises when information is costly to acquire. Suppose Aumann is correct and you can extract information so fully that expectations converge. Convergence means that it is *as if* you and your opponent shared the same information (following Harsanyi). But if it is costly to acquire information, why would anyone ever acquire information? Why not free-ride on other people's efforts? However, if everyone does this, then neither agent will have a reason to revise their beliefs when a disagreement is revealed because the disagreement will not reflect differences in information (since no one has acquired any). The only way to defeat this logic is by assuming that information is not transparently revealed through actions, so there is still possibly some gain to an individual through the acquisition of information rather than its extraction from other agents. But if this is the case, then expectations will *not* converge because agents will always hold information sets which diverge in some degree.

Thus to summarise, the case for rational agents having the same information is not robust, if either information is costly to acquire or games are played only once. Furthermore, even when rational actors have access to the same information, there can be no general presumption that they will draw the same inference from this information.

We conclude this discussion of Nash with a final puzzle. Suppose there are multiple Nash equilibria in pure strategies. What should rational actors who believe in unique solutions do? (See Box 2.5 for a famous interaction in political philosophy where there are multiple Nash equilibria in pure strategies.) If there is a unique solution, then we agree that it must be a Nash equilibrium, but this is not very helpful insight when there are multiple Nash equilibria. Rational agents must be able to draw some further inference if they are to distinguish the uniquely rational way to play the game. So what is this further inference? It needs to be specified. Otherwise the claim that rational agents will draw the same inference about how a game should be played will sound hollow.

Box 2.5

INELIMINABLE UNCERTAINTY AND ROUSSEAU'S STAG HUNT

J.-J. Rousseau (1762) was concerned with the problem of collective production where each member of a community of producers must choose between different degrees of commitment to a common good whose eventual value will be proportional to the effort of the *least committed team members* (who can, thus, 'let the side down'). His point was that, in such cases, whether the common cause will be served or not depends on the degree of optimism among team members, or citizens. If they believe that it will be served, then it will. And if they are pessimistic, the common cause is doomed. Rousseau's own parable for making this point revolved around a team of hunters who could either join forces in order to catch a stag, so that all can eat well (a feat depending on the commitment to the task of each and every member; as opposed to the average commitment), or abscond and hunt separately for smaller prey (e.g. rabbits) to be eaten individually. One way of encapsulating Rousseau's argument game theoretically is by amending slightly the game in Box 2.3.

As in Box 2.3, player i's payoff equals $P_i = 1\text{-}3\mu+2X_i$, except that here instead of μ being the *average* private appropriation of the public good (i.e. the *average* value of X chosen within the population of players), it is the *maximum* private appropriation of the public good (i.e. the maximum value of X chosen by someone in the group). In keeping with Rousseau's parable, setting X=0 is equivalent to spending all your time pursuing the common cause: the stag. As X rises, you are abandoning the common cause and diverting your energy to hunting rabbits. The pay-off function captures Rousseau's conjecture that the probability of catching the stag diminishes even if one person chooses a high X (presumably because it takes a single lax hunter to let the stag escape).

Notice how our small emendation altered the game's strategic complexion radically; in the sense that the natural tension between the private and the collective interest, which we encountered in Box 2.3, has now disappeared. To see this, suppose that i expects every one else to respect the common good/cause to the full and set their private greed X=0. If i does likewise, i will collect the maximum payoff $P_i = 1$ (since μ=max (X) = 0 and i's own X is also zero). Does i have an incentive to 'cheat' by 'depleting' the common good (i.e. set his/her greed X_i>0) when others abstain from so doing? The answer is clearly negative since if i sets X_i>0, then she/he will have boosted μ (the maximum value of X) single-handedly and caused his/her own payoff P_i to fall (note that with every increase in μ, due to a rise in X_i, P_i declines by a factor of 3 and increases by a factor of 2; overall it falls). So, if i believes

that no one will grab bits of the common good, nor will *i*. Indeed, whenever *i* expects that μ will equal, say, μ^*, *i*'s best reply strategy is to set her/his own $X_i=\mu^*$. In simpler terms, the optimal strategy here is to choose a value of *X* equal to what you think the maximum choice of *X* amongst the rest of the group will be: *to be as committed to the common good as you think the* ***least*** *committed person will be.*

The problem here, which Rousseau was keen to bring to the surface, is that the outcome depends on the degree of optimism within the group. If it is high (i.e. players expect a low μ), their optimism will be confirmed as each will choose a low *X* (and, therefore, μ will turn out to be low). But if they are pessimistic, anticipating a high μ, their fears will be realised when individual players, fearing that someone else will choose a high *X*, choose a high *X* themselves. This dependence of the outcome on average expectations can be thought of as the *Power of Prophesy*.

In analytical terms, we have two results: (a) there exists an infinity of possible Nash equilibria and (b) which of those will emerge depends, largely, on the average degree of optimism within the group. To see (a), recall that your best reply to the prediction that $\mu=\mu^*$ is to select $X_i=\mu^*$. That choice corresponds to a pure strategy Nash equilibrium because: When you believe that some one in the group will expect the maximum *X* to equal μ^*, ***and*** others believe that you believe that, ***and*** you know that they believe this, and so on and so forth, ***then*** everyone is expecting everyone else to choose μ^* ***and*** the best reply of each to his/her expectations is to choose $X_i=\mu^*$. By definition, $X_i=\mu^*$ is a Nash equilibrium. But this logic applies whatever the actual value of $X_i=\mu^*$ within the interval [0,1]. In other words, we have ended up with a continuum of Nash equilibria on the interval [0,1].

Turning to result (b), the likelihood of a mutually advantageous equilibrium (i.e. one with a low μ) depends not on how rational the players are but on how optimistic they are (as Rousseau would have it). One might argue that $X_i=\mu^*=0$ stands a better chance because it is clearly mutually advantageous as it yields the largest payoff for all (or, in the language of economics, it is Pareto optimal). But there is another argument favouring a value of that $X_i=\mu^*$ closer to 1: Choosing a high X_i (e.g. putting all your energy into catching a stag communally) is particularly risky since, if *one* person absconds by choosing his or her X=1 (e.g. abandons the stag hunt and catches a rabbit instead) then you are left with a nasty payoff equal to $P_i = 1-3 = -2$ (i.e. in Rousseau's terms, an empty stomach as the stag will escape and you will have nothing to eat). Thus, all it takes to wreck the *stag hunt,* and send each player hunting puny rabbits, is a whiff of pessimism regarding the chances of sticking together like a solid team. This is not too dissimilar from John Maynard Keynes' view of the determinants of an economy's aggregate investment: At the first whiff of an impending recession, investors go on an investment strike and the recession occurs (the *Power of Prophesy* again).

2.6 NASH EQUILIBRIUM IN MIXED STRATEGIES

2.6.1 The scope and derivation of Nash equilibria in mixed strategies

In this section we take up the issue of indeterminacy with respect to the Nash equilibrium concept as a solution to all games. We have noted one difficulty in this respect: the existence of multiple Nash equilibria in pure strategies. Games 2.13-16 are some of the more famous games that exhibit this problem. There is another kind of difficulty. Some games do not have a Nash equilibrium in pure strategies. Game 2.19 is an example. Lastly, 2.20 is the prisoner's dilemma (the most famous game featuring a unique pure strategy equilibrium) which is reproduced here for the purpose of comparison.

	C1	C2	
R1	0,0	$^{+}$3,1^{-}	¾
R2	$^{+}$1,3^{-}	0,0	¼
	¾	¼	*NEMS*

Game 2.13
BOS: battle of the sexes

	C1	C2	
R1	-2,-2	$^{+}$2,0^{-}	⅓
R2	$^{+}$0,2^{-}	1,1	⅔
	⅓	⅔	*NEMS*

Game 2.14
Hawk-Dove or Chicken

	C1	C2	
R1	$^{+}$1,1^{-}	0,0	¾
R2	0,0	$^{+}$3,3^{-}	¼
	¾	¼	*NEMS*

Game 2.15
Pure coordination

	C1	C2	
R1	$^{+}$1,1^{-}	2,0	½
R2	0,2	$^{+}$3,3^{-}	½
	½	½	*NEMS*

Game 2.16
Stag Hunt or Common Assurance

	C1	C2	
R1	$^{+}$1,0	0,1^{-}	½
R2	0,1^{-}	$^{+}$1,0	½
	½	½	*NEMS*

Game 2.17
Hide and Seek

	C1	C2	
R1	$^{+}$1,1^{-}	$^{+}$4,0	*1*
R2	0,4^{-}	3,3	*0*
	1	*0*	

Game 2.18
The Prisoner's Dilemma

Nb. Nash equilibria in pure strategies are highlighted; e.g. (R1,C2) and (R2,C1) in Game 2.15. Nash equilibria in mixed strategies (NEMS) are denoted by the probabilities on the margin of each game. E.g. in Game 2.17, NEMS suggests that R1 is played with probability ½.

So, what should rational agents, who have swallowed the doubts of the previous section and remain wedded to the Nash equilibrium concept, do here? One answer turns on the use of *mixed strategies*. In Section 2.2.1 we defined two types of strategies: *pure* and *mixed*. Whereas the former concerned specific moves (e.g. 'Play R1'), a *mixed strategy* boils down to acting *as if* by tossing a suitably weighted coin between your available pure strategies (e.g. 'Play R1 with probability *p* and R2 with probability 1-*p*'). It may sound bizarre at first, but on second thoughts there are many occasions when randomisation makes perfect sense for two interrelated reasons. First, when you have no idea what to do, you do choose (perhaps subconsciously) *as if* at random; and, secondly, perhaps even more significantly, there are situations in which there is a strategic advantage to be had from keeping the opposition guessing (see Box 2.7 for some examples).

But how do people *actually* randomise? Surely they do not go around tossing coins and throwing dice! Randomisation can be implicit. Player R may choose R1 if the first person that

walks through the door is a smoker, or if the first car she sees out of the window is red, white or blue. The randomisation can even be subconscious, as when we somehow choose to locate ourselves in one position on a railway platform despite the fact that we do not know which is the optimal spot. We do not need to be specific about the exact mechanism by which agents randomise. For instance, government ministers often form committees whose job is to make a recommendation to the minister on some sensitive issue. This is not unlike choosing a policy via randomisation where the politician determines the probability of each possible conclusion by handpicking the committee's members and then allowing it to work independently to a conclusion.

The question thus becomes: Given that we can conceive of randomising, is there any way we can deduce logically the *probabilities* with which rational players might choose between their pure strategies? And if there is, does this help with the problem of indeterminacy in games like those above?

The answer to the first of these questions is yes, if we use the Nash equilibrium concept.The answer to the second question is also affirmative as the associated Nash equilibria in mixed strategies (NEMS) can help with the indeterminacy problem.

Box 2.7

WHY USE MIXED STRATEGIES?

The idea of mixed strategies often strikes people as a little strange. Do people really ever behave in this way by mixing probabilistically some pure strategies, by deciding what to do on a metaphorical toss of a coin? We shall discuss the possible foundation for this type of behaviour in more detail later. For now, some examples may help to bring the general idea to life. They turn on the strategic advantage of being unpredictable.

Consider a bowler in cricket or a pitcher in baseball. Both types of player can have a fast and a slow ball in their repertoire and each type of ball is most effective when it is not expected. For instance, the slow ball in cricket, when slipped in among a string of fast balls, is often the one which gets the batsman out. How, then, is a bowler to achieve this unexpected effect? If the bowler always selects a pure strategy (a type of ball) in some predictable way then this will not cause surprise because the batsman will learn how the bowler decides on the type of ball; whereas if the bowler mixes the strategies by selecting the type randomly (by, say, a mental toss of a coin) then the batsman can never be quite sure which ball is coming down the pitch (just as no one can be sure how the toss of a coin will turn out).

Should you bluff in poker? If you were always to bluff, then the bluff would not work because it would have been anticipated and your opponents would call you all the time. On the other hand, if a bluff were never a good pure strategy, then it would *not* be anticipated; but if it were *not* anticipated, then a bluff would *always* work! This obvious contradiction leads to a simple conclusion: Rational players must mix, with a certain probability, their pure strategies. Put simply, they bluff at random.

In a similar fashion, have you ever wondered why airlines are reluctant to tell you how many stand-by seats are available? Presumably they want to encourage marginal travellers but they do not want at the same time to encourage any of their regular passengers to switch to stand-by tickets, as they might if they knew with certainty whether they can pick up a stand-by ticket.

As an illustration, consider Game 2.13, with its two Nash equilibria in pure strategies (R1,C2) and (R2,C1). The rationale for NEMS and its derivation can be set out in the following fashion.

(1) Indifference: Neither player has any reason to think that one of the available pure strategy Nash equilibria is more likely to emerge than the other. *They are both equally plausible* and, therefore, players have no reason to choose one pure (Nash) strategy over another. They are *indifferent* between the two.

(2) Randomisation: When indifferent between different pure strategies, players choose between them random. Moreover, *the chosen randomisations must be consistent with* **(1)** *above.*

Proposition: The NEMS of Game 2.15 is one where R chooses R1 with probability ¾ and R2 with probability ¼. Similarly, C chooses C1 and C2 with probabilities ¾ and ¼ respectively.

Proof: Assuming that players are maximisers of expected utility, axiom **(1)** above means that, for R to be indifferent between R1 and R2, R's expected utility returns from R1 must be identical to those from R2, or ER(R1)=ER(R2).[12] For exactly the same reasons, axiom **1** implies that C's expected utility returns from C1 must be the same as his expected utility returns from C2; that is, ER(C1)=ER(C2). Turning now to axiom **2** above, the indifference of R and C toward their available strategies means that each must randomise. Suppose that R expects C to select C1 and C2 with probabilities q and 1-q respectively. Meanwhile C expects R to select R1 and R2 with probabilities p and 1-p respectively. Can we say anything about these probabilities? We can do more than that: We can compute the precise values of p and q which are *uniquely* consistent with axiom **1**! This is how:

Our players' randomisation probabilities (p,q) must be such that ER(R1)=ER(R2) and ER(C1)=ER(C2). Otherwise, they will not be consistent with axiom **1**. From here on, it is a matter of simple arithmetic to show that p=q=¾. First, we note that when choosing R1, player R will collect either payoff 0 or 3; the former with probability q (i.e. the probability with which C will select strategy C1) and the latter with probability 1-q. Were she to choose R2, she would anticipate payoff 1 with probability p and 0 with probability 1-p. All in all,

$$ER(R1) = 0\times q + 3\times(1-q) = 3-3q \qquad (2.1)$$

$$ER(R2) = 1\times q + 0\times(1-q) = q \qquad (2.2)$$

As we have already proven, R's randomisations are consistent with axiom **1** only when ER(R1)=ER(R2), or 3-3q= q; that is, when q = ¾. It is now straightforward also to show that p = ¾.[13]

In summary, if *both* players are to have no clue as to which of their pure strategies they ought to play, each must expect the other to play her/his first strategy (R1 for and C1 for C) with probability ¾. These expectations will, of course, only obtain if R and C opt for mixed strategies p = ¾ and q = ¾ respectively. All we now have to do is to prove that these mixed strategies correspond to a Nash equilibrium.

To see that they do, recall the definition of a Nash equilibrium: it comprises a set of strategies, say, s_R for R and s_C for C, such that when R expects C to choose s_C she has *no* incentive to choose some strategy other than s_R, while if C expects R to choose s_R he has *no* incentive to choose some strategy other than s_C. When this applies, the combination of strategies (s_R,s_C) constitutes a Nash equilibrium. When we deal in mixed strategies, strategies (s_R,s_C) are the probabilities with which our players will choose each pure strategy. Is the combination of mixed strategies p = ¾ and q = ¾ a Nash equilibrium (a NEMS)?

Consider what happens when R expects that C will play q = ¾. Does she have an incentive to play a strategy *other than* her p = ¾? No, she does not. By definition, when R anticipates mixed strategy q = ¾ by C, ER(R1)=ER(R2) and so R is well and truly indifferent between all the strategies (pure and mixed) available to her.[14] In other words, when she expects C to set q = ¾, R cannot do better by abandoning her strategy p = ¾. And vice versa. Meanwhile when C expects R to set her p = ¾, ER(C1)=ER(C2), C is indifferent and thus he has no incentive to abandon his strategy q = ¾ either. Thus, the combination of mixed strategies p = ¾ and q = ¾ is a Nash equilibrium; a *Nash equilibrium in mixed strategies* or NEMS.

The same argument can be made to resolve the indeterminacy in Games 2.14, 2.15 and 2.16 (and the associated NEMS are set out in those tables). The argument here turns on a reflection, courtesy of the *Harsanyi-Aumann doctrine* that, since these games are symmetrical, one should not expect that rational players will come to different conclusions with respect to how play this game. Unlike the Nash equilibria in pure strategies in these games, the NEMS is symmetrical. In

other words, there is no difference from a strategic point of view between players R and C and so we should not expect rational players, knowing this, to draw different conclusions about how each will play. Thus, we have only one symmetrical Nash equilibrium: the NEMS.

NEMS also helps resolve the indeterminacy of the type found in Game 2.17 because although there is no Nash equilibrium in pure strategies, a NEMS does exist. Indeed, it is one of the important theorems in game theory that every game has a Nash equilibrium in either pure and/or mixed strategies. So the Nash equilibrium concept can be applied to all games in the sense that it will always provide a candidate solution(s) for the actions taken by rational players. This is one of its strengths as a solution concept.

NEMS also helps in some degree with the problems noted in Section 2.5.3 with respect to the Nash equilibrium in pure strategies in Games 2.9 and 2.12. There is a NEMS in 2.9 which dissolves the worry associated with the Nash equilibrium in pure strategies: It has each playing randomising between their 1st and 3rd strategies with probability ½ and thus avoids the unattractive Nash equilibrium in pure strategies altogether.[15] In Game 2.12, there is a NEMS that mixes the 2nd and 3rd strategies for each player with probabilities of 2/3 and 1/3 and so brings into play what seemed like an attractive but non-Nash (in pure strategies) courses of action, R3 and C3. In so far as NEMS does help in this respect, of course, it adds to the overall *indeterminacy* as there are already two Nash equilibria in pure strategies in this game. Nevertheless, as the illustrations demonstrate, the idea of NEMS adds considerably to the power of the Nash equilibrium concept .

There is, however, an important difference between a Nash equilibrium in pure strategies and a NEMS which will have been apparent in the derivation above. There is a sense in which *NEMS* are *weaker* because, whereas a Nash equilibrium in pure strategies can be built upon a layer of strong preference, a NEMS is founded on indifference. To see this, let us consider a game with a single pure strategy Nash equilibrium; e.g. Game 2.12. In that game R has a strong preference to play R1 if she expects C to play C1, while C has a strong preference for C1 if he expects R1; put differently, if they anticipate the Nash equilibrium they will move resolutely toward it.[16] By contrast, as we have seen above, a NEMS feeds on indifference: if, in Game 2.13, one anticipates that the other will stick to NEMS, one has no incentive to stick to it too. It is a Nash equilibrium only to the extent that one has *no* reason *not* to stick to it. As we shall see below, this weakness causes NEMS to be rather 'unstable'.

Nash Equilibrium in Mixed Strategies (definition)

(1) A NEMS comprises a set *M* of mixed strategies, one for each player. A mixed strategy is a well-defined probability distribution (assigning a precise probability to each available pure strategy of the player such that all probabilities lie within the interval [0,1] and their sum equals 1). These mixed strategies are in a Nash equilibrium when no player can improve her/his expected utility by opting for a strategy other than her/his mixed strategy in *M*.
(2) All finite *normal (or matrix) form* or *static* games have at least one NEMS.
(3) If a 2x2 game has more than one pure strategy Nash equilibria, then there is a unique NEMS that assigns a positive probability to each player's pure strategy.

NEMS in Games 2.13 to 2.17

(1)&(2): Game 2.17 has no pure strategy Nash equilibrium; in such games according to (1) and (2) above, there exists a NEMS and it assigns a positive probability (not necessarily the same) to every strategy of both players.

(4): In Games 2.13-2.16, there are two Nash equilibria in pure strategies and each pure strategy corresponding to such a Nash equilibrium is assigned a positive probability. By contrast, in Game 2.12, where there are also two Nash equilibria in pure strategies, the pure strategies corresponding to the latter are assigned positive

probabilities (½ each) but strategies R3 and C3 are assigned zero probabilities as they do not correspond to some pure strategy Nash equilibrium.

2.6.2 The reliance of NEMS on CAB and the *Harsanyi doctrine*

In the previous section we derived the NEMS concept by suggesting that players might randomise when the notion of a pure strategy Nash equilibrium offers no helpful advice on what to do. The key to NEMS was that *there is only one pair of randomisations which is consistent with each **not knowing** what will happen in a game when each **knows** what the other will do*. The last sentence holds the key to NEMS and so requires careful examination.

Starting with the last part (*each **knows** what the other will do*), it refers to the axiom of CAB and the *Harsanyi doctrine*: If both players have the same information, they will come to the same conclusion as to what each should do. Thus, whatever it is that R will do in a game like, say, Game 2.15 C will anticipate it; and vice versa. So, NEMS is founded on the assumption that R knows with certainty what (mixed) strategy C will select and vice versa.

Turning to the first part of that convoluted sentence (*each **not knowing** what will happen*), how can it be consistent with the *Harsanyi doctrine* which, the last paragraph, told us that informational equity (or symmetry) must mean that "R knows what strategy C will follow and vice versa"? The answer is simple: *Players randomise!* Therefore, it is perfectly plausible for R not to know exactly which outcome will obtain when C randomises between his pure strategies while *simultaneously* she has perfect knowledge of the probability with which C will randomise. E.g. it is like you know with certainty that I shall choose strategy "Toss a fair coin" (and thus that a head will come up with probability ½) without knowing in advance whether a head will come up. (Box 2.8 illustrates another rationalisation of NEMS that again relies on the CAB)

So, as it turned out in the previous section, when (a) neither knows what to do (i.e. which pure strategy to opt for) and, simultaneously, (b) each knows what the other will do (i.e. they know one another's mixed strategies) there is only one mixed strategy per player: their NEMS! This is a fascinating result which, once more, displays the brilliance of Nash's idea: It is akin to saying that, *because it is impossible to know what rational people will do in these games, we know what they will do*! Put differently again, we may not know the pure strategy that they will choose but, the fact that we know that we cannot know this, makes it possible for us logically to derive the *probability* with which they will choose one pure strategy rather than another.

Box 2.8

HOW *CAB* UNDERPINS *NEMS*

Consider Game 2.13 (it could be any of those lacking a unique Nash equilibrium in pure strategies) and imagine that C is *convinced* (for some unspecified reason) that R will choose R1 with a probability p' such that $1>p'>¾$ (nb. $p=¾$ is the one given by NEMS). Moreover, suppose that the above is *common knowledge*: namely, C *believes* that R knows that C is convinced etc. etc. that R will choose R1 with probability $1>p'>¾$.

Question: Is this a belief that C can rationally entertain under CKR?

Answer: Under CAB it is not. But, if C has doubts about CAB (that is, if C thinks there is a chance that his estimate of p will *not* be the same as R's estimate of q) there is no reason why the expectation that p will lie between 1 and ¾ is irrational.

Proof: What is C's best reply to the expectation that C anticipates R's mixed strategy to be $1>p'>¾$? C's expected returns from playing C1 equal $3(1-p')$ whereas his expected returns from playing C2 equal p'. Let us denote the difference between the two d^C_{12} = ER(C1)-ER(C2). Thus, $d^C_{12} = 3(1-p') - p' = 3-4p'$. (Note that when $d^C_{12} > 0$, player C expects to get more on average if he chooses C1 instead of C2.) From this expression, it is clear that if C expects $p'>¾$, difference $d^C_{12}>0$, and thus C has a dominant strategy: *Play C1!* We notice immediately that this is inconsistent with CAB. For CAB means that players choose the (mixed) strategies that are best replies to the *actual* strategies of their opponents. So, if C is about to choose C1 because he expects $1>p'>¾$ (see above conclusion), then R should not be choosing $1>p'>¾$ but, instead $p'=0$. Thus, C's belief that p might exceed its NEMS value of ¾ is inconsistent with CAB. Similarly with any belief he might entertain that p is less than ¾. In fact the only belief that

is *not* inconsistent under CAB is the one associated with the NEMS: $p=q=¾$. Of course there is nothing 'wrong' with beliefs other than $p=q=¾$ if we introduce the possibility that players will not take it for granted that all beliefs (theirs as well as their opponents') must be consistently aligned. It turns out that we have a simple choice: Either we assume CAB, and end up with NEMS, or we do not, accepting that this type of game is *indeterminate*.

2.6.3 Aumann's defence of CAB and NEMS

The connection between CAB and the derivation of NEMS is intimate but it is not obvious that CAB answers the troubling question regarding the latent instability of a NEMS. The source of the instability, noted above, is the apparent absence of a strong incentive to play according to NEMS once one expects one's opponent to do so. As mentioned above, the best that can be said is that one does *not* have an incentive *not* to play NEMS. But does this lend NEMS sufficient pulling power? Does it render it *stable*? Or will players, once they have computed their NEMS probabilities, drift away and mix their pure strategies with probabilities different to those recommended by NEMS? After all, once the NEMS obtains, a player's expected returns are the same *whatever* she or he does. So, why stick to NEMS?

The presumption of CAB does not seem to answer this question. Aumann (1987), however, rebuts the thought that NEMS are weak equilibria prone to instability by interpreting the probabilities of a NEMS differently. They are not to be interpreted as the individual's probability of selecting one pure strategy rather than another. Rather, probability p, with which R plays R1, should be thought of as the *subjective belief* which C holds about what R will do; not R's *actual* randomisation. Likewise probability q (the probability that C will choose C1) reflects the *subjective belief* of R regarding what C will do. So players will do what players will do. And probabilities p and q (provided by NEMS) simply reflect a consistency requirement with respect to the subjective beliefs each holds about what the other will do. The requirement is the following:

(1)	Given R's beliefs about C (q), then C, when forming an assessment about R (p), should not believe that R will play any strategy which is not optimal relative to those beliefs (q).
(2)	Given C's beliefs about R (p), then R, when forming an assessment about C (q), should not believe that C will play any strategy which is not optimal relative to those beliefs (p).

In Game 2.13, R1 and R2 are equally attractive strategies, each corresponding to a Nash equilibrium in pure strategies. They must, therefore, both be rationally playable. But, Aumann notes, there is only one value for q (=¾) which could rationalise the play of *both* R1 and R2. For if $q>¾$, then R1 should never be played. And if $q<¾$ R2 should be eliminated. Thus $q=¾$ is uniquely compatible with both R1 and R2 remaining in contention and making the assessment of p by C something different from either 0 or 1. Similarly, there is only one value for p (=¾) which would keep C1 and C2 in play (and so make the assessment of q by R something other than either 0 or 1). [Note the similarity between this argument and the one in Box 2.8.]

The crucial question, however, which this defence of NEMS overlooks, as it stands, is how each player comes to know the beliefs that the other holds about how he or she is going to play. For instance, in (1) how does C come to know what are R's beliefs (q) about how she will play? Of course, he can work it out from (2) provided C's beliefs about R (p) are known to R. But this merely rephrases the problem: How are C's beliefs about R known to R?

The answer Aumann offers to this conundrum turns again on the *Harsanyi doctrine* and the CAB assumption. In our game player C will choose either C1 or C2 and one can think of some kind of event pushing C in one direction or the other (the event can be anything, it is simply whatever psychologically moves one to action in these circumstances). So there is an event of

some sort which will push C in one direction or another; and following the *Harsanyi doctrine*, it is argued that both players must form a common prior probability assessment regarding the likelihood of the event yielding C1 or C2. So, both R and C must entertain the same belief regarding how C will act. Of course, both players also know that there is no event which could occur which would make C take an action *sub*-optimal relative to his beliefs. Thus, the value of q in (1) must also satisfy the condition set on q in (2). In other words, the q in (2), which comes from recognising that C's behaviour is optimal, must be the same as the q in (1), because otherwise the two players would not be drawing the same inference from the same information set. Likewise the beliefs that C holds about R (p) must be the same as the beliefs that R holds about herself and they both know that any admissible belief must be consistent with each maximising their expected utilities.

In many respects this is an extraordinary argument. As Bob Sugden (1991) remarks

> …by pure deductive analysis, using no psychological premises whatever, we have come up with a conclusion about what rational players must believe about the properties of a psychological mechanism. (p.798)

To summarise, while NEMS *always* depends on CAB and the Harsanyi-Aumann doctrine, the Nash equilibrium concept in pure strategies does not. As we saw in Sections 2.3 and 2.4, some Nash equilibria in pure strategies can be derived through dominance reasoning alone. The fact that NEMS *always* relies on CAB and the Harsanyi-Aumann doctrine is a weakness in our view for the reasons that were set out in Section 2.5.3.

Perhaps intriguingly, NEMS could receive support from another quarter that was foreshadowed in the earlier discussion of section 2.5.3: Kantian thinking! Such an admittedly idiosyncratic argument in favour of NEMS would draw on the part of Kant which demands that agents only hold beliefs that they know can be held by all without generating internal inconsistency. As already argued in Section 2.5.3, this might license all Nash equilibria including NEMS. In effect under this interpretation, CAB is explained by a Kantian 'universal principle' which in turns engenders the NEMS as the only set of beliefs which is both mutually consistent and consistent with both players being uncertain about what action will be undertaken.[17] By contrast, a Nash equilibrium in pure strategies is both mutually consistent and consistent with each player knowing for certain which action will be undertaken.

Box 2.9

EVIDENCE IN FAVOUR OF *NEMS* FROM THE BASEBALLGROUND AND WIMBLEDON

Have you ever heard the argument that footballers or tennis players of the past were better than today's crop? Stephen Jay Gould was quite fed up with the argument that baseball players of yesteryear were so much better than contemporary ones that he did some research. There is, he found, no doubt that best batting averages have been declining for the past seventy years (see Gould, 1985). However, if we look not only at the best averages but at all the averages, a quite different picture emerges. What seems to have happened is that, though the best averages have declined, the worst averages have improved! In other words, the data points to a general decline in *variation*. Why? Gould's argument is that, in previous epochs, the difference between players' strategies and skills was much greater than today. Professionalisation and improved training meant the elimination of players whose strategy and skills fell very short of those required to play well against a good player. In the language of game theory, players converged towards something like a Nash equilibrium, where one's strategy/skills is the best (or a pretty good) reply to another's. And since in ball games one is forced to adopt mixed strategies (e.g. a striker always taking a penalty to the keeper's right hand side will be ineffective), Gould's argument offers some support to the NEMS concept.

Turning to tennis, two researchers from the University of Arizona studied the service and return strategies of top-notch players at Wimbledon.[18] Their theory was that if the server manages to wrong-foot the receiver, his chances of winning the point are increased. And vice versa. Hence the nature of the game is such that players must choose mixed strategies (see Box 2.7). But do they choose *Nash equilibrium mixed strategies* (NEMS)? If they do, then the server should win with the same probability whether he serves to the receiver's left or right. The researchers' data comprised the results of serves in ten matches; each match comprised more than 100 points involving the same players. Clearly there was plenty of data to test the proposition. The conclusion was

unequivocal: the probability of winning a serve was almost entirely independent of whether the server served on the receiver's forehand or backhand. The mixed strategies of good players had converged on the unique NEMS randomisation.

Though supportive of NEMS, the above raises an important question that will occupy us for the rest of this book: The data from baseball and tennis shows that NEMS may become established as a result of an evolving process (involving learning) in *historical time*. However, NEMS was introduced in this Chapter in the context of one-shot games in which players must work out their strategies in the pristine isolation of logical time. Will historical time offer NEMS the support that it now only gets from CAB, or will it throw out new challenges for game theory? Read on!

2.7 CONCLUSION

The central solution concept in game theory is the Nash equilibrium. We hope to have shown that there are games in which a Nash equilibrium is the obvious 'solution' as long as players are rational and have sufficient trust in each other's rationality (CKR). For this to be the case, the Nash equilibrium must be unique (in pure strategies) and be reachable through a stepwise elimination of strategies that rational players would discard (i.e. dominated strategies). However, we have also shown that there are numerous interesting interactions in which the Nash equilibrium cannot be reached logically by appealing to the assumptions of rationality and CKR. Something more is required.

That *something* might be supplied by extra data on the social context in which the game is played, or the history of the people who play it (since the games in this chapter had no history themselves; they are of the one-shot type) or/and a different conception of rationality. However, game theory strives for a solution that draws on what is in the game allied with the assumptions that agents are rational and operate under CKR. The *something* extra that keeps faith with such a purely rational approach is the assumption of 'common priors' and it guarantees the Nash equilibrium in all cases. It is based on the Harsanyi doctrine that *people must form the same (probabilistic) assessment of what is likely to happen when they go to work with the same information* and incorporates Aumann's argument that rational people could not '*agree to disagree'*. The assumption of 'common priors' thus delivers the *consistently aligned beliefs* (CAB) which characterise all Nash equilibria.

When it is made clear that the general claim made on behalf of Nash's equilibrium is founded on CAB, it becomes obvious that some of the debates at the foundations of game theory touch on matters regarding the treatment of uncertainty which have always been central to debate in political economy. Thus the debates about what R thinks that C believes that R will do in some silly game are not as frivolous as they might seem. At their heart lie some profound philosophical, even political, disagreements regarding human reasoning, psychology, history and the relation between private motives and public norms.[19] In particular, we doubt that CAB can generally hold because both the Harsanyi and the Aumman parts of their combined defence of the CAB (or common priors) assumption are only likely to hold in special cases. For instance, there does not seem to be a general model of rationality which supports the *Harsanyi doctrine*: that is, a model which can be applied to all possible settings. Likewise, it is difficult to see how the *Aumann conjecture* can hold when information is costly to acquire.

Setting the doubts aside for the time being, a game theory where the Nash equilibrium is the key concept is still going to have some difficult questions to answer if it focuses only on equilibria in pure strategies. This is because there are many games where there are either no Nash equilibria in pure strategies or there are many. We have seen how the idea of NEMS can be helpful in this context. Moreover, there is always a NEMS even when there is no Nash equilibrium in pure strategies. In symmetrical games, there is a unique NEMS that fully reflects this symmetry and which, it can be argued, ought to commend itself to rational players.

Even setting the doubts about the Nash equilibrium concept aside in this way, problems still remain because not every game with multiple Nash equilibria in pure strategies is a 2×2 symmetrical interaction, and not every game with multiple Nash equilibria has a unique NEMS. So indeterminacy is still plaguing us even if doubts over NEMS are set aside. The next chapter presents the game theorists' attempts to deal with this *Indeterminacy*. Its theme is known in the literature as the project of 'refining' Nash's equilibrium (or the *Refinement Project*). The purpose of these 'refinements' is to increase the sophistication of the analysis so much so that we can discriminate between more and less 'likely' Nash equilibria. The great hope is that we shall be able to select a single one, among the many Nash equilibria, as the most plausible.[20]

PROBLEMS

2.1 Find all the Nash equilibria (in pure as well as mixed strategies) in the two 3X3 games below.

	C1	C2	C3
R1	5,0	-1,-5	10,-1
R2	-1,-1	0,5	-1,-2
R3	-1,1	-2,-1	6,6

	C1	C2	C3
R1	3,3	-100,-100	9,0
R2	-100,-100	3,3	0,0
R3	0,9	0,0	8,8

2.2 Consider the following *N*-person game: Each player chooses any real number between (and including) 1 and 100. The player whose choice of number is closest to *the average choice of integer* (among all *N* players) divided *by 2* will win €1000. Find all the Nash equilibria. What if the winner is the player whose choice of number is closest to *minimum choice multiplied by 2*?

2.3 Apply the *Successive Elimination of Strictly Dominated Strategies* to the following game.

	C1	C2	C3	C4
R1	100,97	98,98	98,100	1,99
R2	99,100	97,99	97,98	0,-1
R3	98,97	100,99	100,98	-1,100
R4	-1,-1	99,0	99,1	2,2

2.4 Which statement is correct with regard to a static game?

(a) For a strategy to be rationalisable, it must correspond to a Nash equilibrium outcome
(b) A Nash equilibrium in mixed strategies (NEMS) takes the form of a lottery amongst rationalisable and non-rationalisable pure strategies
(c) Suppose there exists no pure strategy Nash equilibrium. There will at least exist two rationalisable strategies available to each player

2.5 (*For students of economics only*) Suppose a duopoly where two firms, 1&2, select their output q_1 and q_2 once and independently of one another (i.e. non-co-operatively) in order to maximise their respective profit. Assuming that (a) the price is set automatically by the market according to (the inverse demand) function $p = 1000 - q_1 - q_2$, and (b) each firm has the same fixed cost (F) and constant marginal cost equal to 10, then show that there exist a pair of quantity strategies (q_1, q_2) which constitute this game's a unique Nash equilibrium in pure (quantity) strategies. Moreover, show that these strategies are also the game's unique *rationalisable* strategies.

CHAPTER 2 NOTES

[1] Indeed whenever we refer to 'strategies' we shall imply 'pure strategies'; mixed strategies will return later and will always be referred to in full.

[2] Note how in Game 2.6 the ($^+$) and the ($^-$) markings all fall in the matrix' second row and the second column; thus, R2 and C2 constitute R's and C's strictly dominant strategies.

[3] Even though 1st order CKR suffices in order to bring about the Nash equilibrium outcome, it does not suffice to inspire confidence in the players that the equilibrium will materialise with certainty. The reason is that C chooses C2 only because he does not expect R to go for R2. Still, he does not know, so far, whether R knows that he knows that she is rational and therefore C is not sure whether R will opt for R1 or R3 (note that C2 is a best reply to both). If, however, we assume 2nd-order CKR, suddenly it becomes clear to him that R will choose R1. For if C believes that R believes C to be instrumentally rational, then C does not expect R to expect him to play C3, in which case he does not expect R to play R3. It is clear that 2nd-order CKR fixes C's beliefs on the unshakeable expectation that R will choose R1. Notice that R's beliefs are still not that precise. She knows (through 1st-order CKR) that C3 is not on the cards, but is not sure as to whether C1 or C2 will be played. This uncertainty makes no difference to her strategy since, in either case, her best reply is R1. But before R can form a certain view on whether C will go for C1 or C2, we need 3rd-order CKR. That is, R must know that C's thoughts are subject to second-order CKR, or to put it differently R must expect C to expect R to play R1 before she can be certain that C will choose C2. This is the same as assuming 3rd-order CKR.

[4] Consider, for example, the following game:

	C1	**C2**	**C3**	**C4**
R1	$^+$5,10	0,11	1,10	10,20$^-$
R2	4,0	$^+$1,0	2,-1	20,1$^-$
R3	3,2	0,4$^-$	$^+$4,3	50,1
R4	2,93$^-$	0,90	0,91	$^+$100,92

The process of elimination begins with C3 (dominated by C2) and throws out, in succession, strategies R3 (dominated by R4 now that C3 is out) and C2 (dominated by C4 now that R3 is out). This leaves R with R1 and R4, and C with C1 and C4. However, no more strategies of either player can be eliminated further. Each player can rationally play one of his/her two remaining strategies and the game is indeterminate.

[5] This is the very meaning of weakly dominated strategies. Though they are bad replies to a range of the opponent's choices, there exists at least one choice by the opposition such that the weakly dominated strategy does not do worse than the player's other strategies.

[6] Note once more that, because SEDS is an iterative algorithm, it is often referred to in the literature as *iterated (strict and/or weak) dominance reasoning*.

[7] Note that each row in Game 2.9 contains a ($^+$) mark and each column a ($^-$) mark.

[8]Nash proved that a *Nash equilibrium* exists in all finite games. However, in many games a Nash equilibrium exists only in terms of a mixed strategy (we shall return to this in Section 2.6.1). For now it suffices to state that, although Nash cannot prove that all finite games have a Nash equilibrium in pure strategies, he *did* prove that every finite game features at least one Nash equilibrium *in mixed strategies* (see Section 2.1.1 for a reminder of mixed strategies).

[9] Of course, the operative word here is *almost*. For there is a fundamental difference between a Socratic or Hegelian dialectic and the Nash equilibrium's mutual-confirmation of predictions regarding your opponent's strategic choice. In the former, one is ill-defined *a priori* as a *person* until she/he reflects either (a) through incessant conversation with others, or (b) infinitely into the eyes of another human being. In Nash's case, on the other hand, the person is perfectly well defined *a priori* (as long as she or he has a well defined expected utility function) and the infinite reflection that occurs when one meets another person helps fix not her or his personality but, less grandly, his or her beliefs about the game's outcome.

[10] This term was first used by Bob Sugden – see Sugden (1991a)

[11] The fact that Games 2.11 and 2.12 share the same strategic structure can be spotted immediately by noticing that the ($^+$) and ($^-$) markings are located in precisely the same cells in both games.

[12] Note, for example, that when ER(R1)>ER(R2), R1 does better, on average, than R2 and ought to be preferred, at least by expected utility maximisers.

[13] ER(C1) = $0\times p + 3\times(1-p) = 3-3p$ and ER(C2) = $1\times p + 0\times(1-p) = p$. Hence, equality ER(C1)=ER(C2) yields p=3/4.

[14] Indeed, she expects the same utility on average regardless of whether she plays R1 with certainty, R2 with certainty, or mixes R1 and R2 with any probability imaginable

[15] The NEMS in Game 2.10 was deemed unattractive in the same way. Since the two games are strategically identical, the puzzle remains concerning why the equilibria that seem plausible in each case are different.

[16] Note that this is not always so. For a Nash equilibrium in pure strategies (just like NEMS) may also rely on weak preference even if it is unique – E.g. outcome (R1,C1) in Game 2.11.

[17] The reason why most game theorists' would rather pass on this Kantian helping hand is that the Kantian 'universal principle' which can potentially ground the alignment of beliefs does not stop there; it extends to an alignment of actions too, such that players can overcome their hypothetical reasoning (i.e. if C plays R1 I am better off playing C2) and replace it with categorical reasoning (i.e. whatever C does it would be better if we were both to choose R2 and C2; thus, I shall do my duty and play R2). Such reasoning simultaneously dissolves the prisoner's dilemma and most game theory with it.

[18] The researchers were Mark Walker and John Wooders and their work is referred to in p.215 of Dixit and Skeath (1999)

[19] Varoufakis (2002/3) connects the critique of CKR and CAB with the debates in political philosophy regarding the nexus between *Rationality* and *Liberty*.

[20] As mentioned above, the leading lights of this *Refinement Project* were John Harsanyi and Reinhart Selten. In recognition of their work, they were awarded jointly with John Nash, the Nobel Prize in Economics in 1994.

3
BATTLING INDETERMINACY
Refinements of Nash's equilibrium in static and dynamic games

3.1 Introduction

3.2 THE STABILITY OF NASH EQUILIBRIA
- 3.2.1 Trembling hand perfect Nash equilibria
- 3.2.2 Harsanyi's Bayesian Nash equilibria and his defence of NEMS

3.3 DYNAMIC GAMES
- 3.3.1 Extensive form and backward induction
- 3.3.2 Subgame perfection, Nash and CKR
- 3.3.3 Sequential equilibria
- 3.3.4 Bayesian learning, sequential equilibrium and the importance of reputation
- 3.3.5 Signalling equilibria

3.4 FURTHER REFINEMENTS
- 3.4.1 Proper equilibria
- 3.4.2 Forward induction

3.5 SOME LOGICAL OBJECTIONS TO NASH, PART III
- 3.5.1 A critique of subgame perfection
- 3.5.2 A negative rejoinder (based on the Harsanyi-Aumann doctrine)
- 3.5.3 A positive rejoinder (based on sequential equilibrium)
- 3.5.4 Conclusion: Out-of-equilibrium beliefs, patterned trembles and consistency

3.6 CONCLUSION
- 3.6.1 The status of Nash and Nash refinements
- 3.6.2 In defence of Nash
- 3.6.3 Why has game theory been attracted 'so uncritically' to Nash?

3
BATTLING INDETERMINACY
Refinements of Nash's equilibrium in static and dynamic games

3.1 INTRODUCTION

The Nash equilibrium concept will be close to vacuous in games featuring multiple Nash equilibria because it will offer no determinate guide to action. This chapter is concerned with how game theory has attacked this apparent indeterminacy. In these circumstances rational players, it seems, will need some way of discarding equilibria if they are to 'operationalise' the Nash equilibrium solution. This sets the agenda for the Nash *Refinement Project* which began life with a notion of 'stability' the purpose of which was to help players discard some Nash equilibria as 'unstable'. We begin the next section with an investigation of this notion of 'stability'.

3.2 THE STABILITY OF NASH EQUILIBRIA

3.2.1 Trembling hand perfect Nash equilibria

A potential refinement was flagged in the previous chapter when we noticed that NEMS are not as 'stable' as the pure strategy Nash equilibria. Stability seems a good way of distinguishing between equilibria for the same reason that it is used in physics: We do not expect unstable equilibria to survive in a world that is even slightly imperfect. Consider Game 3.1 below.

	C1	C2	C3
R1	$^{+}$50,0	$^{+}$5,5^{-}	$^{+}$1,1
R2	$^{+}$50,50^{-}	$^{+}$5,0	0,-1

Game 3.1 – *Two Nash pure strategy equilibria, only one trembling hand equilibrium*

By inspection, we note that there are two pure strategy Nash equilibria in Game 3.1: (R1,C2) and (R2,C1) [again see how the ($^{+}$) and ($^{-}$) markings coincide in these cells]. Strategy C3 is ruled out of any equilibrium since it is strictly dominated by C2 while R2 is weakly dominated by R1 (since, from R's perspective, R1 is just as good as R2 when C plays either C1 or C2 but it is preferable when C plays C3). In this sense (see previous chapter), under 1st-order CKR there is no fear that C3 will ever be played. Thus R is utterly indifferent between R1 and R2 since both yield the same pay-off when C opts for C1 or C2 and so we cannot apparently predict whether R will play R1 or R2.

Most people, in R's shoes, would probably prefer R1 in case some lapse of C's reason causes him to play C3. We are humans after all and if R1 guarantees the same pay-offs to R as R2, but in addition offers her a guarantee against *any* blunder of her opponent (however improbable), surely R would have a slight preference for R1 over R2. However slight that preference, it eliminates one of the two Nash equilibria and, with it,

the embarrassing indeterminacy. This is our first example of a 'refined' Nash equilibrium and it is due to Reinhart Selten (1965) who named it a *trembling hand perfect Nash equilibrium*

More precisely, as long as the probability of a lapse by C is non-zero (that is, an imaginable, though infinitesimal, magnitude), R is better off opting for R1. If C knows this, C's best move is C2 (a best reply to C1) and (R1,C2) appears as the only Nash equilibrium consistent with such a non-zero probability of an accidental selection of (strictly dominated) C3. Game theorists refer to unpredictable lapses of reason, accidental mistakes, tiny errors of judgment, etc. as *trembles*. Interactions in which trembles are possible are known as *perturbed games*. Finally, a *trembling hand* equilibrium is a Nash equilibrium not undermined by infinitesimally small trembles or perturbations.

Perturbed games and trembles (definition)

A *perturbed* version of a game is a version of the game played with *trembles*; namely, games in which there is always some positive, however tiny, probability ε that every strategy will be played by each player. The corollary of this is that no strategy can be chosen with probability 1 (or 0) due to these 'trembles'.

Trembling hand perfect equilibrium (definition)

A *trembling hand perfect equilibrium* is the limit of the sequence of Nash equilibria in *perturbed* versions of the game as the trembles go to zero (i.e. $\varepsilon \rightarrow 0$)

Turning again to Game 3.1, and according to the above definitions, were we to 'perturb', the game through the introduction of trembles, only one of the two pure strategy Nash equilibria would survive *even when the trembles tend to zero*. To see this, let p and q be, respectively, the probabilities with which C will select C1 and C2 ***intentionally*** (with the probability of C3 equalling $1-p-q$). By *intentionally*, we mean that although players *intend* to choose C1 with probability p (and C2 with probability q), inadvertedly, they may choose among their strategies with slightly different probabilities due to 'trembles' (that is, errors). Without any trembles, R's expected returns from R1 and R2 are:

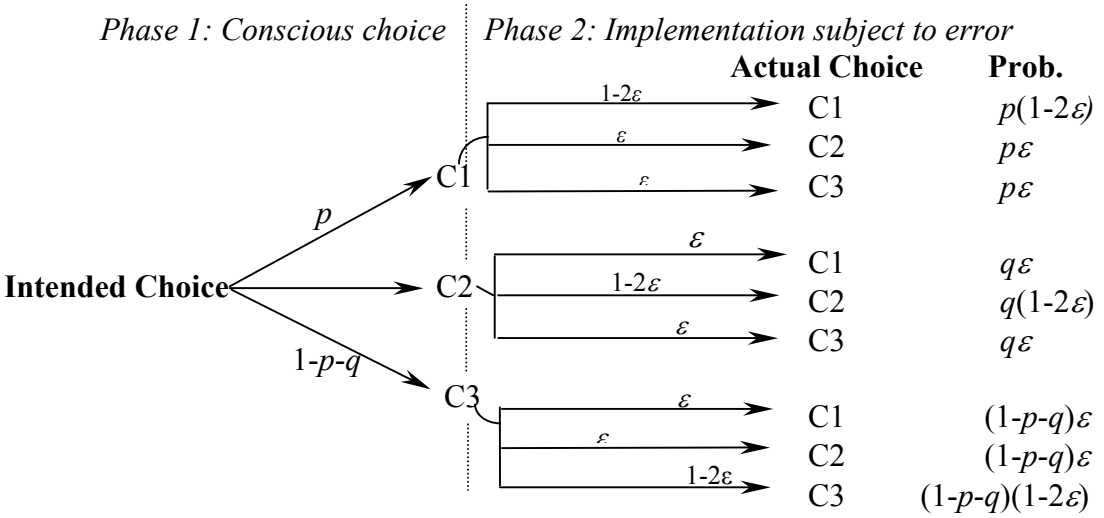
$$ER(R1) = 50p + 5q + (1-p-q) \text{ and } ER(R2) = 50p + 5q$$

Subtracting ER(R2) from ER(R1), we derive R's expected net gain from choosing R1 rather than R2.

$$d^{R}_{12} = ER(R1) - ER(R2) = (1-p-q)$$

Now, suppose that players are prone to small errors or 'trembles' *after* they have formed their strategic intentions. One way of modelling these trembles is to assume two phases in a player's decision-making process. In *Phase 1* the player chooses his or her mixed strategy, spins the mental coin, so to speak, and as a result *forms an intention* to play one strategy. For example, in the game above, player C chooses mixed strategy $(p,q,1-p-q)$ over pure strategies (C1,C2,C3). Suppose now that this strategic randomisation has been played out and it yielded C1, we say that C has formed the intention of playing C1. In *Phase 2* the player implements this choice. It is in this second phase that errors happen. C

may have formed the intention of playing C1 but ends up, accidentally, playing C2 (with probability ε) or C3 (again with probability ε). The probability that he will enact his rationally chosen pure strategy becomes 1-2ε.



The same applies for any choice and, as the above tree-diagram illustrates, the overall probability of actually playing, say, C1 equals: (a) the probability that C will select it intentionally (i.e. p) in *Phase 1* times the probability that there will be no error causing C to deviate unintentionally toward either C2 or C3 in *Phase 2* (i.e. 1-2ε), *plus* (b) the probability that he will choose C2 in *Phase 1* but in *Phase 2* a 'tremble' will force him to err toward C1 (i.e. $q\varepsilon$), *plus* (c) the probability that C will opt for C3 in *Phase 1* but in *Phase 2* a 'tremble' will cause him to select C1 accidentally (i.e. $p\varepsilon$). Summing up,

$$\Pr(C1) = p(1-2\varepsilon) + q\varepsilon + (1-p-q)\varepsilon = p + \varepsilon(1-3p)$$

Similarly, $\Pr(C2) = p\varepsilon + q(1-2\varepsilon) + (1-p-q)\varepsilon = q + \varepsilon(1-3q)$

and $\Pr(C3) = p\varepsilon + q\varepsilon + (1-p-q)(1-2\varepsilon) = (1-p-q) + \varepsilon[(3p+q)-2]$

In view of the above, what should R do? Committed to maximising her expected pay-offs, she will opt for R1 if d^R_{12} = ER(R1) - ER(R2) > 0. This time, in computing her expected pay-offs, she must take account of C's potential trembles. Clearly,

ER(R1) = $50[p + \varepsilon(1-3p)] + 5[q + \varepsilon(1-3q)] + \{(1-p-q) + \varepsilon[(3p+q)-2]\}$
ER(R2) = $50[p + \varepsilon(1-3p)] + 5[q + \varepsilon(1-3q)]$
And, therefore, $d^R_{12} = (1-p-q) + \varepsilon[(3p+q)-2]$.

For $d^R_{12} > 0$, it suffices that the probability of trembles is positive, however small, even under the assumption of CKR. To see this, recall that CKR in this case means that: (a) C will set 1-p-q=0 (i.e. he will never play the strictly dominated C3), *and* (b) R will know this, insert 1-p-q=0 in her d^R_{12} function above, and resolve to play R1 only if she expects $d^R_{12} = \varepsilon[(3p+q)-2] = \varepsilon > 0$. So, if R expects trembles with positive probability ε (even if ε is tiny), she will have a preference for R1 over R2. Lastly, we note that the above holds

as $\varepsilon \to 0$ and so conclude that (R1,C2) is the game's unique *trembling hand perfect Nash equilibrium* (or, for short, the *trembling hand equilibrium*).

Why have these trembles knocked out one of the two pure strategy Nash equilibria? Looking at Game 3.1 once more, it becomes apparent that it happened because one of the pure strategy Nash equilibria (R2,C1), was supported by the *weakly dominated* strategy (R2). Given that it seems sensible (under any view regarding trembles) to allow for at least the slightest smidgen of an execution error, the *trembling hand perfect equilibrium* concept provides secure grounds for eliminating any Nash equilibrium which is formed by a *weakly dominated strategy*.

The strength of the *trembling hand* refinement of Nash's equilibrium is that it constitutes the least restrictive concept involving trembles that we could use to reduce the number of Nash equilibria. It is least restrictive in the sense that it does not require us to assume anything specific about the nature of the trembles. All we had to do is assume that they occurred with small but positive probability.

Its weakness, on the other hand, is that it only helps eliminate some, but not all, implausible Nash equilibria. For instance, the unreasonable equilibrium in Game 2.11 (R1,C1) is as much *trembling hand perfect* as the more plausible one (R2,C2). The reason it fails in this game is that neither Nash equilibrium here is supported by a weakly dominated strategy. Similarly with the rejection of the players' third strategies in Game 2.14: though most observers would agree that reasonable players may well choose them, the fact that the pure strategy Nash equilibria of that game do *not* rely on weakly dominated strategies means that *trembling hand perfection* does not expose the implausibility of insisting that R3 and C3 are not rationally playable.

One possible route for boosting the power of 'tremble' based refinements would be to say something more about the 'trembles'. What else could we say about them? Perhaps that ε, the probability of the tremble, might be a function of the cost of erring. If C errs toward C3 in Game 3.1, he forfeits pay-offs ranging between 1 and 50 (depending on R's choice). What if, however, C's pay-offs in column C3 equalled –1 million (as opposed to a mere –1)? Would we expect him to play C3 mistakenly with the same probability given the dire consequences of such a tremble?

We shall return to the question of what trembles can be 'reasonably' assumed later on in this chapter. For now the example serves to flag a potential weakness with all refinements based on 'un-theorised' trembles: they need a plausible theory of trembles to go with them *and* one that players share. For now we conclude this section by extending the analysis to Nash equilibria in mixed strategies (NEMS).

Since the notion of the trembling hand makes a difference only when some pure strategy Nash equilibria rely exclusively on weakly dominated strategies, it follows that in games featuring multiple pure strategy Nash equilibria, which are not reliant on weakly dominated strategies, the corresponding NEMS also passes the test of trembling hand perfection. To give examples, let us juxtapose Game 2.8 against Games 2.13, 2.14, 2.15, and 2.16.

In Game 2.8 Nash equilibrium (R1,C1) is supported by R1, a weakly dominated strategy. Thus, any NEMS which assigns a positive probability to R1 is *not* a trembling hand perfect NEMS. By contrast, in Games 2.13 to 2.16 none of the pure strategy Nash equilibria are formed by some weakly dominated strategy. Therefore, their NEMS *are* trembling hand perfect. Again we see the weakness of this refinement because not all the

NEMS in these games are equally plausible. Why would players choose to go for the mutually most advantageous pure strategy Nash equilibrium in Game 2.15 with probability only ¼?

3.2.2 Harsanyi's Bayesian Nash equilibria and his defence of NEMS

The idea of trembles introduces a new kind of *uncertainty* into games which is helpful in the sense that it can reduce the number of Nash equilibria. There is another kind of uncertainty that players might suffer from: Players may not know each other's pay-offs. After all, who knows exactly what motivates another person. At best when one knows someone well, one is making an educated guess at what makes them tick (and hence how they value any particular outcome).

To put this point slightly differently, so far we have been examining games of *complete information*: that is, interactions in which players have complete information over every possible outcome and its associated utility pay-offs for each player. It may be more realistic to assume that interactions are best characterised as games of *incomplete information*: that is, players do not have perfect knowledge of what a certain outcome is worth to their opponent(s). Game theorists recognised early the need to offer useful analyses of *incomplete information games* and John Harsanyi extended Nash's analysis to these type of uncertain interactions.

We set out his approach in this section with two objectives in mind. One is to see how Nash's original idea actually transfers quite naturally to uncertain settings. This is important because games of incomplete information are quite likely and if the basic analysis from games of complete information cannot be applied to these settings, then much of what we have discussed so far in this book would be of limited value. The other is to build up some of the tools that are important for further *refinements* of Nash's equilibrium.

Let us consider Game 3.2 below.

	C1	C2
R1	0 or 3, -1	2 or 5, 0
R2	2,1	3,0

Game 3.2 – *A game of incomplete information concerning R's pay-offs*

R's pay-offs are unclear to C (if she chooses R1) here. Thus player C is uncertain as to what outcomes (R1,C1) and (R1,C2) are worth to his opponent (R). In the case of outcome (R1,C1) R collects either 0 utils or 3. Another way of putting the same case is to say that (R1,C1) yields 0 to R if she is of type R_a and 3 if she is of type R_b. Similarly in the case of (R1,C2) in which R collects either 2 or 5 (depending on her 'type'). Meanwhile, whatever her 'type', R (who knows her own type) knows precisely what each of the four potential outcomes is worth to C. For obvious reasons, this category of game is also known as one of *asymmetrical information*.

In this example C can easily work out his best reply to R1 and R2. They are respectively C2 and C1. He just can't predict whether R will choose R1 because of the uncertainty over what type of R he is playing. If C is playing an R of type R_a, then C

knows that R has a strictly dominant strategy (R2). But, if C is playing a type R_b, then C knows that R has a strictly dominant strategy (R1).

Harsanyi's (1967/8) suggestion in such cases is that we should assume that C holds *some* probability assessment regarding the likelihood of R being one type or the other. C attaches probability p to being engaged in a contest with type R_a and 1-p to type R_b. C now works out his best actions given: (a) the utility payoffs in each of the two contests, and (b) his probabilistic beliefs (p,1-p) concerning the relative likelihood of encountering a R_a and R_b type.

What Harsanyi did, in effect, was to convert a game of incomplete information to one where there is complete information, albeit one where some of this information is probabilistic (i.e. with respect to the type of one of the players). To see this more clearly, we re-write Game 3.2 as follows:

		Strategy C1		Strategy C2	
Strategy R1	Type R_a (p)	0	-1	2	0
	Type R_b (1-p)	3		5	
Strategy R2	Both types of R	2,1		3, 0	

Game 3.2 *re-written in the spirit of Harsanyi's conversion*

Harsanyi's recommendation for C is: ***if p> ½ then C will choose C1; otherwise choose C2***. *Proof*: Since R2 is a strictly dominant strategy for type R_a, C expects that R will play R2 with probability p. And since R1 is the strictly dominant strategy for type R_b, C anticipates R1 with probability 1-p. Now, in view of the above, if C plays C1, he expects to receive on average –1 with probability 1-p and 1 with probability p. So, ER(C1) = -(1-p)+p = -1+2p; and if he were to play C2, he would receive 0 whatever the type of R (or, indeed, R's choice) (ie ER(C2) = 0). Subtracting ER(C2) from ER(C1) we derive C's expected net gains from preferring C1 (over C2): d^C_{12} = -1+2p. Thus C will prefer C1 when d^C_{12} = -1+2p > 0, or when p > ½.

In summary,

> Type R_a will always play R2 (since it does not matter what action C takes as R2 is strictly dominant for Type R_a).
> Type R_b will always play R1 (since R1 is strictly dominant for Type R_b).

> Player C (who is of only one possible type in this game) will play according to: (a) his understanding of the different strategies that the two types of R will adopt (see above), and (b) his expectations regarding which of the two types of R is more likely. More precisely, C will choose C1 if p>½, C2 if p<½ and will be indifferent between his two strategies if p=½.

The above strategies for each type of R and C constitute a Nash equilibrium in the sense that both types of R are adopting a best reply to C's strategy and C is adopting a best reply to the strategy of R *given his expectations about her type*. This type of Nash equilibrium is known as a *Bayesian Nash equilibrium*.

Bayesian games and Bayesian Nash equilibria (definition)

Consider an *N*-person game. A *Bayesian* version of this *game* consists of: (a) a finite set of potential types for each player i=1,…,N, (b) a finite set of perfect information games, each corresponding to one of the potential combinations of the players' different types, and (c) a probability distribution over a player's types (reflecting the beliefs of her opponents about her true type). A *Bayesian Nash equilibrium* is the Nash equilibrium of the Bayesian version of the game; that is, the Nash equilibrium which obtains once we take into consideration not only the strategic structure of the game but also the probability distributions over the players' different (potential) characters or types. In this sense, the Bayesian Nash equilibrium has the following properties:

(A) It specifies a strategy s_i (pure or mixed) for each player i=1,…,N conditional on (i) perfect knowledge of her *own* type, and (ii) her probabilistic expectations [see (c) above] regarding the type of her opponent.

(B) The strategy of player i must be a best reply to the strategy of player j for all i,j=1,…,N given their expectations about one another.

In practical terms, the computation of a *Bayesian Nash equilibrium* involves three steps: (a) Note the prior beliefs of each player concerning their opponent and propose a strategy per player; (b) calculate expectations generated by these strategy combinations regarding expected payoffs; (c) check that each player's strategy choice is the best reply to that of all others.[1]

Box 3.2

A BILATERAL MONOPOLY GAME UNDER ONE-SIDED UNCERTAINTY

Suppose Game 3.2 captures the interaction between a monopoly supplier of oil (R) and a monopoly producer of electricity using oil-fired generators (C). Suppose the cost of oil has risen and the oil company must raise its price in order to maintain its monopoly profits. Its strategic choices are: to raise the price of oil (C1) or not to raise it (C2). The electricity company in turn has a choice between building (R1) and not building (R2) a new power plant (strategy R1) which utilises another energy source (e.g. gas), so as to sever its dependence on the oil company. The oil company's pay-offs are fairly predictable in the sense that it will receive no utils from this interaction if it does not raise its price (presumably because its profits will be zero if it continues to sell oil to the electricity producer at the low price; and it will also be zero if it sells no oil following the construction of the gas-fired generator). But if it raises its oil price and the electricity company chooses to build the new plant, it makes a hefty loss.

Turning to the electricity company's pay-offs, they shall depend on the cost of building the new plant. If they are excessive, its dominant strategy is not to build. Otherwise, its dominant strategy is to go ahead with the gas plant. However, if only the electricity company knows the true cost of building the new plant, it alone will know its own type. In the context of Game 3.2 we can capture the above by identifying type R_a with the case where the new plant is high-cost and R_a with the low cost scenario. It turns out that if the oil producer thinks that there is a better than even chance that the cost of a new gas-powered station will be high, the price of oil will rise. Otherwise, it will remain unchanged.

As we have suggested, a major benefit of Harsanyi's approach is the extension of the Nash equilibrium concept beyond a world of transparent motives.[2] A further benefit is

that it offers a new interpretation, and defence, of NEMS (the Nash equilibrium in mixed strategies) concept.

In an influential paper Harsanyi (1973) attempted to deflect the criticism that NEMS is an unstable equilibrium by radically re-interpreting it. His first point was to concede that it makes little sense to imagine that players randomise between their pure Nash equilibrium strategies with the probabilities provided by NEMS. In fact, Harsanyi acknowledged that players are unlikely to choose mixed strategies at all. Instead, each player chooses a firm pure strategy given her beliefs about her opponents' motivation. However, the fact that the latter cannot be known with certainty adds a new dimension. The players are engaged in a game of incomplete information and NEMS emerges, in the limit, as a Bayesian Nash equilibrium.

As an illustration of how this argument works, consider Game 3.3 borrowed from Myerson (1991). It is chosen because there is no Nash equilibrium in pure strategies.

	C1	C2
R1	$0,0^{-}$	$^{+}0,-1$
R2	$^{+}1,0$	$-1,3^{-}$

Game 3.3 – *No pure strategy Nash equilibria, one NEMS*

The NEMS of this game is given by p (= probability of R1) = ¾ and q (= probability of C1) = ½.[3] Suppose now that each player is drawn from a population of types, along Harsanyi lines. From C's perspective, player R can be any type α drawn from a population which is uniformly distributed across the interval (0,1). Likewise, in R's eyes, player C can be any type β drawn from a population which is uniformly distributed across the same interval. Of course R knows her value of α but knows not C's β; and likewise C knows his β, but R does not. The values of α and β affect the players' pay-offs in a small way which we capture by re-writing Game 3.3 as Game 3.3b below.

Assuming that ε is a suitably small number close to zero (similar in concept to the 'trembles' introduced earlier), it can be thought of as an index of how the return to a player is affected by the differences between his or her potential 'types'. So, when ε vanishes to zero there is no important difference between types and the uncertainty in the game shrinks to zero. As we shall show below, when $\varepsilon \to 0$ the Nash equilibrium of Game 3.3b is none other than the NEMS of the original Game 3.3. To see this explicitly, let us compute the 'perturbed' game's expected returns per strategy:

$$ER(R1) = \varepsilon\alpha;\ ER(R2) = 2q\text{-}1 \text{ and } ER(C1) = \varepsilon\beta;\ ER(C2) = 3\text{-}4p$$

Or, $d_{12}^{R} = ER(R1)\text{-}ER(R2) = \varepsilon\alpha - (2q\text{-}1)$ and $d_{12}^{C} = ER(C1)\text{-}ER(C2) = \varepsilon\beta - (3\text{-}4p)$

	C1	C2
R1	$\varepsilon\alpha,\varepsilon\beta$	$\varepsilon\alpha,-1$
R2	$1, \varepsilon\beta$	$-1,3$

Game 3.3b – *Game 3.3 re-written to accommodate random variations in the player's perception of how her/his opponent values a zero pay-off*

Thus, R will choose R1 with certainty if and only if $d_{12}^R > 0$, which means that p = Pr(R will play R1 with certainty) = Pr[$\varepsilon\alpha > 2q$-1]= Prob[$\alpha > (2q$-1)/ε]. Similarly, C will choose C1 iff and only if $d_{12}^C > 0$, which means that q = Pr (C will play C1) = Prob[$\beta >$ (3-4p)/ε].

To complete our proof we need to use a standard result of probability theory. If x is a random variable which is uniformly distributed within the interval [0,1], then Pr($x>X$) (= the probability that $x>X$) = 1-X. Therefore, the above expressions can be re-written:

$$p = \text{Prob}[\alpha > (2q\text{-}1)/\varepsilon] = 1\text{-}[(2q\text{-}1)/\varepsilon]$$
$$q = \text{Prob}[\beta > (3\text{-}4p)/\varepsilon] = 1\text{-}[(3\text{-}4p)/\varepsilon]$$

Solving the above system of equations for p and q, we obtain the probabilities that R will play R1 and C will play C1 as functions of the trembles (ε). Lastly, and this was Harsanyi's analytical triumph, it is simple to show that, as ε tends to zero, we are back to the NEMS, i.e. p=3/4 and q=1/2.[4]

In conclusion, Harsanyi's clever extensions of Nash's equilibrium to games of incomplete information succeeded in bringing asymmetrical information games into the fold of game theory as well as offering an alternative way of understanding the idea of mixed strategy Nash equilibria. Like Aumann's defence (see section 2.6.3), it depends on an assumption of CAB and makes 'mixing' an attribute of this consistency between beliefs. There is only one set of beliefs regarding the likelihood of each action which can be consistently held by all. The only real difference is that Aumann's selection of a pure strategy turns on a psychological twitch whereas Harsanyi's selection depends on selection of the type of player. The key question remains, however: Since rationality cannot in general be counted upon to generate the required consistency of beliefs (see section 2.5.3), why do we assume CAB?

This question will plague us to the book's end. We set it aside, again for now, and turn instead to the most significant refinements of the Nash equilibrium concept: the ones based on the introduction of *Sequence* and *Time*.

3.3 DYNAMIC GAMES

3.3.1 Extensive form and backward induction

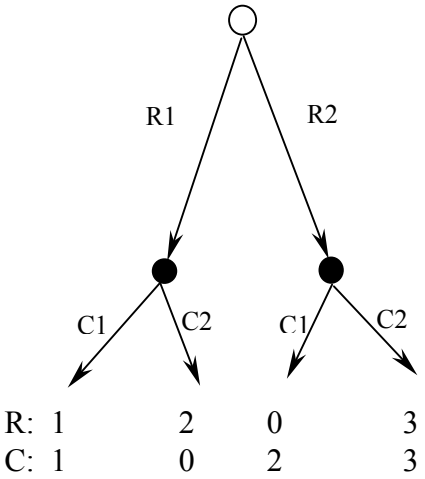
The second crucial refinement of Nash's analysis is based on the study of a game's dynamic structure and it is this refinement to which we shall devote the rest of this chapter.

The games so far lack a dynamic structure. Players choose simultaneously in a time vacuum and their thought processes unfolded solely in logical time. We now introduce a possible temporal structure to the game. Of course, many games are played simultaneously, but some have an explicit order of play and this allows some refinements of the Nash equilibrium concept which can help remove indeterminacy. We begin with Game 2.16, the so-called *Stag Hunt* (which we re-produce below).

	C1	C2	
R1	$^{+}1,1^{-}$	2,0	½
R2	0,2	$^{+}3,3^{-}$	½
	½	½	*NEMS*

Game 2.16 (re-produced) – *The Stag Hunt game*

When play is simultaneous, there are three Nash equilibria: two in pure strategies {(R1,C1), (R2,C2)} and the game's NEMS (play each strategy with probability ½). Now consider a version of the game in which R plays first and C second (once he has observed R's choice). Utilising the extensive form representation (see Section 2.2.2), the game changes substantially:



Game 2.16 with R moving first

The game in extensive form with R moving first (choosing between R1 and R2), C observing R's choice and then selecting either C1 or C2.

Under 1st-order CKR, R will expect C to choose rationally once he observes her choice. So, R thinks to herself: "If I were to select R1, C will play C1 (preferring a pay-off of 1 to 0), leaving R also with pay-off 1. Then she asks: "And what if I play R2?" In that case, C (preferring a pay-off of 3 to 2) will play C2 and thus R gets a pay-off of 3. Therefore under CKR of only 1st order R plays R2 and this followed by C2 yielding a unique outcome: (R2,C2).

The above analysis shows how introducing a *sequence of moves* can cut through the indeterminacy as two of the three Nash equilibria (the 'inferior' pure strategy Nash equilibrium (R1,C1) as well as the counter-intuitive NEMS) are discarded. However, introducing a *sequence of moves* does not always have this effect. In fact, it can also create wholly new equilibria.

Recall Game 2.1. When players choose simultaneously, the game has a unique Nash equilibrium: (R1,C2). The reason was that, though (R2,C2) was mutually advantageous, R1 was a strictly dominant strategy for R (and C's best reply to R1 was C2). However, were R to choose her strategy before C (and in full view of her opponent), the unique Nash equilibrium would no longer hold. This is clear under the extensive form (or tree diagram) representation of the same game (see Game 2.2a, in Section 2.2.2). R predicts that, if she chooses R1, they will end up at the same outcome (R1,C2) as the Nash equilibrium of the simultaneous-move version of the game (or its normal form). However, her first-mover advantage creates a new possibility. When she chooses R2, she will have effectively restricted C to choose between pay-offs 9 (if he chooses C1) and 3 (if he responds with C2). Thus, R has a way of forcing C to play C1. Does she want to? Of course, since in that case she also receives 9 as opposed to the miserly 1 associated with outcome (R1,C2). In conclusion, when R moves first, the unique normal form (or simultaneous play) Nash equilibrium vanishes and is replaced by a new type of Nash equilibrium; one that can only be discerned if we take seriously the sequence of moves. The game now consists of two *subgames*: one in which R plays (under the eyes of C) and a second one in which C responds. This type of Nash equilibrium is known as a *subgame perfect Nash equilibrium* and is due to Reinhart Selten (1965).

It is worth reflecting on the structure of the argument in both these cases. It embodies the so-called *logic of backward induction*. We concluded what the player moving first will do by considering what the player moving second would do: we assumed in other words that she would reason backwards. In this sense, players work out their best strategy *backwards*; they induce their beliefs about what constitutes the wisest choices by starting at the end and then moving to the beginning. The logic of this may seem impeccable, but as we shall see, it can be controversial in some instances.

Box. 3.3

PARLOUR GAMES AND BACKWARD INDUCTION

Before game theory was invented, backward induction had already been known in the context of the parlour games that the better off indulged in. Two such games (*Nim* and *Marienbad*) are discussed below: Suppose that there are two piles of matches on a table: Pile 1 (P1) and Pile 2 (P2). Two players (A&B) visit the table in sequence (A first, B second, then A again and so on) and remove any number of matches from either pile. The rules specify that each player *must* remove *some* matches if either pile has matches remaining and *only* remove matches from one pile at a time. In *Nim* the player who collects the last match wins. In *Marienbad* the player who is left with the last match loses. What is the best strategy in these games? Backward induction delivers the answer. Let us first concentrate on *Nim* and assume x to be the number of remaining matches in pile P1 with y the remaining number of matches in P2. If players reach a stage of the game where $(x,y)=(1,1)$, and it is A's turn to play, then in *Nim* B can win with certainty (since A is forced to collect one of the two matches, thus leaving the last one to B). Similarly if they reach a stage of the game where $(x,y)=(2,2)$: In that stage, if it is A's turn to play, A will either remove 1 or 2 from one of the two piles. In the former case, B removes the other two and wins whereas, in the latter case, we are back to $(x,y)=(1,1)$ with A playing first (and therefore delivering victory to B). Notice that the same applies if we start with $(x,y)=(N,N)$: If A plays first, B wins. We might call this a decisive *second-mover advantage* and summarise the wining strategy thus:

Best strategy in Nim when the piles contain the same number of matches: To win A must not remove the last match from one of the two piles. To prevent this from happening, she must ensure that there are matches in each pile every time she leaves the table. However, since A moves first, she cannot but leave unequal piles behind her. Thus on his visits, B can always remove matches from the pile with most matches so as to keep the number of matches in each pile equal. That way, they shall end up with one match per pile when A visits the table for the last time and B will win. (The opposite is of course true in *Marienbad* as it is the first mover has the opportunity to force the second mover to collect the last match).

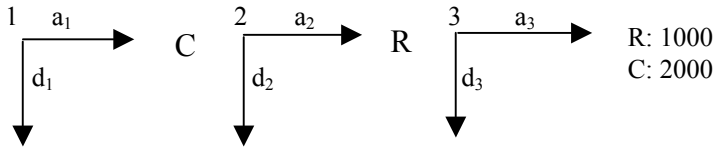
Best strategy in Nim when the piles are unequal. In this variant of the game it is the first-mover who has the opportunity to keep the piles equal after she leaves the table. Thus, they shall reach the stage of the game in which a single match remains in each pile when it is the second-mover's turn to play. Thus the first-mover is guaranteed victory. (And defeat in *Marienbad*.)

Conclusion: As long as players use *backward induction* to work out their best strategies, the first-mover (second-mover) is guaranteed victory in *Nim* when the piles are initially unbalanced (balanced). The opposite is true in *Marienbad*.

3.3.2 Subgame perfection, Nash and CKR

Selten (1965,1975) blended backward induction with the Nash equilibrium to produce the *subgame perfect Nash equilibrium* or SPNE in dynamic games. In conjunction with Harsanyi's Bayesian extension of Nash (see Section 3.1.3), Selten's contribution was largely responsible for the revival of game theory in the 1970s (after languishing in the backwaters for most of the post-war period). This is why the 1994 Nobel Prize in Economics was awarded jointly to Nash, Harsanyi and Selten. We now look more closely at the SPNE concept.

Consider the following game in which two players take turns to play either *across* (strategy *a*) the tree diagram (or extensive form) or down (strategy *d*). R kicks off and can end the game on node 1 by playing d_1 or pass the baton to C by choosing a_1. Then C has a similar choice between d_2 and a_2. If he chooses a_2, R completes the game with a choice between d_3 and a_3.



Game 3.4 The Short Centipede[5]

Backward induction can be used here as follows. If the game is to reach node 3, R will have a clear choice between 2000 and 1000 utils. She will thus choose d_3, collecting her 2000 utils and leaving C with 800. If the game is to reach node 2, by then C will have worked out our previous sentence's conclusion and predict that, were he to choose a_2, he will end up with 800 utils (while R will receive payoff 2000). Clearly, he would rather have the 1000 utils that d_2 guarantees him. This brings us to node 1 and R's initial choice problem. Does she play d_1 or a_1? The latter offers the prospect of rich pickings if node 3 is reached. But R reasons, as above, that by playing a_1 she would be enticing C to play d_2,

thus leaving R with nothing. Though a measly payoff, 1 util is better than none and therefore R plays d_1 at the outset.

The above logic is the foundation of the game's SPNE that recommends players end the game (by playing *d*) whenever it is their turn to play. It involves not just the principle of backward induction but also CKR. This is not always the case. Sometimes backward induction alone is enough because it does not matter what an opponent does when called upon to play (see the games of Marienbad and Nim in Box 3.3). This is not so in Game 3.4 where backward induction *only* works when combined with CKR. The point here is that CKR is necessary before we conclude that, in node 1, R is *convinced* that, in node 2, C will be *convinced* that, in node 3, R *will* play d_3. The SPNE is, therefore, supported by a blend of (a) backward induction *and* (b) Nash's assumption that players' strategies will be best replies to one another *at each node*. We call this blend *Nash backward induction* in order to differentiate it from simple backward induction.

Backward induction and Nash backward induction (Definition)

The difference between *backward induction* and *Nash backward induction* turns on the use of CKR assumptions. The former does not require CKR. Backward induction without CKR carries no implication for the mutual consistency of each player's beliefs. In games like ***Marienbad and Nim*** (see Box 3.3) this is not necessary since what one thinks that the other thinks is irrelevant and the application of backward induction is analytically equivalent to strict dominance reasoning (that is, the first player has a strictly dominant strategy, discerned backwards from the game's last node). *Nash backward induction*, on the other hand, *does* require CKR. For example, in Game 3.4 a 'solution' cannot be reached because backward induction alone does not yield a strictly dominant strategy for R. In this game (as in many others), only by blending backward induction with CKR (what we call *Nash backward induction*) does a *subgame perfect Nash equilibrium* (SPNE) emerge.

Before we define the *subgame perfect Nash equilibrium* (SPNE) we shall clarify the meaning of a *subgame*.

Subgames, information sets and singletons (definitions)

A *subgame* is a segment of an extensive (or dynamic) game, i.e. a subset of it. To qualify as a *subgame*, a subset of the extensive game's tree diagram must have four properties: (a) it must start from some node, (b) it must then branch out to the successors of the initial node, (c) it must end up at the pay-offs associated with the end nodes, and finally (d) the initial node (where the *subgame* commenced) must be *a singleton in every player's information set*.

Parts (a), (b) and (c) of the definition of a *subgame* are straightforward. But what is the meaning of a *singleton*, or a player's *information set* in (d)? The information set refers to what a player knows. Recall how a player may not know exactly where he or she is in the tree diagram. In Game 2.2a (see Section 2.2.2, Chapter 2) player C knows which branch of the tree diagram he is in when his turn comes to choose between C1 and C2. His information set is a *singleton* when he is called on to play because he will know whether he is at the left or the right decision node. However, in Game 2.2b the broken line linking the two nodes of player R indicates that, when it is her turn to play, R does

not know which node she is at. Her information set, when called upon to play, is *not* a singleton as she could be at either of two possible decision nodes.

Thus a *singleton* is an *information set* which contains *only one node*; that is, when in this *information set* (and about to choose), the player has no doubt at all as to where in the tree diagram he or she is (which means that the player knows all the previous moves in the game).

We can now decipher part (d) of the definition of a *subgame*: Its purpose is to say that a *subgame* must start at a stage of the game where the player whose turn it is to act knows what has happened previously. From that moment onwards a new 'chapter' in the game (that is, a new *subgame*) begins which we can analyse separately.

Here are some examples. In Game 3.4 players observe fully the history that has gone on before they are called to play. Thus every node coincides with a player's information set. In other words, each of the three information sets (i.e. moments when some player can, potentially, be called upon to play) is a singleton (that is, contains a single node). In conclusion, Game 3.4 comprises four subgames: the whole game, and the ones commencing at nodes 1,2 and 3 respectively.

In Game 2.2b (Section2.2.2) the only subgame is the whole game, since R's information set (that is, at the stage where R comes into the game) contains more than one node. (Note that the broken line implies that R does not know for certain which node she is at, the left or the right.) For this reason the game has only one singleton (that is, the initial node at which C makes a choice) and thus only one subgame: the whole game.

Game 2.2a, by contrast, has three subgames and three singletons: there is the game as a whole which starts from the initial decision node; there is the game which starts at C's node when R has chosen R1 (a singleton since it is an information set comprising a single node); and there is the game which starts at C's right hand side node when R has chosen R2 (another singleton). Lastly, the extensive form of Games 3.5 and 3.6 below feature only one singleton and thus only one subgame each.

The intuition behind the subgame perfect Nash equilibrium (SPNE) concept is that we discard a strategy which specifies actions in its subgames which are not best replies to each other *in that subgame*. Otherwise it seems we will be entertaining behaviour which is inconsistent with instrumental rationality and CKR at some stages of the game. Thus in the game of Game 2.16 when R moved first (see Section 3.2.1 above), the Nash equilibrium (R1,C1) in the normal form suddenly looks untenable when the game is analysed in extensive form. This is because it specifies an action, at the subgame where C decides, that is not the best reply to what has gone before. As it turns out, the only equilibrium outcome which passes this test of an analysis of the subgames is (R2,C2). Game theorists accordingly call it a *subgame perfect Nash equilibrium (SPNE)*.

Subgame Perfect Nash Equilibrium – SPNE (definition)

Strategies are in *a subgame perfect Nash equilibrium (SPNE)* in an extensive form game when they constitute a Nash equilibrium in *each* of the game's subgames. Alternatively, an SPNE is a *strategy profile* (that is, a set of strategies one for each player and for each of the game's nodes) for which each player's strategy is her best one given the strategies for the others at *any* history after which it is her turn to play, whether or not the history occurs if the players follow their strategies. For example, the SPNE strategy profile of

Game 3.4 specifies that each player chooses *d* when it is her or his turn to play and regardless of what has happened before them. (As we shall see below, this is a rather controversial property of SPNE).

Some clarification of the reference to Nash in this definition may be helpful. In the case of Game 3.4 (but also of Games 2.2a and 2.2b, where one player moved first) the equilibrium seemed to emerge because (a) we reasoned backwards, and (b) we applied CKR. Where does Nash fit in this? In the last chapter we emphasised that, in the context of static games, CKR may *not* lead to a Nash equilibrium *unless CAB is also assumed*. (Recall Game 2.9.) With dynamic games there is *no* need to assume CAB *directly* before Nash equilibrium beliefs and actions are forged. In fact, CAB can be induced indirectly simply by combining CKR with backward induction. This combination has already been mentioned in this chapter under the name *Nash backward induction.*

Let us see how *Nash backward induction* plays in dynamic games the roles played by CAB in static ones: Take, for example, Game 3.4. In its earlier analysis, we showed that an R who assumes CKR *and* reasons backwards will play d_1 at node 1. Why? Because she believes that, were she to give him a chance to move at ode 2, C would play d_2. And why does R believe this? Because she believes that, if C were to find himself at node 2, he would come to believe that, were R to be given a chance to play once more at node 3, she would have chosen strategy d_3. In other words, R knows that if she were ever to reach node 3 she would, indeed, play d_3. Since this is common knowledge, she knows that C will prefer to play d_2 in node 2. And given that this is common knowledge too, R prefers d_1 in node 1. Moreover, since all these thoughts, or knowledge, are common, C expects that this is indeed what R will do.

It is thus not difficult to see that the combination of (a) assuming CKR, *and* (b) reasoning via backward induction, *forces players to hold CAB (consistently aligned beliefs).* Indeed, it is a combination which guarantees that players' beliefs will spawn, as Nash would have it, actions that confirm those very beliefs. This is why we call it *Nash backward induction*.

Box 3.4

PATIENCE AS AN IRRATIONAL VIRTUE

Two players, Anna (A) and Bill (B), are invited to play the following game (without communicating with one another). A is offered a choice between \$10 and passing. If she takes the \$10, the game is over: she collects \$10 and B receives half of that sum minus \$1 (that is, \$5-\$1=\$4). On the other hand, if she passes, B is offered double the money that A was offered previously; that is, \$20. Now it is B's turn to 'take' or 'pass'. If B takes the money, the game is over: he collects his \$20 and A receives half of that minus \$1 (\$9). If however he passes, a double amount (\$40) is offered to A who, once more, must choose between taking the money (in which case she collects \$40 and B collects exactly half of that minus \$1, i.e.\$19) and passing (in which case we have a new round during which B is offered twice as much, namely \$80). And so on. The game's organisers tell our players that, as long as they keep passing, the amount offered will be doubled in each round until one of them has collected no less than \$100000 (thus leaving the other with a healthy guaranteed pay-off of at least \$49999). What is this game's SPNE?

Clearly, in this game patience on the part of both players pays handsomely. If (between the two of them) they refuse the offered money for fifteen consecutive rounds (allowing it to be doubled each time), A and B will collect \$163640 and \$81819 respectively. However, in a manner reflecting the Short Centipede, or Game 3.4, the SPNE has A taking the measly \$10 at the very beginning and ending the game. This is the inescapable outcome of applying what we called earlier *Nash backward induction*: B predicts that, if the game were to reach its last node (i.e. node 15), he will receive \$81819 (that is, half of A's \$163640 pay-off

minus $1). However in the penultimate node (i.e. node 14), he is offered $81820. Clearly, to the extent that an extra dollar is desirable (and thus pay-off $81820 is preferred to $81819), B will prefer to end the game in node 14 by accepting the $81820 offer. Thus the game will *never* reach node 15. By the same token, it will never reach node 14, since A will end it in node 13 (preferring $40960 to the $40959 she is bound to get in node 14). And so on, until we reach the conclusion that A will end the game at the very beginning by accepting the initial offer of $10.

Is this the uniquely rational way of thinking about this type of game?

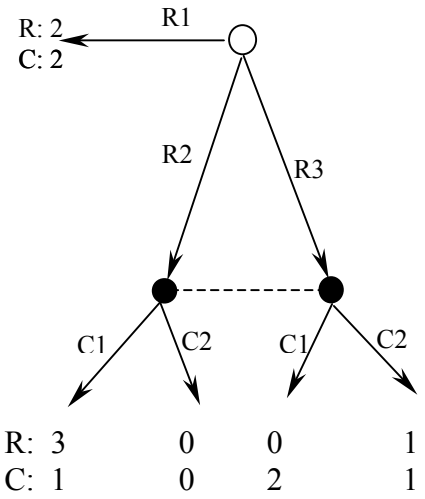
3.3.3 Sequential equilibria

	C1	C2
R1	2,2^{-}	$^{+}$2,2^{-}
R2	$^{+}$3,1^{-}	0,0
R3	0,2	1,1

Game 3.5

Consider static Game 3.5 and its two Nash equilibria in pure strategies: (R1,C2) & (R2,C1). Suppose the game has the following dynamic structure (or, in more formal game theoretical language, *extensive form*): *R moves first*. If she chooses R1, then (as is also evident from the adjacent pay-off matrix) the outcome is not influenced by anything C might do later as each receives 2 utils. However, if R chooses either R2 or R3, then C's turn to play comes and he must choose between C1 and C2 *without having observed whether R chose R2 or R3*.

The game's dynamic structure (i.e. that R moves first) and an appeal to SPNE does not immediately help the refinement of the Nash equilibrium analysis here.



Game 3.5 in extensive form with R moving first (choosing between R1 and R2), and C playing only if R selects either R2 or R3. Note the broken line connecting C's nodes. Its meaning is that if C gets to play at all he does not know which of the two nodes he is at (since he has not observed whether R chose R2 or R3). Nodes connected with such broken lines are called a player's *information set* – see the earlier definition

This is because C's decision over whether to play C1 or C2 no longer forms a subgame because he does not know in which part of his information set (or, more loosely, of the game tree) he is when called upon to play. Alternatively, there is no unique route from his C1 or C2 decision back to the original decision node of the game and, thus, we cannot solve backwards from this node as we did before. The result is that there is only one subgame for this game: *the whole game* with its familiar two Nash equilibria (R1,C2) and (R2,C1) (see the earlier static *normal form* or *matrix representation*). And since the game as a whole is a subgame, both Nash equilibria are subgame perfect. In short the SPNE does not refine the analysis in this instance.

Nevertheless, we notice that despite the fact that we may not observe R's move, the moment we are informed that R is to move (or has moved) first, there is something decidedly fishy about *one* of the two Nash equilibria. In particular, C's best reply is to play C1 whenever called upon to play *regardless of whether R chose R2 or R3*! So, R can be confident that by playing R2 she will secure 3 utils as opposed to the 2 utils that R1

guarantees for her. Clearly, of the two Nash equilibria one [i.e. (R2,C1)] makes a lot more sense than the other [(R1,C2)].

How did we get to this useful conclusion in spite of subgame perfection's inability to come to our help? We did it by amending slightly the concept of subgame perfection. Before we explain the amendment, let us recall the definition of a subgame perfect Nash equilibrium (SPNE). It is defined as the set of strategies for all players and for each node of the game (also known as a *strategy profile*) such that: (a) player *i*'s strategy is a best reply to the strategies of the others at any *point of the game* after which it is *i*'s turn to play, and (b) this is so independently of what has happened beforehand.

Now, the problem in adapting this equilibrium to Game 3.5 is that C does not know which *point of the game* he is at.[6] But he *does* know which *information set* he is at courtesy of the fact that he has been asked to play. It turns out that, in this game, knowledge of which *information set* he lies at is enough in order to make C decide that he wants to play C1 and not C2. From a theoretical viewpoint, we succeeded in eliminating (R1,C2), a Nash equilibrium in pure strategies, by defining another type of equilibrium similar to SPNE in every regard except one: We replace part (a) of SPNE's definition in the previous paragraph with the phrase "player *i*'s strategy is a best reply to the strategies of the others at any *information set of the game* after which it is *i*'s turn to play."

This new concept is known as a *sequential equilibrium* and is due to Kreps and Wilson (1982a). Applying it to Game 3.5, it turns out that Nash equilibrium (R2,C1) is a *sequential equilibrium* because *if* C's information set is reached *for whatever reason* (i.e. *whatever* the reason that C thinks caused R to ignore R1) C2 strictly dominates C1 and R2 is a best reply to that conjecture.

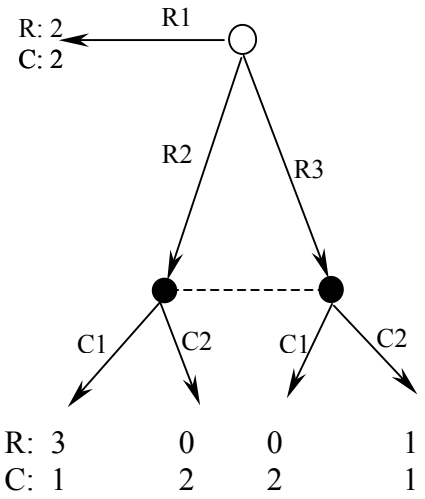
The basic idea, then, behind the *sequential equilibrium* concept is exactly the same as *subgame perfection*. They both use *backward induction* and require that strategies be rational in the sense of being best replies at each stage of the game. The only difference concerns how each stage of the game is defined. *Subgame perfection* presumes that at each stage of the game players' strategies are best replies to one another when players know the *precise node they are at*. In contrast, *sequential equilibrium* comprises strategies which are best replies to one another in stages of the game where players are uncertain regarding their precise location (or node) on the tree diagram.

Game 3.5 is useful for introducing the idea of a sequential equilibrium but its simplicity obscures the full significance of this new equilibrium concept. Consider a subtle change to Game 3.5 which yields Game 3.6.

If C is called upon to play, he has no dominant strategy now. Thus, what he will do depends on his beliefs regarding whether R played R2 or R3 before him and these will have to be specified before backward induction can be used. The *sequential equilibrium* now (and in general) comprises both a *strategy profile* and a *belief system,* where the latter specifies for each *information set* of each player a belief held by the player who is about to act at that *information set* regarding what has happened in the past (or, equivalently, which precise node she/he at).

Once these beliefs are specified we can proceed to solve the game using the process of *Nash backward induction*. Thus, if C believes that R chose R2 with probability greater than ½, he will play C2. Otherwise he will either play C1 (or randomise). This being the case, we can see that R would only choose R2 provided C believed that the

probability of R doing this is less than ½ because in this case C chooses C1 and R obtains a pay-off of 3 which is better than what can be obtained by playing R1 or R3.[7]



	C1	C2
R1	2,2^{-}	$^{+}$2,2^{-}
R2	$^{+}$3,1	0,2^{-}
R3	0,2	1,1

Game 3.6

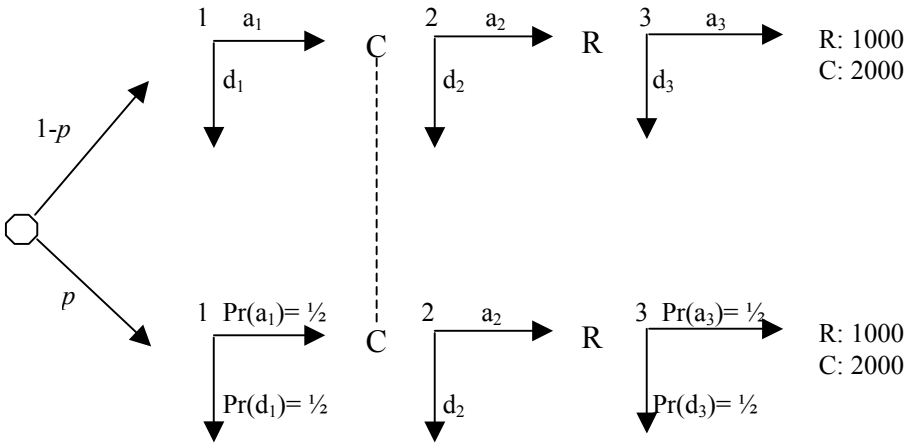
In summary, in this game C's beliefs are what they are and he acts on them. However, there are other, longer games in which we wish place constraints on the kind of beliefs that rational players can hold. In particular, we want them to be consistent with the observed behaviour of their opponents. In such cases, the sequential equilibrium concept relies on *Bayes's rule*.

3.3.4 Bayesian learning, sequential equilibrium and the importance of reputation

In Chapter 1 we introduced *Bayes's rule* as a consistent method for updating our predictions regarding uncertain events as new information is made available. In game theory uncertainty comes in two guises. There is uncertainty about an opponent's earlier move; e.g. Game 3.6 in which, prior to making a choice, C did not know whether R had chosen R2 or R3. In addition, there is uncertainty about an opponent's character or, equivalently, pay-offs (recall Sections 3.1.3 and 3.1.4).

When players do not observe their opponents' behaviour, they must act on the beliefs they already have and those they derive from studying carefully the game's structure (e.g. Game 3.6). But there are other dynamic games (e.g. Game 3.4) in which players *do* get an opportunity to witness the behaviour of their opponents and therefore to gather information about them which can be used to sharpen their predictions about what will happen in the future. It is in these cases where the observations of others' behaviour helps players update their beliefs by means of *Bayes's rule*.

Consider Game 3.4 again. Given CKR and backward induction (or *Nash backward induction*) we arrived at a unique equilibrium: the game's SPNE which has each player choosing strategy d_i at each node of the game, thus ensuring that R will 'kill' the game immediately. Suppose now that we relax CKR, allowing for C to believe that there is a small probability p that R is irrational. How is this going to alter the game's equilibrium? Let us re-write the game's extensive form as Game 3.7 in order to capture this initial uncertainty about R's rationality. If C is ever called upon to play, he does not know whether he is at the upper centipede or the lower centipede. He merely estimates that he is at the upper branch with probability 1-p and at the lower with probability p.



Game 3.7 The short centipede with one-sided uncertainty regarding R's rationality

In summary, Game 3.7 is identical to Game 3.4 barring one difference. Before players act, nature and/or nurture has determined whether R is rational or not. If R was 'created' rational, players find themselves in the original short centipede game at the upper branch of the tree diagram. But, if R was 'created' irrational (an event whose probability equals p), the bottom branch applies. In that branch, an irrational R simply acts without rhyme or reason; that is, unpredictably. When it is her turn to play she is equally likely to play down (d) or across (a) [i.e. $Pr(a_1)= Pr(d_1)= Pr(a_3)= Pr(d_3)= ½$]. The presumption here is that a rational R knows with certainty whether she is rational (and therefore which branch of the game she is at) but C does not. This is why we say that in this type of game we have one-sided, or asymmetrical, uncertainty. Diagrammatically this is captured by the broken line that joins up C's nodes to create one information set at which C finds himself when, and if, called upon to play.

The uncertainty about R's rationality (or, equivalently, the relaxation of CKR) means that the game's SPNE no longer holds as there are more than one end-nodes to this game (it is *not* a singleton) and backward induction cannot be applied by starting at some specific end-node (c.f. the case in Game 3.4). Here we need to employ the logic of sequential equilibrium and study (backwards) what happens in C's player's multi-node information set. So, let us begin by working out what a rational R and a C (whom we

presume to be rational with probability 1) will think will happen if the game reaches its 3rd stage and R is called upon to play again. Then we shall investigate what C will do if his information set is ever reached (in stage 2). Finally, we shall return to the very beginning in order to assess what a rational R would do. (Note that an irrational R is assumed to choose at random and, therefore, we need no complicated theory in that case.)

Stage 3: If this stage is ever reached, it will be R's turn to choose. If R is rational, she will always play down (d_3). If irrational, she will choose *as if* at random. In the notation of conditional probabilities, where Pr(A|B) denotes the probability that event A will occur given that even B has already occurred (or conditional on B), we have:

$\Pr(a_3|\text{R is rational}) = 0$ $\Pr(d_3|\text{R is rational}) = 1$

$\Pr(a_3|\text{R is irrational}) = ½$ $\Pr(d_3|\text{R is irrational}) = ½$

Stage 2: At this stage, if it is ever reached, it is C's turn to play. What will he do? It depends on which branch of the tree diagram he thinks he is in. If he thinks he is at the upper one, he will definitely play down, since he predicts that a rational R (who appears in the upper branch) will play down, given a chance, in the next stage. On the other hand, if he thinks that he is in the lower branch, he has good cause to play across (a_2) as the irrational R occupying the lower branch will err toward a_3 with a probability of ½ and this translates into good odds for C collecting 2000 utils at stage 3 rather than 1000 at stage 2. Of course, the danger always looms that he will end up with pay-off 800 if he gives R the chance to play again in the last node and she shrewdly plays down.

What C will do in node 2 will, therefore, depend on his beliefs regarding R's rationality: his estimate of probability p. Let C's estimate of p at stage 2 equal p_2. C will play across if, given his estimate p_2, he feels that there is more to gain on average from risking to play across than from playing down and collecting the safe 1000 utils. In other words, C will choose a_2 only if C's expected returns from a_2 [$ER^C(a_2)$] exceed 1000 utils. Now, those expected returns equal

$$ER^C(a_2) = 2000[½\, p_2] + 800[½\, p_2 + (1\text{-}p_2)]$$

since, if C plays a_2 the game moves into stage 3 where C will collect either 2000 or 800 utils, depending on whether R plays a_3 or d_3. The probability that R will play a_3 equals the probability that R is irrational (p_2) times the probability that an irrational R will play a_3 [i.e. $½p_2$]. And the probability that R will play d_3 equals the probability that she is irrational and simply 'trembles' toward d_3 *plus* the probability that she is rational and merely collects her best pay-off [i.e. $(1\text{-}p_2)$].

So, for C to lean toward a_2 the latter's expected returns $ER^C(a_2)$ must exceed the sure 1000 pay-off from d_2: i.e. $2000[½\, p_2] + 800[½\, p_2 + (1\text{-}p_2)] \geq 1000$ or, $p_2 \geq ⅓$. In conclusion, C will play across at stage 2 as long as he thinks that there is at least a chance of one in three that R is irrational (and therefore that C finds himself in the game tree's lower branch in probability at least ⅓).

Stage 1: An irrational R will choose *as if* at random between a_1 and d_1. A rational R will have to consider the possibility that, were she to play a_1, C will play a_2 for the reasons outlined above (that is, because C might think that R is irrational with probability at least equal to ⅓). In this sense, R may have a good reason to bluff (pretending that she is irrational by choosing a_1 instead of the SPNE strategy of d_1): namely, a reason to think that if she plays a_1, then C will form an estimate $p_2 \geq ⅓$. Using the notation of conditional probabilities, a rational R's initial choice will hinge on her subjective conditional

probability q=Pr($p_2 \geq$ ⅓| a_1). If q is large enough, this means that R feels that the prospects of a successful bluff are good. The question then is, how good are these prospects?

Suppose that, at the outset, C's subjective estimate that R is irrational is p_1. If R plays across then C will receive some extra evidence that R is irrational. Of course C will always suspect that R is bluffing (which means that he will not immediately update his belief from p_1<1 to p_2=1). By how much will his belief that R is irrational increase? This is where *Bayes's rule* comes in. C will effectively be trying to compute the conditional probability that R is irrational *given* that R played across (a_1) at stage 1. *Bayes's rule* gives this conditional probability as follows:[8] p_2 = Pr(R is irrational | a_1) =

$$\frac{\Pr(a_1 \mid R\,is\,irrational)\times \Pr(R\,was\,irrational)}{\Pr(a_1 \mid R\,is\ irrational)\times \Pr(R\,was\ irrational)+\Pr(a_1 \mid R\,is\ rational)\times \Pr(R\,is\ rational)}$$

where Pr(a_1|R is irrational) is the probability that an irrational R would have played across in the game's first stage, and equals ½; Pr(R was irrational) is C's *initial* belief that R is irrational, and equals p_1; *and* Pr(a_1|R is rational) is the probability that a rational R would bluff at stage 1 in order to masquerade as an irrational player. Let us denote this last conditional probability with *r*. *Bayes's rule* can then be re-written as

$$p_2 = \frac{\tfrac{1}{2}\,p_1}{\tfrac{1}{2}\,p_1 + r(1-p_1)}.$$

In a sequential equilibrium players must discern strategies that are best replies to one another after reasoning backwards. Moreover these strategies must be consistent with the thought that, were players to *know* their opponents' mixed strategies, they would not regret adopting them (Nash's original idea). Notice that, in a sequential Nash equilibrium, if at the beginning p_1<⅓, the best R can hope to achieve through bluffing is to push p_1 up to *exactly* p_2=⅓ by stage 2. Why?

Suppose that, through sheer bluffing (i.e. the employment of some mixed strategy) in stage 1, R has pushed his reputation for irrationality to some value p_2>⅓. This is tantamount to saying that R managed to force C to play pure strategy 'across' (a_2) at stage 2 of the game (N.B. we deduced in the analysis of Stage 2 above that when p_2>⅓ C always plays a_2). But how can this be? For if R could do this simply by bluffing at Stage 1, C would know it in advance and would expect a rational R *always* to bluff at Stage 1 (that is, always to play a_1 at stage 1). But this means that observing R play a_1 at Stage 1 contains no useful, new information for C. C expects a_1 to be selected as a matter of course. Thus, C's original subjective belief p_1<⅓ will remain totally unchanged after C has observed a_1 at stage 1. But this contradicts the initial hypothesis that R can succeed in pushing p from a level below ⅓ in Stage 1 to one above ⅓ in Stage 2. In summary, the idea that bluffing can push p_2 above ⅓ *is incompatible with the notion of Nash equilibrium*.

Nor can rational bluffing procure a value of p_2 below ⅓ in any Nash equilibrium. To see why, we shall show by contradiction that there exists no (Nash) equilibrium logic according to which R might bluff (that is, play a_1 at stage 1) with positive probability (r>0) if her bluff is going to procure a value of p_2 below ⅓. To set up the logical contradiction, suppose that the following were to occur: At stage 1 R bluffs with

probability $r>0$. Suppose now that this randomisation (or mixed strategy) results in her choice of the bluff strategy a_1 at Stage 1. C observes this and by Stage 2 has updated his belief from $p_1<⅓$ to $p_1<p_2<⅓$. In other words, the observation of 'irrational' behaviour by R at Stage 1 has strengthened C's estimate that R might be irrational but not to the extent that would compel him to choose strategy d_2 at stage 2. Having worked all this out (as is our players' wont in equilibrium), a rational R would never risk playing a_1 knowing that her bluff would always backfire (that is, C will always play d_2 at Stage 2 as p_2 would fail to rise above ⅓). But then C would also know this. And having worked out that no rational R ever bluffs, he would assign zero probability to a rational R playing a_1 (i.e. bluffing). So, if C were to observe an R playing a_1, he would immediately conclude that R is definitely irrational. [Check from *Bayes's rule* above that if $r=0$, and C observes a_1, then C sets $p_2=1$.) But then, a rational R would know that by playing a_1 her bluff would *always* work – a conclusion which contradicts the earlier conclusion that it would *never* work.[9]

We conclude that, in equilibrium, p_2 can be neither greater to nor less than ⅓. It must, by simple deduction, equal *precisely* ⅓. We can now input this value in *Bayes's rule* and solve for r:

$$⅓ = \frac{½\, p_1}{½\, p_1 + r(1 - p_1)} \quad \Rightarrow \; r = 2p_1$$

Hence, in a *sequential equilibrium* (also known, in this case, as a *Bayesian perfect equilibrium*[10]) we have the following possible scenaria:

(1) $p_1 \geq ⅓$, in which case at Stage 1 a rational R always bluffs (i.e. plays pure strategy a_1), and an irrational R randomises between a_1 and a_2; at Stage 2 C always plays a_2 (as long as he gets a chance), and at Stage 3 a rational R plays d_3 whereas an irrational R randomises between d_3 and a_3.

(2) $p_1<⅓$, in which case at Stage 1 a rational R bluffs (i.e. plays a_1) with probability $r=2p_1$, while an irrational R randomises between a_1 and a_2; at Stage 2 C randomises between a_2 and d_2 (as long as he gets a chance), and at Stage 3 a rational R plays d_3 whereas an irrational R randomises between d_3 and a_3.

We conclude with three observations. First, even a tiny initial doubt in C's mind regarding R's rationality (i.e. $p_1=\varepsilon$, where $\varepsilon>0$ but close to zero) suffices in order to motivate a rational R to escape the SPNE (i.e. to 'tremble' toward a_1) with positive probability (equal to $r=2\varepsilon$). Second, this initial doubt can be thought of as R's reputation for irrationality. Clearly, *the greater R's reputation for irrationality the greater her own benefits from this game as well as her willingness to preserve this reputation.*

This type of model was developed by Kreps and Wilson (1982a) and has come to be associated in the literature with *reputation building.*[11] Third, the sequential equilibrium above (as well as the preceding two observations) resulted from the relaxation of CKR and its substitution with another type of common knowledge: the *common knowledge of R's initial reputation* (p_1) *as well as the deduced probability with which a rational R will bluff* (r). As we shall later see, this new common knowledge assumption is controversial.

3.3.5 Signalling equilibria

The reputation preserving behaviour modelled by the sequential equilibrium above is one of a kind. Player R played Game 3.7 in the manner specified by sequential equilibrium in order to maintain C's expectation that she is irrational and thus reap benefits from this initial reputation for irrationality. Of course, there may be actions that can be taken *outside* the game and which have similarly profitable effects on the beliefs of others. This is known as *signalling* behaviour and is considered briefly in this section to round out the discussion of reputation within a sequential Nash equilibrium framework. It is of potential relevance not only to dynamic, but also to one-shot (or static) games.

A famous illustration comes from Spence (1974). Let us suppose that out of N people who apply for managerial positions, half of them are of high ability while the rest are of low ability. Suppose that employers have no direct way of identifying worker quality. Suppose also each employee has the option to undertake a Masters in Business Administration (MBA) at considerable personal cost. Spence shows that it may make sense to do the course because it signals that the employee is of high quality *even if the MBA is useless from an educational point of view and employers are fully aware of this!* To simplify the problem, suppose that:

(i) High ability employees generate 5 units of 'value' for the firm that employs them
(ii) Low ability employees generate 3 units of 'value' for the firm that employs them
(iii) Doing the MBA course costs high quality employees less than low quality employees (1.25 as opposed to 2.5 units of value). (The assumption here is that high quality is correlated with the capacity to survive more easily the strains of an MBA course; e.g. pass examinations more speedily, fewer hours of study etc.)
(iv) Competition between employers forces them to pay the same wages so that no firm ends up extracting more value from a certain type of worker than its competitors. For simplicity we shall assume that

One Nash equilibrium in this labour market has all employers offering a wage equal to 4 units of output per period to all employees and no employee enrolling on an MBA course. An MBA is known to be useless and will have no effect on salaries since the employer does not think that an MBA is indicative of high or low ability. The probability remains ½ for each type (as half of the prospective employees are of high quality and the rest are of low quality). This is referred to as a *pooling* (or non-revealing) Nash equilibrium because there is no distinction between employees.[12]

To see why it is a Nash equilibrium, first notice that assumption (iv) above is satisfied because the expected benefits from employees [½(3) + ½ (5)] minus the wage cost [4] is the same for all firms. Further, when prospective employees shun MBA courses, everyone's best response is to ignore MBAs: Individual employees of either type have no incentive to enrol, because it is costly and it does not affect their wage, and employers have no incentive to link offered wages to possession of an MBA (as they know that it has no educational value *and* have no reason to expect high quality candidates to possess an MBA). Moreover note that, because no one acquires an MBA, these beliefs (both of employers' and of employees') are never tested.

Suppose, however, that employers (for some reason) believe that an MBA signals high ability. Then another Nash equilibrium exists which is referred to as *separating (or revealing)* because it separates employees between those who receive a high and those who receive a low wage on the basis of a signal. In particular, suppose that those who hold an MBA are paid 5 and those without are paid 3. Moreover, the high ability employees enrol at their nearest business school on an MBA course while low ability ones do not.

Again it is easy to check that this is *also* a Nash equilibrium. When employers face these wage differentials conditional on having or not having an MBA, enrolling on an MBA is a best reply *only* for high ability employees (they receive a wage rise, due to their MBA, equal to 2 units while the MBA costs them only 1.5 units; as compared to a low ability employee for whom an MBA costs 2.5 units thus making its acquisition uneconomic). And vice versa: When *only* high quality employees acquire MBAs, it is a best reply by employers to offer higher wages to MBA holders.

Thus the employers' beliefs are confirmed by the actual behaviour of employees *even though it is common knowledge that undertaking an MBA degree has no positive effect on productivity*. Given this behavioural pattern (of both kinds of employees and of employers) condition (iv) is satisfied exactly as the difference between the workers' value to the firm and the wages it pays them are the same for all workers and in all firms. The Nash loop between actions and beliefs is thus complete: The employers' choices (to pay MBA holders 5 units and non-MBA holders 3 units) are a best reply to the choices of employees (that is, read for an MBA only when you are of high ability and vice versa). Put differently, we started with the assumption that employers (somehow) have come to think that an MBA is a signal of ability and, once this belief was in place, we discovered that there exists a set of actions which (a) confirm it, *and* (b) are motivated by it.

The example is interesting for at least two reasons. First, it reveals yet again the importance of which beliefs agents are assumed to hold; and there is no obvious reason for preferring one set to another. After all, what is wrong with the pooling Nash equilibrium in which employers believe that an MBA signals nothing about its holders ability? Did the model not assume initially that this type of education really does not have any effect on ability?

Secondly, it provides an interesting possible explanation for the current high correlation between managerial salaries and MBAs. The normal explanation revolves around business education making people more productive (an investment in human capital, no less). Against this, we now see that business schools need not contribute to productivity and yet their qualifications may be associated with high earnings. They could signal something which is valuable given the beliefs of the employers. The latter are plainly crucial as we have seen and it is not difficult to construct more complicated beliefs which could also explain other aspects of the income distribution (see Box 3.6 below).

Box 3.6

SELF-FULFILLING SEXIST BELIEFS AND LOW PAY FOR WOMEN

Suppose that high ability women find it more costly to acquire MBAs than high ability men because they live in a world where they are expected to do more in the household than men: so an MBA costs them, say, 2.1 units (compared with 1.25 for high quality men and 2.5 for low ability men). Further we assume that men are as likely to be highly able as women and that half the labour pool is men and half is women. With the earlier assumption that half the labour pool has high ability, we have a new separating (or revealing) Nash equilibrium: Men with high ability acquire MBAs and are paid 5 while women of high ability do not acquire an MBA and receive the same low wages, equal to 3 units, as low ability men and low ability women!

To check that this is a Nash equilibrium, note that when MBA graduates are rewarded with wages of 5 units and non-MBA holders with wages of 3, the wage boost bestowed by an MBA equals 2. Men of high ability can complete an MBA degree at a cost of only 1.25 and thus it pays them to do so. However, high ability women must fork out 2.1 units to acquire an MBA. Clearly, it is uneconomic for them to do so. Thus, women end up without an MBA regardless of ability and are paid the same amount as workers (men and women) of low ability. Note too that this equilibrium is consistent with condition (iv) in the sense that all firms are forced, through competition with one another, to offer the same 'deal' to workers.

The political significance of this separating Nash equilibrium is that, when employers harbour sexist beliefs (e.g. 'all women are of low ability and this is why they do not succeed in receiving MBAs'), it gives women strong incentives to make choices which sustain the employers' sexist beliefs: Even though the employers' beliefs are false (since, by assumption, the proportion of high ability women is the same as the proportion of high ability men), all able men acquire MBAs while not even the most able of women ever acquire MBAs in this equilibrium and, therefore, the employers' sexist outlook is never challenged. Consequently, they carry on thinking (mistakenly) that men monopolise the business schools because they are inherently more able!

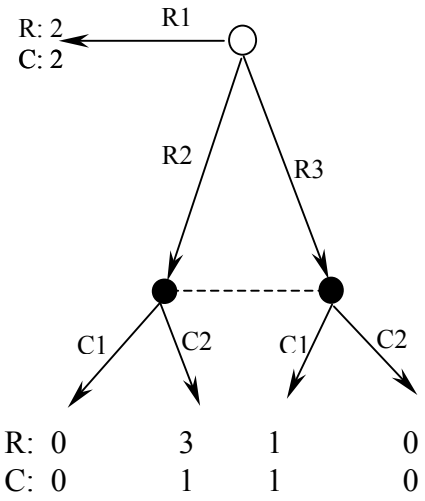
A further wrinkle to this tale of sustainable discrimination is that employers profit from these untested and 'erroroneous' beliefs; a fact that reinforces the latter further (see Chapter 6 for evolutionary reinforcement mechanisms). The reason is that, unlike high ability men, or low ability people of both genders, high ability female employees are *underpaid systematically* (e.g. they produce value equal to that of MBA-holding high ability men but receive the lower wages that all other types of employee receive). Interestingly, this underpayment is not intentional: it reflects the employers' *genuine* (albeit false) belief that there is no such thing as a highly productive woman!

The gist here is that MBAs work as signalling devices only to the extent that employers *believe* (rightly or wrongly) that ability is *uniquely correlated* with low costs in studying for an MBA. As before (see main text), there exists a simpler (pooling) Nash equilibrium in which all workers receive the same wage and business schools close down! However, if a separating Nash equilibrium emerges, it is bound to be sexist as long as MBAs cost more for women. The heart of the matter is that, in the above example, the cost of an MBA is not uniquely correlated with ability – in spite of employers' beliefs. Indeed, employers neglect an important determinant of an MBA's cost: MBA costs are generally higher for women (independently of their ability) simply because they have to do more housework than men *after class*. As long as employers do not factor this fact into their belief system, women have an incentive to behave in a manner that precludes any subversion of employers' sexist equilibrium beliefs.

3.4 FURTHER REFINEMENTS

3.4.1 Proper equilibria

Subgame perfection and sequential equilibria help tackle some kinds of indeterminacy, but there are many cases where these refinements fail to shrink the set of Nash equilibria. Consider Game 3.8, a further variant of Games 3.5 and 3.6.



	C1	C2
R1	$^{+}2,2^{-}$	$2,2^{-}$
R2	0,0	$^{+}3,1^{-}$
R3	$1,1^{-}$	0,0

Game 3.8

The game's *normal* (or *matrix*) *form* reveals two Nash equilibria (R1,C1) and (R2, C2) and both are *sequential equilibria*. To see this, let p be C's subjective probability that, if he gets a chance to play, this has happened because R chose R2 (with 1-p being C's expectation that R opted for R1). At his information set, C's expected returns from C1 and C2 are $ER^C(C1) = 0\times p + 1\times(1-p)$ and $ER^C(C2) = 1\times p$ respectively.

Clearly if $p>½$, $ER^C(C1)<ER^C(C2)$ and the following is the game's sequential equilibrium: R anticipates that, given a chance to play, C will choose C2 and so R plays R2. Thus, as long as it is common knowledge that $p>½$, (R2,C2) is the game's sequential equilibrium, yielding pay-offs (3,1). On the other hand, if $p<½$, we have a different sequential equilibrium: (R1,C1); i.e. R will expect C to play C1 and his best response to that thought is to play R1, since the (2,2) outcome is better for her than anything she might get by letting C play. Concluding, we find that in Game 3.8 the sequential equilibrium refinement does not reduce the number of equilibria, just as subgame perfection failed to in Game 3.5.

The problem really is that the *sequential equilibrium* concept has nothing to say about the origin of C's subjective probability belief p. The latter is what it is and the sequential equilibrium is predicated upon it. If we are to make headway in discerning which of the two equilibria [(R1,C1) or (R2,C2)] is more sensible, we need to tell a story regarding this p value. In game theoretical jargon, the sequential equilibrium idea actually imposes very little on *out-of-equilibrium beliefs* and so we have no guide as to whether C might think it slightly more likely that R 'trembles' toward R3 rather than R1.

One response to these difficulties has been to consider reasons for placing constraints on the type of 'trembles' which are allowed. Thus, for example, one might

argue that R is less likely to play R3 than R1 because R3 is strictly dominated by R1 (see the normal form representation). By contrast, R2 is not dominated by either R1 or R3. Thus C might well expect that p>½ in which case (R2,C2) emerges as a more plausible equilibrium than (R1,C1). Again in game theoretical language, a tremble toward R3 seems less likely than one to R2.

Indeed, if one thinks of 'trembles' occurring because players experiment, there would be no point in experimenting with R3. Alternatively, Myerson (1978) has suggested that an assessment of the cost of trembles should determine their likelihood. So people's attention is better focussed on the task and the probability of an error is lower when there is more to lose from an error. In this example, there are two equilibria but the loss to R from trembling away from one is larger than the cost of trembling away from the other. Thus, Myerson suggests that one equilibrium, (R2,C2) is more likely than the other. To see this in more detail, consider the table below.

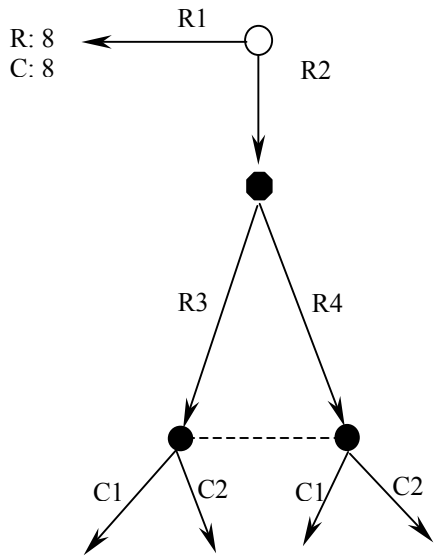
R's Expectation	**R's best reply**	**R's loss from trembling to the dominated strategy R3:**
C1	R1	2-1=1
C2	R2	3-0=3

When R anticipates that C will play C1 (if he gets a chance), she knows that her best reply is R1. If she accidentally selects R3 instead, her utility loss is 1 util (2-1=1). Had she anticipated C2, her best reply would be R2 and her utility loss from a tremble to R3 equals 3 utils (3-0=3). Thus, if Myerson is right, the greater consequence of an error when she is expecting C2 (than C1) means that the probability of trembling to R3 when expecting C2 will be considerably lower than the probability of a similar error if she expects C1. In other words, the likelihood of deviating from equilibrium (R2,C2) is smaller than (R1,C1). In Myerson's terminology (R2,C2) is, as a result, a *proper equilibrium*.

3.4.2 Forward induction

All the refinements that have been considered so far work within the tradition of *Nash backward induction*. In this section we investigate an alternative type of reasoning: the principle of *forward induction*. Nash backward induction looks to future stages of the game in order to instruct a player on what to do at present under the assumption that CKR is preserved at each node or information set. The idea behind *forward induction* is that a player takes steps now in order to cause her opponents to draw inferences on how she will play in the future. From an analytical viewpoint this is interesting because, while preserving CKR, it uncouples it from backward induction and, in so doing, brings to light some interesting issues. To illustrate this idea, consider Game 3.9.

Player R kicks off and either ends the game immediately by playing R1 or gives it another lease of life by selecting R2. In the latter case, R plays again (choosing either R3 or R4) and C makes his only choice in ignorance of whether R chose R3 or R4 (notice that broken line again). So, when and if C gets to play, he has observed R2 but not R3 or R4.



From the game's *normal*, or *matrix*, *form* we find that there are two Nash equilibria: (R1,C1) and (R2,R4,C2). Of those two it is the former that passes the test of *Nash backward induction* in the context of the tree diagram and, therefore, it is the equilibrium of this *extensive form* of Game 3.9; at least as far as the hitherto analysis of the *extensive form* goes.

Recall if we are to reason backwards from the game's last information set to its first (as we have done consistently so far), we must start with the moment when C is about to choose, without worrying about what has happened before. So, at the moment of C's choice (and *without* any information of past moves) C1 and C2 seem equally attractive to player C. Why? Because C has *no dominant strategy*. If he thought he is on the left hand side of his information set, he will want to play C1 and if he thinks that he is on the right hand side of his information set he will want to play C2.

But do R's pay-offs in *that same information set* reveal any useful information to C regarding R's strategy that led C to this information set of his? No, is the answer since R does not have a dominant strategy either. If R thought that C would verge toward C1, R would choose R3; and if she had anticipated C2, she would have selected R4. So, with no guidance from the careful study of this *particular* information set, there is no way that, if/when called upon to play, C will think it more likely that he is on the left hand side node rather than on the right hand side one. Thus C thinks that his chances of being in either of the two branches are the same and therefore randomises between C1 and C2 with probability ½.

In equilibrium, R predicts this and thinks "if I play R1, I shall receive a safe 8, otherwise, the strategy combination (R2,R3) will give me an average pay-off of [½×2+½×0] = 1 while the combination (R2,R4) will result, on average, in pay-off 5. 'I might as well choose R1." This is the unique conclusion arrived at by *Nash backward induction*.

Let us now consider an alternative type of induction, put forward by Kohlberg and Mertens (1986) and known as *forward induction*; a logic pointing to the other of the game's two Nash equilibria. Kohlberg and Mertens begin by raising a good, subversive question against the earlier conclusion of *Nash backward induction*: Why would R ever play R2 if she intended to play R3 at her second decision node?

The most R could ever expect to gain from R3 is pay-off 2 and thus a rational R would not shun R1 for R3. Indeed, the only logical explanation for passing on R1 and choosing R2 instead is that she has been lured by the prospect of pay-off 10; a prospect that remains alive only if R chooses R4. Under CKR, player C must thus interpret R's R2 choice as a firm intention to play R4, in which case his best reply is C2.

In this way, the logic of *forward induction* leads to the opposite conclusion to that of *Nash backward induction*: the latter points to Nash equilibrium (R1,C1), the former to (R2,R4,C2). The analytical difference between the two is that, whereas *Nash backward induction* draws conclusions about what will happen at the game's final information set from the pay-off structure in *that* information set *exclusively* (that is, without consulting the pay-off structure in previous information sets), *forward induction* takes into account the pay-off structure in earlier information sets and permits players to send meaningful messages (e.g. R shunning a safe 8 pay-off at the outset) concerning their future intentions (e.g. that she does *not* intend to choose R3 next).

Box 3.7

SEQUENTIAL EQUILIBRIUM, TREMBLES AND NASH BACKWARD INDUCTION

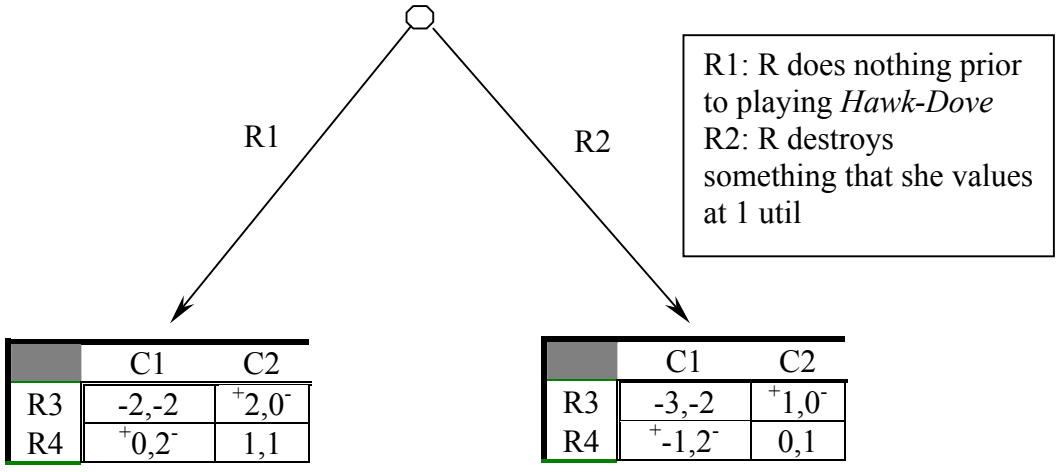
An alternative defence of equilibrium (R1,C1), and the use of *Nash backward induction* to support it, is to treat it as a *sequential equilibrium* which results from R's inability to control her 'trembling hand'. Suppose that R is intent on opting for R1. In equilibrium this must be common knowledge. What should C think, in this case, if he witnesses an R2 choice by R? Clearly, that it was an error; a tremble. In short, C's belief about the reasons he has found himself at his information set arises from a small tremble to R2. Using Bayes's rule together with the possibility of another small tremble, this time away from R4 towards R3, he will form a small probability assessment that he is at the information node on the left hand side of the tree diagram; in which case his best reply is to play C1: a conclusion reinforcing R's resolve to stick to R1 and therefore a conclusion that supports the Nash equilibrium (R1,C1).

Of course it is not surprising that a sequential equilibrium argument can be made in favour of Nash backward induction, since the latter is a constituent part of the logic behind the former. What is, however, interesting to note here is the role of trembles: whereas *forward induction* allows players to send signals regarding their future behaviour (often by deviating from some equilibrium), *Nash backward induction* assumes that any such signal is going to be dismissed as an error or tremble.

The tension between backward and forward induction becomes even more intriguing when we peruse Game 3.10. Here we have a *Hawk-Dove* game (recall Game 2.14 from Chapter 2) with a twist: one of the two players (R) has the opportunity to inflict some damage *on herself* prior to playing the familiar (static) *Hawk-Dove* game against C. Let us suppose that this damage equals one lost util for R.[13]

Why would an instrumentally rational agent choose to hurt herself? The only plausible explanation consistent with rationality (CKR to be precise) is that R may destroy something she values in order to get something she values even more. E.g. that the act of self destruction is intended to signal to an opponent that she 'means business', causing him to panic and so deliver to R more overall utility than the utility she sacrificed initially. Indeed, the expression 'burning one's bridges' refers to the military strategy of

cutting off one's *own* escape routes in order to signal to the enemies that one has no intention of retreating, and thus causing *them* such consternation that *they* flee.



	C1	C2
R3	-2,-2	$^{+}$2,0^{-}
R4	$^{+}$0,2^{-}	1,1

	C1	C2
R3	-3,-2	$^{+}$1,0^{-}
R4	$^{+}$-1,2^{-}	0,1

Game 3.10 Strategic self-punishment

To see this in more detail, let's begin with the *Nash backward induction* approach. Here, it does not make sense to play R2. Just as in any dynamic game, Nash backward induction recommends that we start at the last node(s) and work out R's optimal strategy backwards. The last node of Game 3.10 are the two static games which are strategically equivalent: They both feature the same three Nash equilibria, two in pure strategies [(R3,C2),(R4,C1)] and one NEMS (with Pr(R3)=Pr(C1)= ⅓). The only difference is that, courtesy of the destroyed 1 util, R's average returns from her NEMS is less in the game that corresponds to R2 than in the standard one corresponding to R1. To be precise, NEMS yields an average pay-off of 2/3 in the left hand side game and -2/3 in the other. So, since R2 leads to lower average pay-offs than R1, there is no point in selecting R2.

In comparison, *forward induction* treats this as commonly known and so poses the question of what C is meant to infer from the observation that R chose R2. "Why did R, whom I know to be rational, play R2 given that this condemns her to a static game whose NEMS yields a lower average pay-off for her?" asks C upon observing R2. "Obviously, she is signalling to me that she is *not* in the business of sticking to NEMS; that she intends to play R3 in order to get 1 util, as opposed to the lower NEMS average pay-off!" concludes C. And since C2 is C's best rely to R3, forward induction recommends Nash equilibrium (R3,C2). [14]

In summary, *forward induction* arguments are founded on the idea that, under CKR, things we do now *against the current of conclusions reached by backward induction* may be successful at signalling to others our intentions about future moves. Intriguingly, in both of our examples (Games 3.9 and 3.10) *backward induction* and *forward induction* pull in opposite directions and clever arguments can be put in defence of either. For instance, let us say that the forward induction argument in Game 3.10 has convinced us and so R should play R2 and follow this up with R3. Meanwhile C predicts R3 and responds with C2. Granted that it seems like a plausible strategy, is it a uniquely rational one? It cannot be. For if it were, under CKR both R and C would take it for

granted that they will end up at the pure strategy Nash equilibrium that favours R. In that case, R would not need to destroy one util to get there, since it would be common knowledge that she always gets the pure Nash equilibrium of her choice *courtesy of her opportunity to destroy a util*. So, R's non-action (i.e. her choice of R1) should still be interpreted as some sort of signal that she intends to play R3. In which case, it is never uniquely rational for R to destroy a util in order to signal an intention that her opponent knows already. And here is the rub! For we have reached the conclusion that non-action (R1) packs the same information as the choice of the self-punishing strategy (R2); a conclusion that undermines the foundation of *forward induction*, namely the supposition that R2 sends a message to C which influences the game in a manner that suits R.

If this kind of argument counts against *forward induction*, then it is not decisive because as we shall see in the next section, there are arguments against *Nash backward induction* as well. Thus we are not persuaded by the claim of one kind of induction over another in all cases and this is troubling as it is plain from examples like these that the selection of an equilibrium can turn on which of these extra 'principles' of reason (e.g. backward as opposed to forward induction) agents share.

3.5 SOME LOGICAL OBJECTIONS TO NASH, PART III

3.5.1 A critique of subgame perfection

Nash backward induction provides the foundation for both SPNE and sequential equilibria. Forward induction arguments (see previous subsection) challenged the idea that CKR must be invariably combined with backward induction. In this section we shall consider a more radical type of forward induction which strikes at the heart of Nash backward induction: its logical cohesion.[15] For the purposes of developing this argument we shall concentrate on the Short Centipede (Game 3.4 in Section 3.2.2).

Nash backward induction yields an algorithm which rational players *must* adopt. Subgame perfect Nash equilibria (SPNE) are the outcomes of beliefs produced in this manner. It is instructive to study this algorithm carefully along with its instructions to both players on how to work out their best strategies for playing Game 3.4:

The algorithm underpinning SPNE in Game 3.4

STEP 1	At node 3 compute P_3 as your maximum pay-off in the following manner: if you are player R, choose P_3 as the largest pay-off; if you are player C, choose P_3 as the pay-off you will collect when player R chooses her largest pay-off
STEP 2	Compute P_2 as your pay-off at node 2 if the game is ended there
STEP 3	If you are player R go to **STEP 6**; otherwise continue
STEP 4	At node 2, if $P_2 < P_3$ play a_2; if $P_2 > P_3$ play d_2
STEP 5	STOP
STEP 6	At node 1 compute P_1 as your pay-off if the game is ended there
STEP 7	At node 1 play a_1 if either (a) at **STEP 4** the decision is to play a_2 *and* $P_1<P_3$, or (b) at **STEP 4** the decision is to play d_2 *and* $P_1<P_2$ Otherwise play d_1

When applied to Game 3.4 our algorithm above yields the unique SPNE with R playing down at the first decision node. A rational R understands this algorithm perfectly well and, under CKR, knows that C expects her to know it with probability 1. Thus R is fully convinced that C expects him to play down at node 1. But at this point a subversive thought enters R's mind: "What will happen if I do *not* follow this algorithm, *because* C expects me to follow it? Would he not be terribly confused if I acted in a way which was totally unanticipated by him? Could there be something in it for me if I confuse him through such 'deviant' behaviour?"

The important point to note here is that Nash backward induction assumes players who (a) can discern the equilibrium path carved out by Nash backward induction, and (b) have no interest in, or capacity for, questioning what will happen if they were actually to deviate from that equilibrium path. They just follow the instructions constituting the uniquely 'good' algorithm that takes them along the equilibrium path to the SPNE (or to some sequential equilibrium when player types or prior moves are shrouded in uncertainty). However, while it is true that algorithms have no capacity for deviant

thought (i.e. for questioning the programming that has gone into them), humans are notorious for precisely that capacity. So, R may think to herself the following thoughts:

Deviant thoughts by player R in Game 3.4

There is a theory (Nash backward induction) that tells me that I should *always* play down at node 1 of Game 3.4. I know this theory, C knows it too and, moreover, C knows that I know it. So, I am convinced by this theory that C expects me, without any doubt to play down at node 1. *But what will he think if I were to play across instead, against the theory's prescription?* C will be startled that he got to play at all and must come up with an explanation of why *I* played across. One thing is certain: he cannot use Bayes's rule in order to produce that explanation. When people observe zero-probability events, Bayes's rule is of no use to them.[16] There are two possibilities. One is that he will think that I might, after all, be irrational for not doing what the theory prescribes. If this is so, he may change his game plan and play across himself (at node 2), on the off-chance that I shall make a mess of my choice again at node 3 (in which case he stands to gain pay-off 2000 instead of 800). Of course, there is the other possibility that I must reckon with. He may realise that this is exactly what I am thinking and refuse to believe that I am irrational simply because I have chosen 'irrationally' at node 1. Or he may rationalise my weird choice as a 'tremble'; that is, a momentary lapse of reason that will recur at node 3 (if we ever get that far) with an infinitesimal probability. Is it worth my while playing across in order to cause such puzzlement in C's mind? The answer is yes as long as I think there is at least a 1 in 2000 chance that, by violating the theory and playing across at node 1, C will come to believe that there is at least a 1 in 3 probability that I am an irrational person who chooses between her strategies without rhyme or reason.[17]

Are such thoughts of deviating from the equilibrium path irrational? It all hinges on what we think is rational for players to suppose when an opponent steps out of the equilibrium path. If CKR were *not* necessary in order to generate the equilibrium, and backward induction sufficed on its own (as in the ***Nim and Marienbad*** games, for example – see Box 3.3), observations of *out-of-equilibrium behaviour* would naturally lead us to conclude that the deviating player is irrational. But when backward induction is *not* enough and CKR *is* essential in order to discern the equilibrium path, things are not so simple. In those cases (e.g. in Game 3.4) CKR is an irreplaceable assumption that supports the equilibrium path. However, in the absence of a dominant strategy per node,[18] the proof that there exists such a unique path under CKR does not, in itself, provide rational players with good reason to stick to it come what may. They need to be told a story of why stepping out of equilibrium is, on balance, bad for them.

The problem is that we cannot come up with such a story which is consistent with CKR. Why? Because according to CKR (and Nash backward induction in particular) no rational player ever exits the equilibrium path. So, CKR can never tell a credible story about why a rational player would *wilfully* find herself out-of-equilibrium in the first place. How can it then explain to a rational player why she should not be 'out there'? In effect, this criticism suggests that, at least in games like this one, backward induction is logically incompatible with CKR.

To recap, under CKR players who think backwards will recognise that there are parts (or 'branches') of the game-tree which reason rules out. Since it rules them out, reason cannot 'reach' these branches [to borrow a line from a famous British advertisement]. By definition, therefore, there can be no uniquely rational theory of what one ought to do if one finds oneself, sometime in the future, in one of these branches (which, recall, rationality cannot reach). Under CKR, the probability of getting to these branches is exactly zero! But this means that, under widely held conditions, rational

players might consider taking the game to those ruled-out branches in order to subvert the course of the game in their favour (recall R's subversive train of thought above). They may want to feign irrationality, to throw dust on their opponents' eyes, to bluff. In other words, they may have a perfectly legitimate motive, as well as a decent plan, for undermining CKR. Is this really such an irrational ambition?

The following two subsections offer two rejoinders to this criticism. The first one is of a negative type: it argues that rational people will *never* deviate rationally in the manner we just discussed. The second one (Section 3.5.3) is a positive response. It admits that rational deviance is not only possible but also likely but then tries to show that these deviations are not at all in conflict with the logic of Nash backward induction. Once these rejoinders have been considered, we take stock in Section 3.5.4.

3.5.2 A negative rejoinder (based on the Harsanyi-Aumann doctrine)

One rejoinder to the above critique of *Nash backward induction* (and, consequently, of *subgame perfection*) is the re-introduction of the *Harsanyi-Aumann doctrine* which we discussed extensively in Chapter 2. According to it, equally rational players sharing the same information about the game's moves (and pay-offs) must reach identical conclusions about what they shall do at each node. To the extent that this prerequisite for consistently aligned beliefs (CAB) is a sound axiom, no room is left for rational bluffing, or any sort of deviance. For if R were to take it for granted that C will anticipate her train of thought (as well as her choice of mixed strategies), any thought of bluffing (or feigning irrationality) will vanish as bluffs succeed only when they remain, at least partially, unanticipated. So, the only behaviour that remains consistent with rationality is that *on* the equilibrium path. But, is CAB a sound axiom?

We seem to have returned to the discussion of CAB that was first aired in Section 2.5.3 of the previous chapter. However, dynamic games make CAB even harder to digest. The reason is that when players play sequentially, they observe one another's behaviour and have the opportunity to surprise their opponents before the latter play again (something that was impossible in the last chapter's static games). So, even if we accept CAB in principle, we still need to give players a credible reason why they should avoid straying off the equilibrium path. The only reason that can be given in the context of the *Harsanyi-Aumann doctrine* involves, yet again, the idea of *trembles*. Just as trembles were used earlier (see Section 3.1.3) to support NEMS, they are essential for any rejoinder to the criticism in Section 3.5.1 above.

The idea of trembles props up *Nash backward induction* by offering an explanation, consistent with CKR, of what happens out of equilibrium. Harsanyi and Aumann would tell R that the reason why she should never play across in Game 3.4 is this:

> If you do, C will interpret your choice as a result of one of those lapses of rationality that occur very infrequently but occur nevertheless. He will never believe that it contained any meaningful strategic information. In technical terms, his estimate that you shall tremble again at node 3 will be unaffected by his observation that you trembled at node 1 as he takes it for granted that errors, or trembles, as uncorrelated between nodes.[19] Thus, you shall achieve nothing as C will ignore your 'tremble' and play down at node 2. So, you are better off taking the 1 util available at node 1; that is, stick to the game's SPNE.

Is this sound advice? It is definitely not unsound, since there is no doubt that C might very well interpret any deviation at node 1 as a mere, inconsequential error of the type

that even hyper-rational people make occasionally. This is indeed a simple explanation of out-of-equilibrium behaviour which is consistent with CKR and keeps rational players on the equilibrium's straight and narrow. Without upsetting CKR, it sprinkles a little irrational dust over the game thus causing the occasional random, and thus strategically uninteresting, deviation. As long as the assumption that the randomness (and lack of correlation across nodes) of these trembles is commonly known, *Nash backward induction* is secure.

The above, unfortunately, re-phrases the problem without solving it. Granted that common knowledge of the strategic irrelevance of deviations from the equilibrium path supports *Nash backward induction* and the resulting SPNE, the question as to whether it is sensible to assume such common knowledge remains. To put it slightly differently, granted that Harsanyi's and Aumann's advice to R is not unsound, why should we assume that it is uniquely sound? Why do we rule out the possibility that alternative advice (e.g. "go ahead; bluff!") may also prove sound sufficiently frequently? And if there is no good reason for ruling this out, then there is no good reason as to why C should *always* dismiss a deviation by R at node 1 as a strategically meaningless move.

There is another way of putting this argument. As we have seen, CKR rules out certain branches of some games, like the *Centipede* (Game 3.4), and so assigns them a zero probability of being reached in SPNE. The analysis which supports this conclusion is based on an understanding of what would happen if players actually reached these out-of-equilibrium nodes. However, CKR also insists that this 'understanding' assumes that players are commonly known to be rational. But why should we assume that players are rational when they get to what is a putative out-of-equilibrium node for rational players?

The only way to avoid a possible internal inconsistency here is to come up with an explanation of why players might be at an out-of-equilibrium node which does not undermine the assumption that they are nevertheless commonly known to be rational. Trembles (that is, momentary random lapses) do the trick, of course, because they allow agents to deviate from what rationality requires without bringing into question their rationality. It is 'trembles' to the rescue, so to speak.

However, the longer the game the less plausible that explanation is. For example, in the longer version of the *Centipede* (see Problem 3.4 at the chapter's end) the SPNE is supported by a long string of out-of-equilibrium beliefs about what would happen at later decision nodes if they were reached. To keep this string consistent with CKR, these stages of the game could only be reached via independent, random trembles. But how plausible is it to assume that a sequence of such trembles could take players to the last decision node?

Trembles in Game 3.4 are one thing, but to get to the last potential decision node in games like the longer *Centipede* of Problem 3.4 (or of the game described in Box 3.4), it will surely seem increasingly plausible that trembles could be a more systematic part of the player's behaviour. And if they are, then any opponent should want to take account of them in deciding what to do for the best; and once this happens trembles will no longer be random events free of strategically important meaning. The moment trembles acquire strategic meaning, all sorts of (equally rational) behavioural patterns become possible in games such as the *Centipede*.

3.5.3 A positive rejoinder (based on sequential equilibrium)

The second rejoinder acknowledges the above point: the need to consider systematic trembles; namely, bluffs and all sorts of other *deliberate* steps out of SPNE. Its analytical foundation is the *sequential equilibrium* concept of Section 3.3.4. In the latter we saw that, if we permit for a small initial probability p that R is systematically irrational, a rational R will begin to contemplate bluffs and stratagems very similar to R's subversive logic as presented in Section 3.5.2 (that is, the train of thought toying with the idea of taking the game to its last node by convincing C that R may indeed be systematically irrational).

In short, if the critique of Section 3.5.2 confirms that CKR is perhaps an inappropriate axiom for a finite dynamic game like Game 3.4, then let us relax it by acknowledging that there is *always* a probability p that player R is systematically irrational. In that case, as was demonstrated in Section 3.3.4, Game 3.4 becomes Game 3.7 and the deviance merely shifts us from *Nash backward induction* to *sequential equilibrium* as the appropriate solution concept. And since the latter is itself founded on *Nash backward induction* (recall how in Section 3.3.4 the sequential equilibrium of Game 3.7 was derived backwards and consistently with CKR), deviance is explained fully as part of some richer and more complex equilibrium (as opposed to out-of-equilibrium) behaviour. Nothing satisfies a game theorist more than turning a criticism of equilibrium into a demonstration of its superior analytical power.

There are, unfortunately, two major difficulties in accepting this as the final word on the subject. The first one is that, although CKR may have been relaxed (with the introduction of $p>0$), it was surreptitiously replaced by another type of common knowledge that seems even more controversial than CKR: common knowledge of two probabilities; of the initial probability p that R is systematically irrational, and the probability r that a rational R will bluff at node 1 (pretending to be irrational). Note the gravity of this common knowledge assumption. It is one thing to say that two people commonly know that 1+1=2 or that they are both rational. It is quite another *commonly* to know the probability that one of them is irrational or the probability with which she will bluff if rational.

The second difficulty is subtler. The sequential equilibrium is an extension of subgame perfection. It is founded on the latter's logic which it extends to cases in which there is uncertainty about the character of one's opponent (or about unobserved earlier moves). Of course, modelling complex situations requires simplifying assumptions. It made perfect sense to establish the SPNE equilibrium *in the absence of uncertainty* before proceeding to model rational behaviour under uncertainty (in the form of a sequential equilbrium). Physicists do the same all the time. For example they first worked out of the laws of mechanics by assuming the absence of friction and then relaxed this assumption in order to generalise and complicate their theories. However, before generalising and re-admitting friction into their model, they had absolutely no doubt that their theories were correct in a world without friction.

This is not so with game theory. *Nash backward induction* did point to a theory of rational action (for Game 3.4) in the absence of uncertainty about pay-offs and rationality. Unfortunately, it was not entirely convincing (as explained in Section 3.5.1). So, the more complex sequential equilibrium story is now enlisted. This is quite unlike

the normal move in physics. It is as if physics had failed to discern the laws of mechanics when there is no friction and introduced friction in order to explain this failure. One worries that game theory elevates the original, simpler problem (e.g. the analysis of Game 3.4) to a higher level of abstraction (by turning it to Game 3.7) without solving it (e.g. replacing CKR with the more problematic common knowledge of probabilities p and r). For if the foundation (SPNE) is problematic, what should we expect of the edifice (sequential equilibrium) built upon it?

3.5.4 Summary: Out-of-equilibrium beliefs, patterned trembles and consistency

In the preceding sections we confronted some difficult philosophical questions. Is instrumental rationality, when commonly known, logically coherent in the context of dynamic games? We encountered powerful arguments on both sides. On the one hand there was the line that most game theorists will adopt: there is no fundamental problem with instrumental rationality that cannot be solved by accepting the presence of independent or uncorrelated errors. On the other hand, there is the minority view within game theory (but one that more often appeals to philosophers and other social theorists) that instrumental rationality falls short of its own ambition in finite dynamic games as it struggles to impose its own narrative on parts of the game that it must rule out in order to procure a solution.

The crux of the problem facing game theory is that it has to introduce the possibility of some lapse of rationality to explain what rationality demands. This is because what rationality demands is often determined in dynamic games by a consideration of what would happen if *rational* players actually end up in what turn out to be out-of-equilibrium decision nodes. But why should one assume that players behave rationally when they find themselves at an out-of-equilibrium decision node? Surely if the analysis of SPNE is correct, then rational players should not reach these out-of-equilibrium nodes.

The only way to avoid a troubling inconsistency here, in the way that SPNE is constructed, is to allow for random trembles that take rational players to out-of-equilibrium nodes. To the extent that trembles remain random, infrequent, and thus meaningless, they can be interpreted as rare 'excursions from reason' which do not affect what reason prescribes. In this sense, players are advised to ignore the trembles, once they occur, and carry on assuming that, in all probability, no further trembles will occur any time soon (since they are so rare). In effect, this presumption allows theorists (and players alike) to continue to assume that all players will think and act rationally at the next node – including those who have recently strayed from the equilibrium path!

Once 'trembles' have become central in this way to dynamic games, the key question becomes whether it is reasonable to assume that 'trembles' are always just random events. If trembles are patterned in some ways, then rational players should take this into account. This may not only change what rational players do in response to trembles, it may even lead rational players to decide to tremble *purposefully*. What happens will depend on how 'trembles' are interpreted, but it is not difficult to see how alternative equilibria to the one proposed by SPNE could arise. In the *Centipede*, it could become rational to play across and in Games 3.9 and 3.10 (among others) alternative *proper equilibria* might obtain. In short, the quest for determinacy that has fuelled much

of the Nash refinement project has apparently and paradoxically recreated indeterminacy because there is more than one plausible way to refine this equilibrium concept.

There is, however, one way in which such patterned trembling can be accommodated within the conventional analysis and that is by allowing for players commonly to know that one of them might be playing 'irrationally', where what it is to play irrationally is also commonly known. The game can then be analysed using the sequential equilibrium concept and this will allow rational players to behave in an irrational way (i.e. 'tremble') in some cases (see Section 3.5.4). Two doubts remain over whether this will dissolve the worry. First the demands of CKR and CAB in such games seem especially onerous. Secondly, this approach still turns on a form of backward induction and so will not appeal to those who are persuaded by a form of forward induction in the interpretation of trembles (as in the concept of *proper equilibria*).

The issue here is closely related to the earlier one (see Section 2.5.3 of Chapter 2) regarding CAB in static games. The construction of subgame perfection assumes that no player can believe that someone will play in a way which they actually would not, which is exactly the point of CAB. So an implicit assumption of CAB is at work. The difference here, however, is that this projection from CKR looks rather more controversial in extensive form games when these are beliefs which need not be tested in equilibrium. The comparison with the role of CAB in the construction of the Nash equilibrium concept in static games is instructive in this regard (see Sections 2.5.2 and 2.5.3).

Most game theorists expect Nash equilibrium behaviour among stable populations who are experienced in the type of game under study. Should players use non-Nash strategies in a static game because they hold *in*consistently aligned beliefs, then the inconsistency would be revealed the moment the moves had been made. Players will thus learn from their mistakes and when they play the same game against some other opponent (always as a one-shot game) their behaviour and beliefs ought to be better equilibrated. However, in extensive form games such as Game 3.4, players who follow their equilibrium strategy of playing down, supported by the belief that C will also play down at node 2, will never be putting this belief to the test because C is never called upon to play. So it seems that rational curiosity pulls players toward deviating from the equilibrium path if only in order to test their beliefs. In conclusion, if we had a problem accepting CAB in Chapter 2, we have good reason to be doubly wary of Nash backward induction (CAB's dynamic version) in this chapter.

3.6 CONCLUSION

3.6.1 The status of Nash and Nash refinements

We conclude this chapter by bringing together some of the arguments which have surfaced over the Nash equilibrium concept. First, this concept depends in general on CAB and not just the assumptions of rationality and CKR. This is the tricky epistemological problem at the foundations of game theory to which we referred in Chapter 1 and which we have followed through various twists and turns in the last two chapters. Something else seems to be required and the best game theory has come up with so far is the *Harsanyi-Aumann doctrine* which is used to underpin an assumption of common priors. This has the effect of making rational players believe that there is a uniquely rational way to play a game *because rational players must draw the same inferences from the same information*.

Once this is conceded, then indeed it follows from the assumptions of instrumental rationality and CKR that rational players must hold consistently aligned beliefs and the way to play must constitute a Nash equilibrium. Therefore, our criticism of Nash effectively boils down to a challenge of the Harsanyi-Aumann argument (recall Sections 2.5.3, 3.1.4 and 3.5). But even if this controversy is set on one side (and we shall say more about how this might be done below), there remains a difficult question which game theorists in the Nash tradition must answer. How is one Nash equilibrium selected when there are many?

Most answers to this question have relied on three components (in varying degrees): the existence of *trembles*, the use of *backward induction* (in dynamic games) and a *Bayesian consistency* between beliefs and the chosen strategies (in games of incomplete or asymmetrical information). Refinements in this tradition either explicitly or implicitly require that agents hold mutually consistent beliefs (CAB). Naturally there are reasons for doubting this in the context of refinements of Nash just as there were in connection with the Nash equilibrium concept itself. In addition, there are special reasons for doubting this in dynamic games because of the difficulty of accounting for out-of-equilibrium beliefs by appealing to trembles alone. In some games, it seems more natural to relax CKR and hence CAB. However, this means that we are moving further away from pinpointing a definitive solution with which to defeat indeterminacy.

Suppose we set this new difficulty on one side as well. Still there are problems. There are, for instance, games with multiple sequential equilibria (the refinement which uses all three elements: trembles, backward induction and a Bayesian consistency between beliefs and strategies). To narrow down the equilibria, yet again something more must be added. In this instance, more needs to be said about those 'trembles'. The difficulty, however, is to know quite what might be said without relaxing CKR and thereby recreating the problem with the introduction of further potential equilibria.

There have been various attempts at this, but none is especially or generally convincing. Indeed, some of these attempts (like the use of *forward induction* arguments for instance) are difficult to reconcile with other refinement principles (like *backward induction*). Perhaps all that can be said is that none of these further ideas regarding trembles can be derived in any obvious way from the assumptions of instrumental rationality and CKR. Hence these refinements (e.g. *proper equilibria*), like the Nash

equilibrium project itself, seem to have to appeal to something *other* than the traditional assumptions of game theory regarding rational action in a social context.

In other words, the attempt to bring greater determinacy to games by refining the Nash equilibrium concept is not always successful, partly part because it often relies on a form of CAB which is even more difficult to accept (as a general proposition) in dynamic games and in part because there are various, plausible but competing types of refinement (e.g. those based on forward as compared with backward induction). Of course, there will be some games where CAB, 'or the 'principle of rational determinacy', is more plausible than others and some games where one refinement seems quite natural. But we seem to be someway short of a *general* theory of rational action in games which is built exclusively around the Nash equilibrium concept.

3.6.2 In defence of Nash

What further assumptions might one have to make to render the Nash equilibrium concept (and its refinements) more appealing? The question is interesting partially as a matter of intellectual curiosity. If we are to use Nash, what else must we assume? But it is also interesting because the answer might help reveal those circumstances in which the Nash equilibrium concept (and its refinements) can be appropriately applied. We mention two possibilities here. The first is for game theory to become more thoroughly Humean.

The Humean turn

In our introduction (see Chapter 1) we emphasised that game theory adopts a version of David Hume's model of human agency, albeit one which relies more on the power of reason than Hume did. For example, Hume did not believe that reason offers a complete guide to action. On the contrary, Hume often remarked that, if reason is not provided with sufficient raw materials, it could offer no guide at all. In other words, preferences alone do not necessarily guide action. To use the metaphor of a pair of scales for reason, it is as if we place two equal weights on each side of the scales; we can hardly blame the scales for not telling us which is heavier!

What happens when preferences are such that reason cannot distinguish the uniquely rational action? According to Hume, it is then that custom and habit (or in more modern terms, *conventions*) fill the vacuum and allow people to act consistently and, with luck, efficiently. If game theory were to become more thoroughly Humean in this sense by allowing for the role of convention, then it might have an answer both to the question of 'Why Nash?' and to the question of how to select between Nash equilibria when there are many.

For instance, without enquiring too deeply about how customs and conventions are constituted at this stage, it seems quite plausible to conjecture that they must embody behaviour consistent with the Nash equilibrium. Otherwise at least some people who reflected (in an instrumentally rational fashion) on their custom-guided behaviour would not wish to follow the custom or convention. Thus in the absence of clear advice from reason, if agents appeal to custom as a guide to action then this might underwrite the Nash equilibrium concept. Likewise with the problem of Nash equilibrium selection: if reason cannot tell us which of the many equilibria will materialise, and we come to rely

on custom, then we have our explanation. For example, Game 3.9 can be resolved if we happen to know that people subscribe as a matter of convention, say, to the principle of forward induction *á la* Kohlberg and Mertens (1986).

The introduction of custom and convention can be helpful to game theory in these ways, but it is also a potentially double-edged contribution. First, there is a potentially troubling question regarding the relation between convention-following and instrumental rationality. The worry here takes us back to the discussion in Chapter 1 where for instance it was suggested that conventions might best be understood in the way suggested by Wittgenstein or Hegel. In short, the acceptance of convention may actually require a radical reassessment of the ontological foundations of game theory. Secondly there is a worry that while conventions may answer one set of questions for game theory, they do so only by creating another set of problems since we shall want to know how conventions become established and what causes them to change. There is an ambitious Humean response to both worries that treats conventions as the products of an evolutionary process and which we shall delay discussing until Chapter 6.

The Kantian move

The second move is to appeal to a part of the Kantian sense of rationality: that part which requires that we should act upon 'universalisable' rules; that is, rules which can be acted upon by everyone. In this context, the 'best reply to another's action' rule is one which generalises to form a Nash equilibrium when 'best' is understood in an instrumentally rational fashion. Of course there may be other demands which Kantian reason makes, but taken in isolation, the *universalisability* condition might provide an alternative foundation for Nash. However, the *universalisability* requirement will not help with the problem of Nash equilibrium selection because every principle of refinement has the principle of *universalisability* built into it by construction. To answer this question, it seems that Kantians, like Humeans, will have to appeal to something outside preferences and calculative beliefs (e.g. something like conventions or an 'objective' sense of 'virtue' or of 'right').

For the most part game theorists have not made either move and we examine why this is the case below. For now, it is worth recording that there is a third move which could be made.

Abandon Nash

There is, of course, a third possibility. Instead of seeking ways of making sense of the use of Nash, one could relinquish the concept. Why not give up on the Nash concept altogether? This 'giving up' might take on one of two forms. First, game theory could appeal to the concept of rationalisable strategies (recall Section 2.4 of Chapter 2) which seem uncontentiously to flow from the assumptions of instrumental rationality and CKR. The difficulty with such a move is that it concedes that game theory is unable to say much about many games (e.g. Games 2.9 to 2.17 and Game 3.4, among many). Naturally, modesty of this sort might be entirely appropriate for game theory, although it will diminish its claims as a solid foundation for social science.

What would such an admission mean for social scientists? Either they could make the Humean (or a Kantian) move as discussed above, or alternatively they could opt for a more radical break. Both the Humean and Kantian critiques recognise the ontological value of the essential elements of instrumental rationality. What they do deny is that instrumental rationality is all that governs human action. Many social scientists would want to go further and to reject that a proper analysis of society can have instrumental rationality at its core. In this case, the whole approach of game theory is rejected and the problem of justifying Nash does not arise.

For example, Hegelians evoke an historical perspective from where the observer sees society as a constantly flowing magma: people's passions and beliefs reach violent contradictions; social institutions clash with community or group interests and are reformed as a result; desires remain unfulfilled while others are socially created; everything is caused by something and gives rise, through contradiction, to something else. Yet this is not an anarchic process.

The Marxist interpretation of this Hegelian move portrays the reason of men and women maturing as a result of their historical participation. It is an evolving reason, a restless reason, a reason which makes a nonsense of an analysis which starts with fixed preferences and acts like a pair of scales. Unlike the instrumentally rational model, for Hegelians and Marxists action based on preferences feeds back to affect preferences, and so on, in an ever unfolding chain. Glimpses of how this complicates the task of game theory will be had in Chapter 6. For now we merely note that even in a simple game (e.g. Game 3.4), instrumental rationality hits the brick wall of having to provide a narrative of irrational behaviour before it can discover itself.

Hegelians would not hesitate to see this as one of those significant occasions where an established way of thinking (instrumental rationality) encounters its contradiction (out of equilibrium behaviour) and, in the process, transcends itself into something richer and more wholesome. Of course, yet again, the price instrumental rationality must pay for overcoming itself in this manner is more indeterminacy as the historical process cannot fit neatly into some pre-ordained equilibrium path. Likewise some social psychologists might argue that the key to action lies less with preferences and more with the cognitive processes used by people; and consequently we should address ourselves to understanding these processes.

We divide as two authors at this point. For SHH, there are major difficulties with a purely instrumental account of reason (see Hargreaves Heap, 1989), but it seems undeniable that there are important settings where people do have objectives which they attempt to satisfy best through their actions (i.e. they act instrumentally). In such settings game theory seems potentially useful both when it tells us what might happen and when it reveals that something more must be said about reason before we can know what will happen. YV also recognises this but insists that the social phenomena which *need* to be explained if we are to make sense of our changing social world, cannot be deciphered in terms of a model of instrumentally rational agents (see Varoufakis, 1991). Quite simply, the lens of instrumental rationality fails to reveal the cogs and wheels of the significant social processes which write history. This destines game theory to a fascinating footnote in some future text on the history of social theory. We let the reader decide.

3.6.3 Why has game theory been attracted 'so uncritically' to Nash?

Whatever your view on this last matter, it is a little strange that game theorists have remained so committed to the Nash equilibrium concept. It seems that either they should address the difficulties by taking one of the, at least, two positive and more expansive philosophical moves identified above; or they should junk the enterprise and recommence the analysis of social interaction using a different tack. In other words, why has game theory been content to use a series of concepts based on Nash (that is, the Nash equilibrium, Nash backward induction, SPNE, sequential equilibria etc.), which do not seem warranted by their foundational philosophical assumptions (instrumental rationality and CKR) but depend on the controversial assumption of CAB (or, equivalently, the 'principle of rational determinacy')? Even more puzzlingly, why do so few game theorists discuss these matters at all? In a sense, this is a question in intellectual history (or perhaps the sociology of knowledge) and we have no special qualifications to answer it. Nevertheless, we believe that a variety of contributory factors can be identified.

First, one possible way to understand the reluctance of game theory to confront its reliance on the Nash equilibrium concept is to see game theory as essentially a child of its times. Its origins belong firmly in the project of 'modernity' and like all modern thinking, it has unreflectingly assumed that there is a uniquely rational answer to most questions. This perhaps explains the commitment to Nash and perhaps why the problems with Nash (which actually have a long history in game theoretical discussions) are only now beginning to worry game theorists in a serious way. The critical momentum now is itself part of the new contemporary zeitgeist and we can expect a much greater receptivity to the idea of conventions (which can vary with time and place) playing a significant role in social interactions once the ideas of postmodernity have seeped further into the consciousness of economists (see Box 3.8).

Box 3.8

MODERNITY UNDER A CLOUD: LIVING IN A POST-MODERN WORLD

One of the quests within modernity has been to find ways of resisting the tendency towards the relativisation of all values and claims to power by grounding knowledge on objective truths and legitimating authority through an appeal to expertise and the common good. According to postmodern philosophers, such as Jean Francois Lyotard, this legitimation crisis has been solved through the invention of what he calls 'the great meta-narratives' of the modern period. By this they mean all those overarching belief systems originating in the *Enlightenment*: from the belief in rationality, science and causality to the faith in human emancipation, progress, democracy and class struggle. These grand stories have been used over what Lyotard calls the past 'two sanguinary centuries' to legitimate everything from war, revolution, nuclear arsenals and concentration camps to Taylorism, Fordist production models and the gulag. The collapse of faith in these meta-narratives heralds what Lyotard calls the *postmodern condition*. Postmodern narratives, by contrast, are eager to eschew any kind of metanarrative "…regardless of what mode of unification it uses, regardless of whether it is a speculative narrative of a narrative of emancipation" (Lyotard, 1984, p.37). In this sense, the game theorists' struggle against indeterminacy would cause postmodernists to smile at the folly of yet another, particularly ambitious, offshoot of modernity.

It is therefore clear that postmodernism would be scornful of the theoretical struggles against indeterminacy which we summarised in this chapter. For it is founded on a radical aversion to: (a) *any* group of 'experts' (e.g. game theorists) who claim privileged access to objective truths; and (b) the model of rational men and women at the heart of modernist theory (especially the rigid, humourless *homo economicus*). It scoffs at grand stories (or metanarratives) about how the world functions (or ought to) and it discerns, behind any such story, a bid by the aforementioned 'experts' to gain, exercise and preserve social power at the expense of 'others' (e.g. *you*, dear reader) who are not privy to the metanarrative's

intricacies. Postmodernity's problem, however, is how to avoid sliding into an intellectually and politically disabling relativism; one in which nothing meaningful can be said about the systematic aspects of social and economic life.[20]

Secondly, it is also possible that the strange philosophical moorings of neoclassical economics and game theory have played a part. They are strange in at least two respects (see also Mirowski, 1989, 2002, for a different account of this strangeness). The first is a kind of amnesia or lobotomy which the discipline seems to have suffered during the post-war period regarding most things philosophical. As evidence of this, one need only reflect on the incongruity of the discipline's almost wholesale methodological commitment to one form of empiricism. This was doubly incongruous not only because most philosophers of science have been agreeably sceptical about the claims of such a method during this period, but also because this methodological commitment has been almost completely at odds with the actual practice of economists (see McCloskey, 1983). The second is the utilitarian historical roots of modern economics. This is important because it perhaps helps explain why the full Humean message has not been taken on board by the discipline. Indeed, had Hume unreservedly been the philosophical source for the discipline, then it is more than likely that conventions would have occupied a more central place in economics.

Thirdly, the sociology of the discipline may provide further clues. Two conditions would seem to be essential for the modern development of a discipline within the academy: (a) the discipline must be intellectually distinguishable from other disciplines, and (b) there must be some barriers to the amateur pursuit of the discipline. (A third condition which goes without saying is that the discipline must be able to claim that what it does is potentially worth while.) The first condition reduces the competition from within the academy which might come from other disciplines (to do this worthwhile thing) and the second ensures that there is no effective competition from outside the academy. In this context, instrumental rationality has served economics very well. It is the distinguishing intellectual feature of economics as a discipline and it is amenable to such formalisation that it keeps most amateurs well at bay. Thus it is plausible to argue that the success of economics as a discipline within the social sciences has been closely related to its championing of instrumental rationality.

Consequently, to venture outside instrumental reason, CKR and CAB (by introducing conventions or, even worse, to make half-disguised invitations to Wittgenstein, Kant or Hegel) is a recipe for undermining the discipline of economics (as distinct from, say, sociology). Of course, intellectual honesty might require such a move but it would be foolish to think that the academy is so constituted as always to promote intellectual development *per se*. It is often more plausible to think of the academy as a battleground between disciplines rather than between ideas. In this struggle, the disciplines which have good survival features (like the barriers to entry identified above) are the ones that prosper. In this vein, the determination of which features help a discipline survive depends less on intellectual criteria and more on the social and political imperatives of the times.

To put the point more concretely, individual economists may find that it is fruitful to explain the economy by recourse to sociological concepts like conventions. Indeed this seems to be happening. But such explanations will only prosper in so far as they are both superior and they are not institutionally undermined by the rise of neoclassical economics

and the demise of sociology. It is not necessary to see these things conspiratorially before discerning the point of this argument. All academics have fought their corner in battles over resources and they always use the special qualities of their discipline as ammunition in one way or another. Thus one might explain in functionalist terms (see Box 3.9) the mystifying attachment of economics and game theory to Nash, as well as the reluctance to discuss the philosophical problems openly.

Box 3.9

FUNCTIONAL EXPLANATIONS

Functional explanations are sometimes regarded as peculiar because they appear to explain something by its beneficial effects. It is the effect of an action rather than an intention which lay behind the action which is used to explain why the action was taken. Such explanations have the following form (see Elster, 1983).

(1) Y is an effect of X.
(2) Y is beneficial for some agent Z.
(3) Y is unintended.
(4) The causal relation between X and Y is unrecognised.
(5) Y maintains X by a causal feedback loop through Z.

In this way the behaviour X of some agent Z is explained by its function Y for this agent. Thus one might argue that economists (Z) utilise CKR/CAB (X) which in turn have the unintended consequence of erecting barriers to entry (Y) because CKR/CAB based narratives are so amenable to intricate formalisation (e.g. mathematical models) and this keeps amateurs at bay. Of course, this is unintended because most economists and game theorists *believe* in the virtue of the equilibrium analysis founded on CKR/CAB (rather than valuing the fact that it can be so readily formalised to keep amateurs out). The existence of barriers to entry (Y) in turn is beneficial for the 'community' of economists/game theorists (Z) in the competitive battle for resources within the academy and so maintains their position in the academy and their use of the CKR/CAB-based models (X).

In other words, we might explain in part the apparent reluctance of economists to go beyond Nash by introducing conventions, notions of human rights and capabilities etc. because this threatens a break with a type of model which holds the key to the success of economics in the academy… well, it is only a story!

We have no special reason to prioritise one strand of our proposed explanation regarding the reluctance of economists to scrutinise the logical and philosophical foundations of Nash. Yet, there is more than a hint of irony in the last suggestion (Box 3.9) because Jon Elster has often championed game theory, and its use of the Nash equilibrium concept, as an alternative to functional arguments in social science. Well, if the use of Nash by game theorists is itself to be explained functionally, then…!

PROBLEMS

3.1 Find the Bayesian Nash equilibria in the static game below when (α,γ) are equal to (0,3) with probability π or (3,1) with probability $1-\pi$, and β equals –1 with probability ρ or 1 with probability $1-\rho$.

	C1	C2
R1	α,β	γ,0
R2	2,1	2,0

3.2 Prove that the Bayesian Nash equilibrium of Game 3.3b is given by the following:

R plays R1 when $\alpha > (2 + \varepsilon)/(8 + \varepsilon^2)$

R plays R2 when $\alpha < (2 + \varepsilon)/(8 + \varepsilon^2)$

and

C plays C1 when $\beta > (4 - \varepsilon)/(8 + \varepsilon^2)$

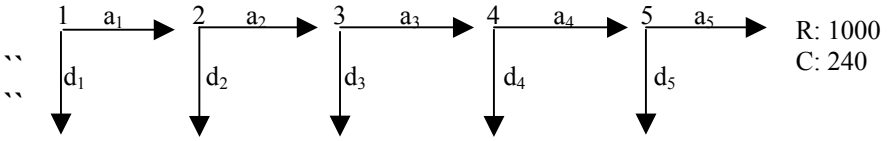
C plays C2 when $\beta < (4 - \varepsilon)/(8 + \varepsilon^2)$

3.3 Re-write the game below in extensive form and show that who plays first determines which of the original Nash equilibria will be selected.

	C1	C2	C3
R1	5,0	-1,-1	10,-1
R2	-1,-1	0,5	0,0
R3	-1,10	0,0	6,6

A battle-of-the-sexes game augmented with a non-equilibrium co-operative strategy

3.4 In the extended centipede below, show that the SPNE is a strategy profile involving the choice of the *down* (d_i) strategy at every node i of the game. Show also that, in the context of a sequential (or Perfect Bayesian) equilibrium, an initial probability that R chooses irrationally (that is, with equal probability between strategies d_i and a_i, for all i) of about 0.0022 suffices to cause a rational R *always* to play across at node 1.



Extended Centipede - Description: R kicks the game off at node 1. If R plays a_1 then C gets a chance to play at node 2. (Otherwise the game ends and they collect 1 and 0 utils respectively.) If C plays a_2 then R gets a chance to play again at node 2. And so on.

3.5 Consider the following game: The Executive (e.g. President) is trying to push through the legislature (e.g. Congress) a series of bills that the latter is unsympathetic towards.

Pay-offs	Congress	President
Congress does not amend	0	3
Congress amends and President acquiesces	½	2
Congress amends and President fights	-½	1

The President proposes legislation and Congress must decide whether to make amendments. If it decides to amend, then the President must decide whether to fight the amendment or acquiesce. Looking at the President's pay-offs it is obvious that, even though he or she prefers that the Congress does not amend the legislation, if it does, he or she would not want to fight on the floor of the House. The SPNE in the one-shot version of this game is simple enough: *Congress amends and the President gives in*. Now, suppose this game begins at time t=0 (e.g. the President's inauguration) and will end once and for all at time T (e.g. the compulsory end of his or her term). Moreover assume that, from the beginning, Congress entertains probabilistic doubt p_0 that the President is dogmatically unbending and would thus 'fight' (irrationally) for his or her legislation regardless of pay-offs (that is, p_0 is the probability that the President would fight even in the one-shot version of the game).

Show that there exists a sequential equilibrium under which Congress will not amend the President's legislation for a period of time k which is proportional to p_0.

NOTES

[1] For future reference, it is worth noting that the tie-in between beliefs and strategies contained in steps (b) and (c) in the definition above is a characteristic move in the so-called Nash refinement project. We will come across it frequently throughout this chapter. It will also be plain that, by construction, these steps impose CAB and so the tie-in sits quite comfortably with an equilibrium concept (that is, the Nash equilibrium) which is already premised on the move to CAB.

[2] This has been cited as the main reason why John Harsanyi was awarded, jointly with John Nash, the Nobel Prize in Economics. The third co-recipient, Reinhart Selten, was responsible for two other major advances: One was the notion of a trembling hand equilibrium (see the previous section), the other the idea of *subgame perfection*, to which we shall turn in the next section.

[3] This can be easily checked as follows: ER(R1) = 0, ER(R2) = q-(1-q) and thus d_{12}^R = ER(R1)-ER(R2) = -*q*+(1-*q*) which equals zero only when *q*=1/2. Similarly, d_{12}^C = ER(C1)-ER(C2) = *p*-3(1-*p*) = 0 only when *p* = ¾. Thus, the game's NEMS is given by probabilities (*p*,*q*) = (3/4,1/2).

[4] Our system of equations $\begin{cases} p = 1 - \dfrac{2q-1}{\varepsilon} = \dfrac{\varepsilon - 2q + 1}{\varepsilon} \\ q = 1 - \dfrac{3-4p}{\varepsilon} = \dfrac{\varepsilon - 3 + 4p}{\varepsilon} \end{cases}$, when solved for *p* yields the equation

$p = \dfrac{\varepsilon - 2\left(1 - \dfrac{3-4p}{\varepsilon}\right) + 1}{\varepsilon} \Rightarrow p = \dfrac{\varepsilon^2 - \varepsilon + 6}{\varepsilon^2 + 8}$. As $\varepsilon \to 0$, $p \to 3/4$. Meanwhile, the limit of *q*, once we substitute $p = \dfrac{\varepsilon^2 - \varepsilon + 6}{\varepsilon^2 + 8}$ in the expression for *q* above and let $\varepsilon \to 0$, is ½.

[5] A longer version of the Centipede is examined at the end of the Chapter – see Problem 4.

[6] In the context of the earlier definition of information sets and singletons (see previous subsection) the same point can be expressed thus: C's information set is not a singleton as it contains more than one nodes.

[7] Recall that the SPNE concept depends on similar beliefs about what will happen in the future. The difference here is that that these beliefs are generated endogenously by the dynamic structure of the game through the process of *Nash backward induction*.

[8] Bayes' rule was first introduced in Chapter 1. As an example consider the case where X is the *observed* event 'cloud in the morning' and Y is the possibility of event 'rain in the afternoon'. If we have just observed a cloudy morning sky, what is the chance of rain in the afternoon? Suppose we know that the probabilities of (i) cloud in the morning when it rains in the afternoon, (ii) cloud in the morning, (iii) rain following a sunny morning are $\frac{3}{4}$, $\frac{1}{3}$ and $\frac{1}{4}$ respectively. Bays' rule is given as

$\Pr(X \mid Y) = \dfrac{\Pr(Y \mid X)\Pr(X)}{\Pr(Y \mid X)\Pr(X) + \Pr(Y \mid not\ X)\Pr(not\ X)}$. Substitution of the given probabilistic beliefs in Bays' rule yields a conditional probability of Pr(X|Y) = $\frac{3}{5}$. This means that, following the observation that the morning was cloudy, the probability with which one should expect rain in the afternoon is $\frac{3}{5}$. Learning here takes the form of using observation in order to form a better probability estimate of the uncertain phenomenon one is interested in. In Bayesian language this is referred to as converting, by means of empirical evidence, prior beliefs into posterior beliefs.

[9] Furthermore, C would also know this and would not believe, in equilibrium, that a choice of a_1 means anything about R's rationality; in which case R would never bluff but, knowing that C knows this and does not expect her to bluff if rational, she would always bluff etc. etc.

[10] A Bayesian perfect Nash equilibrium is a special case of sequential equilibrium. It applies to games in which the source of uncertainty concerns not the moves that players made previously but, instead, the character (or utils) of one's opponents when earlier moves are public knowledge.

[11] In reality, this is a model of *reputation preservation* (as opposed to *reputation building*) since players are motivated to bluff with positive probability (while on the brink of indifference) in order to maintain a pre-existing reputation.

[12] This is similar to case (1) of the sequential equilibrium of Game 3.8 in the previous section (when R's initial reputation p_1 exceeded 1/3) because there too R's high initial reputation meant that, in a bid to preserve it, she would always behave as if totally irrational. This is also a pooling equilibrium in the sense that rational and irrational Rs act in an identical manner.

[13] Notice that in this type of game we have given a static game a dynamic structure by incorporating a prior move that might or might not be taken by one of the players before the two choose among the static game's strategies simultaneously. Once the original static game has been transformed into a dynamic one, the analysis of this chapter holds.

[14]More precisely, this strategy will have succeeded as long as the choice of R2 signals to C that the probability that R will select R3 is a smidgen over ⅓. To confirm this, recall the stability analysis of the NEMS in the context of the *Hawk-Dove* game – see Section 3.2.1 of this chapter.

[15] See Binmore (1987), Pettit and Sugden (1989), Reny (1992) and Varoufakis (1991,1993) for more on this debate.

[16] Recall Bayes' rule from note 8 above. If event Y had been assigned probability zero and is observed, then the conditional probability that X will occur given that Y was observed is not defined, as the denominator equals zero.

[17] To arrive at these probabilities we first assume that an irrational R chooses between his strategies, at each node, at random. So, if p is the probability that R is irrational in that manner, C's expected returns at node 2 from playing across $[p\times½\times2000 + p\times½\times800 + (1-p)\times800]$ exceed his certain pay-off of 1000 utils from playing down if p is at least equal to ⅓. So, R's bluff will be worth it if, by playing across at node 1, R thinks that there is a good chance p will exceed ⅓. What counts as a good chance? At node 1 R is guaranteed pay-off 1 if she plays down immediately. If she bluffs (by playing across) and the bluff fails (that is, C plays down at node 2), she collects zero. If she bluffs and succeeds, she will receive 2000 utils. In this sense, as long as the probability that her bluff, or deviance from SPNE, will succeed is at least 1 in 2000, R will have good reason to deviate from SPNE (i.e. to bluff). In summary, as long as a deviant move has a 1 in 2000 chance of making C think that there is a 1 in 3 chance that C is irrational, a rational R has every reason to violate the theory.

[18] Recall how in ***Nim*** and ***Marienbad*** players had a dominant strategy at each node of the game. See Box 3.3. In Game 3.4, by contrast, neither R nor C have dominant strategies in the first two nodes. Only R has one at node 3 (play down).

[19] By this we mean that the observation of a tremble at node 1 does not affect C's subjective probabilitsitc estimate of a tremble at node 3. If the probability of a tremble at node 1 equals ε, his estimate of another tremble at node 3 will still equal ε (that is, after a tremble was observed by C at node 1).

[20] See Varoufakis (1993) for an hypothetical postmodern attack on Nash backward induction. Also see Hargreaves-Heap (2001) for an attempt to infuse some game theoretical ideas into postmodern thinking. Lastly, Varoufakis (2002) makes the controversial claim that instrumental rationality and postmodernity may be accomplices rather than foes.

4

BARGAINING GAMES

Rational agreements, bargaining power and the Social Contract

4.1 INTRODUCTION

4.2 CREDIBLE AND INCREDIBLE TALK IN SIMPLE BARGAINING GAMES

4.3 JOHN NASH'S GENERIC BARGAINING PROBLEM AND HIS SOLUTION
- 4.3.1 The bargaining problem
- 4.3.2 Nash's solution – An example
- 4.3.3 Nash's solution as an equilibrium of fear
- 4.3.4 Nash's axiomatic account
- 4.3.5 Do the axioms apply?
- 4.3.6 Nash's solution – a summary

4.4 ARIEL RUBINSTEIN AND THE BARGAINING PROCESS: The return of Nash backward induction
- 4.4.1 Rubinstein's solution to the bargaining problem
- 4.4.2 A proof of Rubinstein's theorem
- 4.4.3 The (trembling hand) defence of Rubinstein's solution
- 4.4.4 A final word on Nash, trembling hands and Rubinstein's bargaining solution

4.5 JUSTICE IN POLITICAL AND MORAL PHILOSOPHY
- 4.5.1 The negative result and the opening to Rawls and Nozick
- 4.5.2 Procedures and outcomes (or 'means' and ends) and axiomatic bargaining theory

4.6 CONCLUSION

4

BARGAINING GAMES

Rational agreements, bargaining power and the Social Contract

4.1 INTRODUCTION

Liberal theorists often explain the State with reference to some state of nature. For instance, within the Hobbesian tradition there is a stark choice between a state of nature in which a war of all against all prevails and a peaceful society where the peace is enforced by a State which acts in the interest of all. The legitimacy of the State derives from the fact that people who would otherwise live in Hobbes's state of nature (in which life is "brutish, nasty and short") can clearly see the advantages of creating a State. Even if a State had not surfaced historically for all sorts of other reasons, it would have to be invented.

Such a hypothesised 'invention' would require a co-operative act of 'coming together' to create a State whose purpose will be to secure rights over life and property. Nevertheless, even if all this were common knowledge, it would not guarantee that the State will be created unless a tricky issue is resolved: The people must agree on the precise property rights which the State will defend. This is tricky because there are typically a variety of possible property rights benefiting different people differently. The manner in which the benefits of peace will be distributed depends on the precise property rights which are selected and, therefore, there will be great disagreement on which property rights the State must enforce (see Box 4.1).

In other words, the common interest in peace cannot be the only element in the liberal explanation of the State, as *any* well-defined and policed set of property rights will secure the peace. The missing element is an account of how a *particular* set of property rights are selected and this would seem to require an analysis of how people resolve conflicts of interest. This is where bargaining theory promises to make an important contribution to the liberal theory of the State because it is concerned precisely with interactions of this sort.

To be specific, the bargaining problem is the simplest, most abstract, ingredient of any situation in which two (or more) agents are able to produce some benefit through co-operating with one another, provided they agree on a division between them. If they fail to agree, the potential benefit never materialises and both lose out (a case of conflict). State-creation, in Hobbes's world, provides one example (which interests us especially because it suggests that bargaining theory may throw light on some of the claims of liberal political theory with respect to the State), but there are many others.

Box 4.1

PROPERTY RIGHTS AND SPARKY TRAINS

How should people decide how to share the use of the 'commons'? This is a classic example where the introduction of some property rights is potentially beneficial to all because without such rights, and even when there is no fighting over use, there is likely to be overgrazing. Dividing the land into little bundles, one for each person, is one solution, but where exactly will boundary lines be drawn? A few feet further in one direction or another will not upset the general advantage any one person has in avoiding overgrazing but it will benefit one person to the detriment of his or her neighbour. Even when the boundary lines have been drawn and the fences have been erected, there are always further tricky issues which property rights do not settle fully. For instance, to quote a rather famous example from economics, the boundary between the farmer and railroad owner might be clear on the map, but when the sparks from the railroad set fire to the farmer's crop, whose fault is it? Is it the railroad owner's because the railroad was the source of sparks? Or was it the farmer's for planting his or her crops so close to the railway line? In other words, there are a variety of external effects associated with the economic activity and a full set of property rights will also have to assign liability for those external effects.

For instance, there is a gain to both a trade union and an employer from reaching an agreement on more flexible working hours, so that production can respond more readily to fluctuations in demand. The question then arises of how the surplus (which will be generated from greater flexibility) is to be distributed between labour and capital in the form of higher wages and/or profits. Likewise, it may benefit a couple if they could rearrange their housework and paid employment to take advantage of new developments (e.g. a new baby, or new employment opportunities for one or both partners). However, the rearrangement would require an 'agreement' on how to distribute the resulting burdens and benefits.

Thus the bargaining problem is everywhere in social life and the theory of bargaining promises to tell us something, not only about the terms of State creation in Liberal political theory, but also about how rational people settle a variety of problems in many different social settings. And yet the two examples in this paragraph seem to warn that the study of the bargaining problem cannot be merely a technical affair as it involves issues of social power and justice. Indeed there are many alternative accounts of how conflict is resolved in such settings. For example, Box 4.2 sketches two different approaches to the analysis of State formation which have little in common with the liberal voluntarist conception.

Box 4.2

MARXIST AND FEMINIST APPROACHES TO THE STATE

"Hitherto men have constantly made up for themselves false conceptions about themselves, about what they are and what they ought to be" (Preface to the *German Ideology*, p.37). According to Marx and Engels, one of these fictions is the idea that the State under capitalism can be thought of as the product of negotiation between agents under conditions of equality. In reality, "*the State is the form in which individuals of a ruling class assert their common interests…[It] follows that the State mediates in the formation of all common institutions and that the institutions receive a political form. Hence the illusion that law is based on…free will."* (p. 80) Yet "*in the State personal freedom has existed only for the individuals who developed within the relationships of the ruling class, and only insofar as they were individuals of this class*" (p. 83). But Marx was not implying that the State is a machine which serves the ruling class directly and unambiguously. Indeed he criticised those on the Left and on the Right who did not recognise the contradictions within the State. In a famous passage he asserts that the State can act independently of the interests of the ruling class. Indeed the ruling class often benefits when it does not control the State fully: "*in order to save its purse it must forfeit the crown, and the sword that is to safeguard it must at the same time be hung over its own head as the sword of Damocles"* (*The Eighteenth Brumaire of Louis Bonaparte*, in Marx and Engels, 1979). [Notice that this is a functional argument of the type discussed in Box 3.9, see Chapter 3.)

Feminists adopt a similarly radical rejection of the fiction of the State as a 'coming together' between free agents. Carole Pateman (1988) contends that the original contract envisaged by liberal theory is both social and sexual. Through it men transform their 'natural' freedom (recall Hobbes's state of nature) into the security of civil freedom. They do this with the help of the implicit sexual contract as they transform their 'natural' right over women into the security of civil patriarchal right. Thus only men 'bargain' and the contract they forge reflects a civil freedom which is masculine and depends upon patriarchal rights. Catharine MacKinnon (1989) takes up this point and applies it to the State. "*Women are oppressed socially, prior to law, without express state acts, often in intimate contexts. The negative (i.e. liberal) state cannot address their situation in any but an equal society – the one in which it is needed least*" (p. 165). "*The liberal state coercively and authoritatively constitutes the social order in the interests of men as a gender – through its legitimising norms, forms, relation to society, and substantive policies*." (p. 62)

Granted these arguments, it may still be worth reflecting on whether the claim of bargaining theory (to provide an analysis of conflict resolution between individuals with well defined interests) has potential, if more limited, relevance to these non-liberal perspectives. After all, however 'unfree' people may be for one reason or another, they typically still have some choices to make and these often explicitly entail conflicts with other 'unfree' people.

The basic elements of the bargaining problem will remind some readers of the *Hawk-Dove* game (Game 2.14, Chapter 2) as players there have an incentive to co-operate but also an incentive to oppose each other, and this explains why it is often taken to be one of the classic games in social life. The problem with such games is, as we have already noted, that they feature multiple equilibria. This means, in effect, that there are apparently many ways of resolving a conflict (each one consistent with one Nash equilibrium) but none that all can agree with because different

agreements (or Nash equilibria) favour different parties. In short, bargaining games would appear to be difficult to 'solve'.

It is in this context that we discuss in this chapter two important but very different types of 'solution' to the bargaining problem. The first is due to John Nash and appeared in 1950. [It was the solution that the young Nash presented, in manuscript form, to his supervisor in the Hollywood hit movie *A Beautiful Mind*]. Nash's brilliant solution inaugurated a 'tradition' in bargaining theory which has come to be known as the *axiomatic approach*. We call it *axiomatic* because Nash based his 'solution' not on the analysis of actual bargaining but, instead, on some principles (encoded in axioms) which he suggested any agreement between rational agents should satisfy. Once his 'audience' agreed with the stated principles (or axioms), Nash pulled his incredible rabbit out of the hat: He proved that there exists *only one* possible agreement that satisfies these conditions (or axioms).

Following Nash's 'solution', other game theorists followed his footsteps and derived alternative 'solutions' to the bargaining problem by imposing slightly different axioms to those chosen by Nash (see, for instance, Kalai and Smorodinski, 1975). Despite its unquestionable appeal, it is not always clear how the axiomatic approach of the bargaining problem is to be interpreted. Indeed, it is sometimes, somewhat misleadingly, referred to as the 'co-operative' approach to the bargaining problem. In fact, in Section 4.5 we suggest that it is best understood as a framework which can be used to address certain problems in moral philosophy and we provide some illustrations of how it can be put to work in this way. But more on this later.

The second approach, which is considered in Section 4.4, treats the bargaining game as a dynamic non-co-operative game: that is, the bargaining process is modelled step-by-step as a dynamic non-cooperative game, one similar to those examined in the previous chapter. Negotiations are modelled explicitly. Unlike the *axiomatic approach*, the dynamic approach focuses on the process, hoping that its careful study will give us the agreement at the end of it. In this tradition, game theorists model bargaining as follows. Player A makes an offer to B and B accepts (agreement) or rejects that offer, at a cost to both. If he rejects A's offer, then he makes a counter-offer which she either accepts or rejects; and so on. The techniques and concepts utilised to analyse such a *dynamic* interaction are those of the previous chapter; e.g. backward induction, SPNE etc. In this way, this chapter tackles a key type of interaction where it seems there is liable to be indeterminacy using the techniques developed in the last chapter. In turn, this supplies a concrete illustration of some of the problems which were discussed in that chapter.

Of these two approaches, the second seems to have gained an upper hand in the literature.[1] There are two reasons for this. First, the axiomatic approach yields different 'solutions' when different axioms (or conditions that the rational solution/agreement must satisfy) are chosen, and thus leaves a lacuna regarding which of the axioms apply. By contrast, the step-by-step analysis of negotiations promises to deliver useful insights on this. Secondly, the axiomatic approach takes for granted the institutions for enforcing agreements (like the State). That is, it pre-supposes that all parties to some rational agreement will respect it *ex post* (even when some have an incentive to renege). But where do these enforcement mechanisms (e.g. the Law) come from? Are they not *also* the result of society-wide negotiation and bargaining? Precisely because the answer is affirmative, it seems that bargaining theory ought to provide an all-encompassing theory of bargaining: an analysis, that is, not only of what we shall agree upon but, also, on how we get to agree on the social institutions which will guarantee that we shall all live by our agreements. And this requires a dynamic (non-co-operative) analysis of bargaining.

At this stage it may be helpful if we recall the basic distinction between *co-operative* and *non-co-operative game theory* from Chapter 1. The *axiomatic approach* (also known as *co-operative game theory*) assumes that agents can talk to each other and make agreements which are binding on later play. However, it offers no analysis of what they say and how they come to an

agreement. It simply compares different agreements and selects the one which is more in tune with specific conditions (or axioms).

Non-cooperative games, in contrast, make no room for binding agreements. For example, recall the *Short Centipede* (Game 3.4) from Chapter 3; a non-co-operative dynamic game which we analysed exhaustively: In that game, suppose player R had a chance to speak with C beforehand. Suppose further that R were to tell C: "If you promise to play a_2 at node 2 I shall play a_1 at node 1 *and* a_3 at node 3." Even though both would be better off if this agreement is struck (compared to the unique SPNE outcome that has A playing d_1 at node 1), B has no reason to believe that A will keep his promise viz. playing a_3 at node 3. So, game theorists tend to assume that, in the absence of a mechanism enforcing agreements, communication is as good as no communication at all.

Thus, in non-co-operative game theory (i.e. the games examined in Chapters 2 and 3) players can say whatever they like, but since there is no external agency which will enforce that they do what they have said they will do, communication is just unenforceable 'cheap talk'. It is, in short, as if there is no communication whatsoever. Since one might suppose that the negotiations associated with bargaining involve quite a bit of talk, it is as well to treat verbal exchanges as explicit moves (or strategies) before studying which type of 'talk' are strategically significant. We do this next, in Section 4.2.

4.2 CREDIBLE AND INCREDIBLE TALK IN SIMPLE BARGAINING GAMES

We begin with two examples.

Example 1 *Suppose players R and C (we retain their labels for continuity even though they will not always choose between row and column strategies) are offered a chance of splitting $100 between them in any way they want. We empower player R to make C an offer that C may accept or reject. If he accepts, then we have agreement on a division determined by R's offer. If he rejects the offer, we take away $99 and let them split the remaining $1. Then player C makes an offer on how to do this. If R rejects it, each ends up with nothing. Finally, assume that players' utilities are directly proportional to their pay-offs (that is, no sympathy or envy is present and they are risk neutral).*

What do you think will happen? What portion of the $100 should R offer C at stage 1? Should C accept? Using backward induction, suppose C rejects R's initial offer. How much can he expect to get during the second stage? Assuming that the smallest division is 1c, and given that the failure to agree immediately loses them $99, the best C can get is 99c (that is, once there is only $1 to split, R will prefer to accept the lowest possible offer of 1c rather than to get nothing). C knows this (and R can deduce that C knows this) right at the beginning. Therefore, R knows that C cannot expect more than 99c if he rejects her offer during the first stage. It follows that C must accept any offer just above 99c, say $1. Backward induction concludes that, at the outset, R proposes that she keeps $99 with C getting a measly $1. Since C knows that he will not be in a position to improve upon this terrible offer, he will accept.

Notice that the above case of backward induction requires first-order CKR (so it is a form of *Nash backward induction*) as it turns on R knowing that C is instrumentally rational. In fact, the equilibrium so derived is *subgame perfect*, i.e. an SPNE (see Section 3.3.2 of the previous chapter).

At this point we must define a notion that we have come across before in the discussion of subgame perfection and which is at the centre of bargaining theory: that of *credibility*. Suppose that agents can talk to each other during the various phases. What if, just before player R issues her offer of $1, player C threatens to reject any offer that does not give him at least, say, $40. He may for instance say:

> We have $100 to split. You have a first-offer advantage which, quite naturally, puts you in the driving seat. I recognise this. On the other hand I do not recognise that this advantage should translate into $99 for you and $1 for me. Thus, I will not accept any offer that does not give me at least $40.

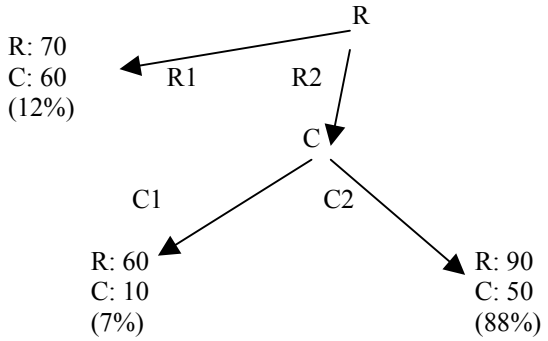
Pretty reasonable, don't you think? No, according to game theorists. For this is a threat that should not be believed by player R. Why not? Because it is a threat such that if C carries it out he will lose more than if he does not carry it out. Thus, it is a threat that an instrumentally rational C will *not* carry out. It is, in other words, an *incredible threat*.

Incredible threats and promises (definition)
A threat or promise which, if carried out, costs *more* to the agent who issued it than *not* carrying it out, is called an ***incredible threat or promise***.

Game theory assumes that agents ignore *incredible threats*; analytically speaking, they resemble the dominated strategies in Chapter 2. Such *cheap talk* should not affect the strategies of rational bargainers. This seems like a good idea in a context where what is and what is not credible is obvious (although see Box 4.3 for some mixed evidence).

Box 4.3
INCREDIBLE THREATS?

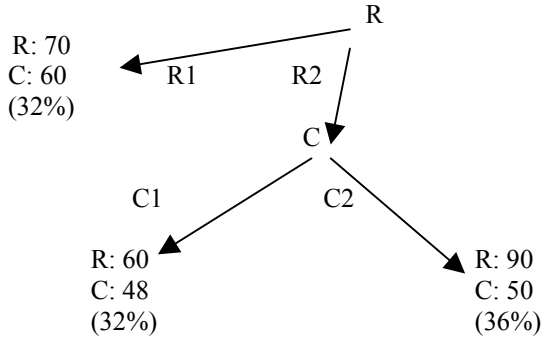
Goeree and Holt (2001) report on the following experiment which gives scope for players to make incredible threats. There are two Nash equilibria in both versions of a static game that is played once anonymously: (R1,C1) and (R2,C2). However, when R chooses first (and the games acquire the dynamic structure on the left hand side; see below), only outcome (R2,C2) is a SPNE as the play by C of C1, although tempting as a form of punishment for R in this game, is inconsistent with backward induction (and, thus, according to game theory, not rational). It is, in effect, an incredible threat which if believed would lead R to prefer R1 and so generate for C a higher pay-off than obtains under the SPNE of (R2,C2).



	C1	C2
R1	$^{+}$70,60^{-}	70,60^{-}
R2	60,10	$^{+}$90,50^{-}

Nash equilibria are shaded. Only (R2,C2) an SPNE in the adjacent tree diagram

The results are reported in parenthesis at each terminal point and the SPNE results in almost all cases, close to 90%. The second version preserves the same strategic structure while altering the pay-offs.



	C1	C2
R1	$^{+}$70,60^{-}	70,60^{-}
R2	60,48	$^{+}$90,50^{-}

Same Nash equilibria as above. Again, only (R2,C2) an SPNE in the adjacent tree diagram

The frequency of the SPNE now drops dramatically, while the Nash equilibrium associated with the incredible threat rises to almost the same frequency as the SPNE. Here it seems that 'cheap threats' should be and are quite often believed; or to put this slightly differently, it seems SPNE does not always well describe behaviour in games (see Box 3.4, also).

Example 2: *There are two people R and C to whom we give $7000. We tell them that one of them must get $6000 and the other $1000. However, we will pass the money over only if they agree on who gets the $6000 and who the $1000 (let us assume for argument's sake that they cannot renegotiate and redistribute the money later). If they fail to agree, then neither gets anything. To give some structure to the process, we arrange a meeting for tomorrow morning during which each will submit a sealed envelope to us including a note with either the number 1 or the number 6 on it. (These numbers convey their claims to $1000 and $6000 respectively.) Finally, we tell them that if both envelopes contain the same number neither gets anything. (Again we assume that the pay-offs are equivalent to utils.) The pay-outs will follow only if one envelope contains number 1 and the other number 6.*

The two bargainers have all night to come to an agreement as to what they will bid for in tomorrow's meeting. According to standard game theory, whether they talk to each other, make promises or issue threats, or even remain silent, there is no difference. For none of these messages are credible and, thus, it is *as if* there was no communication. The reason can be found in the following matrix representation of this bidding game.

	Bid for $6000	Bid for $1000
Bid for $6000	0,0	$6000,$1000
Bid for $1000	$1000,$6000	0,0

Game 4.1 – The *bidding game*

Suppose that during the night, R calls C and declares pompously that she will certainly claim the $6000. Should C take notice? No, because C ought to know that, when it comes to the crunch, an empty threat does not change anything. It is not that one does not expect the other to go for the $6000, but rather that no one can threaten credibly to do so with certainty since it is plain that if R believes C will go for the $6000 then her best action is to bid for the $1000. Game theory's conclusion is that, if a binding agreement is not reached, it makes no difference whether agents can or cannot communicate with each other prior to playing the game.[2]

What matters here is that it is very difficult to make people believe your intentions when you have an incentive to lie. If so, there is nothing new in Game 4.1. A brief comparison of this game with Game 2.13 reveals that our bidding game above is no more than a variety of *battle of the sexes*. Once this is noted, we need not go into a great deal of detail concerning the problems that such a game presents when treated non-co-operatively. Chapters 2 and 3 have, indeed, analysed these sufficiently. The root problem is that this game has no unique Nash equilibrium in pure strategies [each strategy is perfectly rationalisable and both (R1,C1) and (R2, C2) are pure strategy Nash equilibria].

There is one slightly strange inference that can be drawn from the analysis of the bargaining problem. Chapter 2 showed how a unique mixed strategy solution to games such as Game 4.1 (which appears here as a primitive bargaining problem) can be built on the assumption of CAB (that is, that the *beliefs* of agents are always *consistently aligned*): the Nash equilibrium in mixed strategies (NEMS). One might be inclined to think, therefore, that when bargaining problems do have unique solutions, then either the latent conflict of the situation is never manifest (as in the case of *Example 1* above, where R takes almost $99 and C accepts the remainder) or the conflict does not teach players anything they did not know already (as in Example 2, Game 4.1, when players follow their NEMS and each claim the $6000 with probability 6/7). Even though the probability of conflict (i.e. both claiming the $6000) is high, nothing is learnt after such a conflict since these NEMS-based strategies are compatible with CAB from the beginning.

But this line of thought plainly runs counter to the kinds of conflict in the real world that are commonly observed because people do appear to change their views (and positions) afterwards. Of

course such change might be explained within mainstream theory by the argument that conflict only ever arises when players have different information sets (i.e. a state of asymmetric information, see Section 3.2.2). After all, in game theory it is the differences in information which explain (recall the *Harsanyi-Aumann doctrine*) how people come to hold different and conflicting expectations about how to play the game. In other words, it seems we are, in effect, asked to think of the 1984 miners' strike in the UK either as the result of irrationality by the bargaining sides, or as the consequence of insufficient information.

However, matters are not so simple. In fact, we doubt that either the NEMS or asymmetric information explanation of conflict is entirely satisfactory. Not only is this due to the problems that we have already rehearsed with respect to concepts like NEMS, it also results from our belief that many conflicts are initiated because matters of justice, equality, honour or principle are at stake and these are not well captured by the instrumental model of action. Moreover, such concerns can develop a momentum of their own precisely because actions tend to feed back and effect desires. We are running ahead of ourselves here as Chapter 7 pursues this line of argument in more detail.

4.3 JOHN NASH'S GENERIC BARGAINING PROBLEM AND HIS AXIOMATIC SOLUTION

4.3.1 The bargaining problem

We begin with a warning. When we refer to Nash's *solution to the bargaining problem*, we are talking about something quite different to the Nash equilibrium. So don't confuse the Nash *equilibrium* concept with Nash's *bargaining solution*. In this subsection, we shall set up a generic bargaining problem and then follow this with a discussion of Nash's axiomatic solution of it.

The bargaining problem to be examined here has the simplest possible form. Imagine two persons, Jack and Jill, who have the opportunity to split between them a certain sum of money (say, $1) provided they can agree on a distribution (or 'solution', or 'agreement'). They have a certain amount of time during which to discuss terms and, at the end of that period, they are asked to submit independently their proposed settlement (say, in a sealed envelope). If bids are compatible (that is, they sum up to no more than the size of the 'pie'), an agreement is reached and neither party have the opportunity to go back on their word. In other words, agreements are enforceable by some outside agency (e.g. the courts, the social environment etc.)

Bargainers care only about the utility they will get from the agreed settlement. Considerations such as risk aversion, envy, sympathy, concern for justice, etc., are all supposed to be included within the function that converts pay-offs into utilities (the utility function). Exactly as in the earlier games, bargainers in the present chapter play for utilities rather than for the dollars and cents that generate these utilities.

In Chapter 1, we examined the connection between utility functions and risk aversion: In Box 1.4 we saw a simple case in which a player's monetary pay-offs translate linearly into utils: a case of *risk neutrality*. Each additional dollar gives the player the same amount of 'satisfaction' regardless of how many dollars she has already. We also illustrated the case in which a player's utility is a non-linear function of his or her share of the pie. In particular, as we move to the right, the slope of the utility function diminishes. This, as noted in Box 1.4, is a case of *risk aversion*. The reason for linking the slope of a player's utility function with the degree of her risk aversion is simple. If the player values an extra dollar less and less the more dollars she has, she may well prefer a smaller certain pay-off to an expected but uncertain larger one because the prospect of something higher does not compensate for the possible losses that come with uncertainty. In other words, the more she has the less she is willing to risk in order to gain a little more.

Jill and Jack's utility functions are $u_L = f(x)$ and $u_K = g(y)$ respectively, where x and y are, respectively again, the portions of the pie that Jill and Jack will receive as a result of their final agreement. To avoid too much formalism, we shall focus on some poignant examples. Let $u_L = x$ and $u_K = y^n$, where n is some constant. So, while Jill is risk neutral (courtesy of her linear utility function), Jack's fear of risk and conflict (i.e. disagreement) depends on the value of n. When $n<1$ the slope of his utility function declines with y and this means that his willingness to risk disagreement declines the more he expects to get from some proposed agreement. And vice versa: When $n>1$ the better the offer he expects from Jill the more he is willing to 'take' the risk of disagreement. In this case, therefore, the value of n captures the players' *relative* risk aversion. (Once again, consult Box 1.4.) As Jill is assumed to be risk neutral, when $n<1$ ($n>1$) Jill is less (more) risk averse than Jack. Since any agreement between Jack and Jill will mean that $x+y = \$1$ (i.e. if they agree, their respective shares will sum to the $1 pie) we can substitute $u_L = x$ in Jack's utility function to get:

$$u_K = y^n = (1-x)^n = (1- u_L)^n \text{ ---------- (4.1)}$$

Thus, we have derived a relationship between Jack's and Jill's utility functions that must hold for any agreement between them which does not waste any part of the pie (i.e. for any x and y such that $x+y=1$). Agreements here that utilise the whole pie are *efficient* (see below).

The *Utility Possibility Frontier* (UPF) and *efficient bargaining agreements* (definitions)
The *Utility Possibility Frontier* (UPF) of some bargaining problem depicts the combinations of bargainers' utilities that are possible *if they come to some efficient agreement*. An *agreement* is *efficient* if there is no reallocation of the pie that can improve one bargainer's utility without red

(a) $n=1$ Jack and Jill are equally fearful of disagreement (i.e. conflict)

(b) $n=2$ Jill fears conflict more than Jack

(c) $n=½$ Jack fears conflict more than Jill

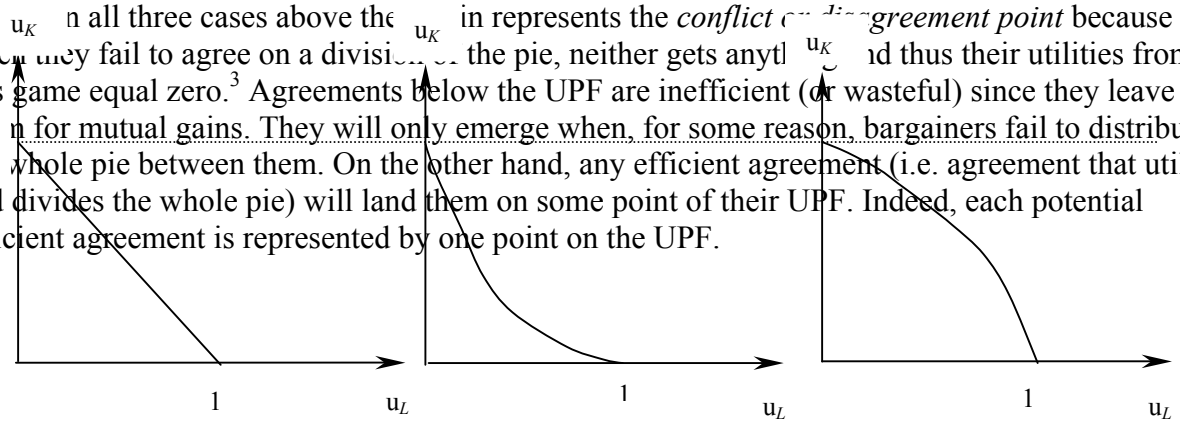
The
neu

Figure 4.1 The Utility Possibility Frontier (UPF)

below. The UPF is a straight line intersecting the horizontal axis at 45 degrees. As Jill's share increases, her utility (u_L) rises and Jack's utility (u_K) falls by an equal amount. The UPF is the game's utility frontier simply because there exists no feasible agreement that raises players utilities above (and to the right of) their UPF.

The UPF looks slightly different when Jack is less fearful of disagreement than Jill; e.g. when $n=2$. In this case, equation (4.1) yields a new UPF [$u_K = (1- u_L)^2$] whose diagrammatic equivalent is Figure 4.1(b). We note that, in this case, the UPF is convex to the origin. By contrast, when Jack fears conflict more than Jill does, e.g. when $n=½$, the UPF [$u_K = (1- u_L)^{1/2}$] becomes

concave [see Figure 4.1(c)].



In all three cases above the origin represents the *conflict or disagreement point* because when they fail to agree on a division of the pie, neither gets anything and thus their utilities from this game equal zero.[3] Agreements below the UPF are inefficient (or wasteful) since they leave room for mutual gains. They will only emerge when, for some reason, bargainers fail to distribute the whole pie between them. On the other hand, any efficient agreement (i.e. agreement that utilises and divides the whole pie) will land them on some point of their UPF. Indeed, each potential efficient agreement is represented by one point on the UPF.

4.3.2 Nash's solution – an example

The object of bargaining theory is to find some division which lies either on or below the players' UPF. Can we pinpoint a solution? Is there a theory capable of predicting how rational bargainers will split the pie? The general difficulty with supplying an answer can be readily seen because *any* proposed division will constitute a Nash equilibrium (note: *not* a Nash solution). To see this point, suppose Jill is *convinced* that Jack will submit a claim for 80% of the prize. What is her best strategy? It is to submit a claim for 20% (since more than that will result in incompatible claims and zero pay-offs for both). It follows that the strategy 'I shall ask for 20%' is rationalisable conditional on the belief 'he will ask for 80%'. Indeed any distribution (x%, 100-x%) or, more simply, (x,1-x) is rationalisable given certain beliefs (see the definition of *rationalisability* in Section 2.4, Chapter 2). If it so happens that Jill's beliefs are identical to Jack's, then we have a case of Nash equilibrium. The following trains of belief illustrate a Nash equilibrium in this bargaining game:

Jill thinks: 'I shall ask for x because I think that Jack expects me to do this and therefore intends to ask for only 1-x for himself.'

Jack thinks: 'Jill is convinced that I shall ask for 1-x and, therefore, intends to claim x for herself. Consequently, my best strategy is to confirm her expectations by claiming 1-x.'

The problem here is that the above equilibrium is consistent with *any* value of x between 0 and 1. We are back to the problem with indeterminacy due to multiple equilibria; the very problem that was plaguing us in Chapters 2&3. Can it be solved? Can the range x=(0,1) of potential equilibria (i.e. bargaining agreements) be wilted down to a single value of x? Nash thought so. He proposed the 'solution' as the value of x which *maximises the product of the utilities* enjoyed by each person. In this section, we illustrate and discuss this solution, leaving its derivation for later.

Nash's Solution

Formally, $x=x^N$ is the Nash solution if it maximises product $u_L(x) \times u_L(1-x)$ or, equivalently, if $u_L(x^N) \times u_K(1-x^N) \geq u_L(x) \times u_K(1-x)$ for any possible agreement x.

Jack and Jill's UPF is given by equation (4.1), which we reproduce here for convenience:

$$u_K = y^n = (1-x)^n = (1-u_L)^n \qquad (4.1)$$

Algebraically, Nash's solution, x^N, was defined as the value of x which maximises utility product

$$P = u_L(x) \times u_K(1-x) = x(1-x)^n$$

It is a matter of simple calculus to show that P is maximised when x equals $1/(1+n)$.[4] In other words, the Nash solution of this bargaining problem, or game, is given as $x^N = 1/(1+n)$. In Figure 4.1 we had looked at three cases: (a) n=1, (b) n=2, and (c) n= ½. In (a) both Jill and Jack are risk-neutral; in (b) Jill is relatively more risk averse than Jack; and in (c) the opposite holds. The Nash solution awards Jill, in these three cases respectively, (a) half the pie [x^N = ½], (b) one third of the pie [x^N = 1/3], and (c) two thirds of the pie [x^N = 2/3].

Figure 4.2 depicts the Nash solution in case (c). To make sense of the diagram, begin with the game's UPF as it has already been depicted in Figure 4.1c. It captures all the combinations of Jill's and Jack's utilities corresponding to all efficient divisions of the $1 pie (that is, all settlements which waste no part of the pie). As we slide down and to the right of the UPF, Jill gets more pie (and thus more utility) and Jack less.

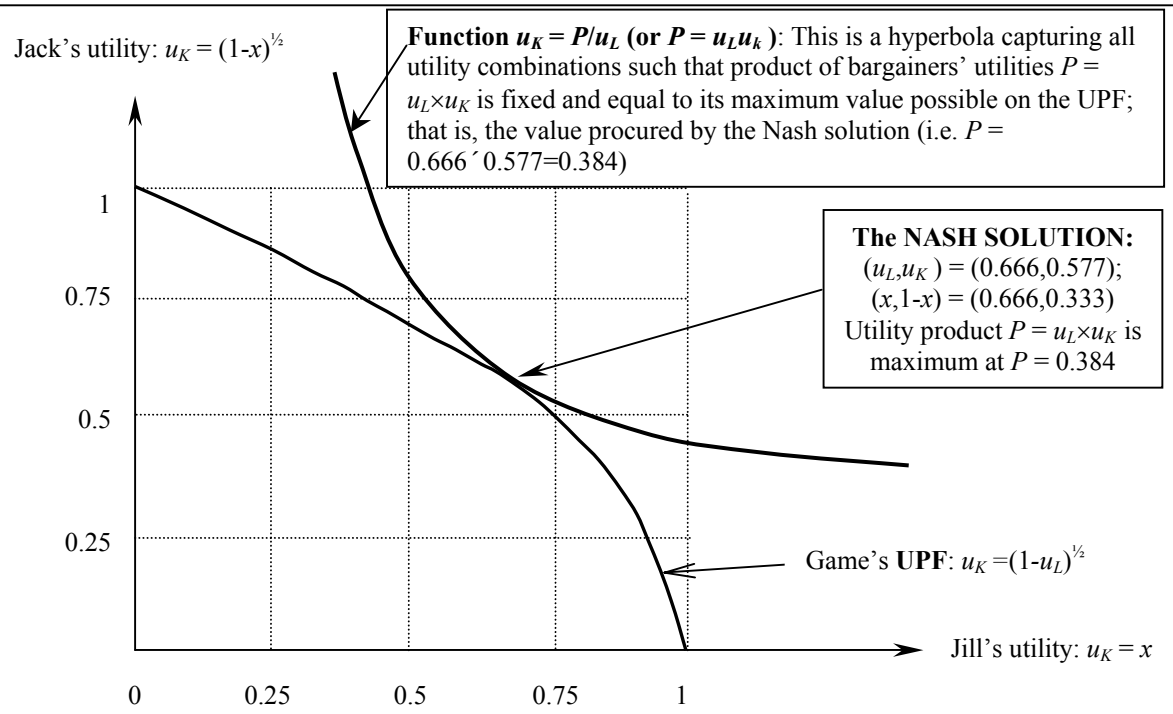


Figure 4.2: The Nash solution in a bargaining game between Jill and Jack whose utility functions are given by $u_L=x$ and $u_K = (1-x)^{0.5}$ respectively. Recalling that the Nash solution maximises the product of the bargainers' utility functions ($P= u_L u_K$), we observe that, geometrically, Nash's solution corresponds to the point where hyperbola $P= u_L u_K$ reaches its largest P value while on the boundary of the game's UPF. That is, the Nash solution is given by the point of tangency between the game's UPF and the hyperbola lying furthest up and to the right of the diagram.

x	$u_L=x$	$u_K = (1-x)^{0.5}$	$P =$ u_L´u_K
1	1	0	0
0.75	0.75	0.5	0.375
0.66	0.66	0.577	0.384
0.5	0.5	0.707	0.354
0.33	0.33	0.816	0.272
0.25	0.25	0.866	0.217
0	0	1	0

The Nash solution is the point on the UPF that maximises product $P = u_L \cdot u_K$. The latter can be re-written as $u_K = P/u_L$; the functional form of a rectangular hyperbola. To find the Nash solution, therefore, what we need to do is to plot the rectangular hyperbola that is: (a) furthest away from the origin (thus maximising $P = u_L \cdot u_K$), but (b) at the same time corresponds to a feasible agreement (i.e. belongs to the UPF). This is achieved geometrically when the hyperbola is *tangential* to the UPF. This is precisely the hyperbola that appears in Figure 4.2. The point of tangency is (u_L,u_K)= (0.666,0.577); a distribution of utilities corresponding to the agreement that Jill gets two thirds and Jack one thirds of the pie.

It will be evident from these calculations that the Nash solution rewards the *less risk averse* bargainer more generously. Or to put this slightly differently the Nash solution assumes (implicitly) that the player with the greatest bargaining power is the one who is least risk averse. In one sense this is not surprising. The Nash solution works only with utility information and the only difference between the utility information of different players relates to the shape of their utility functions (i.e. their respective degrees of risk aversion). Thus if there is to be a difference in the allocation of the pie, it could only be related to differences in the degree of risk aversion. In another sense it is, nevertheless, not so obvious why the solution should favour those who are relatively more risk loving. We shall want to recover the intuition behind this result and check whether it accords with experience in bargaining.

Box 4.4

ACCORDING TO NASH, RELATIVE RISK AVERSION TRANSLATES INTO BARGAINING WEAKNESS

Note in the above example that Jill gets most of the pie. This is so because she is less risk averse than Jack. We know this because the rate of increase of her utility as her share rises (that is, the slope of her utility function) is larger than Jack's equivalent rate of increase. This is a general result that is inbuilt in Nash's bargaining solution (see Section 4.3.2 for the relevant analysis). To recognise the generality of the inverse relationship between a player's 'Nashian' bargaining power and her relative risk aversion, we may use some simple calculus: Let there be two bargainers with utility functions u_1 and u_2 respectively. Nash's solution maximises product $u_1 \cdot u_2$. This maximisation is achieved when the first order derivative of $u_1 \cdot u_2$ (subject to the amount that player 1 will get) equals zero. That is, the condition for maximisation is: $u_1 \cdot u'_2 + u'_1 \cdot u_2 = 0$ (where u'_i denotes the first order derivative of u_i). Re-writing, we have $\frac{u_1}{u_2} = \frac{u'_1}{u'_2}$.

On the left hand side of this equation we have player 1's utility relative to player 2. We may think of this ratio as a proxy for 1's bargaining power; for the larger a player's power the greater her final utility relative to her opponent's. The right hand side is the ratio of the slopes of their utility function. As we have seen a number of times so far, these slopes (or rates of increase) reflect a player's risk aversion. The larger the rate of increase in one's utility as one's share rises (at the expense of one's opponent), the less risk averse one is *and* the greater (courtesy of the last equation) one's relative bargaining power.

4.3.3 Nash's solution as an equilibrium of fear

In this subsection we offer an interpretation of Nash's solution quite different to that in his original 1950 paper. (Nash's own axiomatic account follows later – see Section 4.3.4 below.) Before proceeding with our presentation of Nash's remarkable idea, we feel compelled to remind the reader that, when Nash set out to solve the bargaining problem, no one thought it could be solved. All previous attempts to analyse the process of negotiations (in such a way as to narrow down the range of rational agreements) had proven impossible.[5] Analysts had resigned themselves to the fact that the bargaining problem has no solution. As we saw above, (and in Nash's own language, as developed in the previous chapters) each and every point on the players' UPF is a Nash equilibrium that resolves the conflict and settles the issue. It took considerable audacity on Nash's behalf even to imagine that *one* of those potential agreements (or Nash equilibria) could be isolated as *the* rational agreement.

Perhaps the most crucial analytical move by Nash was his decision *not* to model the bargaining process as such. He, like all other theorists before him, could see that the actual process of offers and counter-offers was not amenable to determinate mathematical modelling. So, instead, Nash decided to take a long look at the relevant game's UPF (see the figure above) and ask himself: Of all these available agreements, could we discard some as unreasonable? That is, as agreements that fail to pass some test that a final agreement between rational bargainers ought to pass? To cut a long story short, not only did he answer his own question in the affirmative (there were, indeed, many points on the players' UPF that could be discarded as agreements that rational bargainers would not reach) but, astonishingly, he concluded that all but one of those could be discarded. The one left standing was, inevitably, hailed as the *Nash solution to the bargaining problem*.

Let us attempt one approach to the *essence* of Nash's solution which also manages to give us a whiff of the *bargaining process*: Suppose Jack makes Jill an offer of x^* (that is, Jack proposes that Jill keeps x^* portion of the \$1 pie). If she accepts, the negotiations are over and the solution is given as the combination $(x^*,1-x^*)$, while Jill's and Jack's utilities are given as $[u_L(x^*), u_K(x^*)]$. If she rejects Jack's offer of x^* it must be because she wants to demand x (such that, $x>x^*$). Indeed, if such a rejection came at no cost, no bargainer would ever accept an offer that does not give her the whole pie. So, of course, a rejection of an offer comes at some cost: the *risk of conflict* which will, we assume, reduce Jill's utility from this bargaining game to zero.

To recap, when Jill rejects Jack's offer of x^*, her rejection may be 'defined' in terms of two parameters: (a) her own counter-demand ($x>x^*$), and (b) the probability with which this rejection will lead them to conflict and mutual loss ($1-p$).

Credible rejections (definition)

A rejection by a player of some offer (x^*) is defined as a combination of a counter-offer ($x>x^*$) and a probability of conflict ($1-p$) engendered by this rejection. [More formally, for offer x^* to be rejected, at the risk of conflict given by probability ($1-p$) and in order to exact a better offer x from one's opponent, it must be the case that the rejecting player prefers px to x^*; that is, she prefers to get x with probability p (and nothing with probability $1-p$) than to collect x^* with certainty. In utility terms, $pu(x)>u(x^*)$.] It is further assumed that the player rejecting x^* selects both her new demand (x) and the probability of conflict ($1-p$); e.g. by threatening to take steps that will lead to a breakdown in negotiations unless her new demand (x) is met.

To give a flavour of the rhetoric behind a player's rejection, imagine the following situation: Jack offers Jill x^* but, instead of taking it, Jill turns around saying:

Jill: No way! I demand x (rather than your offer of x^*) and, unless you yield, I am prepared to act in a manner that will kill off our negotiations with probability $1-p$. Moreover, this is no cheap talk since I can assure you that I prefer to stick to my guns, in favour of my demand x, rather than settle for your x^*; even if this means that I shall collect zero with probability $1-p$. To put it technically for your benefit, in my estimation of the situation, I value product px more highly than I value your offer of x^* [that is, $pu_L(x)>u_L(x^*)$].

Suppose the above is true. That is, Jill can issue this threat credibly (e.g. there is a third party who can guarantee the end of the negotiations with probability $1-p$ once Jill makes this threat). The question then is: Does Jack have what it takes to counter her threat? To do so, Jack needs to be able to tell Jill that her threat does not frighten him sufficiently; that it does not have the power to shift him from his original offer of x^*. Let's see an answer that would convince Jill that her (credible) rejection of his offer above would *not* force his hand:

Jack: Be warned Jill that, even if you do as you say, and take steps to destroy our negotiations with probability $1-p$ in order to claim x portion of the pie (instead of the portion x^* which I have just offered you), I shall not be moved. The reason is that, even under this threat, I am still better off insisting that you accept x^* than succumb to your

pressure and increase my offer to x. Do take seriously what I am telling you. For, as in your case you prefer px to x^*, I prefer $p(1-x^*)$ to $1-x$ [that is, $pu_K(x^*)>u_K(x)$].

Clearly, if Jack cannot muster such an answer to Jill's threat, he will have to concede. But if he can, then he has the power to reject Jill's rejection of his x^* offer! Is it not intuitively plausible that an agreement between Jack and Jill will be reached when neither can make the other concede by means of a credible threat? For this balance of fear (of disagreement) to be achieved, it must be true that an offer has been made such that, if one of them rejects it, then the other can reject the rejection. Indeed, this is consistent with the simple thought that agreement is reached not necessarily when both parties are 'happy' but, rather, when neither feels that a rejection will lead the other side to concessions worth the extra risk of conflict.

Equilibrium Fear Agreements - EFA (definition)

Consider a point or agreement *on* the bargainers' UPF with the property that, when offered to *any one* of the bargainers, a (credible) rejection can be counter-rejected (credibly) by the offering party. We define EFA as the set comprising *all* such points or agreements.

So, set EFA is the subset of UPF to which the agreement will belong, courtesy of the fact that agreements belonging to EFA, once offered, cannot be rejected without having their rejection counter-rejected. To get a feel for the sense of wonder that Nash's solution inspires, consider this: Set EFA can be shown to contain *a single point or agreement*: the Nash solution! In other words, there is only *one* division of the pie that if Jack offers to Jill, Jill's credible rejection can be, in turn, rejected by Jack (credibly); and, at the same time, if Jill offers it to Jack, she can credibly reject any credible rejection by Jack. To the extent that we agree that this condition, i.e. the existence of an 'equilibrium of rejection-proofness' (due to an underlying balance of fear), is a prerequisite for the final agreement, the fact that only one point on the UPF satisfies this condition must surely mean the end of our quest: *that* point on the UPF must surely be the unique solution of the bargaining problem! The fact that this special point on the UPF is no other than the Nash solution explains the latter's hold over game theorists' imagination.

Nash's Solution (Theorem)

Set EFA contains a single point or agreement on the bargainers' UPF. That unique agreement maximises the product of the bargainers' utility functions (or utility pay-offs) and is known as *Nash's solution to the bargaining problem*. [Formally, $x=x^N$ is the Nash solution if it maximises product $u_L(x)\cdot u_L(x)$ or, equivalently, if $u_L(x^N)\cdot u_K(x^N) \geq u_L(x)\cdot u_K(x)$ for any possible agreement x. The theorem implies that x^N is the only EFA agreement.]

Nash (1950) told the world that the bargaining problem has *a* solution and that it happens to be the one which maximises the product of bargainers' utility functions. The above theorem was not actually stated by Nash but represents a powerful argument in favour of Nash, pointing out that the only agreement which *cannot* be rejected credibly (without the rejection itself being rejected) is Nash's proposed solution: the one maximising the bargainers' utility product.

The proof of the theorem involves two steps. **Step 1**: We show that, if x^N maximises the product of the bargainers' utilities [i.e. $u_L(x)\cdot u_L(x)$], then x^N is an EFA agreement. **Step 2**: We demonstrate that, if x^N is an EFA of agreement, then x^N necessarily maximises the product of the bargainers' utilities. Having completed these two steps, and noting that there exists only one agreement (x^N) on the players' UPF that maximises the product of the bargainers' utilities, it becomes evident that set EFA contains a single agreement: the Nash solution! Let us now explain each of the proof's two steps in some detail. [Although we do not recommend this, the reader may skip this proof without loss of continuity.]

Step 1: Proof that the agreement which maximises the product of the bargainers' utilities [i.e. Nash solution x^N such that if $u_L(x^N) \cdot u_K(x^N) \geq u_L(x) \cdot u_K(x)$ holds for all x, x^N] is an EFA agreement; that is, neither Jack nor Jill can reject agreement Nash's proposed solution x^N without having their rejection rejected

Consider Nash's proposed agreement x^N, defined as the one maximising utility product $u_L(x) \cdot u_K(x)$. Equivalently, $u_L(x^N) \cdot u_K(x^N) \geq u_L(x) \cdot u_K(x)$ for any possible agreement x (on the game's UPF). Let us now focus on the point of the negotiations at which Jack proposes to Jill (or vice versa) that she keeps x^N portion of the pie. At that point she has the option of rejecting his offer of distribution $(x^N, 1-x^N)$ in favour of some other agreement, say $(x, 1-x)$, that Jill threatens to back up by taking steps which will bring impasse about with probability $1-p$. What we can now show is that, when Nash solution x^N is on the table, Jack and Jill will take two *symmetrically* uncompromising stances:

Jack's stance: Jack is prepared to reject *any* demand by Jill for a portion of the pie greater than x^N even if she backs this demand with a threat of otherwise ending negotiations with probability $1-p$.
Jill's stance: Jill is prepared to reject *any* demand by Jack that she should accept a portion of the pie less than x^N even if Jack backs this demand with a threat of otherwise ending negotiations with probability $1-p$.[6]

Let us re-write the above 'stances' in terms of expected utilities:

$$pu_K(x^N) > u_K(x) \text{ for all } x \qquad (4.2)$$

i.e. Jack prefers to insist on agreement $(x^N, 1-x^N)$ rather than settle for any alternative agreement $(x, 1-x)$ giving Jill more (i.e. one where $x^N > x$) even if the probability of avoiding conflict is only p.

$$pu_L(x^N) > u_L(x) \text{ for all } x \qquad (4.3)$$

i.e. Jill prefers to insist on agreement $(x^N, 1-x^N)$ rather than settle for any alternative agreement $(x, 1-x)$ giving her less (i.e. one where $x^N < x$) even if the probability of avoiding conflict is only p.

Let us now prove by means of some simple algebra that the above stances [inequalities (4.2) and (4.3)] are uniquely compatible with one another; that when Jack is uncompromising in this manner against deviating from Nash's solution, so is Jill.

At the Nash solution (or agreement) we have (by definition) $u_L(x^N) \cdot u_K(x^N) \geq u_L(x) \cdot u_K(x)$.[7] Jack, as we saw, can only credibly adopt his uncompromising stance if condition (4.2) holds; i.e. if $pu_K(x^N) > u_K(x)$ for all x. Multiplying both sides of inequality (4.2) with Jill's utility from the Nash proposal $u_L(x^N)$, we get: $pu_K(x^N) \cdot u_L(x^N) > u_K(x) \cdot u_L(x^N)$. But, have we not assumed that (since x^N is the Nash solution) $u_L(x^N) \cdot u_K(x^N) \geq u_L(x) \cdot u_K(x)$? Given this assumption, the above inequality becomes $pu_K(x^N) \cdot u_L(x^N) > u_K(x) \cdot u_L(x^N) \geq u_L(x) \cdot u_K(x)$ and hence,

$$pu_L(x^N) \geq u_L(x) \qquad (4.3)$$

The required proof is complete: Starting with inequality (4.2), we proved that inequality (4.3) must hold at the same time. So, starting with the condition for Jack's uncompromising insistence on the Nash solution [inequality (4.2)], we ended up with inequality (4.3) which reports that Jill will also defend this agreement uncompromisingly. We have, therefore, just proven that, when Nash's solution has been proposed by one of the bargainers, both players are prepared uncompromisingly to reject any inferior offer when facing the same threat of conflict (probability $1-p$).

At the risk of repetition, the meaning of this is that when Jack (Jill) suggests to Jill (Jack) that they agree on x^N, Jack (Jill) is *not* willing to abandon his (her) offer of x^N in favour of Jill's (Jack's) demand for a different agreement (x). Indeed, Jack (Jill) is not about to yield in this manner even if Jill (Jack) credibly threatens to cause impasse with probability $1-p$. This is *the equilibrium of fear* (of conflict) that typifies Nash's solution to the bargaining problem.

Summarising Step 1, we just proved the proposition *that if x^N maximises the product of Jill and Jack's utilities, then x^N represents some equilibrium of fear* (of conflict); what we have already defined as an EFA agreement. With Step 2 we shall now show the opposite: that an EFA agreement *must* maximise Jill and Jack's utility product.

Step 2: Proof that an EFA must maximise the product of bargainers' ilities.

Recall the definition of EFA agreements. They are agreements (*on* the bargainers' UPF) with the property that, when offered to *any one* of the bargainers, a (credible) rejection can be counter-rejected (credibly) by the offering party. Now let us suppose that x^* is such an EFA agreement. Our present task is to prove that x^* maximises the product of the bargainers' utility functions. So, to prove that x^* maximises the product of the bargainers' utility functions is, effectively, to prove that $x^*=x^N$!

Since we have assumed that x^* is an EFA agreement, this means that, if Jack proposes to Jill that she keeps portion x^* of the pie, he must be ready and willing to counter-reject all her demands for portions greater than x^*. More precisely, if Jill rejects x^* and demands instead portion x ($x> x^*$), threatening to end the negotiations (with probability 1-p) if Jack does not yield, he must remain unmoved. Indeed, he must be prepared to say: "And I tell you Jill that, unless you accept my offer of x^*, I am going to be the one who ends the talks with probability 1-p."

Algebraically, this translates into inequality (4.4) for Jack (i.e. upon offering Jill portion x^*, Jack does not mind opposing any of Jill's demands for more than x^*; thus Jack insists that Jill accepts his offer of portion x^*) and inequality (4.5) for Jill (i.e. the condition under which she is willing to reject Jack's offer of x^* in pursuit of a higher portion, x, instead):

$$pu_K(x^*) \geq u_K(x) \text{ for all } x \qquad (4.4)$$

$$pu_L(x) > u_L(x^*) \text{ for all } x \qquad (4.5)$$

Inequalities (4.4) and (4.5) simply confirm our assumption that Jack's offer (to Jill) of portion x^* of the pie is an EFA agreement. It is now easy to show that x^* is *the* Nash solution to the bargaining problem; i.e. that $x^*=x^N$. Dividing (4.4) with (4.) and re-arranging we find that:

$$u_K(x^*)\times u_L(x^*) \geq u_K(x)\times u_L(x) \text{ for all } x \qquad (4.6)$$

Inequality (4.6) completes the proof, since it tells us that x^* is such that the product of the bargainers' utilities is greater than that procured by *any* alternative agreement. That is, *because* x^* is an EFA, it maximises the bargainers' utilities (on their UPF). But we know that only one agreement does this: the one corresponding to Nash's solution to the bargaining problem ($x^* = x^N$). *QED*.

In summary, Step 2 started with the assumption that x^N is an EFA agreement and showed that it must *necessarily* maximise the product of bargainers' utilities. Step 1 had already demonstrated that an agreement that maximises this product of utilities is *necessarily* of the EFA sort. Additionally, we know that only one agreement does this: the Nash solution. Thus, we conclude that an agreement corresponds to an equilbrium of fear (EFA*) if and only if* it coincides with the Nash solution! In other words, the equilibrium of fear which, intuitively, characterises rational agreement, only occurs when players agree to divide the pie according to Nash's solution.

Does this remarkable theorem settle the issue? It seems so. However, as we have already discovered many times in this book, the existence of some equilibrium (even when it is unique) does not *per se* mean that all the mysteries of the strategic situation at hand have been unravelled. The same qualms that manifested themselves in Chapters 2 and 3 return here to challenge Nash's claims regarding his bargaining solution. As before (see Sections 2.5.3, and 3.5), the hidden assumption of *consistently aligned beliefs* (CAB) will be the bone of contention.

To see how the critique of Nash and CAB re-appears here, recall the definition of EFA agreements (which proved to be but a single agreement: Nash' solution). A pivotal aspect of this definition was the assumption *that a rejection is credible only to the extent that it is backed by a threat to cause conflict with probability* 1-p. Our theorem identified EFA agreements with the Nash

solution and portrayed the latter as a unique equilibrium of fear that one's (credible) rejection will be rejected (credibly). However, for this equilibrium to come about in practice, it must be the case that, while bargaining, Jill and Jack have *common knowledge* of the true value of p every time an offer is made or is turned down. But this is a tall order.

As explained in previous chapters, for two people to labour under common knowledge of the outcome of 1+1 is one thing; but to entertain commonly known subjective probabilities is quite another. To have common knowledge that Jill will go to the movies tonight with probability 46.52% means not only that Jack predicts with 100% certainty that Jill will go to the movies with probability precisely equal to 46.52% but, also, that Jill is 100% sure that Jack is 100% certain that Jill will go to the movies with probability 46.52% etc. etc. If such common knowledge sounds a little extreme, common knowledge of probability p in our analysis of bargaining above is utterly absurd.

For we know that Jill has good reason to under-play the true value of her p every time she rejects Jack's offer (since $1-p$ is the threat of conflict with which she is trying to extract a concession from him). So, in a strategic environment in which players have strong incentives to shroud their p-choices in mystery, the assumption that these probabilities can be commonly known is impossible to digest.

Interestingly, this assumption is identical to that which (in the past two chapters) we have been referring to as CAB (consistently aligned beliefs). As we have already written long tirades against CAB, we shall confine ourselves here to remarking that the extent to which one believes that Nash solved the bargaining problem coincides with one's readiness to accept CAB in settings in which more than one set of beliefs are rationalisable.

4.3.4 Nash's axiomatic account

Nash derived his solution axiomatically. He proposed certain axioms (conditions) that any solution should satisfy and then showed that there was only one solution (the maximising utility product rule) which did. Nash's four main axioms have come to be known as: (i) *Independence of Utility Calibrations*; (ii) *Symmetry*; (iii) *Pareto Optimality*; and (iv) *Independence of Irrelevant Alternatives*. The plausibility of the Nash solution to the bargaining problem then depends on the reasonableness of these axioms. Are these axioms, or conditions, ones that all rational players would agree to? Below, we examine them one by one.

Box 4.5

NASH'S AXIOMATIC PROOF: WHY IS IT SO REMARKABLE?

Nash (1950) begins by assuming that we are looking for a *rule* which will identify a particular agreement for a bargaining game defined by its conflict point (i.e. the players utilities if no agreement is struck) and the game's utility possibility frontier; the UPF. The remarkable theoretical feat was that Nash did *not* contrive a unique solution. Although he was looking for rules which specified unique outcomes for different sets of axioms, he did not assume that there was only one such rule. You only get the unique solution to the bargaining problem when you combine the fact that there is only one rule with the fact that the rule specifies a unique outcome. Had there been many rules, each specifying a unique but different outcome, then there would have been many solutions, one for each rule. Yet Nash showed that there is only one rule that satisfies all four axioms.

Axiom 1: Independence of Utility Calibrations (IUC)

It will be recalled from Chapter 1 that the utility function representation of an individual's preferences under uncertainty is arbitrary up to any linear transformation. Thus an individual's preferences which can be represented by $u(x) = x$ could as well be represented by the function $u(x) = 2x$ or $4x$, etc. This axiom specifies that the solution should not be sensitive to the 'arbitrary'

choice of which precise utility function is used to represent an individual's utility function.[8] In practice, it is often convenient to choose for each player a normalised utility function such that the worst outcome ('no pie') has a value of zero and the best possible outcome ('all the pie') has a value of 1.

Axiom 2: Symmetry (S)

Symmetry requires that when two players have identical utility functions, the solution must respect this symmetry by giving each player the same share of the pie.

Axiom 3: Pareto Optimality (Pareto)

This axiom requires that a solution should lie *on* the UPF frontier. In other words a rule should not involve any waste, or inefficiency, as rational bargainers would never agree to throw part of the 'pie' in the dustbin, rather than distribute it amongst themselves.

Axiom 4: Independence of Irrelevant Alternatives (IIA)

This axiom entails a consistency requirement across different bargaining problems and requires that a solution to a bargain should not depend on the set of excluded outcomes. It is perhaps best understood through an illustration. Imagine the following situation. Jack and Jill have to divide a pie and there is a rule of rational bargaining behaviour which specifies that there are three possible divisions {x, y and z} and that, of these, distribution y should be agreed upon. Suppose now that the bargaining problem remains exactly the same (i.e. same people, same utility functions, etc) except for one difference: x is now precluded by law. Axiom IIA necessitates that the rule should have the following property: It should still recommend agreement y. This is because outcome x was irrelevant to the solution in the first instance (in the sense that it was not to be chosen) and so the fact that it is now precluded should not affect the solution in what is, to all intents and purposes, an 'identical' bargaining problem.

Nash's axioms are uniquely compatible with Nash's solution (Theorem)
Only one solution (or agreement) satisfies all four of Nash's axioms (IUC, S, Pareto and IIA) at once. It is the Nash solution already defined in Section 4.3.2. Furthermore, Nash showed that his solution applies to bargaining situations in which more than two players are involved. As in the two-person game, the Nash solution with N>2 players is the division (or distribution) of the pie that maximises the product of their N utility functions. In short, the Nash solution specifies a distribution ($x_1, x_2, \ldots, x_N$) such that $x_1 + x_2 + \ldots x_N = 1$ and the values of ($x_1, x_2, \ldots, x_N$) maximise the product $f_1(x_1) \times f_2(x_2) \times \ldots \times f_N(x_N)$, where $f_i(x_i)$ is the utility function of bargainer i (=1,…,N) which relates the utility of player i from having received x_i portion of the 'pie'.[9]

4.3.5 Do the axioms apply?

Since it would be difficult to see how a solution which is sensitive to the arbitrary aspects of the choice of a utility function could be justified, we pass over the first axiom. Equally, it is not obvious why rational players would object to the condition of Pareto optimality.

The axiom of symmetry may also seem entirely plausible at first. After all, if the two agents are one another's mirror image (that is, they have the same motives, the same personality, etc.), should we not indeed expect a totally symmetrical outcome: a 50-50 split? From a normative perspective, this seems unobjectionable. If two people are identical, why should one get more than the other? This sounds plausible until we ask the question: 'What does it mean to say that two

agents are identical?' *Can* two agents be identical? The answer within Nash's framework is that they have identical utility functions and the nagging question from a practical point of view regarding what is likely to happen in actual bargains is whether bargainers will *always* agree that people are identical when their utility functions have the same shape.

Utility information actually ignores many features of the bargaining situation which one might suspect many agents might regard as relevant. For example, utility representations are gender blind. A man and a woman with the same utility representations are treated identically by game theory (and so they should be), but is this a plausible assumption about *actual* behaviour in all settings? In a sexist society, is it not more plausible to assume that the 'convention' operating in that society may actually treat men and women differently even when their utility information is identical?

Independence axioms, like IIA, are often thought to be requirements of individually rational choice (see Chapter 1) on the grounds that consistency requires that if you prefer A to B when C is available, then you should *still* prefer A to B even when C is *not* available. Nevertheless, experimental work on expected utility theory has shown that such consistency may be violated by perfectly rational people (see Hargreaves Heap *et al*., 1992, Chapter 3). Are real people less consistent than the theory expects because they are less rational? Or has the theory missed something out? To give a simple example that the problem does not always lie with the subject's rationality, imagine that A = croissant, B = bread and C = butter. You may prefer A to B in the absence of C (i.e. you prefer a plain croissant to a piece of plain bread) but your preference may be reversed when C is available (i.e. you prefer a buttered piece of bread to a croissant, buttered or plain). Such complementarities have been used to explain paradoxes like that of Maurice Allais - see Box 1.5.

In the case of bargaining the potential for violations of independence axioms, such as IIA, is enhanced. This is so because it is another person who sets your constraint (through his or her demands). Therefore what you cannot have depends on what the other person thinks you will not ask for. The greater the interaction the more problematic it is to assume independence. Consider for instance the illustration used earlier: Imagine that Jill were about to settle with Jack on the basis of a 60%-40% split. Just before they agree, the government legislates that Jack cannot get anything less than 40%. Will you expect Jack to see this as an opportunity to up his claim? IIA *assumes* that you will not expect this; that Jack will not do this; and that Jill will also not expect him to do this! But why would it constitute a violation of Jack's rationality to see this piece of 'legislation' as a new twist to the game that he may benefit from? And why is it irrational of you (and/or Jill) to think that Jack may contemplate using this new twist in his favour?

At best then it seems that IIA is no more than a convention bargainers may or may not accept as a condition which agreements (as well as demands) will satisfy. The problem is that there are other, equally plausible conventions to which rational bargainers may follow. For example, the convention that when an external agency (such as the State) underpins the bargaining position of one party, this will benefit the pay-off of that party, even if the intervention is mild. Industrial relations experience is one source of rich insights in this regard. For it shows that the bargaining position of trade unions is improved when a minimum wage is introduced. Moreover, and this is important here, this improvement is not restricted to bargains which involve workers at the bottom of the pay scale; indeed there are spill-over effects to other areas in which the minimum wage would not apply and yet the union position (and thus the negotiated wage) improves as a direct repercussion of the minimum wage legislation. This experience contradicts directly the axiom of IIA.

Indeed it is possible to devise explicit alternatives to the IIA axiom. These alternative conventions play the same role as IIA (that is, they provide a consistent 'link' between the outcomes of different bargaining games), albeit lead to different bargaining solutions. For instance, a *monotonicity* axiom has been proposed by Kalai and Smorodinsky (1975) whereby when a

bargaining game is slightly changed such that the outcomes for one person improve (in the sense that, for any given utility level for the other player, the available utility level for this person is higher than it was before), then this person should do no worse in the new improved game than he or she did before. This might seem more plausible because it embodies a form of natural justice in linking bargaining problems. However, the substitution of IIA with this 'monotonicity' axiom yields a different bargaining solution to that of Nash: one where there is an *equal relative concession from the maximum utility gain*. Indeed some moral philosophers have argued that this is the result that you should expect from instrumentally rational agents (see Gauthier, 1986).[10]

Box 4.6

SOME VIOLATIONS OF NASH'S AXIOMS

Imagine that two bargainers A and B have identical utility functions. They are walking together when A spots a $100 bill lying on the ground. As they are on a Greek island (on holiday), they want to exchange it for euros at the local bank. However, A does not have her passport with her, but B does (we assume a passport is required at the Bureau de Change). The Nash solution predicts that they will share the proceeds. Fair enough. What if, however, A and B come from a place where 'finders' are thought to be more deserving 'keepers' than 'non-finders'? Of course B's passport is important in this instance. Yet both A and B may entertain expectations that the person who *actually* found the $100 deserves more than 50% - in which case they might violate Nash's symmetry axiom and agree on an asymmetrical distribution (e.g. 60%-40%).

Here is another example. Suppose that A and B, while holidaying in Thailand, have won a sum of Bhat in some lottery. A is due to fly home hours later while B will stay on for another fortnight. If they split the sum of Bhat in half, then A will be disadvantaged since she will have to change it immediately into her country's currency and will, thus, forfeit a significant amount in bank fees. Nash's IUC axiom assumes that the two friends will take this into account and will agree to divide the Bhat in an asymmetrical fashion so that their utility gains are identical (to reflect the fact that their utility functions are identical). But it seems at least possible that A and B might share the expectation that each would demand an equal division of Bhat.

The point of the above examples is that we cannot rule out the possibility that bargainers will, quite rationally, act on a basis of some convention different to those behind Nash's axioms. One final example: A and B are about to decide that they want to split the Bhat they just won on a 60%-40% basis (perhaps for the reason offered in the previous paragraph). Before they do, they find out that the lottery rules specify that, in the case of joint winning tickets, no partner should get less than 40% of the winnings (some Thai law)! Nash's IIA axiom insists that this rule should not change their mind since the 60%-40% split which they were going for is (just) legal. But it is possible that A will not expect B to settle for the bare minimum of his legal entitlement. Equally, it is possible that B (recognising this) will demand more than 40%.

We conclude from this discussion that although Nash's axioms have some appeal, it is difficult to believe that they will *always* reflect the bargaining conventions followed by *all* rational individuals. As a result, the axiomatic approach does not seem to provide a compelling foundation for the Nash solution to bargaining problems. We turn therefore now to the most ambitious attempt to put the matter beyond contention, and vindicate Nash totally: a theorem by Rubinstein (1982) followed by another remarkable result in Binmore, Rubinstein and Wolinsky (1986).

4.4 ARIEL RUBINSTEIN AND THE BARGAINING PROCESS: THE RETURN OF NASH BACKWARD INDUCTION

4.4.1 Rubinstein's solution to the bargaining problem

So, why should rational bargainers select the Nash solution? In the 1980s, game theorists came up with an answer to the question in two important papers. First, Ariel Rubinstein showed in 1982 that when offers and demands are made sequentially in a bargaining game, and if a speedy resolution is preferred to one that takes longer, *there is only one offer that a rational bargainer will want to make at the outset of negotiations*. Moreover, *the rational bargainer's opponent* (if rational) *has no* (rational) *alternative but to accept it immediately.* Secondly, in 1986 Binmore, Rubinstein and

Wolinsky proved that *this unique settlement is equivalent to Nash's bargaining solution*. If all this is correct, then John Nash's solution has received the most spectacular boost because it can be shown to result from a bargaining process where players make sequential offers. Indeed this is the only outcome that rational players with CKR will settle on.

We begin with a quick sketch of Rubinstein's argument before presenting the unabridged story. Recall *Example 1* in Section 4.2. Jill was asked to suggest to Jack how to split $100 between them. If Jack rejected her suggestion, the $100 shrank to a measly $1 and it was Jack's turn to offer Jill a portion of the remaining $1. If she rejected it, no one won anything. Backward induction, coupled with 1st order CKR, led us to the conclusion that Jill would make Jack an offer he could not refuse: "You take $1 and I keep $99!"

Now consider a richer setting. Again Jill and Jack are given the opportunity to split $100 with Jill making the first 'move'. Jack either accepts her offer or counter-proposes an alternative settlement. However, to add some urgency to the proceedings, let us imagine that, if Jack rejects Jill's initial offer, a timer starts ticking and, with every second that passes without agreement, 1c is taken off the $100 prize. That is, if they take M minutes to reach agreement, the $100 will, by then, have shrunk to $(100-0.6$M$).

How should one play this game? Jill must now balance the urge to make Jack an offer that he will not refuse (so as to avoid 'shrinkage' of the prize) against the worry that she might end up offering him too much. Recall that in all bargaining games, *any* outcome is rationalisable (moreover, any outcome is a Nash equilibrium). If for example Jill expects Jack to accept 40% and thus issues a demand for 60%, while Jack anticipates this, then a 60%-40% split is an equilibrium outcome (as it confirms each bargainer's expectations). And since any outcome is rationalisable, the theory offers no guidance to players. To the rescue comes *Nash backward induction*.

Consider the following strategy that Jack may employ in his negotiations with Jill: 'I shall refuse any offer that awards me less than 80%.' This may be rationalisable (and a Nash equilibrium) when we look at the final outcome independently of the bargaining process, but it may not be if we examine the various alternative strategies against the background of the actual bargaining process. Why? Because such a strategy may be based on an *incredible threat* (recall the definition of such threats in Section 4.1). This is why:

Suppose Jill offers Jack only 79.9%. Were Jack to stick to his 'always demand 80%' strategy, he would have to reject the offer. However, this rejection would cost him as the prize shrinks continually until an agreement is reached. Even if his defiant strategy were to bear fruit soon after the rejection of Jill's 79.9% offer (i.e. if Jill were to succumb and accept Jack's 80% demand M minutes after her 79.9% offer was turned down), Jack will only get 80% of a smaller prize. How much smaller the prize will be depends, of course, on M; i.e. on how long it will take Jill to accept Jack's demands. If it takes more than 12.5 seconds, Jack will be worse off than he would have been had he accepted her offer of 79.9%![11] Thus, if it is commonly known that it takes well over ten seconds for bargainers to respond to an offer, Jack has no incentive to stick to the strategy '*always* demand 80%'. And so, if during negotiations Jack threatens to reject *any* offer less than 80%, Jill should take this threat with a pinch of salt; and a very large one if it takes more than about 10 seconds to make a response to any offer.

The above is an important thought. By means of Nash backward induction (which in Chapter 3 emerged as the backbone of the SPNE), we can discard a very large number of possible negotiating strategies on the basis that they will not work if the agents' rationality is commonly known. Ariel Rubinstein (1982) used this SPNE-based logic to prove a remarkable theorem: There exists only *one* SPNE that does *not* involve use of incredible threats. The brilliance of this thought matches that of John Nash's original idea for solving the bargaining problem and, what is even more extraordinary, yields a solution analytically equivalent to that of Nash as the time delay between offers and demands tends to zero (the latter was shown by Binmore, Rubinstein and Wolisnky, 1986).

4.4.2 A proof of Rubinstein's theorem

The precise bargaining process examined by Rubinstein is very similar to the preceding example. There is a prize to be distributed and Jill kicks the process off by making a proposal. Jack either accepts or rejects it. If he rejects, it is his turn to make an offer. If, in turn, Jill rejects that offer, the onus is on her to offer again, and so on. Every time an offer is rejected, the prize shrinks by a certain proportion which is called the *discount rate*. Analytically it is very simple to have different discount rates for each bargainer and this allows one to introduce differences between the bargainers, differences that are equivalent to the differences in the rates of change of utility functions (or *risk aversion*) discussed earlier in the context of the Nash solution. Rubinstein's theorem asserts that rational agents will behave as follows: ***Jill will make Jack an offer that he cannot refuse*** (or, more precisely, *does not want to refuse irrespectively of how much he likes it*).

Thus, there will be no delay and the prize will be distributed before the passage of time reduces its value. Moreover, the settlement will reflect two things:

(a) Jill's first-mover advantage, and

(b) Jill's relative eagerness to settle (i.e. their relative discount rates).

By (a) we imply that Jill (who makes the first, and allegedly, final offer) will retain (other things being equal) a greater portion than Jack courtesy of the advantage bestowed upon her by the mere fact that she offers first (something like the advantage of the white player in chess). [Note, however, that if offers can be exchanged *very* quickly, the first-mover advantage disappears (in the limit).[12]] By (b) it is meant that eagerness to settle is rewarded with a smaller share. If Jack is more eager to settle than Jill, then he must value a small gain now more than Jill does, as compared with a greater gain later.

This result is perfectly compatible with Nash's solution which, as we showed, penalises *risk aversion*. To the extent that *risk aversion* and an *eagerness to settle* are similar, the two solutions (Nash and Rubinstein) are analytically interchangeable. This is Binmore et al's (1986) contribution: they prove that, when agents exchange offers at the speed of light, and their discount rates reflect their risk aversion, Rubinstein's solution is identical to that of Nash.

That the Nash solution could be given an SPNE-based defence was known for some time before Rubinstein presented his remarkable theorem. Take for instance a game like that in Box 4.7 below. In that game there exists a unique SPNE which coincides with the Nash solution. However, the problem with these pre-Rubinstein proofs was that they pertained to bargaining processes that lacked realism. For instance, the game in Box 4.7 features three stages only. Why not four? And who imposes the rules in such a way that Jill cannot return after the third stage (at, say, t=4) with a fresh offer? Vindications of Nash's solution by means of finite dynamic games are interesting but unconvincing in a world where bargaining processes seldom feature exogenous time limits. [Readers may have no trouble, after Chapter 3, discerning the reason why dynamic bargaining games (e.g. Box 4.7) need a *finite* number of stages: without a definite end to the bargaining process, Nash backward induction cannot be applied (as it needs a last stage on which to anchor itself before unfolding backwards). And without *Nash backward induction*, no SPNE can be induced.]

Box 4.7

A THREE-STAGE DYNAMIC BARGAINING GAME IN WHICH NASH'S SOLUTION IS THE UNIQUE SPNE

Consider the following three-stage game whose purpose is to help Jill and Jack come to an agreement on how to divide between them some 'pie'.

t=1 Jack chooses agreement y on their UPF. (Think of y as his proposal to Jill.)

t=2 Jill chooses agreement x on their UPF and, in addition, a probability 1-p with which the game will end forewith. (If $x=y$ and $p=1$, we say that agreement has been reached instantly. But if $x \neq y$ and $p<1$, we may think

of x as her counter-claim and of 1-p as the probability of conflict she threatens Jack with in order to impose her will.)

At this stage, if $p<1$, the game ends (and players collect nothing) with probability 1-p. *Otherwise (with probability p) it proceeds to stage $t=3$.*

t=3 Jack chooses between certain agreement (accepting Jill's demand of x for herself and 1-x for Jack) or a lottery that will force upon Jill his initially preferred agreement (y,1-y) with probability p but lead to a collapse of the game (with neither player getting anything) with probability 1-p.

Howard (1992) devised this dynamic three-stage game and proved that it possesses a single SPNE: At t=1 Jack will offer Jill (and she will accept it at t=2, by setting $x=y$ and p=1) the *Nash solution*. In other words, Jack will select the y-value that maximises the product of their utility functions. And Jill will accept it immediately.

Rationale: Note that the structure of this game reminds us of the analysis in Section 4.3.3. It is set up so as to ensure that Jill's rejection (at t=2) of Jack's (t=1) offer comes complete with a credible threat of conflict (probability 1-p). Once Jill sets $p<1$, there is no way she can take it back (and hence the threat's credibility). In equilibrium (NEMS) at t=2 Jill must make an offer that will leave Jack indifferent between accepting it and opting for the more risky option. Jack predicts this at t=1 and makes her an offer that gives her no incentive to reject it (or setting $p<1$). Howard (1992) then proves that the only offer respecting these equilibrium conditions (that is, backward induction coupled with CKR) is the Nash solution.

The extension of this insight to games without a potential end is Rubinstein's important contribution and his proof is a gem. We propose to sketch it here utilising only high-school algebra. However, the logic is quite tortured. Although we feel that the non-technical reader will benefit from perusing the proof, the rest of this section can be skipped without loss of continuity. We begin with a definition of the player's discount rates and a statement of the theorem. Then, we present a 'simple' proof.[13]

Discount rates and their relation to risk aversion (Definition)

Every time an offer is rejected, Jill's valuation of the prize loses a proportion given by 1-α (where α lies between 0 and 1). It is as if, in her eyes, portion 1-α of the pie has been lost. Similarly, with every rejection that occurs, Jack's valuation of the prize diminishes by 1-β. For example, if $\alpha = \beta =$ 0.8, then, when an offer is rejected, only 80% of the prize is preserved in the next round. Thus, if Jill and Jack come to an agreement at $t = 3$, the prize they will be splitting will have shrunk twice; the extent of the 'shrinking' depends on α and β.[14] These parameters (α and β) are known as the bargainers' *discount rates*. They are closely related to the player's *risk aversion*. To see this consider the position of the person deciding at some stage whether to accept the other's offer. The player has a choice between accepting a share of the cake now or rejecting the offer and bargaining over the pie in the next time period. The outcome of the bargain in the next time period is uncertain whereas acceptance of the offer now yields a known quantity. The extent of the perceived shrinkage of the pie will then reflect the person's perception of the risk associated with this uncertainty.

Discount rates α and β are also sometimes known as the bargainers' *time preferences*; referring to their capacity to capture the players' valuation of a larger pay-off tomorrow compared to a smaller one today. By comparing discount rates we gauge the bargainers' relative urgency to settle. E.g. if $\alpha > \beta$, Jill is clearly less eager to settle than Jack (as the prize shrinks, with every failed offer, (relatively) faster for him than it does for her). In any case, it is evident that the ratio α/β is a good proxy for Jill's relative fear of disagreement (as compared to Jack's). For if $\alpha/\beta > 1$, Jill loses less from each rejection (and, thus, from each delay in reaching agreement) than Jack. Other things being equal, we might therefore expect Jill to be less acquiescent to Jack the higher the value of α/β. In this dynamic sense (that is, when time comes into the bargain), ratio α/β is the equivalent to the relative risk aversion that determined the outcome in Nash's 1950 solution.

Rubinstein's solution to the bargaining problem (theorem)

The setting: Suppose Jill and Jack issue alternating offers on how to split a \$1 pie. Jill begins at t=1 and then, if Jack has rejected her offer, he issues a counter-offer at t=2. If Jill does not accept it, then she issues a counter-counter-offer at t=3. And so on until agreement is reached. With every offer that is turned down, the pie is discounted. Jill's and Jack's *discount rates* (see definition above) are, respectively, α and β (where $0<\alpha,\beta<1$).

The theorem: At t=1, Jill will suggest to Jack the following division and Jack will accept it immediately: $\left(\frac{1-\beta}{1-\alpha\beta}, \frac{\beta(1-\alpha)}{1-\alpha\beta}\right)$. Moreover, it constitutes the unique SPNE of this dynamic bargaining game.

So, how does Rubinstein manage to induce an SPNE from a dynamic game lacking a definite end? By means of a *hidden assumption*, is our answer. While postulating an indefinite bargaining horizon, Rubinstein presumes (without stating this explicitly in his paper) that there shall come a stage, call it round k, at which Jill's and Jack's beliefs (regarding the maximum share they can each expect to get) will have converged. Furthermore, Rubinstein's hidden assumption has it that Jill and Jack have common knowledge (even at time t=1) of that distant round k. This is the hook on which Rubinstein secures the logic of *Nash backward induction*. Thus the latter unravels, beginning at $t=k$, then moving to $t=k$-1 and finally to t=1 where Jill's unique SPNE offer to Jack is computed. And since it is a unique equilibrium offer, Jack simply accepts it. In summary, our proof will involve four steps:

(A) We state the *Hidden Assumption* that gives Nash backward induction its foothold.

(B) We consider the minimum offers Jill and Jack will issue at $t=k$, $t=k$-1, $t=k$-2,…, t=1. [We also prove that they have no incentive to offer less than these minimum offers (namely, that their minimum offers equal their maximum offers).]

(C) Once the unique offer at t=1 is derived, we utilise (again) the *Hidden Assumption* to argue that Jill must propose a division at t=1 identical to that which they would agree on (at much greater cost) at $t=k$. Once this assumption is made, Jill's offer to Jack at t=1 will be computed.

(D) We prove that the actual value of k does *not* matter, as far as the bargaining solution is concerned. All that matters is that, in accordance to the *Hidden Assumption*, we presume that at t=1 bargainers entertain common knowledge of k (whatever its value might be).

Step A: The Hidden Assumption

There are two parts to this assumption. (a) There exists some round $t=k$(>2) in which bargainers' beliefs will have become consistently aligned on distribution (V,1-V). (b) Jill and Jack have *common knowledge* of k at t=1.

The *Hidden Assumption* is, of course, a reincarnation of the CAB assumption (consistently aligned beliefs) that has occupied us repeatedly in the last two chapters. Part (b) is familiar territory. If there exists an unknown parameter (k in our case) and players are equally rational and well informed, their estimates of k must be common (and it must be commonly known that they are common). Part (a) is another application of CAB. Players assume that their beliefs about the outcome will be, at some stage (k), consistently aligned.

Thus Rubinstein has the anchor for the logic of Nash backward induction that he needed: the k^{th} round of the bargaining process. From there one simply moves backwards through rounds k-1, k-2,… and, finally, to the first round.

Step B: Computing bargainers' offers backwards

Consider the value k=3. Why k=3? Because it is a small number of stages which will help us keep the proof simple. Do we not lose generality by assuming such a small number of stages? No, because as we shall see in Step D, the actual choice of k makes no difference to the proof. And since it makes no difference to the proof, or to the bargaining solution, whether k equals 3 or 3000, we might as well keep things simple.

So, suppose that Jill and Jack commonly know at t=1 that by stage t=k(=3) they will discern the *same* solution, say (V,1-V), to the negotiations over the distribution of the pie. This does not, of course, mean that at t=1 they know the value of V which will surface at t=k=3. All it means is that at t=1 they have common knowledge of the 'fact' that, come t=k=3, there will be *some* portion of the pie, say V, which (in the eyes of both Jack and Jill) Jill will not be able (or willing) to improve upon (be means of more bargaining).

The following table acknowledges this in its first row, which depicts the bargainers' offers and demands in round t=k=3. Once more, let us remind the reader that neither we (as theorists), nor our bargainers, have any way (at this stage of the theorem's proof) of knowing the value of V. All that is known at t=1 is that, come t=3, both will have in their minds *some* value V which will be a commonly held estimate of Jill's final share of the pie. The task is to compute this value.

Round	Proposer	Proposed share for Jill	Proposed share for Jack
t=k=3	**Jill**	V	1-V
t=2	**Jack**	αV	1-αV
t=1	**Jill**	1- β(1-αV)	β(1-αV)

The *Hidden Assumption* asserts that there is some commonly known round (t=k) during which Jill and Jack will believe that the pie will be divided in portions (V,1-V). Round k is commonly known at all stages. The path of the Nash-backward-induction is depicted by the direction of the arrows. Supposing that t=k=3, at t=3 (if the negotiations last that long) Jill will receive portion V. Thus at t=2 Jack must offer her αV to avert conflict, keeping 1- αV for himself. Thus at t=1 must offer him at least β(1-αV) so as to avoid a rejection that will delay agreement. If she does this, she keeps 1-β(1-αV).

Figure 4.3 *The backward induction of optimal offers based on the Hidden Assumption*

Let us now investigate Jack's situation at t=2. It is his turn to accept or reject the offer Jill made him at t=1. Should he reject Jill's t=1 offer (and come up with a counter-offer)? Or should he accept it? If he rejects it, what counter-offer should he make? He knows (from the *Hidden Assumption*, with k=3) that, if the negotiations proceed to t=k=3, Jill can expect to get V. So, Jack knows that if he were to offer her, at t=2, portion αV of the pie, she has no reason to turn it down: indeed, she cannot reasonably expect (given the *Hidden Assumption*) to do better. The reason, of course, is that Jill's t=2 valuation of V at t=3 equals α (her discount rate) times V. Put differently, Jill must (by definition) be indifferent between portion αV at t=2 and V at t=3. Any offer by Jack (to Jill) at t=2 below αV will spark off a rejection.

Thus we computed Jack's minimum offer to Jill at t=2 *if he wants to avert a rejection by Jill*: It is an offer of portion αV. The question now becomes: Will Jack want to avert a rejection at t=2? If he offers Jill anything below αV, Jill reject it and bargaining will continue until t=3 where Jill will receive portion V and Jack 1-V. Jack knows this at t=2. Thus he knows that, if he offers Jill

less than αV, the most he can get at t=3 is 1-V. What is 1-V at t=3 worth to Jack at t=2? Since his discount rate is β, 1-V at t=3 is worth to Jack β(1-V) at t=2.

In short, Jack has a stark choice at t=2: Offer Jill αV immediately (at t=2); an offer that we know she will accept, therefore leaving him at t=2 with 1-αV of the pie. Or offer her less than αV; a move that will lead him to a pay-off whose value to him, at t=2, equals β(1-V). Clearly, he will induce rejection at t=2, by offering Jill less than αV of the pie, only if β(1-V)>1- αV. However, as all these parameters (α, β and V) lie between 0 and 1, this inequality is *never* satisfied. Which means that, at t=2, Jack will *never* offer Jill less than αV. And since (as we have already shown) an offer of αV is the minimum she will accept, Jack has no reason to offer her more than that. Thus we have proved that Jack's maximum offer at t=2 will be the minimum that Jill will accept and we have a uniquely rational offer: *At t=2 Jack offers Jill portion αV and she accepts it.* We make a note of this result in the second row of the table above.

With the analysis of round t=2 complete, we now turn our attention to what happens during t=1 at which point it is Jill's turn to make an offer. She knows (see above) that if her offer is turned down, bargaining will proceed to round t=2 where she will be offered portion αV; an offer that she will accept. By the same token, she knows that Jack knows that he can expect, at t=2, a sure portion of 1-αV. What is his valuation at t=1 of portion 1-αV at t=2? Given his discount rate of β, at the outset (t=1) Jack's valuation of his share of the second round (1-αV) equals β(1-αV). Thus we (and Jill along with us) know that if Jack is offered portion β(1-αV) at t=1, he will accept it. Any offer less than that will cause disagreement and a counter-offer by Jack at t=2.The question then, predictably, becomes: Does Jill want to settle with Jack at t=1?

If she offers Jack less than portion β(1-αV), he will reject her offer and he will return to the bargaining table with the suggestion that Jill keeps portion αV; an offer that, as we have already proven, Jill will accept. So, it all comes down to whether Jill prefers 1-β(1-αV) immediately (i.e. at t=1) or the prospect of a certain portion equal to αV in the next round (t=2)? Since the later is valued by Jill at t=1 at $\alpha^2 V$, she will opt for immediate settlement (at t=1) if 1-β(1-αV)>$\alpha^2 V$.

It is easy to show that this inequality holds *always*! This means that Jill will *always* prefer to offer Jack portion β(1-αV) at t=1; an offer that he has *no* incentive to reject and which Jill prefers to offer over any alternative offer that will cause Jack to turn it down. We have, consequently, reached the conclusion that at t=1 Jill will offer Jack division [1-β(1-αV), β(1-αV)]. And Jack will accept it.

Step C: Consistent preferences over time

Before we investigate further, consider any case of conflict between two persons, countries, firms etc. If they both knew at the outset *how* the 'war' between them would be settled, would it not be rational to agree at the very beginning to settle it in that manner while skipping the costly fighting? In our case this would mean that Jill tells Jacks: "We know (recall the *Hidden Assumption*) that if we wait till t=k, I shall receive V portion of the pie. Why wait until then? Let me have portion V now and, in this manner, no part of the pie will be lost (through delay in reaching agreement) for either of us."

Rubinstein assumes that an instrumentally rational Jack has no reason to disagree; and we call this the assumption of *consistent preferences over time*. The only problem is that they do not know the precise value of V. However, there is a way of discovering it now. We have concluded above that both will entertain the same expectation of what Jill will get if they ever reach t=3: Jill will get V. At the same time, we have concluded that, at t=1, Jill will demand 1- β(1-αV) for herself and Jack will let her have it. So, why not say that she will demand now [1- β(1-αV) at t=1] the same share (V) that she will get if she were to hold out until t=k=3? In other words, the assumption of consistent preferences over time, implies that 1- β(1-αV) = V. Solving this simple equation for V, we find $V = \frac{1-\beta}{1-\alpha\beta}$.

This completes the proof of Rubinstein's theorem according to which Jill and Jack will settle at t=1 on portions $\frac{1-\beta}{1-\alpha\beta}$ and $\frac{\beta(1-\alpha)}{1-\alpha\beta}$ respectively. Why at t=1? Because, as rational people, they recognise that the equilibrium division will be the same whether they settle immediately or much later and, therefore, conclude that there is nothing to gain (and much to lose) from delaying the agreement.

Step D: k does not matter

The proof above relied on the assumption that k=3; namely that it is common knowledge to Jill and Jack (at t=1) that in the space of merely three periods their beliefs on the outcome will have become consistently aligned. We shall now show that this was assumed only for convenience as the proof of Rubinstein's theorem holds for any finite value of k. To get a flavour of why this might be so, let us consider the case k=5. Figure 4.3 is now replaced by Figure 4.4 in which there are an extra two rows and backward induction begins at t=k=5. Otherwise, the three first rows of Figure 4.4 are identical to Figure 4.3.

Round	Proposer	Proposed share for Jill	Proposed share for Jack
$t=k=5$	**Jill**	V	$1-V$
$t=4$	**Jack**	αV	$1-\alpha V$
$t=3$	**Jill**	$1-\beta(1-\alpha V)$	$\beta(1-\alpha V)$
$t=2$	**Jack**	$\alpha[1-\beta(1-\alpha V)]$	$1-\alpha[1-\beta(1-\alpha V)]$
$t=1$	**Jill**	$1-\beta\{1-\alpha[1-\beta(1-\alpha V)]\}$	$\beta\{1-\alpha[1-\beta(1-\alpha V)]\}$

We begin at the last stage ($t=5$) where Jill gets V; then we move to $t=4$ where Jack offers her αV, keeping $1-\alpha V$ for himself; then to $t=3$ where Jill must offer him $\beta(1-\alpha V)$, claiming $1-\beta(1-\alpha V)$ for herself; then to $t=2$ where Jack offers Jill $\alpha[1-\beta(1-\alpha V)]$ to induce a settlement, claiming $1-\alpha[1-\beta(1-\alpha V)]$ for himself; and, finally, to $t=1$ where Jill offers Jack $\beta\{1-\alpha[1-\beta(1-\alpha V)]\}$, demanding $1-\beta\{1-\alpha[1-\beta(1-\alpha V)]\}$ for herself. Setting $V=1-\beta\{1-\alpha[1-\beta(1-\alpha V)]\}$ and solving for V yields the same solution as in Figure 4.3

Figure 4.4 *The case of k=5*

Applying the logic of consistent preferences over time (as we did in the previous stage of the proof),[15] Jill is assumed to make a demand at $t=1$ equal to the share of the pie she can expect to get were she to hold out until $t=k=5$. In other words, $V=1-\beta\{1-\alpha[1-\beta(1-\alpha V)]\}$. Solving for V we find Jill's optimal opening demand of $V=\frac{1-\beta}{1-\alpha\beta}$; precisely the same offer as we had when $k=3$. More generally, for any value of k, Step D yields the following equation:

$$V=1-\beta\{1-\alpha[1-\beta(1-\alpha(1-\beta(1-...\alpha V))...)]\}$$

Solving for V yields, as before, $V=\frac{1-\beta}{1-\alpha\beta}$ independently of how many extra 'stages' the dots (…) entail. In conclusion, as long as the *Hidden Assumption* holds, the actual value of k does not make a difference to the Rubinstein solution.

Recapping, the above proof shows that, in the context of the assumptions made, there is only one rational bargaining strategy that does not involve incredible threats: that is, there is one SPNE (subgame perfect Nash equilibrium). Of course, there are logical difficulties not only with the extra assumptions made (primarily the *Hidden Assumption*) but also with the use of Nash backward induction (the combination of CKR and backward induction) in the construction of the SPNE; see the extensive critique in Section 3.5 of the previous chapter. Let us rehearse our objections to SPNE in general and to Rubinstein's use of it in particular.

Objections to Rubinstein

Rubinstein's solution to the bargaining problem depends on the SPNE concept allied to three important, albeit potentially controversial, assumptions:

(a) A further application of the CAB (or common priors) assumption according to which both players know that, at a commonly known date k, they would settle for V and 1-V respectively.

(b) The assumption of consistent preferences over time.

(c) The assumption that the rate of discount remains the same for each player *over time*.

We have already expressed our objections to CAB and the common priors assumption (see Sections 2.5, 3.5 as well as the preceding pages of the present chapter). Perhaps the only thing we need add here is a comment on the innovative manner in which CAB was utilised by Rubinstein. By assuming consistently aligned beliefs on the number of rounds (k) it would take our bargainers to form consistent estimates on how the pie will be distributed between them (i.e. of division V and 1-V), Rubinstein cunningly introduces a 'final' stage of the bargaining game (stage k) which gives *Nash backward induction* the foothold it needs in some future date before it starts unfolding backwards (from t=k to t=k-1, to t=k-2,… to t=1).

To make the same point slightly differently, the innovation in question is that Rubinstein uses CAB in order to impose a finite end-state to an otherwise infinite-horizon dynamic game. Compare this to the development of SPNE in Section 3.3.2. There, we had assumed an *externally determined end of the game*. Only then did we blend CKR with backward induction in order to procure *Nash backward induction* and the resulting SPNE. Rubinstein, by contrast, treats us to a double dose of CAB: First, he uses CAB in order to render the bargaining process effectively finite (by fixing consistently aligned beliefs on k). Secondly, given the fixity of k, he puts it to work in the traditional manner (see Sections 3.3.2 and 3.5) to constrain the bargainers' beliefs from straying off the equilibrium path. Those sceptical of CAB should take note of the twin use to which it must be put before Rubinstein's solution to the bargaining problem is entertained.

Turning now to the other two assumptions underpinning Rubinstein's solution [(b) and (c) above], while they may sound like straightforward consistency requirements, they abstract from the common human experience of preferences that can be endogenous to bargaining. Thus they ignore the possibility that people pay decreasing attention to material (i.e. money) pay-offs and, instead, as the bargaining process unfolds (especially when their opponents prove more recalcitrant than expected), place more emphasis, for example, on 'beating' them. This psychological interplay is ruled out by Rubinstein, as it is (in all fairness by all game theory). We return, however, to games in which this type of psychological interplay is modelled explicitly in Chapter 7.

To make our critique more concretely, let us use an example which helps bring out the problem of interpreting out-of-equilibrium bargaining behaviour (the very behaviour CAB and the SPNE treatment prohibit). Suppose that Jill gets the bargaining process going and that $V = 0.6$. Jack's best strategy (according to Rubinstein's theory) is to accept 40% of the pie instantly. What will happen if he rejects this and counter-claims, say, 60% at $t = 2$? For this bargaining strategy to make sense, two conditions must hold: (a) there must exist a portion W (>0.4) of the pie which at $t = 2$ is worth more to Jack than 40% of the pie did at $t = 1$; and (b) Jack must have a rational reason for believing that it is possible to get at least W at $t = 2$ if he rejects offer V at $t = 1$ and counter-proposes that he keeps 60%.

Condition (a) is easy to satisfy provided the rate at which the pie is shrinking (in Jack's eyes) is not too high. Condition (b) is far trickier. Specifically, it requires that the experience of an unexpected rejection by Jack may be sufficient for Jill to panic and make a concession not predicted by Rubinstein's model. This development would resemble a tactical retreat by an army which realises that, in spite of its superiority, the enemy may be, after all, determined to die rather

than (rationally) to withdraw; so it is not completely implausible. If Jack's rejection of offer 1-V at t=1 inspires this type of fear in Jill, then she may indeed make a concession beneficial to Jack; and if Jack manages to bring this about by straying purposefully from Rubinstein's SPNE path, then it is not irrational to stray in this manner.[16] Of course, the usual 'trembling hand' explanation of out-of-equilibrium behaviour would preclude this interpretation and so preserve the SPNE outcome. We consider this defence in more detail now.

4.4.3 The (trembling hand) defence of Rubinstein's solution

Suppose for simplicity that $\alpha = \beta = ½$. Then Rubinstein's model predicts that Jill will demand V=⅔ of the pie and Jack will immediately accept this, settling for the remaining one third. Can Jack reject Rubinstein's advice and, instead, reason as follows?

I may do better by rejecting ⅓ of the pie consistently and always insist on a 50-50 split. In this way Jill will eventually understand that I am not prepared to accept less than half the pie. She will then offer me 1-V=½ as this is her best response to the signal I shall be sending.

According to the theory of subgame (Nash) perfection (see Section 3.3.2), the above is wishful thinking. The reason is that the theory assumes that any deviations from the SPNE (i.e. from Rubinstein's strategy) must be due to tiny errors caused by a 'trembling hand'. If this is so, then it is common knowledge that no deviation can be the result of rational reflection; when it *does* occur it is attributed to "some unspecified psychological mechanism" (Selten, 1975, p.35). Moreover, these lapses are assumed to be uncorrelated with each other. If all this were true, then no bargaining move is unexpected since every move has *some* probability of being chosen (mistakenly) by a bargainer. This means that when Jack rejects Jill's offer of ⅓ of the pie, Jill finds it surprising, but not inexplicable. "My rival", Jill thinks, "must have had one of those lapses. I shall ignore it since the probability of a lapse is very small and it is uncorrelated between stages of the process. Next time he will surely accept ⅓, albeit of a smaller pie."

If Jack can predict that Jill will think this way, then he will have to abandon his plan to reject ⅓ of the pie as a signal that he means business. The reason, as explained in the previous paragraph, is that he will know that Jill will not see his rejection as any such signal but only as a random error. Thus Rubinstein (1982) can appeal to Selten's (1975) *trembling hand equilibrium* in order to show that, provided the assumptions of *subgame perfection* are in place, the only rational bargaining strategy is for Jill to demand at the outset a share of the pie equal to V=⅔ (and for Jack to accept the rest, i.e. 1-V=⅓).

The formal trembling hand defence

The complete trembling hand defence of the Rubinstein solution goes like this. Let x ($0<x<1$) be some share of the pie that goes to Jill. Consider the pair of strategies below:

Jill's strategy:	In periods 1, 3, 5,… propose x
	In periods 2, 4, 6,… accept Jack's proposal only if it is no less than x
Jack's strategy:	In periods 1, 3, 5,… accept any demand by Jill if it is not greater than x
	In periods 2, 4, 6,… propose that Jill gets x

These strategies are in a Nash equilibrium (regardless of the value of x) since they are best replies to one another. Underlying them is the threat that any demand by Jill for more than x will be rejected, and that any attempt by Jack to reduce Jill's share to a value below x will be resisted. The question is: *Are these threats credible?*

Rubinstein defends his solution by showing that all x values different to his V are *not* credible (even though they are all potential Nash equilibria). To see this, suppose that the pair of strategies above are in place but that some 'lapse' at t=1 leads Jill unwittingly to propose $x+\varepsilon$ (where ε is some very small positive number) instead of x. If the pair of strategies above are in a *trembling hand equilibrium* (see Section 3.2.1), this means that they can survive small trembles (i.e. small values of ε), in which case Jack will stick to his guns and will not concede to Jill's $x+\varepsilon$ demand. But if the strategies are *not* in a trembling hand equilibrium, they will break down (and be abandoned by bargainers) the moment the possibility of lapses (i.e. $\varepsilon>0$) makes an appearance. Rubinstein argues that a 'good' bargaining solution must be in a *trembling hand equilibrium* and shows that his is the only one that is!

To demonstrate this, following Jill's demand for $x+\varepsilon$ (when she intended to demand only x), Jack can reject it hoping that in the next round (t=2) he will accept Jill's offer of exactly x. Indeed Jack has a good reason to expect this, since the probability of another lapse in Jill's rationality (i.e. of $\varepsilon>0$) is tiny and independent of what happened at t=1. But then again, even if this happens, Jack values $(1-x)$ at t=2 less than he values $(1- x- \varepsilon)$ at t=1 – recall that the pie will shrink if he rejects Jill's proposal at t=1. Thus if ε is sufficiently small, Jack's best reply is to accept Jill's slightly inflamed demand at t=1. Thus Jack's strategy to threaten that any demand by Jill exceeding x will be categorically rejected, is not credible. Thus the pair of strategies above are *not* in a trembling hand equilibrium. Rubinstein's defence concludes by showing that the only pair of strategies that *are* in a trembling hand equilibrium is the one his bargaining solution recommends, i.e. $x = V = \frac{1-\beta}{1-\alpha\beta}$.

Objections to the trembling hand defence of Rubinstein

There is nothing in the above trembling hand defence of Rubinstein which deflects the earlier criticism. It merely demonstrates that the Rubinstein solution is internally consistent provided one assumes that out-of-equilibrium behaviour is explained by random trembles. If any deviation from the SPNE (e.g. rejection of Jill's demand V by Jack at t=1) is so interpreted by Jill, then Jill will take no notice of this rejection. And if it is common knowledge that Jill will take no notice of such a deviation from the SPNE at t=1, then Jack cannot entertain rational hopes that by rejecting offer $1-V$ at t=1 he will bring about a better deal (e.g. $1-W>1-V$) for himself.

But why should one assume this? Why is it uniquely rational for Jill to see nothing in Jack's rejection at t=1 which can inform her about his future behaviour? And why does Jack have to accept that Jill will *necessarily* treat his rejection as the result of a random tremble, rather than as a signal of a defiant, purposeful, stance?

Of course it is entirely possible that Jill will not 'read' anything meaningful in Jack's resistance to V at t=1. It is equally possible that Jack will have anticipated this, in which case he will not reject $1-V$. But equally it seems difficult to rule out, through an appeal to reason alone, the possibility that Jill will take notice of Jack's rejection of $1-V$ at t=1 and to see in it evidence of a 'patterned' deviation from Rubinstein's SPNE. If this happens, she may rationally choose to concede more to Jack. And if Jack has anticipated this, he will have rationally rejected $1-V$ at t=1. In conclusion, an SPNE solution (like that by Rubinstein) may or may not hold… rationally. It seems, therefore, that whether or not it applies is a question to be resolved empirically. We have already reported on some evidence in Chapter 3 with respect to the SPNE concept. Box 4.8 reports on some evidence specifically on the SPNE in bargaining games.

Box 4.8

BARGAINING EXPERIMENTS

The ultimatum game was one of the earliest sequential bargaining games to be tested. In this game, there is only one round. There is a resource ($10) and a proposer must make an offer to give some fraction of this to the second player,

the respondent. If the second player accepts the offer, then they each take their proposed share; but if the second player refuses, then both get nothing. The SPNE of this game has the proposer offering the respondent \$0.01 which is accepted (as this is better than nothing).

Guth et al (1982) report that, in fact, the usual offer is actually close to \$5, the majority of offers were below \$5 but rarely were they close to \$0.01. Rejections were not common but tended to rise with the meanness of the proposer (i.e. with the smallness of the offer).

The interpretation of this and similar results is often that players have regard to behaving 'fairly', in particular that they have a more or less strong preference against inequality which pushes the offer towards the equal share of \$5 (see Fehr and Schmidt, 1999). This could explain, for example, why a different framing for the experiment produces different results. Thus when the experiment is set in market one might argue considerations of fairness are less likely to apply. And indeed when the 'proposer' is asked to set a 'price' for some product which the respondent must either accept or reject, the median 'price' (the equivalent to the offer in the earlier experiment) drops to \$4. Furthermore, when the person playing the role of the 'proposer' earns the right to do this by doing well in a trivia test, the median offer drops even further, to \$3 (see Hoffman et al, 1991). This finding may be due to the fact that 'responders' are happier to accept an unequal division when the 'proposers' have somehow earnt their right to propose.

Goeree and Holt (2001) have recently reported on how a change in the pay-offs affects behaviour in a two-stage bargaining game in a somewhat similar fashion. In this game, when the second player rejects the proposer's offer, he or she gets the chance to make a counter offer which the first player must then accept or reject. However the pie shrinks for this second stage of the game. In the first version that they tested, the pie was \$5 at the first stage and this shrank to \$2 at the second stage. The SPNE of this game has the first player offering \$2 which is accepted (because if the second player rejects any offer, then the best they can expect to get is \$1.99 as the counter offer of \$0.01 is better than nothing and will be accepted by the first player). In the second version, the pie shrinks to \$0.5 in the second stage and the SPNE here has, following the same kind of reasoning, the first player offering \$0.5 which is accepted.

In the first version, the average offer was close to the SPNE at \$2.13. But in the second version, the average offer was \$1.62 which is only slightly lower than before and considerably higher than the SPNE of \$0.5. Again, it is tempting to interpret this result as a reflection of the way that people value fairness. In the first treatment, the SPNE does not generate much inequality, whereas the second does. Thus, the preference against inequality affects behaviour more strongly in the second treatment.

4.4.4 A final word on Nash, trembling hands and Rubinstein's bargaining solution

There is one way to justify the Rubinstein/Nash solution. If we assume that there is a unique solution to the bargaining problem from the beginning, then the case for Rubinstein becomes significantly stronger. This is because a unique solution makes it more plausible to argue that equally well informed players will come to the same conclusion about what this unique way of playing is (i.e. the *Harsanyi-Aumann doctrine*). Thus, they can plausibly assume that (a) players share a belief about k (the stage in the future when players would settle on a commonly known value of V), and (b) that the SPNE concept applies as out-of-equilibrium behaviour will be interpreted as random trembles (since when it is known that there is a unique solution no one could rationally be trying to use wilful trembles to obtain some other outcome when there is CKR).

These are the key building blocks for the Rubinstein solution and so it is the prime candidate once it is assumed that the bargaining problem has a unique solution. Thus, as Sugden (1992a) puts it, Rubinstein's model "…show us what the uniquely rational solution to a bargaining game would be, were such a solution to exist. But we still have no proof that a uniquely rational solution exists" (p. 308). In other words, we still need to know why it is rational to assume that there is a unique solution.

This has become the central issue in our developing discussion of the Nash equilibrium concept and its refinements. There have been other issues (like whether to use forward or backward induction and whether error driven trembles might be related to the costs of trembling) but it is the sometimes explicit, sometimes implicit use of the *Principle of Rational Determinacy* which we have sought to bring out. The Nash equilibrium concept depends (in general) on a presumption that rational players will agree on a uniquely rational way to play any game (see 2.5.2), SPNE and the related sequential equilibrium concept also depend on this (see 3.3.2 and 3.3.3). The *Principle* is

typically justified through an appeal to the *Harsanyi-Aumann doctrine*, but the doctrine is hardly watertight (see Section 2.5.3). This means that any presumption that the Nash equilibrium and its refinements offer the *only* way to analyse interactions between rational players must at best be an act of faith.

For those who make this act of faith, the 'refinement project' remains unfinished, and somewhat pressing, business. It remains pressing because the refinements have not so far produced the necessary amount of determinacy. Most of the games central to social life remain riddled with multiple Nash equilibria and, unless one can explain through an appeal to the assumptions of rationality the theorists' dependence on CKR and CAB, then the faith that there is a unique solution is, at best, stretched. After all, how can one credibly defend the belief that there is a unique solution when there are clearly so many?

None of this should be taken as an argument for ignoring the Nash equilibrium concept and its refinements. We would not have used up so much space explaining these concepts, if this is what we thought. It is an argument about knowing the limits of this analysis and, as a consequence, knowing that something more is needed in most settings. We have hinted here and there at what that 'something' might be. The prime candidate is a concept of convention and a richer concept of rationality to go with this. In particular, in bargaining games it seems that a convention with respect to fairness affects behaviour and the challenge concerns whether this influence can be simply captured within the model of instrumental rationality.

In other words, does the convention just modify the pay-offs in these games (so the SPNE changes but still applies) or does it produce a more fundamental change to the model of rational action and the analysis of social interaction? We shall say more about this in Chapters 6 and 7. For now, we conclude this chapter by casting the absence of a general solution to the bargaining problem in a more positive light: As an invitation to engage with some of key issues in moral and political philosophy which might otherwise be ignored by economics.

4.5 JUSTICE IN POLITICAL AND MORAL PHILOSOPHY

If the Nash solution were unique, then game theory would have answered an important question at the heart of liberal theory over the type of State which rational agents might agree to create. In addition, it would have solved a question in moral philosophy over what justice might demand in this and a variety of social interactions. After all, how to divide the benefits from social co-operation seems at first sight to involve a tricky question in moral philosophy concerning what is just, but if rational agents will only ever agree on the Nash division then there is only one outcome for rational agents.

Whether we want to think of this outcome as a kind of 'natural justice' seems optional. But if we do, and if the Nash solution is the unique outcome of instrumentally rational bargaining, then it will speak to us unambiguously on what justice demands regarding the distribution of benefits in a society inhabited by instrumentally rational agents.

Unfortunately, though, it seems we cannot draw these inferences because the Nash solution is not the only possible one that rational players will converge on. Accepting this conclusion, we are concerned in this section with what bargaining theory then contributes to the liberal project of examining the State as if it were the result of rational negotiations between people.

4.5.1 The negative result and the opening to Rawls and Nozick

Our conclusion is negative in the sense that we do not believe that the Nash solution is the unique outcome to the bargaining game when played between instrumentally rational agents, not even under CKR. This means that game theory is unable to predict what happens in such games. However, this failure to predict should be welcomed by John Rawls and Robert Nozick as it provides an opening to their contrasting views of what counts as justice between rational agents.

Nozick (1974) and entitlements

Nozick argues against *end state theories of justice*, that is theories of justice which are concerned with the attributes or patterns of the outcomes found in society. He prefers instead a procedural theory of justice, that is one which judges the justice of an outcome by the procedure which generated it. Thus he argues against theories of justice which are concerned, for instance, with equality[17] (a classic example of an *end state* or *patterned* theory) and suggests that any outcome which has emerged from a process that respects the 'right' of individuals to possess what they are 'entitled' to is fine. The two types of theory are like chalk and cheese since an intervention to create a pattern must undermine a respect for outcomes which have been generated by voluntary exchange. You can only have one and Nozick thinks that justice comes from a procedural respect for people's entitlements. And, in his view, you are entitled to anything you can get from voluntary exchange (i.e. at the market place). Furthermore, Nozick equates a respect for such entitlements with a respect for a person's liberty.[18]

The importance of the negative result for Nozick's defence of procedural (in preference to *end state*) theories will now be obvious. If each bargain between 'free' and rational agents yielded the Nash solution then it would be a matter of indifference whether we held an *end state* theory or Nozick's *procedural theory* because there would be an *end state* criterion which uniquely told us what we should expect from Nozick's procedure: the Nash solution.

Rawls (1971) and justice

Rawls is concerned with the agreements between rational agents with what he calls 'moral personalities' regarding the fundamental institutions of their society. The introduction of 'moral

personalities' is important for his argument because he suggests that they will want their institutions to be impartial in the way that they operate with regard to each person. In turn, it is the fact that we value impartiality which explains Rawls' particular view on the make-up of our agreements about social arrangements.

Consider how we might guarantee that our institutions are impartial. The problem, of course, is that we are liable (quite unconsciously sometimes) to favour those institutional arrangements which favour us. So Rawls suggests that we should conduct the following thought experiment to avoid this obvious source of partiality: We should consider which institutional arrangement we would prefer if we were forced to make the decision without knowing what position we will occupy under each arrangement. This is known as the *veil of ignorance* device: we make our choice between alternative social outcomes *as if* we were behind a veil which prevented us from knowing which position we would get personally in each outcome.

He then argues that we should all agree on a social outcome based on the principle of rational choice called *maximin*. *Maximin* implies the following procedure: Imagine that you are considering *N* alternative social outcomes (e.g. types of societal organisation, or income distribution). You look at each of these *N* potential social outcomes on offer and observe the person who is worst off in each. Thus you mark *N* persons. Then you make a note of how badly off each of these *N* persons is. Finally, you support the social outcome which corresponds to the most fortunate of these *N* unfortunate persons. That is, you select the social outcome (or, more broadly, the society) in which the well-being of the most unfortunate is highest.[19] (The principle is therefore called *maximin* because it *maximises* the *minimum* outcome.)

Rawls carefully constructs his argument that *maximin* (or the *difference principle* as it is also called) is the principle that rational agents would want to use behind the veil of ignorance. It is not just that they ought to choose it; they will also have a rational preference for it. In other words, we would all choose the social arrangement which secured the highest utility level for the person (whoever it actually turns out to be) who will have the lowest utility level in the chosen society. Thus inequality in a society will only be agreed to (behind the veil of ignorance) in so far as it makes the worst-off person better off than this person would have been under a more equal regime (see Box 4.9).

This is an interesting and controversial result in a variety of respects. We will mention just two before returning to the theme of bargaining theory. First, you will notice that the thought experiment requires us to be able to make what are, in effect, interpersonal comparisons of utility. We have to be able to imagine what it would be like to be the poorest person under each arrangement even though we do not know who that person is (or indeed whether we will be that person). In general we might have to weigh this possibility up with all the other possibilities of occupying the position of each of the other people under some arrangement (although, in fact, the *maximin* rule means we can ignore the latter types of comparisons).

In other words, in general, we have to be able to assign utility numbers to each possible position under each possible arrangement and make a judgement by comparing these utility numbers across arrangements and across positions. As a result, there is a troubling question about where we get these apparently (interpersonally) comparable utility numbers from and why we should assume that all people from behind the veil of ignorance will work with the same numbers for the same positions under the same arrangements.

It is perhaps interesting to note that the *Harsanyi-Aumann doctrine* has been used by some game theorists (see Binmore, 1987) to paper over this problem. The point you will recall is that, according to the *Harsanyi-Aumann doctrine*, rational agents faced by the same information must draw the same conclusions, and this includes assessments of various arrangements from behind the veil of ignorance. Thus given the same information about the institutional arrangements, all rational agents are bound to come up with the same arrays of utility numbers.

Box 4.9

BEHIND THE VEIL OF IGNORANCE

Let us suppose there are three people (A, B and C) in some society. They wish to design the institutions for that society and there are four possible options (I, II, III and IV). They decide in Rawlsian fashion to go behind the 'veil of ignorance' in order to select one of these arrangements. Each arrangement gives the following utility triple for the three possible positions in that society:

Person	*Arrangement I*	*Arrangement II*	*Arrangement III*	*Arrangement IV*
A	5	3	4	1
B	6	5	4	6
C	7	12	4	12

Behind the veil of ignorance, no person knows which position they will occupy under any arrangement, so for instance under *Arrangement I*, each person simply knows they could end up with a '5' or a '6' or a '7' util level. Rawls argues that each person will select *Arrangement I* in these circumstances because it generates the highest utility for the worst-off member of society – Rawls' *maximin* principle (5 is better than 3, 4 and 1, which are the values received by the poorest member under the other arrangements). You will notice this rule does not simply pick out the egalitarian solution (*III*). This is because, by construction, it has been assumed that some inequality in society makes the society as a whole more productive, possibly by providing suitable incentives. So *Arrangement I* yields a total of 18 utils while *Arrangement III* only yields 12 utils.

The arrangement with the highest total utils is *II* (= 20 utils). It is also the one which would be chosen if the people behind the veil of ignorance, rather than using the *maximin* rule, selected the arrangement which offered the highest expected (or average) utility (since the expected utility level under this arrangement with an equi-probability of occupying each position is 6.66 compared with 6 under *Arrangement I*). For this reason a (19th century) utilitarian social philosopher would have recommended *Arrangement II*, as opposed to Rawls' choice of *Arrangement I*. It will also be clear from this example how justice and self-interest need not coincide. If people knew which position they were to occupy, then the middle and best-off people would in all likelihood band together and vote for *Arrangement IV* and this is the arrangement that neither Rawls nor the utilitarians think justice singles out.

Secondly, the *maximin* principle for decision-making is controversial because it is not what economists take to be the general principle of instrumentally rational choice under conditions of uncertainty. The general principle for this purpose in game theory and neoclassical economics is expected utility maximisation (see the relevant boxes in Chapter 1). This has an interesting implication. Suppose (as Rawls asks us to) that behind the veil of ignorance people attach an equal probability to landing in each position under each arrangement. If they select an arrangement on the basis of expected utility maximisation, they will select the arrangement which generates the highest average utility level for that society (see Box 4.9 again). So, expected utility maximisation behind Rawls' veil of ignorance would return us to 19th century utilitarianism; that is, to the principle that the good society is the one which maximises average utility. Of course Rawls rejects expected utility maximisation and argues strongly that rational agents will be using his *maximin* principle behind the veil.

This is enough of the parenthetic comments on Rawls' theory. The general point is that the whole apparatus of the 'veil of ignorance' only fits smoothly into this argument *once we accept that there is no unique solution to the bargaining problem*. After all, if rational agents always reached the Nash agreement, then why do we need to worry about what justice demands when agents contract with each other over their basic institutions? In short, the introduction of 'moral personalities' and the concern with impartiality is a way of selecting arrangements (by appealing in this sense to justice), and this presumes that the bargaining problem is indeterminate. Otherwise why do we need to bring justice into the discussion?

Of course, even if Nash's solution were the unique outcome to the bargaining problem between instrumentally rational agents, then we might still believe that justice has a part to play in the discussion (because, for example, in addition to being instrumentally rational we may also have 'moral personalities'). But this does not avoid a difficulty. It simply recasts the problem in a slightly different form. The problem then becomes one of elucidating the relationship between instrumental reason and the dictates of our 'moral personalities' when they potentially pull in

different directions. Whichever way the problem is construed, it is plain that Rawls' argument is made easier when there is no unique solution to the bargaining problem.

4.5.2 Procedures and outcomes (or 'means' and ends) and axiomatic bargaining theory

One of the difficulties in moral philosophy is that our moral intuitions attach both to the *patterns*, or attributes, *of outcomes* (our *ends*) and to the *processes* (or the *means*) which generate them. These different types of intuition can pull in opposite directions. A classic example is the conflict which is sometimes felt between the competing claims of (a) freedom from interference and (b) equality. We have already referred to this problem when discussing Nozick (who simply finesses it by prioritising freedom from interference, which he identifies with liberty).

Another example in moral philosophy is revealed by the problem of torture for utilitarians. For instance, a utilitarian calculation focuses on outcomes by summing the individual utilities found in society. In so doing it does not enquire about the fairness or otherwise of the processes responsible for generating those utilities with the result that it could sanction torture when the utility gain of the torturer exceeds the loss of the person being tortured. In recent years, and especially since 11th September 2001, we often hear justifications of torture in the cause of "fighting terrorism".

Yet liberally minded people would feel uncomfortable with a society which sanctioned torture on these grounds because it unfairly transgresses the 'rights' of the tortured. Even if it could be demonstrated that average utility would rise if the rights of a 'terrorist' were violated, there is a serious problem: If average utility is the *ultimate* justification for such violations, the source of the utility (or the process that gives rise to it) makes no difference. For example, if torture is justified courtesy of the rise in average utility, then it does not matter whether this rise is due to saving civilians from terrorist attack or allowing many enthusiastic sadists to torture a sole person.[20]

To explore the nature of these conflicts between *means* and *ends*, and advance our understanding of what is at stake when such conflicts occur, it would be extremely helpful if we could somehow compare these otherwise contrasting intuitions by, for instance, seeing how constraints on means feed through to affect the range of possible outcomes. This is one place where axiomatic bargaining theory might be useful. In effect, the rule for selecting a utility pair under this approach is like a procedure because it shows how to move from an unresolved bargain to a resolution, or an outcome. The axioms then become a way of placing constraints upon these procedures which we select because we find them morally appealing and the theory tells us how these moral intuitions with respect to procedures constrain the outcomes. We may or may not find that the outcomes so derived accord with our moral intuitions about outcomes, but at least we will then be in a position to explore our competing moral intuitions in search of what some moral philosophers call a *reflective equilibrium*.

But even those who have little time for moral philosophy or for liberal political theory may still find it interesting to ask: "Granted that society (and the State) are not the result of some living-room negotiation, what kind of *axioms* would have generated the social outcomes which we observe in a given society?" That is, even if we reject the preceding fictions (i.e. of the State as a massive resolution of an *N*-person bargaining game, or of the veil of ignorance) as theoretically and politically misleading, we may still pinpoint certain axioms which would have generated the observed income distributions (or distributions of opportunities, social roles, property rights, etc.) as a result of an (utterly) hypothetical bargaining game. By studying these axioms, we may come to understand the existing society better.

The reader may wish to think in this light about axiomatic bargaining solutions (such as the Nash or the Kalai and Smorodinsky solutions) and the axioms on which they are based. Do they embody any moral or political intuitions about procedures? And if so, how do the Nash or Kalai and Smorodinsky solutions fare when set against any moral or political intuitions that we have

about social outcomes? Rather than pursue these questions here, we shall conclude this chapter with an example based on a different set of axioms.

Roemer (1988) considers a problem faced by an international agency charged with distributing some resources with the aim of improving health (say lowering infant mortality rates). How should the authority distribute those resources? This is a particularly tricky issue because different countries in the world doubtless subscribe to some very different principles which they would regard as relevant to this problem; and so agreement on a particular rule seems unlikely. Nevertheless, he suggests that we approach the problem by considering the following constraints (axioms) which we might want to apply to the decision rule because they might be the object of significant agreement.

(1) The rule should be *efficient* in the sense that there should be no way of reallocating resources so as to raise infant survival rates in one country without lowering them in another.
(2) The rule should be *fair* in the sense (a) of *monotonicity* (that an increase in the agency's resources should not lead to a lower survival rate for any one country) and (b) of *symmetry* (that for countries which have identical resources and technologies for processing resources into survival rates, then the resources should be distributed in proportion to their populations).
(3) The rule should be *neutral* in the sense that it operates only on information which is relevant to infant survival (the population and the technology and resources available for raising infant survival).
(4) Suppose there are two types of resources the agency can provide: x and y. The rule should be *consistent* in the sense that if the rule specifies an allocation $[x^{¢}, y^{¢}]$, then when it must decide how much of x to allocate to countries which already have an allocation of y given by $y^{¢}$, the rule should select the allocation $x^{¢}$. (This means the agency, having decided on how to allocate resources, can distribute the resources to countries as they become available and it will never need to revise its plan.)
(5) The rule should be applicable in *scope* so that it can be used in any possible situation (that is, budget, technologies, etc.).

Each constraint cashes in a plausible moral, pragmatic or political intuition and Roemer shows that only one rule will satisfy all five conditions: It is a *leximin* rule which allocates resources in such a way as to raise the country with the lowest infant survival rate to that of the second lowest, and then if the budget has not been exhausted, it allocates resources to these two countries until they reach the survival rate of the third lowest country, and so on until the budget is exhausted.

4.6 CONCLUSION

The solution to bargaining games is important in life and in political theory. To put the point baldly, if these games have unique solutions, then there are few grounds for conflict either in practice (for example, there will never be a genuinely good reason for any industrial strike[21]) or in theory (when we come to reflect on whether particular social institutions might be justified as products of rational negotiations between individuals). In this context, the claim that the Nash solution is a unique solution for a bargaining game between rational agents is crucial.

Is the claim right? It is at its strongest when it emerges from a dynamic (non-cooperative) analysis of the bargaining process (as in Rubinstein, 1982). The problem with its justification is, however, the same whether we are looking at its static version (e.g. the analysis of Section 4.3) or its dynamic incarnation (see Section 4.4): It relies on the contentious assumptions which support

the Nash equilibrium concept (see Chapter 2), as well as on the extensions of these assumptions which are necessary for the refinements of the Nash equilibrium (see Chapter 3). In brief, we must assume that there is a uniquely rational way to play all games and it is not obvious that this can be justified by appeals to the assumptions of rationality and common knowledge of rationality. Instead, it requires the additional assumption of CAB (or common priors). With respect to solutions based on refinements to the Nash equilibrium (e.g. trembling hand equilibrium, SPNE etc.), what seems to be missing is a generally acceptable theory of mistakes, or trembles, and of how they can be sensibly distinguished from bluffing. Without such an authoritative account, it seems possible to adopt a different view of behaviour which deviates from Nash behaviour, with the result that many potential alternative outcomes to those proposed by the Nash theoretical project remain plausible.

Of all human interaction, bargaining seems to demand the most of game theory. Nothing complicates human behaviour more than face-to-face, open-ended negotiations. Stratagems of mind-numbing complexity are born in such a pressure-filled environment. Bluffs and threats trade places with promises and smiles. Foresight competes with stubbornness and rational assessment of the other side's offers must negotiate successfully a minefield of folly. It is no great wonder that bargaining games expose all of game theory's weaknesses to the most relentless sunlight.

PROBLEMS

4.1 Consider a two-person bargaining game over a pie of size 1. Suppose that, initially, player 1 proposes an agreement which would give her utility u_{11}, leaving player 2 with utility u_{21}. In other words, player 1 proposes point (u_{11},u_{21}) on their utility possibility frontier (UPF). In the meantime, player 2 has other ideas, proposing instead point (u_{12},u_{22}) [where $u_{11} > u_{12}$ and $u_{22} > u_{21}$]. Let us now assume that, for each player, there is a maximum subjective probability of conflict (that is, no agreement) that she can stand (i.e. if her estimate of conflict equals this probability, she is indifferent between acquiescing and holding firm). Moreover, assume that (a) the player with the lower maximum subjective probability of conflict concedes first, and (b) agreement implies that the two players' maximum subjective probabilities of conflict are equal.

Show that the players' subjective probabilities of conflict are equalised only by the agreement corresponding to Nash's solution of the bargaining problem. How do you interpret this result?

4.2 Consider another two-person bargaining game. Player 1 has utility function give by $u(x) = x$ where $x \in (0,1)$ is the share of the pie that 1 will receive, leaving $1-x$ for player 2, whose utility function is given as $v(x) = (1-x)^{\frac{1}{2}}$. Suppose further that an outside body (the Law, an arbitration commission, the mafia etc.) prescribes that in case of non-agreement between players 1 and 2, player 2 will be awarded one third of the pie while player 1 will receive nothing. Find the Nash solution to this bargaining problem.

4.3 The bargaining game above (see problem 4.2) is repeated with one difference: The outside body prohibits any agreement which gives player 2 a share of the pie below one-third. Find the Nash solution to this bargaining problem.

NOTES

[1] Having said that, it is important to note that the axiomatic approach retains its appeal at the philosophical level, since it allows political and moral philosophers to explore the different agreements that might emerge under different social conventions, or under different assumptions about what society deems 'important' values.

[2] What if an agent were to claim during pre-play negotiations that he or she would bid for the $1000? That would indeed be a significant signal. In equilibrium, no one would have an incentive to make such a claim. The reason is that, if there were such an incentive, it should apply to both players; both players would therefore announce an intention to bid for $1000 and, thus, we would be back at square: once each announces an intention to claim $1000, both players have a reason to violate their announcement and claim the $6000. But this cannot be an equilibrium either. James Farrell (1987) challenges this viewpoint, arguing that signalling one's intention to back down *can be credible*. But for this to be so, in equilibrium, a convention must be introduced; namely, that those who announce their intention to settle for the smaller pay-off, do not change their minds later.

[3] It is possible, naturally, to move the conflict point anywhere in the diagram. Suppose for example that disagreement leaves Jill and Jack with utilities equal to 0.1 and 0.4 respectively. Then the conflict point has new co-ordinates: (0.1,0.4). In this case, we would need to draw new axes going through that point and re-draw the game's UPF accordingly.

[4] P is a function of x; that is, it fluctuates as x changes; in particular, for each potential agreement x on the game's UPF, there corresponds one and only one value of P. A simple way of finding the value of a variable (e.g. x) that maximises a function of that variable [e.g. $P(x)$] is by finding the value which sets the slope of that function equal to zero; that idea being that at the function's maximum, its slope is zero. Since the slope of a function is given by its first order derivative, to find x^N [as the value of x maximising $P(x)$] we simply differentiate $P(x)$ subject to x and set this derivative equal to zero. Then we solve for x. The value we derive is x^N. In this example, $P(x) = (x)(1-x)^n$. The derivative of $P(x)$ is given as: $\frac{dP(x)}{dx} = \frac{d(x)}{dx}(1-x)^n + (x)\frac{d[(1-x)^n]}{dx} = 1\times(1-x)^n + x\times n\times(-1)\times(1-x)^{n-1}$ $= (1-x)^{n-1}\big((1-x)-nx\big) = (1-x)^{n-1}[1-x(n+1)]$. This derivative is rendered equal to zero either when x equals 1 (i.e. Jill takes the whole pie) or when $1-x(n+1) = 0$; that is, when x equals $1/(n+1)$. For $n>0$, we note that $P(x)$ is positive when $x<1$ and zero when $x=1$. Hence, of the two values for x which set the above derivative equal to zero, the one that maximises P is $x = 1/(n+1)$. This is the Nash solution.

[5] Zeuthen (1930) managed to provide such a step-by-step account, leading to a unique solution, but to get to it he had to assume that bargainers were myopic. To be precise, he had to assume that every time they change their offer or demand, they fail to recognise that this 'change' in their bargaining position will affect the bargaining position of their opponent. This assumption of bargaining myopia allowed him to define the system of difference or differential equations that defined an agreement. Analytically, this was equivalent to Cournot's (1838) assumption that firms engaged in oligopolistic output-setting games without recognising that changes in their output choices will bring on similar changes in the output choices of their competitors (see Problem 2.5 for an illustration of Cournot's model). In the same way that Nash's equilibrium rationalised Cournot's solution in the context of non-co-operative games, his solution to the bargaining problem rationalised Zeuthen's bargaining solution. However, the price paid by Nash in both cases was that he had to 'lose' time; in other words, unlike Cournot and Zeuthen who were modelling real-time processes (i.e. firms choosing their outputs in real time and successively, or bargainers making a sequence of offers), Nash's models were purely static.

[6] Note that these 'stances' hold for any probability $1-p$, provided it is the same for both bargainers.

[7] Note that, for this inequality to hold, neither Jack's nor Jill's utility from the proposed agreement x^N can equal zero; i.e. $u_L(x^N)>0$ and $u_K(x^N)>0$.

[8] Recall from Chapter 1 that the choice of utility function is, by definition, rather arbitrary. Thus it is important to have a solution which is not affected by different calibrations of the utility function, since no one calibration is better than another. Moreover, if we add a constant to Jill's utility pay-offs, the rate of change (that is, the derivative) in her utility remains unchanged. Similarly, if we multiply her utils with a constant (and do the same with Jack's), again there will be no effect on the ratio of their rates of change (or slopes, or derivatives). And since bargaining power is, according to Nash, a mere reflection of this ratio, we end up with the assumption that the final agreement must *be independent of utility calibrations* (i.e. of linear transformations).

[9] The formal proof of this theorem is beyond the technical scope of this book. However, the idea is quite straightforward. With each axiom that Nash introduces, he narrows down the range of potential agreements. For instance, *Pareto* rules out any outcome not on the UPF. IUC and S impose specific rules that disallow points on the UPF which would entail asymmetries in final pay-offs not reflecting asymmetries in the slopes of the bargainers' utility functions. Finally, IIA disqualifies all remaining agreements (by assuming that what happens in certain parts of the UPF should not affect the agreement) except one: the Nash solution which coincides with the point of tangency between the UPF and some rectangular hyperbola. See Figure 4.2.

[10] Although David Gauthier invoked the Kalai and Smorodinsky bargaining solution in his 1986 book, he has retreated from that position since then. In a more recent book (see Gauthier and Sugden, 1993) he seems convinced by game theorists' criticisms of his espousal of non-Nash bargaining theory: "The argument" writes Gauthier, "in Chapter V of *Morals by Agreement* cannot stand in its present form." (p.178) With this statement, Gauthier 'capitulated' to the argument of many game theorists (Ken Binmore in particular) that bargaining solutions like that of Kalai and Smorodinsky *cannot* be rationalised as the subgame perfect Nash equilibrium (SPNE) of some stage-by-stage model of negotiations. By contrast, Nash's solution can – as we shall see in the next section. However, Gauthier may have been too deferential to his game theoretical critics. For as we shall also see, there are some dark and heavy clouds hanging over the plausibility of the SPNE- interpretation of Nash's bargaining solution (something that should not surprise the reader of this book following our pointed critique of SPNE in Chapter 3).

[11] To see this, recall that acceptance of Jill's 79.9% offer means that Jack can receive \$79.9 here and now. Now suppose he rejects that offer, insisting that he should get 80% of the \$100 (i.e. \$80). If Jill acquiesces, his pay-off will equal 80% of (100-0.6M), where M is the time Jill takes (in minutes) to accept Jack's terms. It is easy to see that if M>0.21 (i.e. 12.5 seconds) Jack is better off accepting the 79.9% offer than holding out for an 80% share (even if Jill is expected to capitulate with certainty).

[12] Suppose that an offer is rejected or accepted instantly. If it is rejected, a counter-offer is issued (again instantaneously) by the rejecting party. However, once a counter-offer is made, there is a fixed time, say ten minutes, the other party replies to this counter-offer. And so on. It is easy to see that this exogenous delay gives the player who kicks off the negotiations a significant strategic (first-mover) advantage: If Jill offers first, a rejection and a counter-offer by Jack will delay agreement by at least ten minutes. She also knows that Jack knows that. In other words, she begins the negotiations under the common knowledge that her opponent can only reject her opening offer by taking a fixed slice of the overall pie (as we have also assumed that with every second that passes without agreement, the pie shrinks by a fixed percentage). Thus, Jill enjoys a *de facto* strategic advantage over Jack, courtesy of her opportunity to issue an offer before the clock starts ticking; an opportunity which means that Jack faces a fixed cost in rejecting Jill;s offer. Now, if the minimum response time (i.e. the cost of delay, or the rate of the pie' shrinkage) is less than ten minutes, Jack's fixed cost of rejecting Jill's opening offer diminishes. As the minimum response time (or the rate at which the pie shrinks) tends to zero, Jill's strategic, first-mover advantage vanishes.

[13] That is, a proof not to be found in Rubinstein (1982) but, rather, one we devised for the purposes of introducing the non-technical reader to the proof's logic.

[14] For example, if α=β=0.8, two rejections will mean the lost of 36% of the pie; while failed offers will 'destroy' more than 50% of the value that was initially available for distribution between the two bargainers.

[15] Before discussing this noteworthy result further, we note that equality $1-\beta(1-\alpha V) = V$, on which Rubinstein's solution rests, demands that our bargainers have the same discount rates at t=1 as they would later on in the game.

[16] Note that this subversive plan on Jack's behalf is analytically equivalent to R's subversive thoughts in Section 3.5.1.

[17] See Varoufakis (2002/3) for a critique of both Nozick's and Rawls' notions of equality.

[18] Of course, there is a great deal of opposition to such identification. For instance, see Varoufakis (1991, pp. 266-8).

[19] This is how Rawls derives his second principle of justice, the so-called *difference principle*. Rawls also argues that agents will agree to prioritise lexicographically his first principle of justice, which only allows arrangements to be considered if they respect each person's basic, formal freedoms.

[20] The idea here is that if we have N sadists, each of whom gain utility W from torturing some unfortunate victim, then this act of multiple torture would boost average utility as long a $NW > D$ where D are the utils that the victim loses as a result of being tortured. This is why we write that as long as N and W are large enough, average utility will increase when sadists are permitted to torture their victim. The poverty of utility is made clear when we realise that the source of W does not matter. Whether it is the result of the joy of N people from escaping a terrorist attach or from sadistic satisfaction by a group of torturers, the calculus of utility is the same, as is its verdict. It is the thought that the source of utility (or process that generates it) *must* count (as part of any decent moral philosophy) that separates *end-state* from *procedural* approaches.

[21] The only explanation for strikes would be that at least one of the parties is irrational, or information is in short supply (and asymmetrically distributed), or the institutional (legal) framework is not well-suited to reaching agreement . In all three cases, industrial conflict is the result of some deficiency. But this only holds if the bargaining problem (at least in its pure, simple form) has a unique solution. See Varoufakis (1991).

5
THE PRISONERS' DILEMMA
The riddle of co-operation and its implications for collective agency

5.1 INTRODUCTION: THE STATE AND THE GAME THAT POPULARISED GAME THEORY

5.2 EXAMPLES OF HIDDEN PRISONERS' DILEMMAS AND FREE RIDERS IN SOCIAL LIFE

5.3 SOME EVIDENCE ON HOW PEOPLE PLAY THE GAME

5.4 EXPLAINING CO-OPEATION
- 5.4.1 Kant and morality: Is it rational to defect?
- 5.4.2 Altruism
- 5.4.3 Inequality aversion
- 5.4.4 Choosing a co-operative disposition instrumentally

5.5 CONDITIONAL CO-OPERATION IN REPEATED PRISONERS' DILEMMAS
- 5.5.1 *Tit-for-tat* in Axelrod's tournaments
- 5.5.2 *Tit-for-tat* as a Nash equilibrium strategy when the horizon is unknown
- 5.5.3 Spontaneous public good provision
- 5.5.4 The Folk Theorem, Indeterminacy and the State
- 5.5.5 Does a finite horizon wreck co-operation? The theory and the evidence

5.6 CONCLUSION: CO-OPERATION AND THE STATE IN LIBERAL THEORY
- 5.6.1 Rational co-operation?
- 5.6.2 Liberalism and the Prisoners' Dilemma
- 5.6.3 The limits of the Prisoners' Dilemma

5
THE PRISONERS' DILEMMA
The riddle of co-operation and its implications for collective agency

5.1 INTRODUCTION: THE STATE AND THE GAME THAT POPULARISED GAME THEORY

In the early 1950s, when Nash's work was known only to a small band of game theorists, Albert Tucker devised a simple game to illustrate and impress an audience with Nash's equilibrium concept. It worked. The game came to be known as the *Prisoner's Dilemma* and, arguably, it has been more responsible for popularising game theory than anything else.

The *Prisoners' Dilemma* fascinates social scientists because it is an interaction where the individual pursuit of what seems rational produces a collectively self-defeating result. Each person does what appears best (and there is nothing obviously faulty with their logic) and yet the outcome is painfully sub-optimal for all. The paradoxical quality of this result helps explain part of the fascination. But the major reason for the interest is purely practical. Outcomes in social life are often less than we might hope and the *Prisoners' Dilemma* provides one possible key to their understanding.

Tucker's original illustration has two people picked up by the police for a robbery and placed in separate cells. The police know that they are the culprits but have no hard evidence on which to found a prosecution. While in different cells, the district attorney sets out what is likely to happen to each of the 'perps'. The conversation goes something like this in our embellished version of the tale.

> If you both 'confess' then the judge, being in no doubt over your guilt, will sentence you, give or take, to 3 years imprisonment. Of course, you know that our evidence against you is insufficient to convict and, hence, if you both deny the charge, I shall have to set you free. However, if you deny the charges *but your friend in the next cell confesses*, I shall make sure that the judge will take a dim view of your recalcitrance and that an example be made of you; let's say an exemplary punishment of at least 5 years. On the other hand, if the situation is the reverse (with you confessing and your next door neighbour denying the charge), I am sure I can intercede with the judge to give you a suspended sentence, on account of your assistance in bringing about a conviction. Not only that but, do you recall that alcohol license which you requested last month? The one that was turned down? I am sure I could swing it for you.

The DA then asks each whether he will confess. The two accomplices are caught in a *Prisoner's Dilemma*. The interaction takes the form of Game 6.1 below, where we have set out the various possible outcomes for one of the prisoners. We encountered the same game in Chapter 2 as Game 2.18, reproduced next to Game 6.1 for convenience. The two games are strategically identical because the best response by the row player is the same regardless of the column player's choice of strategy: In Game 6.1 'you' (i.e. the row player) will be better off confessing both when the 'other' confesses and when he denies the charge.[1] Similarly, in Game 2.18, R1 yields higher utility pay-offs as a reply to both C1 and C2. (Note that in Game 2.18 we have also included the pay-offs of the column player and

appended the common way of referring to the strategy choice as between the 'co-operative' and 'defect' move.)

	Other Confesses	*Other Denies*
You confess	3 yr sentence	You walk *plus* licence
You deny	5yr sentence	You walk

Game 6.1
The prisoner's dilemma

	C1 (defect)	*C2 (co-operate)*
R1 (defect)	$^{+}1,1^{-}$	$^{+}4,0$
R2 (co-operate)	$0,4^{-}$	3,3

Game 2.18
Prisoner's Dilemma in utility pay-offs

The analysis of the game is startling. Each selects their dominant strategy, confess (or defect), and they both go to prison for 3 years when they could have both walked by both denying the crime.

Box 5.1

TOSCA'S DILEMMA

In Puccini's opera, Tosca, there is a police chief called Scarpia who lusts after Tosca. He has an opportunity to pursue this lust because Tosca's lover is arrested and condemned to death. This enables Scarpia to offer to fake the execution of Tosca's lover if she will agree to submit to his advances. Tosca agrees and Scarpia orders blanks to be substituted for the bullets of the firing squad. However, as they embrace, Tosca stabs and kills Scarpia. Unfortunately, Scarpia has also defected on the arrangement as the bullets were real. Thus an elated Tosca, expecting to find her lover and make good their escape, actually discovers that he has been executed; and in one of opera's classic tragic conclusions, she leaps to her death.

It is tempting to think that this catastrophe for the prisoners only arises here because they cannot communicate with one another. If they could get together they would quickly see that the best for both comes from 'denying'. But as we saw in previous chapters, communication is not all that is needed. Promises as well as threats must be *credible*. After they talk, each will still face the choice of whether to hold to an agreement that they have struck over not confessing. Is it in the interest of either party to keep to such an agreement? No, a quick inspection reveals that the game's structure remains intact: the best action in terms of pay-offs is still to confess. As Hobbes (1651,1991) remarked when studying a similar problem "covenants struck without the sword are but words". While together, the prisoners may trumpet the virtue of denying the charges but, if they are only motivated instrumentally by the pay-offs, then it is only so much hot air because each will 'confess' when the time comes for a decision.

What seems to be required to avoid this outcome is a mechanism which allows for joint, or collective decision, making, thus ensuring that both actually deny the charge. In other words, there is a need for a mechanism for enforcing an agreement; Hobbes's 'sword', if you like. And it is this recognition which lies at the heart of a traditional liberal argument, dating back to Hobbes, for the creation of the State (or some enforcement agency to which each individual submits, so it could apply equally to an institution like the Mafia).

In Hobbes's (1651,1991) argument, each individual in the state of nature can behave peacefully or in a war-like fashion. 'Peace' is like 'deny' above because when everyone

behaves in this manner it is much better than when they all choose 'war' ('confess'). The reason is, naturally, that peace allows everyone to go about their normal business with the result that they prosper and enjoy a more "commodious" life. However, bellicosity is the best response to both those who are peaceful (because you can extract wealth and privileges by bullying those who choose peace) and those who are bellicose (because might can only be stopped by might). In short, 'war' is the strictly dominant strategy and the population is caught in a prisoners' dilemma where war prevails and life is 'nasty, brutish and short'.

The recognition of this predicament helps explain why individuals might rationally submit to the authority of a State which can enforce an agreement for peace. People voluntarily relinquish to the State some of the freedom that they enjoy in this 'state of nature' because it unlocks the prisoner's dilemma. Of course this is not to be taken as a literal account of how States, or indeed other enforcement agencies, arise. The point of the argument is to demonstrate the conditions under which a State or enforcement agency would enjoy *legitimacy* among a population, even though it restricted individual freedoms.

While Hobbes thought that the authority of the State should be absolute so as to discourage any cheating on 'peace', he also thought the scope of its interventions in this regard would be quite minimal. In contrast much of the modern fascination with the *Prisoner's Dilemma* stems from the fact that it seems such a ubiquitous feature of social life. For instance, it lies plausibly at the heart of many problems which groups of individuals (for instance, the household, a class, or a nation) encounter when they attempt a collective action.

The next section provides some illustrations of how easy it is to uncover interactions that resemble prisoners' dilemmas. This is important for Liberal political theory because it seems to suggest that the State (or some similar collective agency) will be called upon to police a considerable number of social interactions in order to avoid the sub-optimal outcomes associated with this type of interaction. In other words, the boundaries of the State (or collective action) will be drawn quite widely.

In Section 3, we consider some of the experimental evidence on how people play the one-shot *Prisoners' Dilemma* in laboratory settings. This evidence sets a puzzle for game theory because people decide to 'co-operate' or do the equivalent of 'deny' with a surprisingly high frequency. This (and other) evidence on co-operation is particularly important for debates in Liberal political theory because it suggests that people can overcome the *Prisoners' Dilemma* in some settings 'spontaneously'; that is, without resort to the State. Thus, the fact that the *Prisoners' Dilemma* is so characteristic of social life need not entail that the State is similarly intrusive because people apparently find ways of solving the dilemma on their own. Whether (and when) this is the right conclusion to draw depends on the robustness of this experimental evidence.

For this reason, and since the scope of the State's activities has become one of the most contested issues in contemporary politics, it is important to understand why people might co-operate in this game. We turn to this question in Section 4 where we consider a range of possible explanations (also see Chapter 7). Neither sits well with mainstream game theory nor the evidence on one-shot games (Chapter 7 takes up this issue).

Section 5 considers repeated *Prisoner's Dilemma* games. It contributes both to the above discussion of the boundaries of the State and to the development of game theory. In real life, most interactions are repeated and this section illustrates how game theory analyses them. It also shows that co-operation can emerge as a solution to a repeated *Prisoners'*

Dilemma. Thus game theory can explain 'co-operation' provided the interaction forming the *Prisoners' Dilemma* is repeated. Unfortunately, this result is something of a mixed blessing because repeated prisoners' dilemma games yield a multiplicity of Nash equilibria and 'co-operation' does not characterise all of them. In short, we re-encounter the problem of *Indeterminacy* (or equilibrium selection) in the context of game theoretical explanation of 'co-operation' in prisoner dilemma games.

The last section draws together the threads of the discussion on whether rational co-operation is possible and what this means for the Liberal theory of the State.

5.2 EXAMPLES OF HIDDEN PRISONERS' DILEMMA AND FREE RIDERS IN SOCIAL LIFE

The *Prisoner's Dilemma* may seem contrived (by the cunning of the District Attorney's office) but it is not difficult to find other naturally occurring examples. Indeed, it is not uncommon to find the dilemma treated as the essential model of social life (see Barry, 1965, Taylor, 1976, and Stinchombe, 1979, for a critical review). Here are some examples to convey its potential significance.

Trust

Hume (1740,1888) discusses the problem of *trust* that needs to be overcome by two farmers whose crops ripen at different times before they can help each other with the harvest. This problem of *trust* arises in every elemental economic exchange because it is rare for the delivery of a good to be perfectly synchronised with the payment for it and this affords the opportunity to cheat on the deal. For instance, you may buy a good through the *Internet* and the supplier is naturally attracted by the opportunity of charging your credit card and not posting the goods (or sending you a 'lemon'). You have to trust that he or she won't do such a thing before you are willing to engage in the transaction.

A related problem of *trust* arises when there is imperfect information. For example, you may make the payment for a second-hand car at the same time as you take delivery, but it will only be over a period of time after purchase that you discover the quality of the car (so, you will not know what you have really purchased until some time after you have paid for it, just as in the example of purchases over the *Internet*). This is particularly worrying because the second-hand car dealer often has a much better idea than you about the respective qualities of her cars and what is to stop her selling you a 'lemon'?[2] Likewise, the problem has attracted much attention in labour economics because the typical employment contract, while specifying that a worker be paid $\$x$ hour for being on the factory premises, it fails to specify the effort which is expected during those hours. What then prevents the worker goofing-off during working hours? And what guarantee does the worker have that she will not be harassed by an over-zealous employer to work harder (however hard she is already working)?[3]

These are two-person examples of the dilemma, but it is probably the *N*-person version of the dilemma (usually called the *free rider problem*) which has attracted most attention. It creates a so-called collective action problem among groups of individuals. Again the examples are legion. Here are a few.

The free rider problem and global warming

Suppose you would like to see a less polluted environment and there is an engine conversion which is capable of reducing your car's emissions significantly. Of course, for this to improve local air quality (thus helping with a number of local ailments like bronchitis, asthma etc.) and mitigate the problem of global warming, a large number of car owners must convert their engines. The conversion is costly, but you think it worth the cost if there is the improvement to the environment. The difficulty is that the improvement to the environment only comes when more than *M* people (where *M* is quite large) opt for the costly conversion; the conversion of a single engine (yours, that is) makes no difference one way or another. Indeed, the conversion of fewer than *M* engines would bring about an insignificant improvement.

Consider your decision (to convert, or not to convert) under two possible settings: (a) fewer than *M* other people convert, and (b) more than *M* others do convert. Suppose that your ranking of the outcomes is co-determined by the technological/ecological realities described above as well as by your environmental sensibilities. A plausible depiction of your decision-problem can be based on a slightly amended version of Game 2.18 (the one-shot prisoner's dilemma). You are the row player and strategy R2 is to convert your car's engine (whereas R1 is to do nothing). Imagine now that the rest of the drivers determine together (through their individual decisions to convert or not to convert) the strategy of the column player. Suppose, that is, C1 corresponds to "fewer than *M* others convert" while C2 corresponds to "more than *M* others convert".

The utility pay-offs in Game 2.18 for the row player are a plausible ranking based on the reflection that when many others convert and you do not, you get all the benefits to the environment without any of the cost (recall that one engine conversion, by itself, makes no measurable difference to the state of the environment). Similarly, if you are the only one to convert, you bear a hefty private cost without any social, or environmental, benefit. Nonetheless you still have a strong preference for a situation in which you fork out the cost of the conversion (as long as others do the same) to one in which no one converts. And yet, due to the *Prisoner's Dilemma* structure of the game, you have a strictly dominant strategy: *Do nothing*! When others are similarly placed, they likewise *do nothing*.

This means that in a population of like-minded individuals, all will sit idly by while air quality is deteriorating, even though they would all prefer that *everyone* converts their cars. This is plainly 'sub-optimal'. Under these circumstances, however, might they not also agree to the State enforcing engine conversions by law? Alternatively, it is easy to see how another popular intervention by the State would also do the trick: The State could tax each individual who does not convert her car a sum the equivalent to the utility loss from forking out the cost of the conversion (thus, turning pay-off 4 in the off-diagonal of Game 2.18 into, say, 2). This tax hike would turn engine conversions into the dominant strategy.

Domestic labour

A similar predicament arises within the household. Every member of the household may prefer a clean kitchen to a dirty one. Even though it is costly to clean up one's mess, the individual effort is worth it when you get a clean kitchen. But unfortunately, no individual decision to clean up one's own mess will have a significant influence on the state of the kitchen when the household is large because it depends mostly on what others do, rather

than on what a single person does. Accordingly, since it is also costly to clean up one's mess after a visit to the kitchen, each individual leaves the mess they have created and the result is a dirty kitchen. There is nothing like the State which can enforce contracts within the household to keep a kitchen clean, but interestingly within a family household one often observes the exercise of patriarchal power instead: The woman, or women, of the house clean up! The role of the State has in such cases been captured, so to speak, by an interested party: the patriarch.

Public goods

In fact all 'public goods' set up forms of the free rider problem (see Olson, 1965, for an extended discussion). To see why, notice, by definition, that these are goods which cannot be easily restricted to those who have paid for the service (for instance, like the clean kitchen, street lighting or indeed the defence of a nation which, once it is there, is enjoyed by everyone). Thus there is always an incentive for an individual not to 'pay' for this good because it can be enjoyed without the cost when others do; and if they do not, it is likely to prohibitively expensive for a single individual to bear the whole cost. (See Box 5.3 below for a simple public good game.)

Disarmament

Hobbes's state-of-nature discussion is also often thought to apply with equal force to the community of nations (see Richardson, 1960). Each nation faces a choice between arming (R1 in Game 2.18) or disarming (R2). Each would prefer a world where everyone 'disarmed' to one where everyone was 'armed'. But the problem is that a nation which is instrumentally rational and is only motivated by its own welfare might plausibly prefer best of all a world where it alone is armed because then it can extract benefits from all other nations (in the form of 'tributes' of one kind or another). Since it is also better to be armed than unarmed if all other nations are armed (so as to avoid subjugation), this turns 'arming' into the strictly dominant strategy; thus yielding the now familiar sub-optimal result. This has sometimes been taken as the basis of an argument for some form of world government, at least for the purposes of monitoring disarmament or protecting the environment (see below).

Adam Smith and the invisible hand

Adam Smith's account of how the self-interest of sellers combines with the presence of many sellers to frustrate their designs and to keep prices low might also fit this model of interaction. An individual seller choosing between a 'low' and a 'high' price in a market with many suppliers might well prefer the situation where all charge a 'high' price to one where all charge a 'low' one. But, the individual's best reply strategy when others charge 'high' could be to charge 'low', as this would dramatically increase market share. Equally the best response when others charge 'low' could be to follow suit to avoid losing market share. If this is the case then the dominant strategy is to set a 'low' price and if all are similarly placed the result is everyone setting the 'low' price… but this is worse than when all set the 'high' one. It is as if an *invisible hand* was at work on behalf of the consumers.

Joining a trade union (or similar voluntary organisations)

Suppose you have no ideological feelings about unions and you treat membership of your local union purely instrumentally: that is, you are concerned solely with whether membership improves your take-home pay. Further, let us suppose that a union can extract a high wage from your employer only when a large number of employees belong to the union (because, say, only then will the threat of industrial action by the union worry the employer). Now consider your decision regarding membership under two scenaria: one where everyone else joins and the other when nobody else joins.

To join looks like the co-operative strategy R2 in Game 2.18 (or 'deny ' in Game 6.1) and not joining is the equivalent of defection (R1) because the benefits of the higher wage when everyone joins the union could outweigh the costs of union membership. Moreover when everyone joins and you do not, you enjoy the higher wage and avoid paying the union dues. Perhaps not unsurprisingly, the recognition that this might be a feature of the membership decision has sometimes led to calls for a *closed shop*; i.e. the requirement that all employees are forced to join the union. Alternatively it might be thought to reveal that ideological commitment is an essential constituent of trade union formation.

The shared interest that workers have here is a class-interest because workers as a group stand to gain from unionisation while their employers do not. Hence the prisoner's dilemma/free rider might plausibly lie at the distinction which is widely attributed to Marx (in discussions of class consciousness) between a class 'of itself' and 'for itself' (see Elster, 1985).

Marx's theory of class conflict and capitalist crises

Marx presents a famous example in the first volume of his *Das Kapital* where he discusses the conflict over the length of the working day in England. First, he reviews the Labour Acts beginning with the statute of Labourers which was passed in England in 1349 under the pretext that the plague had so decimated the population that everyone had to do more work. The Act of 1883 created a normal working day in the four textile industries: from 5.30am to 8.30pm (the only restriction being that children between 9 and 13 years of age could 'only' be employed for 9 hours). Since the formation of unions, Marx notes, the length of the working day has been a source of industrial conflict. He then offers the following perspective based, implicitly, on the free rider problem:

> History further shows that the *isolated* 'free' labourer is defenceless against the capitalist and succumbs… Thus the labourer comes out of the production process quite different than he entered. The labour contract was not an act of a *free-agent*; the time for which he is *free* to sell his labour power is the time for which he is *forced* to sell it, and only the *mass* opposition of workers win for them the *passing of a law* that shall prevent the workers from selling, by voluntary contract with capital, themselves and their generation into slavery and death. In place of the pompous catalogue of the inaliable rights of man comes the modest *Magna Charta* of the Factory Act. (*Das Kapital*, Vol. 1, 1973, Chapter 3, section iv)

Marx took his argument further by appealing to another free rider problem that plagued, not the workers, but their employers. A capitalist wishes that all *other* capitalists pay high wages, so that workers at large have money to spend on *her* commodities, but is unwilling to

pay her *own* workers more than a pittance. Thus, each capitalist is caught in this free rider trap, all pay low wages, and the level of aggregate demand is too low to absorb the products that the factories are churning out. Thus, capitalism is prone to crises of under-consumption (or, equivalently, over-production). In Marx's own words,

> Every capitalist knows this about his worker, that he does not relate to him as a producer to a consumer, and he therefore wishes to restrict his consumption, i.e. his ability to exchange, his wage, as much as possible. Of course he would like the workers of other capitalists to be the greatest consumers possible of his own commodity. But the relation of every capitalist to his own workers is the relation as such of capital and labour, the essential relation.' (Marx, 1973, p.420).

Joan Robinson, the Cambridge economist, interpreted this same point in the following manner:

> Each [capitalist] benefits by a rise in the wages paid by his rivals, and loses by a rise in the wages which he must pay himself. Each group has an interest in resisting the particular trade union with which it has to bargain, and it does not follow from the fact that each separately has an interest in low wages that all collectively suffer from a rise in wages. (Robinson, 1966, p.88)

So, in one sense, Marx thinks that, if workers manage to escape one free rider problem (the one that prevents them from banding together in pursuit of common goals[4]) they may help rid society of the irrationalities (e.g. under-consumption) caused by another (that is, the free rider problem between capitalists).

Corruption

The free rider problem might also lie behind a worry that the pursuit of short-term gain may undermine the long-term interest of a group or individual. For instance, it is sometimes argued that every member of a government will face, at some point in their career, a choice between a 'corrupt' and an 'upstanding' exercise of office. Corruption[5] by all reduces the chances of re-election for the government and this undermines the long-term returns from office holding (including the ability to form policy over a long period as well as the receipt/exercise of minor, undetectable bribes or local biases). Thus, it is probably inferior to a situation where all are 'upstanding' and long-term rule is secured. Nevertheless, each member of the government may act 'corruptly' in the short run because it is the best action both when others are 'upstanding' and when others behave corruptly (since a single act of corruption will not affect the party's chance of re-election significantly but it will enrich the individual). Thus each individual finds it in their own interest to pursue the short-run strategy of corrupt practice in government and this undermines the long-term party by shortening its period in office. (For a model of corruption based on the free rider problem but also sprinkled with psychological effects, see Section 7.2.2 in Chapter 7.)

The diminishing effectiveness of antibiotics

It is now quite well understood that the widespread use of antibiotics for relatively mild ailments (e.g. the common cold) in any given population reduces the antibiotics' effectiveness and causes bacteria to evolve in a manner that makes them immune to antibiotics. This is a standard free rider problem. Each would prefer a situation in which no

one uses antibiotics gratuitously. But, just as in the example with the engine conversion above, prescribing or not prescribing antibiotics to one patient makes no discernible difference to the evolution of sturdier bacteria and for the population as a whole. And since antibiotics somewhat reduce the suffering of individual patients *irrespectively of how many are using them* (albeit at varying degrees, depending on how widespread the use is), the dominant strategy of each patient is to pressurise their doctor to prescribe the drug. Taking into account also the interest of pharmaceutical companies to sell more antibiotics, the result is that the effectiveness of antibiotics diminishes, more potent variants are developed by pharmaceutical companies, these are again over-prescribed, even stronger viruses evolve and so on…

Why stand when you can sit?

To end on a lighter note, consider the choice between standing and sitting at a sporting event or open-air concert. Each person's view of the action is the same when either everyone stands or everyone sits, the only difference between these two outcomes is that sitting is less tiring and so is preferred. However, when you stand and everyone else sits, the view is so improved that the fatigue is worth bearing. Of course, the worst outcome is when everyone stands and you sit because you see nothing. Thus standing can be associated with strategy R1 in Game 2.18 and sitting with R2, and the strict application of instrumental logic predicts that we shall all stand, wear ourselves out, and lament our collective irrationality.

5.3 SOME EVIDENCE ON HOW PEOPLE PLAY THE PRISONERS' DILEMMA AND FREE RIDER GAMES

People are not always guided by the apparent logic of the *Prisoners' Dilemma*. To take the last example, in many sporting events in England, or in free classical musical events the world over (such as *Opera at the Park* in Sydney), spectators/audiences typically remain seated when confronted by this dilemma. On the other hand, things are different in countries such as Greece or in different contexts, e.g. rock concerts in the UK. Likewise, the widespread existence of voluntary organisations that depend on public subscription which are individually costly, but which give widespread benefits (e.g. trades unions, public service broadcasters, the British Life Boat Service, the system of voluntary blood donation in the UK) suggests that there are many social interactions in which people are capable of solving the dilemma without recourse to a collective agency like the State. At the same time, however, public spiritedness seems absent when it comes to keeping a public toilet clean (especially those situated on an aeroplane) or maintaining good road manners on clogged up highways.

What is it that determines whether the free rider logic will prevail or not? We review some of the experimental evidence on this question now.

The prisoners' dilemma, and its free rider version, has probably been the topic of more experiments than any other game. Coleman (1983), for instance, lists 1500 experiments. Some of the list refer to repeated versions of the game, which will not be discussed in detail until Section 4. Nevertheless, even those that test the one-shot version constitute an impressive collection and they supply a variety of insights into how people play this game. The *Prisoners' Dilemma* was also probably the subject of the first

experiment in economics. In 1950, two Rand Corporation researchers, Flood and Drescher, asked two friends (Alchian, an economist at UCLA, and Williams, a colleague at Rand) to play the prisoners' dilemma 100 times. Mutual defection [outcome (R1,R1) in Game 2.18] occurred on only 14 of the plays while there was mutual co-operation [outcome (R2,R2)] in 60.

The subsequent experiments on the one-shot version of the game reveal a similar pattern: people co-operate (i.e. choose R2 in Game 2.18) much more than game theory leads one to expect. While defection (strategy R1) is the clear prediction, people choose to co-operate in these experiments between 30% and 70% of the time. To give a flavour of the results in more detail, we describe an experiment by Frank, Gilovich and Regan (1993). They organised the experimental subjects into groups of three and each was told that they would play the game once with each of the other members of the group and that confidentiality/anonymity would be maintained.[6] In one version, the groups could make promises to each other, even though the anonymity of play makes such promises non-credible, unenforceable and thus 'unbelievable'. In the other two versions no promises as such could be made and they varied in the amount of time allowed for pre-play discussion. Their main statistically significant findings were:

(1) the probability of men defecting (i.e. playing R1 or C1) was 24% higher than women;
(2) the probability of defecting was 33% lower in the groups where promises were allowed in the pre-play discussion;
(3) the probability of an economics major defecting was 17% higher than non-economics majors;
(4) when promise-making was not possible, economists defected 72% of the time, compared with 47% for non-economists, whereas in the sessions in which promises were allowed economists defected only 29% and non-economists 26% of the time. [So it seems that the difference in (3) is wholly attributable to the play in the groups where pre-play discussion is constrained and promise-making not possible.];
(5) the probability of defection fell as students progressed through university (a third year student was 13% less likely to defect than a first year one).

The results in (3) and (4) have been replicated in free rider experiments (see Marwell and Ames, 1981) and raise an important question concerning the influence of an economics education on behaviour (see Box 5.2). The proportion who play the equivalent of the dominated co-operative move (R2 in Game 2.18) in these one-shot free rider games is again in the same region, of 40-60% (see Davis and Holt, 1993). The positive influence of pre-play discussion on co-operation also emerges strongly in these free rider experiments, against game theorists' predictions.

One reason why this might be the case is that discussion involves, as Miller (1992) has argued, people trying to persuade one another to each other's view over what should be done for the best. This rhetorical effort encourages appeals to general (that is, impersonal) principles regarding action rather than ones (like game theoretical instrumental reasoning) which are simply based on the individual pursuit of narrow self interest. In the *Prisoners' Dilemma*, or the free rider problem, this seems especially likely as the logic of self-interested dominance-reasoning makes everyone worse-off than they might otherwise be. So it is perhaps not terribly surprising that discussion encourages people to behave differently as co-operation is so blatantly linked to the common good in these games. Alternatively, discussion may help the players to identify with some group

and this may trigger a different group-way of thinking about the problem (see Bacharach, 1999).

Box 5.2

THE CURSE OF ECONOMICS

Is the fact that Economics majors are less likely to co-operate to be explained by the influence that economics has on them or is it that economics tends to attract the less co-operative kind of person in the first place? Frank et al (1993) found no evidence to support the hypothesis that economics attracts misanthropes. Instead it seems that a training in economics significantly increases the probability that a person becomes less co-operative and that they will be more pessimistic about the likelihood of other people co-operating.

Should we, as economists, be pleased with this result? On the one hand, it seems that students of economics have internalised the message of dominance reasoning, but on the other hand they are less likely to enjoy the benefits from social co-operation. One thing is clear, economic ideas have the power to influence the way people think about themselves and act. This means we can quite legitimately wonder not just whether some theory is a good description of how people behave but also whether it is actually desirable.

Orbell, Dawes and van de Kragt (1989) report on a free rider experiment which examines the influence of discussion in more detail. In their experiment, groups of seven subjects were formed and each subject was given $6 and could either keep it or contribute it to the 'public good'. If the $6 were so contributed, then it would become worth $12 and would be shared equally among the other 6 subjects. Thus, every contribution of $6 to the public purse, resulted in $2 being given to each player, regardless of whether he or she had contributed (in the manner of all public goods which are, by definition, non-excludable). The dominant strategy for an individual is to keep the $6 because he or she then receives this plus $2 from any contribution made by any of the others. This is a typical *Public Good Provision Problem* (see Box. 5.3 for the generic case).

Box 5.3

A GENERIC PUBLIC GOOD GAME

Jill is one of N persons belonging to some group or community. They are all asked to contribute amount c to a common fund. If an individual member does contribute as asked, the common fund grows by amount b, where $b>c$ (we also assume that $N>b/c$). When all contributions are made, they will be divided equally among the N members. Thus, if Jill contributes, she will confer benefit b/N to everyone (including herself) at a personal cost of c. The best symmetrical outcome would be for everyone to contribute since, in that case, each would give c and receive b ($>c$) in return. But, the free rider problem undermines this thought by pointing out Jill's dominant strategy: "Avoid any contribution!" Indeed, it is easy to see that, letting x be the number of members who will contribute, Jill can expect $b\left(\frac{x+1}{N}\right)-c$ if she contributes and $b\left(\frac{x}{N}\right)$ if she does not. It is easy to check that, for any value of x, the latter is always larger than the former (given the earlier assumption that $N>b/c$). In conclusion, the dominant strategy is *not* to contribute and the unique Nash equilibrium corresponding to it is zero-contributions all around and, hence, *no* public good provision.

In the Orbell *et al* experiment, there were two treatments. In the first treatment, no discussion was permitted. In the second, each group was given the opportunity to discuss the game prior to making their choices. Each treatment was further divided. In some the contributions to the public good were divided among members of their group. In others they were given to some other group.

In the absence of discussion, only about 30% gave money to the public good. Against the predictions of game theory (according to which talk is cheap unless backed up by credible threats or promises), discussion raised contributions to 70%, but only in those groups which were the beneficiaries of their own contributions. In those groups where there was discussion but where the public purse was given to members of other groups, the contribution rate was less than 30%.

In a related experiment, as reported in Dawes and Thaler (1988), promise-making in groups was once more found to be correlated with co-operation. However, this influence only holds when the whole of the group makes a promise to contribute to the public good. It seems that when there is less than unanimity (by whatever degree), promise-making is no longer related to the extent of co-operative behaviour. Perhaps universal promise-making helps create a 'group' identity and nothing less will do the trick.

Box 5.4

THE GAME THEORISTS' RETORT I: BEST TO ASSUME THE WORST

Some game theorists reject the notion that the experimental results undermine game theoretic insights. Roger Myerson, for instance, has argued[7] that game theory makes pessimistic assumptions regarding the nature of rationality because its role is to study the sort of social institutions that might work well (and engender social well-being) even when peopled by instrumentally rational egotists. He might have added that, if people turn out to be less instrumental, so much the better.

Box 5.5

THE GAME THEORISTS' RETORT II: HAS THE RIGHT GAME BEEN TESTED?

It is sometimes argued that the evidence from so-called prisoners' dilemma experiments do not tell against the predictions of game theory. The predictions of game theory are purely logical and so they cannot be undone by empirical evidence (to imagine the contrary is to confuse the 'analytic' with the 'synthetic', in the jargon that is sometimes used). It follows that when people play a co-operative move in what seems like a *Prisoners' Dilemma*, the fault lies not with game theory but with the person conducting the experiment: the subjects must have thought that they were playing some other game.

This interpretation is, of course, absolutely correct. Either the pay-offs in the experiment were misdescribed or the subjects were governed by some other form of reasoning. If the subjects had been instrumentally rational and these were the pay-offs then defection is the only course of action. Nevertheless, this way of putting things does nothing to dispel the disquiet which these experiments create. It simply means the disquiet needs to be more carefully phrased. The trouble for game theory arises because any theory that is to be potentially useful in the world needs to be capable of being mapped on to the world. One needs to know when and where it applies and these experiments suggest that many situations which look under an ordinary description like prisoners' dilemmas turn out not to be interactions of this sort.

The experiments, therefore, not only exhibit surprisingly high levels of co-operation (although see Boxes 5.4 and 5.5 for contrary views), they also suggest that co-operation is likely to be highest when it is expected to be reciprocated. Since the apparently conditional/reciprocal nature of co-operation proves important in the discussion in the next section of how co-operative behaviour might be explained, we end this one with some complementary evidence from outside the laboratory on the role of reciprocation.

Examples where co-operation depends on reciprocation are not difficult to find. For instance Hardin's(1982) discussion of how the free rider problem is overcome by voluntary organisations in the US emphasises the part played by an American commitment to a form of contractarianism whereby 'people play fair if enough others do'. Likewise Axelrod (1984), building on the work of Ashworth (1980), describes how the combatants during the Great War (WW1) overcame the problem through a reciprocal 'live-and-let-live' norm.

This was a war of unprecedented carnage, both at the beginning and the end. Yet during a middle period, non-aggression between the two opposing trenches emerged spontaneously in the form of a live-and-let-live norm. Christmas fraternization is one well known example, but the live-and-let-live-norm was applied much more widely. Snipers would not shoot during meal times, allowing both sides to go about their business 'talking and laughing' at these hours. Artillery was predictably used both at certain times and at certain locations. So both sides could appear to demonstrate aggression by venturing out at certain times and to certain locations, knowing that the bombs would fall predictably close, but not on, their chosen route. Likewise, it was not considered 'etiquette' to fire on working parties who had been sent out to repair a position or collect the dead and so on.

Of course, both sides (that is, the troops, not the top-brass) gained from such a reciprocal norm; and yet it was surprising that the norm was adhered to because there was an incentive for every individual to behave differently. After all, each individual was under extreme pressure to demonstrate aggression (through, for instance, the threat of court martial if you were caught being less than fully bellicose) and no individual infraction of the norm were likely to undermine the existence of the norm itself. So, adherence to the norm quite plausibly involved solving a free rider problem.

Yet there is little doubt, as the designation 'live-and-let-live' suggests, that the norm would not have survived the failure of one side to reciprocate. Indeed, this is one of the defining characteristics of norms: a norm requires widespread adherence so that an action informed by it entails an expectation of reciprocation based on others following it too. Thus when the norm specifies a form of co-operative behaviour (as with the 'live-and-let-live' norm during WW1), it is implicitly a form of conditional co-operation.

Examples in economics where co-operative norms have been invoked to explain economic performance quickly multiply. For instance, it is sometimes argued that the norms of Confucian societies enable those economies to solve the prisoner dilemma/free rider problems within companies without costly contracting and monitoring activity and that this explains, in part, the economic success of those economies (see Hargreaves Heap, 1991, Casson, 1991, North, 1991). Akerlof's (1984) discussion of loyalty filters, where he explains the relative success of Quaker groups in North America by their respect for the norm of honesty, is another example. As Hardin(1982) puts it: "they came to do good and they did well". As a final illustration of the part played by reciprocation, consider Turnbull's (1961) account of what happens when someone fails to reciprocate.

Turnbull is discussing how the Forest People (the Pygmies of the Congo) hunt with nets in the Ituri forest. It is a co-operative enterprise resembling a cross between the *Prisoner's Dilemma* and the *Stag-Hunt* (Games 2.18 and 2.16 respectively from Chapter 2): it requires each person to form a ring with their nets to catch the animals which are being beaten in their direction. In addition, it is tempting for each individual to move forward from their allotted position because they thereby get a first shot at the prey with their own net.

Such action is, of course, disastrous for the others because it creates a gap in the ring through which the prey can escape and so lowers the overall catch for the group.

Hunting among the Pygmies, therefore, has many of the elements of a free rider problem and yet, almost without exception, the norm of hunting in a particular way defeats the problem. Turnbull witnessed, however, a rare occasion when someone, Cephu, ignored the norm. He slipped away from his allotted position and obtained a 'first bite' at the prey to his advantage. He was spotted (which is not always easy, given the density of the forest) and Turnbull describes what happened that evening.

> Cephu had committed what is probably one of the most heinous crimes in Pygmy eyes, and one that rarely occurs. Yet the case was settled simply and effectively, *without any evident legal system* being brought into force. It cannot be said that Cephu went unpunished, because for those few hours when nobody would speak to him he must have suffered the equivalent of as many days solitary confinement for anyone else. To have been refused a chair by a mere youth, not even one of the great hunters; to have been laughed at by women and children; to have been ignored by men – none of these would be quickly forgotten. Without any formal process of law Cephu had been put in his place...(p.109-10) (emphasis added).

The description is a classic account of how the reciprocal character of norms is informally policed in a group.

5.4 Explaining Co-operation

5.4.1 Kant and morality: Is it rational to defect?

Kant supplies one possible explanation of co-operative behaviour. His practical reason demands that we should undertake those actions which, when generalised, yield the best outcomes. It does not matter whether others perform the same calculation and actually undertake the same action as you. The morality is *deontological* and it is rational for the agent to be guided by a *categorical imperative* (see Chapter 1). Consequently, in the free rider problem, the application of the categorical imperative will instruct Kantian agents to follow the co-operative action (R2 in Game 2.18 or 'deny' in Game 6.1), thus enabling 'rationality' to solve the problem when there are sufficient numbers of Kantian agents.

This is perhaps the most radical departure from the conventional instrumental understanding of what is entailed by rationality because, while accepting the pay-offs, it suggests that agents should act in a different way upon them. The notion of rationality is no longer understood in the means-end framework (as the process by which the agent chooses slavishly the means most likely to satisfy her given ends). Instead, rationality is conceived more as a capacity to transcend one's preferences (as opposed to serve them); as an expression of what is possible. This is not only radical; it is also controversial for the obvious reason that it is not concerned with the actual, direct consequences of an individual action (see O'Neill, 1989, for a defence, however).

It is also rather difficult to reconcile with the evidence above, which seems to suggest that the willingness to co-operate, at least for some agents, depends on whether they have confidence that others will do likewise. This makes co-operation conditional in ways that would be alien to Kantians. Thus while there may be some agents who adopt a high-minded Kantian attitude to the *Prisoner's Dilemma*, it seems that there are other kinds of

motivation at play.[8] Of course, whenever people actual behaviour violates their theories, Kantians can adopt a position similar to that of game theorists: "People 'misbehave' not because our theory of rational action is wrong but because they are not fully rational".

5.4.2 Altruism

Kant's linkage between rationality and morality may seem rather demanding, but there are weaker or vaguer types of moral motivation which also seem capable of unlocking the prisoners' dilemma. For example, an altruistic concern for the welfare of others may provide a sufficient reason for people not to defect on the co-operative arrangement.

Box 5.6

SMITH'S MORAL SENTIMENTS

'How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it.' (Smith, 1759/1976, p.1)

To see how this might work, consider (following Elster, 1989) a commonly understood form of altruism where people act so as to maximise their utility but their utility from a public good is a weighted sum of his or her consumption of the public good *and* everyone else's consumption of it. In this way, although an individual contribution to a public good has a very small effect on the provision of the public good, when this effect is summed across everyone else, as it is in the altruist's utility function, this can tip the instrumental balance towards contribution (i.e. the pay-offs in Game 2.18 associated with 'co-operate' rise and it becomes the dominant strategy). In other words, the recognition of altruistic preferences transforms what appeared to be a *Prisoners' Dilemma* when cast solely in terms of selfish preferences into another kind of game (perhaps into a *Stag-Hunt*, see Game 2.16).

The experimental evidence does not readily fit, however, with this model of altruism. Again it is the apparently conditional nature of people's co-operativeness in the experiments which causes the problems. Indeed, in so far as the altruist's contribution was related to the contribution of others, the relationship is likely to be the reverse of that found in the experiments (and other natural settings, see Box 5.7). This is the case because, if everyone has diminishing marginal returns from consuming the public good, then when an individual altruist expects a higher level of contribution from others, the marginal return from his or her contribution will be lower and so he or she will contribute less.

Box 5.7

EXPERIMENTAL EVIDENCE CASTS DOUBT ON UTILITARIAN ALTRUISM

Sugden (1993) discusses the *British Lifeboat Service*; an institution financed entirely through public donations. "Why do people contribute money to it?" he asks. He points out that the answer cannot lie in utilitarian altruism. For if donors are motivated by an interest in ensuring that the Service has sufficient funds to perform its lifesaving duties, they ought to think of each contributed pound as a perfect substitute for each pound contributed by someone else. Yet the econometric evidence contradicts this hypothesis.[9]

Selten and Ockenfels (1998) make a similar point. They report that, in an experimental setting, winners of a simple lottery proved quite willing to donate a portion of their winnings to the losers but, surprisingly, their donations turned out to be largely *independent* of how much the latter collected from other donors, or even of how the donations were to be divided amongst a number of recipients.[10] This result, just like the econometric evidence reported in Sugden (1993), amounts to a violation of utilitarian

altruism's requirement that donors' valuations of recipients' utility from contributions be symmetrical vis-à-vis the contributors.[11]

5.4.3 Inequality aversion

Another form of moral motivation which has found some support in the experimental literature in other settings (see Box 4.6 on the ultimatum game) is inequality aversion. For example, suppose that a person's pay-offs are one's direct pay-offs minus a psychological utility loss whenever one gets more than the other player. Jill's utility can then be written as: $U_L = \pi_L + \gamma(\pi_L - \pi_K)$ where π_L and π_K are respectively Jill's direct utility from the monetary pay-offs and the indirect utility from the dislike of inequality, and parameter γ is generally zero except when Jill gets more than Jack, in which case $\gamma>0$. We make the assumption about Jack's additional moral motivation and the transformation of the *Prisoners' Dilemma* is profound (see Game 6.2 below).

Assuming Jill chooses among the rows, it is clear that her moral motivation has no effect on her strategic outlook if she is expecting Jack to defect (i.e. to select C1). In this case, defection remains her best reply. But, if she anticipates co-operation from Jack, it is not at all clear that she would wish to defect. As long as $\gamma>1$, Jill's best reply is to co-operate. Indeed, if $\gamma=2$ for both players, then the game is fully transformed into the *Stag Hunt* (Game 2.16). (Sen, 1967, is a good source here.)

	C1 (defect)	C2 (co-operate)
R1 (defect)	1,1	4-γ, 0
R2 (co-operate)	0,4-γ	3,3

Game 6.2
An example of how a form of inequality aversion can transform a Prisoner's Dilemma (Game 2.18; with $\gamma=0$) into a Stag-Hunt Game (Game 2.16 ; with $\gamma=2$)

Like the earlier model of altruistic motivation, this version of moral motivation fails to explain the seemingly conditional nature of people's willingness to co-operate in the experimental evidence. This is a weakness.

Nevertheless, there is an interesting subsidiary issue which is worth brief reflection because it seems that the instrumental model of rational action has been preserved in both these models of moral motivation (unlike the earlier Kantian move). Agents still have utility functions that represent their preferences and which they attempt to maximise. Thus one might expect that game theory would be entirely happy to embrace these kinds of moral motivations (even if they do not quite reconcile the theory with the evidence in the case of the *Prisoners' Dilemma*).

There is, however, a tricky question as to whether a moral orientation can be captured by a set of 'ethical' preferences. On the one hand, the instrumental model is usefully quiet about the specific nature of people's preferences, and so there is no reason to exclude ethical ones of this kind. On the other there are two reasons for suspicion. First, there are well known (at least among non-economists) difficulties associated with any coherent system of ethics (like utilitarianism), and so it seems quite unlikely that a person's

ethical concerns will be captured by a well-behaved set of preferences (see for instance Sen, 1970).

Second, there is some evidence that forms of moral motivation do not quite work in an instrumental manner because they are not all of a piece with other preferences. In particular, whereas it makes perfect sense to say that someone acts to satisfy, say, a preference for hunger, and that one's actions here are sensitive to the prices of different kinds of food, it makes much less sense to talk of acting ethically because the price was right. Indeed it seems that this difference can lead to a form of crowding-out of moral motivation when actions are encouraged through price incentives which appeal to instrumental reason.

Titmuss's (1970) comparative study of blood donation systems is the famous original piece of empirical work on this issue and it has been followed up since, notably by Frey (1995). Titmuss's major finding was that the quantity of blood supplied was higher in countries, such as the UK, where no money was offered to those who gave blood and donation relied solely on the altruism of the population. Why might that be the case?

Surely people in countries like the US, where there seems to be no reason for supposing that they are any less altruistic than the UK, ought to be additionally encouraged by the payment which is offered in those countries. They weren't. One possible explanation of this surprising result is that the moral motivation was 'crowded out' in the US by the instrumental one. This is because, once people are paid for giving blood, the act is no longer uniquely identified with acting morally and so giving it ceases to be a way of expressing an ethical commitment.

To put this slightly differently the message that is conveyed to others when one gives blood becomes 'confused'. It could reflect a simple cost-benefit calculation because the price of blood was right or it could reflect a moral concern for others; and people who want to act morally do not wish their actions to be open to the interpretation of being driven by pecuniary advantage (see Hollis, 1988, for further philosophical discussion and Frey, 2001, for a survey of the evidence).

This is merely an early warning that ethical concerns tend to sit uneasily when cast as just another kind of preference within the instrumental model. We shall return to this issue in Chapter 7 when we discuss some new models of motivation that make reciprocal moral action central. They belong to a class of norm-guided explanations of behaviour and warrant a chapter in their own right.

5.4.4 Choosing a co-operative disposition instrumentally

One line of argument that both explains the reciprocal/conditional character of co-operation and appears to keep faith with the instrumental model comes from Gauthier (1986). He remains firmly in the instrumental camp and ambitiously argues that its dictates have been wrongly understood in the *Prisoner's Dilemma*. Instrumental rationality demands co-operation and not defection, claims Gauthier (1986).

To make his argument he distinguishes between two sorts of maximising dispositions: They can be *straightforward maximisers* (SM) or *constrained maximisers* (CM). An SM invariably defects, following the standard logic of strict dominance. On the other hand, a CM uses a conditional strategy of co-operating with fellow CMs and

defects against SMs. Gauthier then asks: Which disposition (SM or CM) should an instrumentally rational person choose to have? (The decision can be usefully compared with a similar one confronted Ulysses in connection with listening to the Sirens; see Box 5.8.)

It is easy to show that, to the extent that one can recognise another's disposition on sight, it pays to be a CM rather than an SM: Jill and Jack play the *Prisoner's Dilemma* (Game 2.18) and Jill must choose between dispositions SM and CM before they choose a strategy. Suppose further that the proportion of CMs in the population (and thus the probability that her opponent, Jack, is a CM) is p.

If Jill becomes a CM, she will meet another CM with probability p and they will co-operate successfully, netting pay-off 3 each. If her opponent happens to be an SM (an event that will occur with probability $1-p$), they will both defect, collecting pay-off 1 each. Her expected returns from disposition CM will thus equal: $ER(CM) = 3p + (1-p)$. Alternatively, if she opts for disposition SM, she will never co-operate with anyone and no one will co-operate with her, thus leaving her with a certain pay-off of 1 util each time. In short, $ER(SM) = 1$. Clearly, for any $p>0$, it pays to choose a co-operative (CM) disposition. Indeed, even in a world full of SMs (i.e. if $p=0$) there is no cost attached to being a CM.

Of course, the CM disposition becomes riskier if an agent's disposition is less than fully transparent. There are two dangers involved here for players adopting disposition CM: First, they may fail to achieve mutual recognition. Secondly, a CM may mistakenly believe her opponent to be another CM, when this is not so, and thus open herself up to costly exploitation. To see whether adopting disposition CM still makes sense, let r be the probability that CMs achieve mutual recognition when they meet; q be the probability that a CM fails to recognise a SM; and p the probability of encountering a CM. Then, Jill's expected returns from adopting dispositions CM and SM are:

$$ER(CM) = p[3r+(1-r)] + (1-p)[(1-q)+q\times 0] = 2pr + 1 - q(1-p)$$
$$ER(SM) = p(1-q) + 4pq + (1-p) = 3pq + 1$$

Thus the instrumentally rational agent will choose a CM disposition when ER(CM) > ER(SM), or when $\frac{r}{q} > 1 + \frac{1}{2p}$. The result makes perfect intuitive sense. It suggests that provided the probability of CMs achieving mutual recognition (r) is sufficiently greater than the probability of failing to recognise an SM, then it will pay to be a CM. What is a 'sufficient' distance between r and q? This depends inversely on how often you encounter a CM. To put some figures on this, suppose the probability of encountering a CM = ½, then the probability of achieving mutual recognition (r) must be at least twice the probability of failing to recognise an SM (q) [i.e. $r/q > 2$].

Hence it is perfectly possible that the disposition of agents will be sufficiently transparent for instrumentally rational agents to choose CM with the result that, on those occasions when they achieve mutual recognition, the co-operative outcome is achieved. Hence it becomes rational to be 'moral' and the *Prisoner's Dilemma* has been defeated! It is an ambitious argument, and if successful, it would connect rationality to morality in a way which Kant had not imagined. (It has been attempted before in a similar way by Howard (1971), see Varoufakis, 1991.) However, there is a difficulty.

Box 5.8

ULYSSES AND THE SIRENS

Ulysses was approaching the rocks from which the Sirens famously sang. He very much wanted to hear their wonderful voices. Unfortunately he knew that if he sailed close enough to hear, he would not be able to continue his journey because, once heard, the voices would beckon him closer and closer until his ship was dashed on the rocks. He could tell himself in advance not to be tempted in this way, but he knew it would be to no avail: such was the beguiling power of their voices. Ulysses' solution to the predicament was to get his men to tie him to the mast. He ordered them to ignore him until they had passed the Sirens, to put wax in their ears and to row on a course which passed close to the rocks.

The problem is: What motivates the CM person to behave in a co-operative manner once mutual recognition has been achieved with another CM? The point is that if instrumental rationality is what motivates the CM in the *Prisoners' Dilemma*, then she or he must want to defect once mutual recognition has been achieved. There is no equivalent of the rope which ties Ulysses hands and 'defect' remains the *best* response no matter what the other person does. This reality re-surfaces in Gauthier's analysis as an incentive for a CM player to cheat on what being a CM is supposed to entail. In other words, being a CM may be better than being an SM, but the best strategy of all is to label yourself a CM and then cheat on the deal. And, of course, when people do this, we are back in a world of defectors.

The obvious response to this worry, over the credibility of constrained maximisation in Gauthier's world, is to point to the gains which come from being a true CM once the game is repeated. Surely, this line of argument goes, it pays not to 'zap' a fellow CM because your reputation for having a CM disposition is thereby preserved and this enables you to interact more fruitfully with fellow CM players in the future. Should you 'zap' a fellow CM now, then everyone will know that you are a rogue and so in your future interactions, you will treated as an SM.

In short, in a repeated setting, it pays to forego the short run gain from defecting because this ensures the benefits of co-operation over the long run. Thus instrumental calculation can make true CM behaviour the best course of action. This is a tempting line of argument, but it is not one that Gauthier can use because he wants to claim that his analysis holds in *one-shot* versions of the game.

Nevertheless, it is a line we pursue next because it provides a potentially simple explanation of how the dilemma can be defeated by instrumentally rational agents without the intervention of a collective agency like the State; that is, provided the interaction is repeated sufficiently often to make the long term benefits outweigh the short gains. We turn to this now.

5.5 CONDITIONAL CO-OPERATION IN REPEATED PRISONERS' DILEMMAS

5.5.1 Tit-for-Tat in Axelrod's tournament

Box 5.9
AN ANIMAL CAPABLE OF PROMISING
"To breed an animal capable of promising – isn't that just the paradoxical task which Nature has set herself with mankind, the peculiar problem of mankind?" Nietzsche (1887,1956)

It is possible to provide a rational game theoretical explanation of reciprocal co-operation when the *Prisoners' Dilemma* is repeated *indefinitely*. In turn, this may help explain the evidence from one-shot games because it is possible that people in life and the laboratory, who are used to repeated interactions, fail to recognise the discrete nature of these one-shot versions of the games. As a result they behave as if they were engaged in a repeated interaction – a mistake, for sure, but an understandable one. Whatever the strength of this particular argument, the possibility of reciprocal co-operation in repeated games is plainly important in its own right. We begin our discussion of this possibility with an interesting experiment by Robert Axelrod.

In the late 1970s Axelrod invited game theorists to enter a competition. They were asked to submit a programme (a computer algorithm) for playing a computer based repeated prisoner dilemma game. Under the tournament rules, each entrant (program) was paired randomly with another to play the *Prisoner's Dilemma* game 200 times. Fourteen game theorists responded and the tournament was played 5 times to produce an average score for each programme.

Tit-for-Tat, submitted by Anatol Rapport, won the tournament (see Axelrod, 1984). This program starts with a co-operative move and then follows whatever the opponent did on the previous move. It is a simple, reciprocal co-operative strategy that punishes defectors with defection in the next round. Forgiveness is equally simple. A defector need only switch to co-operation, suffer a period of punishment as *Tit-for-Tat* follows the previous period's defection and then the *Tit-for-Tat* program will switch back to co-operation.

A second version of the tournament was announced after the publication of the results of the first one. The rules were basically the same. The only change came with the introduction of a random end to the sequence of plays between two players (i.e. rather than fixing the number at 200). This time 62 programs were entered. Even though many of the submitted programmes were designed to defeat *Tit-for-Tat* (the first tournament's uncontested victor) *Tit-for-Tat* proved, once more, not only the simplest program but the winner as well. (And again only one person submitted it, Anatol Rapoport.) The results were also qualitatively similar in other regards.

This and other contests where strategies like *Tit-for-Tat* do better than defecting ones are interesting because they suggest that a reciprocally co-operative rule does best when people use simple rules of thumb to guide their behaviour, rather than complex game theoretical reasoning. It is tempting then to suppose that evolutionary pressures might as a result favour the spread of such rules. If it did and people used such rules of thumb then we would have an explanation of the evidence on co-operation in one-shot games as well as repeated ones. We consider this line of argument in more detail in the next chapter and turn

now to a formal demonstration that *Tit-for-Tat* can be a Nash equilibrium in an indefinitely repeated version of the *Prisoners' Dilemma*.

5.5.2 *Tit-for-tat* as a Nash equilibrium strategy when the horizon is unknown

We saw in Chapter 3 that dynamic games are quite different to static ones in one crucial sense: *they give players the opportunity to condition their behaviour on what their opponent did earlier.* This is precisely what *Tit-for-Tat* does. It co-operates first and then copies in round t+1 the opponent's behaviour in round t. This opportunity explains why behaviour in the repeated version of a game can be so different to that of the same game's one-shot version. When players can adopt such punishing strategies there is a new potential reason for co-operating now as it will avoid punishment at some future date. In this way, the concern for a long run pay-off from future potential co-operation can outweigh the gain from defection now. In the technical language of Chapter 3, the Nash equilibrium of the original one-shot game may not be the only SPNE (subgame perfect Nash equilibria, see Section 3.3.2) of the repeated version.

In the context of the *indefinitely* repeated prisoner's dilemma, we will now show formally that co-operation, although ruled out as a Nash equilibrium in the one-shot case, is, potentially, a Nash equilibrium in the repeated version. This is good news and bad news. The good news is that co-operation is no longer ruled (and, thus, it is not necessarily the case that the State must intervene in order to prevent defection). The bad news is that the spectre of indeterminacy we encountered in Chapters 2 and 3 returns with a vengeance here (and appears in the guise of the so-called *Folk Theorem* – See section 5.5.5 below).

> ***Tit-for-Tat* co-operation as a Nash equilibrium (theorem)**
> The *Tit-for-Tat* strategy is a Nash equilibrium (and also an SPNE) when the *Prisoner's Dilemma* is repeated indefinitely and the chance of the game being repeated in any round now and in the future is sufficiently high.

Proof: Suppose two persons play Game 2.18 (the *Prisoner's Dilemma*) and that, once the current round is over, another round of the same game will follow with probability p (which means that there will be no further round with probability 1-p). Thus, while they know for sure that they will play at least once, the probability that they will play k times equals $1+p+p^2+p^3+\ldots+p^{k-1}$. Probability p can thus be interpreted as the probability that one of the two players will leave the scene (e.g. a company going bankrupt, a death) or the probability that some external agency will end the game (e.g. an impending law that alters the social setting or changes the rules of play).

Let us use letter τ as a shorthand for *Tit-for-Tat*. The proof will come in two steps. *Step 1* shows that there are only three types of replies a player can choose against an opponent playing τ. *Step 2* completes the proof by showing that τ is the best reply to τ(and thus consistent with a Nash equilibrium).

Step 1: There are only three types of response to someone playing τ

The only three broad types of best reply strategies to an opponent who is playing τ are:

c – Co-operate in all future plays;
λ – Alternate co-operation with defection; and
d – Defect in all plays.

To see why all the best possible replies will fit into one or other of these three types (see also Axelrod, 1984, and Sugden, 1986), notice first that, since your opponent is a τ player, she will either co-operate in the present round or defect *depending on what you did in the last round under your best reply strategy*. In each case, your best strategy will either specify that you co-operate now or that you defect. There are therefore four possible scenaria:

Scenario 1: *Your opponent co-operates and your best reply is to co-operate also.* If this is the case then in the following round your opponent will co-operate under τ and so you will face exactly the same situation. If your best strategy specified co-operation in round t before, it will do so again in round t+1. Thus your best reply strategy will specify co-operation in all periods; the case of strategy c above.
Scenario 2: *Your opponent co-operates but your best reply is to defect.* In this case, your opponent will defect in the following round and what happens next can be studied under the next two scenaria which apply whenever you expect your opponent to defect.
Scenario 3: *Your opponent defects but your best reply is to co-operate.* Then your opponent will co-operate in the subsequent round and, since you defect in response to co-operation, a pattern of alternate defection and co-operation will have been established as the best response; the case of the λ strategy above (see also Problem 5.2).
Scenario 4: *Your opponent defects and your best reply is to defect also.* In this instance, there is defection thereafter; the case of strategy d above.

The above exhaust all possible types of best replies to τ. Faced with a τ-opponent, one must co-operate always [strategy c], or alternate between defection and co-operation [strategy λ], or always defect [strategy d].

Step 2: Proving that τ can be a best reply to τ

If both play according to strategy τ, they will achieve co-operation during each and every round of Game 2.18. Let ER(x,y) be a player's expected returns from playing strategy x when her opponent plays strategy y. Then a τ-player matched with another τ-player expects to collect 3 utils (the co-operative pay-off) from each round. Given that they will play k rounds with probability $(1+p+p^2+p^3+\ldots+p^{k-1})$, each of the τ-players anticipates an average pay-off equal to: $\mathrm{ER}(\tau,\tau) = 3(1 + p + p^2 + p^3 + \ldots) = 3/(1 - p)$.

Now let us compare this with the expected return to strategy λ above; that is, to the strategy of alternating between defection and co-operation: A λ-player, when matched with a τ-player, will enter a cycle whereby in the first round the λ-player defects while the τ-player co-operates (netting the λ-player utility equal to 4), in the next round the λ-player co-operates while the τ-player defects (netting the λ-player zero utility), the next round the behaviour is reversed and so on… Thus, the λ-player should expect average pay-off: $\mathrm{ER}(\lambda,\tau) = 4+0p + 4p^2 + 0p^3 + \ldots = 4/(1 - p^2)$.

Finally, we need to compute the expected performance against a τ-player of the third possible type of best reply; the one which defects *always* (strategy d). In the first

round, the *d*-player, matched with a τ-player, will defect in the face of a co-operative move. Thus, she will collect the highest one-shot pay-off of 4. But thereafter, the τ-player will not co-operate again until and unless the *d*-player does so first. But *d*-players never co-operate. Thus, after an initial pay-off of 4 utils, the *d*-player will receive a string of 1 utils (the pay-off from mutual defection). In summary, $ER(d,\tau) = 4+p+p^2 + \ldots = 4 + p/(1-p)$.

To show that τ is a best reply to itself, and therefore that the strategy combination (τ,τ) is a Nash equilibrium, we need to show that τ is as good a reply to τ as either λ or *d*. In other words, the following two inequalities must hold:

$$ER(\tau,\tau) \geq ER(\lambda,\tau) \textit{ and } ER(\tau,\tau) \geq ER(d,\tau)$$

But *both* inequalities hold whenever $p>\frac{1}{3}$. Accordingly, we can conclude that (τ,τ) is a Nash equilibrium in the repeated *Prisoner's Dilemma* whenever the chance of repetition exceeds one-third. *QED*

Box 5.10

CO-OPERATION IN SMALL GROUPS AND THE OPTIMAL SIZE OF A GROUP

It is commonly observed that small groups seem better able to co-operate amongst themselves than do large groups. For example, Olson (1965, 1982) has made much of the role of small groups in this regard when explaining why nations rise and fall (although see the discussion in the next section for some contrary evidence). The result in this section regarding the possibility of co-operation through a *Tit-for-Tat* Nash equilibrium provides a potential (though by no means unique) explanation of this apparent phenomenon.

Suppose individuals randomly interact with members of their group in a manner which has the structure of a *Prisoners' Dilemma*. By definition, the larger the group the smaller the probability that any one individual will interact with the same member of the group again. Hence the smaller the group the greater the likelihood that the chance of future interaction exceeds ⅓ (which we proved was the condition for *Tit-for-Tat* to be a Nash equilibrium in our case).

On the other hand, there is a contrary argument according to which a large group size is important because it permits greater specialisation. Thus it seems likely that groups might have some optimal size: one which strikes a good balance between the benefits from a division of labour while, at once, maintaining a healthy capacity for trust and co-operation in individual relationships. And indeed there seem to be examples of this. For instance, most modern armies have evolved to a form with companies consisting of three platoons of 30 to 40 people each; and it is a conventional rule of thumb in the management literature that 100 to150 people is the maximum size for a firm run on a person-to-person basis (thereafter, some kind of hierarchic form of management becoming necessary).

5.5.3 Spontaneous public good provision

We generalise the result of the previous section here to the *N*-person *Prisoners' Dilemma game* (i.e. the free rider problem). Our example is borrowed from Sugden (1986).

Assume that a group of individuals live in an environment which exposes them to danger (it could be robbery or illness or some such negatively valued event). The danger is valued at *d* utils by each person and it occurs with a known frequency: It affects one person randomly in any period, i.e. with *N* people the chance of falling 'ill' at any one period is $1/N$.

In this environment, each individual faces a choice between 'co-operating' (that is, helping a member of the group who falls 'ill') at a cost equal to *c* utils and 'defecting'

(that is, not helping a member of the group who falls 'ill') which costs nothing. These choices have consequences for the person who is 'ill'. In particular the 'ill' person obtains a benefit bM from the M group members who contribute ($M \leq N$). Further, we assume that $b>c$ as help is more highly valued by someone who receives it, when 'ill', than someone who gives it, when 'healthy'.

The free rider character of the interaction will be plain. Everyone has an interest in a collective fund for 'health care'. But no one will wish to pay: When others contribute you enjoy all the benefits without the cost and when others do not you will be getting no help when you need it.

Now consider a *Tit-for-Tat* strategy in this group which works in the following way. The strategy partitions the group into those who are in 'good standing' and those who are in 'no standing' based on whether the individual contributed to the collective fund in the last time period. Those in 'good standing' are eligible for the receipt of help from the group if they fall 'ill' this time period, whereas those who are in 'no standing' are not eligible for help. Thus *Tit-for-Tat* specifies co-operation and puts you in 'good standing' for the receipt of a benefit if you fall 'ill' (alternatively, to connect with the earlier discussion, one might think of co-operating as securing a 'reputation' which puts one in 'good standing').

To demonstrate that co-operating (to secure a reputation) could be a Nash equilibrium in this indefinitely repeated game, consider the case where everyone is playing *Tit-for-Tat* and so is in 'good standing' with the rest (i.e. $N=M$). You must decide whether to act 'co-operatively' (that is, follow *Tit-for-Tat* as well) or 'defect' by not making a contribution to the collective fund. Notice your decision now will determine whether you are in 'good standing' from now until the next opportunity that you get to make this decision (which will be the next period if you do not fall 'ill' or the period after that if you fall 'ill'). So we focus on the returns from your choice now until you next get the opportunity to choose.

We assume that the game will be repeated next period with probability p (for instance, because you might die this period or migrate to a different group). So, there is a probability p/N that you will fall 'ill' next period in this group and a probability $(p/N)^2$ that you will remain 'ill' for the time period after and so on. Consequently, the expected return from 'defecting' now is that you will not be in 'good standing' and that you will fall 'ill' next period with probability p/N. Moreover, there is a further probability $(p/N)^2$ that you remain 'ill' the period after while still in 'no standing' and so on. Thus the expected return from 'defecting' now is:

$$\mathrm{ER}(D) = (-d)[\, p/N + (p/N)^2 + \ldots \,] = -d\frac{p/N}{1-(p/N)} = -\frac{dp}{N-p}$$

The above expression gives the expected returns from putting yourself in 'no standing' (by declining to contribute to the collective health fund) until you next get a chance to decide whether to contribute.

By the same kind of argument, if you decide to co-operate now, by adopting *Tit-for-Tat*, or τ, then the expected returns are given as:

$$\mathrm{ER}(\tau) = -c + [(N-1)b - d]\,[p/N + (p/N)^2 + \ldots] = -c + [(N-1)b - d]\frac{p}{N-p}$$

Co-operating is more profitable than defection as long as $\mathrm{ER}(\tau) > \mathrm{ER}(D)$ or when:

$$p > \frac{Nc}{c + (N-1)b}$$

The intuition behind this result is simple. You have more to lose from losing your reputation when there is a high chance of the game being repeated and when the benefit (b) exceeds the cost (c) of contribution significantly. To demonstrate the connection with earlier insights from the repeated prisoners' dilemma, suppose $N = 2$, $b = 3$ and $c = 1$: co-operation becomes possible as long as $p > ½$. (The proof is the same as that of the earlier proposition that, in the two-person *Prisoner's Dilemma*, *Tit-for-Tat* is a Nash equilibrium provided the game will be repeated with probability at least equal to 50-50.)

Box 5.11

THE POWER OF PROPHECY

The result in this section suggests that groups that have some degree of permanency will be more likely to achieve co-operation than those which are transient because the probability of repetition is greater in the former than the latter. Indeed, perhaps this explains why cowpokes are notorious lawbreakers in Westerns while the barber and the Sunday school teacher are upstanding members of the community. On this account, there is nothing morally defective, after all, about the cowpokes. They simply interact with other members of the town with lower frequency than do the permanent residents and so their assessments of the probability of future interactions with members of the town are correspondingly lower. In turn, this suggests a further interesting implication for the power of prophesy when group membership is made endogenous.

Suppose people join a group and stay with it when it is successful, but they leave when it fails because it cannot secure co-operation among its members. In these circumstances, an initial belief regarding the likely success/permanence of the group is liable to be self-fulfilling. When people expect permanence rather than transience, the equivalent probability condition above is more likely to be satisfied and so it is more likely that co-operation is achieved with the result that the group keeps its members (i.e. achieves permanence). By contrast, when transience rather than permanence is anticipated, the probability of repetition is expected to be lower and so the (probabilistic) condition for co-operation is less likely to be satisfied. Thus when the group's members are pessimistic about the possibility of co-operation, their pessimism feeds into the constituents of their decision making and often confirms their expectations. The unfortunate aspect of such a confirmation is that it does not mean that, objectively, co-operation (and thus the success of the group) was impossible or 'irrational. It *becomes* impossible or irrational for individuals to co-operate *because of the cloud of pessimism*; and so perhaps the sort of inspirational rhetoric which encourages optimism among a group does play an important part in the success of groups.

Of course, this capacity for beliefs to become self-fulfilling makes the source of the original beliefs quite crucial; and perhaps in turn this provides a key to our understanding of why co-operation is often achieved among a group which has 'other reasons' for being together. For instance, the fact that there is a family relationship or a shared ethnicity or shared gender or some 'common cause' often seems to provide 'other reasons' for being together and these might be just enough to tilt the original expectation with respect to likely repetition in an optimistic direction. Thereafter, the self-fulfilling character of any belief does the rest. Likewise, perhaps this helps explain why revolutionary change is so difficult to achieve. By definition, revolution involves overthrowing traditional sources of allegiance within the group and, unless it can offer a new reason for the group to exist as a group, it will not be able to create the beliefs of new permanence which secure co-operation within the group on new terms.

5.5.4 The Folk Theorem and Indeterminacy in Indefinitely Repeated Games

Once time is introduced into the analysis and the *Prisoner's Dilemma* (or free rider problem) is repeated, everything seems to change. Co-operation gets a chance and defection is suddenly no longer a dominant strategy as long as the probability of the game being repeated in any time period is sufficiently large. In this section we explain why matters are rather more complicated than this.

It is true that the *Tit-for-Tat* strategy (τ) has been shown to be consistent with a Nash equilibrium. However, this is not to say that we should *expect* such a Nash equilibrium to prevail. This is because it is only one of *many* possible equilibria. What are the others? Let us begin with the most pessimistic of them all: the mutual defection (d,d) outcome.

Consider once more the repeated version of the *Prisoner's Dilemma* (Game 2.18) in Section 5.5.2. It becomes immediately obvious that (d,d) is another Nash equilibrium, in addition to (τ,τ). Although τ may be a better reply to τ than d is, it is always the case that d is a best reply to d. You cannot do any better than defect when your opponent is always going to defect and this response is certainly better than using the *Tit-for-Tat* co-operative strategy (τ).

$$\text{ER}(d,d) = 1 + p + p^2 + p^3 + \ldots > \text{ER}(\tau,d) = 0 + p + p^2 + p^3 + \ldots$$

Thus far we have established two Nash equilibria: *Tit-for-Tat* co-operation and mutual defection. Unfortunately, this is not the end. Let's consider a variant of *Tit-for-Tat* (τ) called *Tit-for-Two-Tats* (τ'). The difference is that, following the opponent's defection, players adopting τ' defect twice in a row (even if the opponent reverts to the co-operative strategy immediately). It follows that, after having defected once against a τ'–player (for whatever reason), one must co-operate twice if they wish to reclaim their 'good standing'. In other words, τ' is a less forgiving version of τ.[12] Now, it is clear that if τ is a best reply to itself, so is τ'. In other words, we have just discovered a third Nash equilibrium: (τ',τ').

Although this latest equilibrium seems very similar to (τ,τ), it is indeed importantly different. The reason is that a co-operative pattern based on τ' has greater built-in resistance to random errors or 'trembles' than one based on the more forgiving τ' (recall the discussion of trembles in Chapter 3). Suppose that players are, at times, susceptible to the odd irrational urge to defect in the middle of a nice string of mutually co-operative moves. These urges are, let us assume, rare and players who succumb to them come quickly to their senses. Nevertheless, even the tiniest such 'tremble' is enough permanently to destabilise the *Tit-for-Tat* equilibrium (τ,τ).

The reason is simple. The moment Jack defects (say in period t), he causes Jill to defect in the following round (t+1). But if Jack immediately regrets his 'misdemeanour' and co-operates at t+1, the players' τ–strategies will set them off on an infinite string of co-ordination failures (with one of them co-operating and the other defecting in every round). By contrast, the 'new' equilibrium (τ',τ') is not susceptible to this problem.[13]

So, it seems that *Tit-for-Tat* is not a single strategy but, rather, a family of potential equilibrium strategies. We just established that, for instance, *Tit-for-Two-Tats* (τ') is not just an alternative equilibrium to *Tit-for-Tat* (τ) but that it is more resistant to trembles as well. However, this is not to say that only reasonable strategies (similar to *Tit-for-Tat* – e.g. τ) are potential equilibria. In fact there is an infinity of variants some of which are, in fact, quite silly. Consider for instance strategy *Tat-for-Tit* ($\tau^{\bullet}$) that recommends the following strange behaviour: "Begin the game by defecting once but then switch to co-operation in the next round, thereafter do at τ+1 as your opponent did at t." While such a strategy makes very little sense, it is easy to establish its credentials as a potential Nash equilibrium.[14]

The point here is that the more we look, the more Nash equilibria we find. Moreover, some are sensible (τ and τ'), others are bordering on the absurd ($\tau^{\bullet}$), many involve co-operative moves (e.g. all the strategies based on a reasonable variant of *Tit-for-Tat*), while an equilibrium of permanent defection (d) always lurks. In short, indeterminacy has returned with a vengeance.

It is helpful to summarise our results so far in matrix form. Consider the four strategies we have studied so far for the indefinitely repeated version of Game 2.18: (a) *defect* in each round – d; (b) *Tit-for-Tat* co-operative behaviour – τ, (c) *Tit-for-Two-Tats* – τ'; and (d) *Tat-for-Tit* – $\tau^{\bullet}$. All four turned out to be (for large enough probabilities of repetition) potential Nash equilibria. Although many more could have been envisaged (e.g. a *Tat-for-Two-Tits*!), it suffices to limit our analysis to these. A player who tries, in the context of this indefinitely repeated game, to work out in advance which of long-term strategy to adopt must realise she is, effectively, playing the following normal form game.

	d	τ	τ'	$\tau^{\bullet}$	…
d	$\frac{1}{1-p}$	$3+\frac{1}{1-p}$	$3+\frac{1}{1-p}$	$3p+\frac{1}{1-p}$	…
τ	$-1+\frac{1}{1-p}$	$\frac{3}{1-p}$	$\frac{3}{1-p}$	$\frac{4p}{1-p^2}$	…
τ'	$-1+\frac{1}{1-p}$	$\frac{3}{1-p}$	$\frac{3}{1-p}$	$\frac{4p}{1-p^2}$	…
$\tau^{\bullet}$	$-p+\frac{1}{1-p}$	$\frac{4(1+p)}{1-p}$	$\frac{4(1+p)}{1-p}$	$-2+\frac{3}{1-p}$	…
…	…	…	…	…	…

Game 6.2 – *A normal form representation of the indefinitely repeated version of Game 2.18 with players selecting long-term strategies in advance; only the expected aggregate pay-offs of the row player are displayed (and 1-p is the probability that the game will cease in this round)*

Suppose you are about to embark on a sequence of *Prisoner's Dilemma* interactions (based on the *per round* pay-offs in Game 2.18) where, regardless of how many rounds have already been played, the probability of an additional round is p. Which long-term (or inter-temporal) strategy should you follow? Game 6.2 helps you explore this question

by asking you to imagine that you are the row player (while your opponent selects among the columns). The crux here is that both of you are required to choose, in advance, one *long-term strategy*. It is, therefore, *as if* you are engaged in a static game where the menu of strategic choices comprises long-term strategies.

Let us take your opponent's four long-term strategies one at a time and compute your best replies to each one (exactly as we did in the static games of Chapter 2). Before doing so, however, note well that there are many more potential strategies (than the four examined here) and this is why we have left room for more rows and columns. Now if, for some reason, you think that your opponent will defect in each round (i.e. you expect her to choose long-term strategy *d*), your compound, long-term pay-offs are given in the first column (depending on your own long-term strategy). For example, if you reply with *d* also, both of you will defect in each round and, thus, you will collect pay-off 1 every time you play. As the probability of playing *k* times equals p^k, your expected payoff equals ER(*d,d*) = 1/(1-*p*). In this manner, we compile the whole table.[15] In effect, a repeated game has been reduced to a timeless, static, matrix and the long-term best replies appear to depend on the value of *p*. To make the example more concrete, suppose that $p = ½$. Then, Game 6.2 turns into Game 6.3 below:

	d	τ	τ'	$\tau^{\bullet}$	…
d	$^{+}2,2^{-}$	5,1	5,1	3½,1½	…
τ	1,5	$^{+}6,6^{-}$	$^{+}6,6^{-}$	2⅔,6^{-}	…
τ'	2,6	$^{+}6,6^{-}$	$^{+}6,6^{-}$	2⅔,6^{-}	…
$\tau^{\bullet}$	1½,3½	$^{+}6,2⅔$	$^{+}6,2⅔$	$^{+}4,4^{-}$	…
…	…	…	…	…	…

Game 6.3

The expected aggregate pay-offs of Game 6.2 for both players when p=1/2. Mark ($^{+}$) *depicts the row player's best reply to a given column strategy and* ($^{-}$) *the column player's best reply to a given row strategy. The shaded outcomes are Nash equilibria* [*i.e. outcomes in which the* ($^{+}$) *and* ($^{-}$) *markings coincide, indicating that the long-term strategies which lead to them are best replies to one another*].

Perusing the shaded part of the game's matrix above, we notice a large number of Nash equilibria along, and around, the diagonal. Of course some (located around the matrix's epicentre) are mutually advantageous. It turns out that, by being repeated indefinitely (though not necessarily infinitely) the *Prisoner's Dilemma* becomes more like Rousseau's *Stag-Hunt* game. When we discussed the latter (see Box 2.5 and Game 2.16 in Chapter 2), we noted that there is no reason to presume that rational agents will, necessarily, move toward the mutually advantageous outcomes (i.e. adopt co-operative behaviour), as opposed to falling in the trap of mutual defection (which is also a Nash equilibrium). Indeed, experiments have shown that, in this type of interaction, mutually advantageous Nash equilibria lose out (to lesser Nash equilibria) as subjects gain experience with games such as Game 6.2 (see Boxes 5.12 and 5.13).[16]

Box 5.12

EXPERIMENTS ON EQUILIBRIUM SELECTION

As argued in Box 2.5, in games of the *Stag-Hunt* variety (Game 6.2 being one of them) two commonly suggested principles for equilibrium selection, *efficiency* and *security*, can pull in opposite directions. In Game 6.2, strategies τ and τ' seem attractive because the Nash equilibria based on them offer the prospect of pay-off 6, when mutual defection promises a miserly 2. (This is the *efficiency principle* working in favour of the co-operative *Tit-for-Tat* type of Nash equilibrium). On the other hand, strategy *d* guarantees you a minimum expected pay-off of 2 if your opponent defects also and a much higher pay-off otherwise, whereas strategies τ and τ' can easily leave you with expected pay-off 1. (The *security principle* favouring the mutual defection Nash equilibrium.)

Van Huyck et al (1990) designed an experiment to test which principle was the most powerful. In this experiment, subjects played a discrete version of the game described in Box 2.5. Players had to choose a whole number between 1 and 7 with individual pay-offs determined by the simple formula: $a\times$MIN - $b\times$OWN, where MIN was the smallest number chosen in the group of subjects and OWN was an individual's own choice. Clearly, there are seven Nash equilibria here: (everyone chooses 1), (everyone chooses 2),…,(everyone chooses 7). Efficiency would point to the selection of the equilibrium *everyone-chooses-7*. However, if security were associated with the choice of the *maximin* action (see Chapter 2), then the *everyone-chooses-1* equilibrium would be selected because the action which maximises the minimum outcome is the choice of '1' for each agent.

After each choice, the minimum number was announced and the subjects calculated their earnings. The choice was then repeated and so on. Subjects were also sometimes asked to predict the choices of the group playing the game. Interestingly (see the discussion on CAB in Chapter 2), the predictions for the first play were widely dispersed and would appear to be inconsistent with the *Aumann-Harsanyi doctrine* that players facing the same information will form the same prior probability assessment of how others will play. The experiment was repeated with several groups, some with slight variations to test particular hypotheses. In the first play of the game, neither principle seems to explain most people's actions, although *efficiency* did much better than *security* with, across all the groups, 31% choosing '7' and only 2% choosing '1'.

Although no group achieved perfect coordination on any integer in the first play, the striking result of repetition is that after 10 plays almost all the subjects in the seven versions of the experiment converged on the *everyone-plays-1* equilibrium. Thus, whereas *security* did not seem to be important in the initial play of the game, it became very important later on. There was, however, one version of the experiment where efficiency held sway with quick convergence on the *everyone-choose-7* equilibrium: Where the number of subjects was reduced to two! So the choice of principle may be both sensitive to repetition and group size.

In summary, indefinite repetition opens up a new vista of possibilities for the *Prisoner's Dilemma*. Co-operation, defection, and a host of other patterns involving both, suddenly become potential equilibria of the new, repeated game. The problem is that we have no idea which one of them is more likely. The problem of *Indeterminacy* which occupied us in Chapters 2, 3 and 4 has not only returned, it has done so with a vengeance. There is a formal result in game theory called the *Folk theorem (*because it was widely known in game theory circles before it was written up) that demonstrates that in infinitely and indefinitely repeated games any of the potential pay-off pairs in these repeated games can be obtained as a Nash equilibrium through a suitable choice of strategies.

The Folk Theorem

Every (individually rational) pay-off profile is a Nash equilibrium in the indefinitely repeated version of a finite normal-form game which, without repetition, features a dominant strategy per player. E.g. in the *Prisoner's Dilemma* there is an infinity of strategies that can be supported in equilibrium by suitable punishment mechanisms (for instance, d, τ, τ',$\tau^{\bullet}$…). For an early discussion of this theorem, see Luce and Raiffa (1957) and Shubik (1959).

This is an extremely important result for the social sciences because it means that there are always multiple Nash equilibria in indefinitely repeated games. Hence, even if Nash's equilibrium is accepted as the appropriate solution concept for games with individuals who are instrumentally rational, and who have common knowledge of that rationality, it will not explain how individuals select their strategies because there are numerous strategy pairs which form Nash equilibria in these repeated games. Of course, we have encountered this problem before in some one-shot games (Chapter 2), in dynamic games (Chapter 3) and in bargaining games (Chapter 4). The importance of the *Folk Theorem* is that it means the problem is always there once *any* game is repeated indefinitely.

Box 5.13

DO MARKETS GENERATE THE EQUILIBRIUM PRICE?

Consider two fish markets located on two islands which are served by many fishers who face the same transport costs for supplying both islands. Will the price of fish be equalised across the two markets? Neoclassical economists typically follow Leon Warlas' idea, speculating that the market will behave *as if* there is some auctioneer who will adjust prices until some 'general equilibrium' is achieved. (This is also known as a Walrasian competitive equilibrium.) But is this *as if* assumption realistic? Would prices converge in practice in the absence of a Walrasian auctioneer? The fishers have to solve what is, in effect, a *Stag-Hunt* problem (similar in structure to Games 2.16 and 6.2).

To see this imagine a rather simple case where there are six fishers who catch the same amount and the price will be the same on both islands when three of them supply each place. There are twenty possible combinations of three fishers that can be selected from a pool of six: Fishers A, B and C going to island 1 and D, E and F going to 2 will do the trick; or A, B, D to 1 and C, E, F to 2; or A, B, E to 1 and D, C, F to 2 and so on. So there are twenty possible equilibrium allocations of the fishers and they need to co-ordinate on one. In practice, do the fishers solve this co-ordination problem and does the Walrasian equilibrium price prevail in both markets?

Meyer et al (1992) ran a simple series of experiments along these lines by asking six subjects repeatedly to decide on which market to supply. After each supply decision, they were informed on the numbers supplying each market and the price in each place (where the price varied inversely with the number supplying that market). They were then asked to make a fresh supply decision and so on for 15 rounds. This experiment was conducted 11 times and co-ordination was usually achieved only 3 to 4 times during the 15 plays. The maximum figure was 7 and the minimum was 2. In other words, the Walrasian equilibrium price was realised much less than the half the time.[17]

5.5.5 Does a finite horizon wreck co-operation? The theory and the evidence

Roger Myerson was one of the game theorists asked by Axelrod to participate in his *Prisoner's Dilemma* tournaments (see Section 5.3.2). He refused to enter the competition because, in his words, "[w]hen Robert Axelrod invited me to participate in his original study of the repeated prisoners' dilemma game, I assumed that he was running *a finitely repeated version of the game, in which defecting at every stage is the unique equilibrium strategy*, and I thought that this solution was so obvious that I did not bother entering his contest" [emphasis added].[18] But why should a finite number of rounds (as opposed to the indefinite number which we have been presuming so far in this section) wreck co-operation and, effectively, ensure that the repeated *Prisoner's Dilemma* is no different to its one-shot version?

Nash backward induction and the associated concept of a *subgame perfect Nash equilibrium* (SPNE, discussed in Chapter 3) supply the answer. The reason why the

unique SPNE of the finitely repeated *Prisoner's Dilemma* is mutual defection in each round is exactly the same as the reason for which Selten argued that the *Centipede* (see Game 3.4 and its longer version in Problem 3.4) must end immediately. To see this it helps to remember that what sustains co-operation in the indefinitely repeated version with a strategy like *Tit-for-tat* is the desire to remain in 'good standing' in future periods so that the co-operative outcome is obtained in those periods and it is this desire which can outweigh the immediate gain from defection now. But, if the number of rounds T is commonly known in advance, everyone will know that in that final round (T) there will be no reason to keep on investing in future rewards from co-operation (since there is no future beyond T). Accordingly, players will plan to defect in the last play and will expect their opponent(s) to do likewise – the last play (at T) is, after all, just a one shot version of the *Prisoner's Dilemma* game (where instrumentally rational players by definition defect).

Once the belief is instilled into players' minds that *any* pattern of co-operation will dissolve in round T, they immediately conclude that, as long as their instrumental rationality is commonly known (recall the CKR axiom), there will be no co-operation in round T-1. The reason is identical to the one in the previous paragraph. Why invest in your 'good standing' at T-1 if this investment cannot pay dividends at T (for we have already established that there will be no co-operation at T). Consequently, no one plans to co-operate (or expect another to do so) at T-1. And so on until *Nash backward induction* destroys all co-operative patterns and restores mutual defection as the game's sole Nash equilibrium (SPNE to be precise).

In Section 3.5 we explained why we remain unconvinced by the logical coherence of this particular equilibrium argument. To recall that argument, the problem lies in the plausibility of combining backward induction with the axiom of CKR. Most game theorists (although not all[19]) have no qualms with this potent brew. We feel it is logically problematic for the reason that, in a finite multi-stage game (such as the *Centipede* or the finitely repeated *Prisoner's Dilemma*), CKR demands a specific analysis of future moves which, in turn, instructs us to do something now that makes these future moves impossible (see Section 3.4 for details). We revisit this particular theoretical debate in Problem 5.5 below. For now, we look at the experimental evidence on finitely repeated games. Did the SPNE emerge? Or did co-operation actually survive the finite horizon?

McKelvey and Palfrey (1992) report on a series of experiments based on the *Centipede* (Game 3.4) in which two players alternately get a chance to take the large portion of a continually escalating pile of money *until some maximum sum is reached*. As soon as one person takes, the game ends with the taker getting the large portion of the pile, and the other player getting the small portion. (See also Box 3.4.)

In this experimental game, 'taking' is equivalent to defecting and 'not taking' is a co-operative move (since it allows the sum of the two players' future pay-offs to increase). Moreover, fixing the maximum sum means that this is a game with a finite (and commonly known) horizon. In this sense, the game in question is strategically similar to a finitely repeated *Prisoner's Dilemma*. To see this, suppose that the two players have resisted 'taking' until the very end. In the last round, the player who gets a chance to take will, of course, take the larger portion, leaving the smaller portion for her opponent. The latter, having anticipated this, will prefer to take the larger portion in the previous round (where he is the active player), even though the sum is smaller in that round. And so on.

In this manner (i.e. *Nash backward induction*) we work out the game's unique SPNE: The player to move first takes the larger portion at the game's beginning and thus ends it (i.e. immediate defection). The experimental results, however, show that this does not occur. Indeed, there was persistent co-operation (i.e. players opting not to take the larger portion until quite a few rounds had passed). So, it seems that subjects tend to co-operate (at least in the early rounds) even when the horizon is finite and the unique SPNE suggests that they defect immediately.

Davies and Holt (1993) and Ledyard (1995) provide surveys of other experiments with finite horizon free rider games and have this summary to offer: "…free riding is common, although not as pervasive as most economists would have expected *a priori*… Free riding is most pronounced in small groups, when the decision process is repeated, when participants are experienced and when the MPCR (marginal per capita return) is low" (Davies and Holt, 1993). Ledyard(1995) suggests that there are three types of player:

(a) Dedicated Nash players who act pretty much as predicted by game theory, with possibly a small number of mistakes ('trembles');

(b) A group of subjects who will respond to self-interest, like Nash players, if the incentives are high enough; but who also exhibit sensitivity to decision costs, fairness, altruism, etc. when the stakes are lower;

(c) A group of subjects who behave in manner inconsistent with instrumental rationality. Game theorists might claim that they are irrational. But, as we have seen, others will find in this group elements of alternative types of reasoning (e.g. Kantian)

The proportions of each type (across many different subject pools) are respectively about 50 per cent, 40 per cent and 10 per cent (see Ledyard, 1995).

To see how experimentalists reach such conclusions, consider a typical public good game where each individual has an endowment which they can either keep or contribute in some proportion towards a public good (recall the game in Box 5.3). The part that is contributed to the public good is multiplied in value by some amount (b>1) and shared among all people (N) playing the game. The MPCR referred to above is the return to the individual from contributing an extra $1 to the public good (and is given by b/N which is always less than 1). The clear game theoretical prediction is that people will contribute nothing to the public good when the MPCR is less than one both when the game is played only once, and when it is finitely repeated.

However, in experiments with both one-shot and finitely repeated versions of such games, people on average contribute around 40% to 60% of their endowment. In finitely repeated games the contribution level often falls with repetition but it never seems to disappear: on average it falls to something around 10-30% after 10 plays. It is possible to interpret this decay as evidence that people 'learn' how to play the game from a game theoretical perspective. For instance, it might be argued that the subjects start out with a variety of 'misperceptions' about such games and it is only with repetition that they learn where their 'true interest' lies. So it is only with repetition that behaviour moves towards what game theory predicts.

Nonetheless, the residual level of contribution might still seem to be worrying. It could be argued, of course, that people always make mistakes and since, in this game the only kind of 'mistake' that one can make is to make a positive contribution, the residual level is not so surprising after all. The evidence from the influence of the MCPR adds

weight to this because the decay often seems less pronounced when the MCPR is high (i.e. when the costs of a 'mistake' are correspondingly lower). For instance, Isaac and Walker (1988a) find that the average contribution over 10 rounds for a group of 4 individuals falls from 57% to 19% as the MPCR falls from 0.75 to 0.3, and in a group of 10 it falls from 59% to 33%.

The difficulty with the learning interpretation comes from a version of the experiment where the subjects play the public good game a finite number of times, and when completed, they re-start: i.e. they play the game again for another finite number of periods (e.g. see Andreoni, 1988). If the subjects were simply learning in the course of the repetition, the decay should carry over to when they re-start. Instead when they re-start, the contribution level initially jumps back up to a high rate and then decays again. In other words, whatever they learnt in the first run seems to disappear the moment they come to do it again.

An alternative explanation of decay which explains this finding is that some people start by expecting that others will contribute at a higher level than they actually do. Indeed, the evidence from one-shot games (recall Section 5.1.3) was that co-operation seems to be conditional or reciprocal. It is then not surprising that these people shade their contribution back as the game is repeated because they discover that others are contributing less than they expected. But this is not the same as having discovered that they erred. Indeed, when a fresh game starts, they do not hesitate to contribute significantly all over again in a bid to give others the opportunity to establish a pattern of conditional, or reciprocal, co-operation.

If this is what is going on, then the evidence from repeated free rider games would seem to add weight to the argument for a more sophisticated model of rational agency: One which makes sense of the normative or reciprocal nature of co-operation. This conclusion receives further support from experiments where pre-play communication is allowed. Such communication might plausibly help create normative expectations or a group identity which would prevent decay (see Issac and Walker, 1988b). On the other hand, one might expect from this perspective that increasing group size would hinder the creation of a group identity (recall Box 5.10) and so it is puzzling to find that contributions appear to rise with group size.

5.6 CONCLUSION: CO-OPERATION AND THE STATE IN LIBERAL THEORY

What are the prospects for co-operation in the *Prisoners' Dilemma* and free rider games? This is the central question of this chapter. The question is pressing in part because of the ubiquity of this kind of social interaction and in part because it lies at the heart of debates in Liberal political theory over the boundaries of the State. In this section, we take stock of the argument so far (a) to recap on what seem the key outstanding issues in game theory in need of further investigation and (b) to see where this leaves the debate in political theory.

5.6.1 Rational Co-operation?

Game theory has difficulty accounting for the extent of co-operation in what are apparently either one-shot or finitely repeated prisoners' dilemma/free rider interactions both in laboratory experiments and in the wider world. None of the simple ways of making instrumentally rational agents act morally by giving them 'ethical' preferences seem to explain well the data on co-operation in these settings. Nor does a fully blown Kantian change in the rationality assumption. The failure here is bound up with the reciprocal or conditional nature of co-operation and this has set the agenda for recent research. In particular, we shall consider in Chapter 7 how some game theorists have responded to this difficulty by developing models of rational action which place the urge to reciprocate right into the heart of the psychological processes that generate our preferences.[20]

In marked contrast, game theory can account for co-operation in indefinitely repeated prisoners' dilemma and free rider games. The moment time comes into the picture (and the interaction ceases to be static), repetition allows the players, in effect, to enforce an agreement *themselves*. Players are able to do this by being able to threaten to punish their opponents in future plays of the game if they transgress now. The *Tit-for-Tat* strategy embodies precisely this type of behaviour. It offers implicitly to co-operate by co-operating first, and it enforces co-operation by threatening to punish an opponent who defects on co-operation by defecting until that person co-operates again.[21]

The demonstration that *Tit-for-Tat* is a Nash equilibrium of the indefinitley repeated free rider problem (see Section 5.5.2) appears to have direct applicability to the social world because there are many examples of social interaction where this type of threat could explain how co-operation is achieved. Plainly it might explain the 'live and let live' norm which developed during the Great War in the trenches (with soldiers refusing to fire at each other during implicitly agreed times of the day/month, *against* their officers' command) since the interaction was repeated and each side could punish another's transgression.

Equally, it is probable that both prisoners in the original example may think twice about 'confessing' because each knows that they are likely to encounter one another again (if not in prison, at least outside) and so there are likely to be opportunities for exacting 'punishment' at a later date. Likewise, internal career ladders in companies are a mechanism for enforcing agreements between employers and employees in this mould as the career ladder both encourages repetition of the interaction (because you advance up the ladder by staying with the firm) and it provides a system of reward which is capable of being used to punish those who do not perform adequately. As for our distrust of

second-hand car salesmen, it may be due to the infrequency with which we interact with him (thus precluding the informal mechanisms for enforcing an implicit agreement to supply us a decent car).

Having said all this, some mysteries remain with our earlier examples of co-operation. For instance, how is it that battalions who were about to leave a particular front (thus discontinuing their long term relationship with the enemy on the other side of their trench) continued to 'co-operate' until the very last moment? There is, in addition, a deeper and more worrying theoretical mystery associated with the *Tit-for-Tat* Nash equilibrium in these indefinitely repeated games. Since it is but one among an infinite number of Nash equilibria (ie the *Folk Theorem* applies), there is a pressing question concerning how it is selected.

This takes us straight back, and adds considerable weight, to what has emerged as *the* weakness of game theory in Chapters 2, 3 and 4: The failure to come up with a well accepted theory of equilibrium selection in the presence of multiple equilibria. We consider one last response to this failure in the next chapter: The rise of an *evolutionary* version of game theory.

5.6.2 The Debate in Liberal Political Theory

Liberalism divides on whether the existence of prisoner dilemma and free rider interactions provide grounds for constraining individual freedom through the creation of a State (or similar collective agency) which substitutes collective action for individual ones. Hobbes supplies the classic argument in favour of constraining individual freedom. It is well known and has informed mainstream Liberal thinking on the State throughout the 20th century. Indeed, as more and more areas have seemed prone to free rider problems, the Hobbesian narrative has plausibly contributed to the growth of the State in this period.

Nevertheless it has been powerfully contested, particularly over the last 25 years, by what Anderson (1992) calls the *Intransigent Right* (see Box 5.14 for another fissure). The rise of the *Intransigent Right*, or the libertarian tradition in Liberal theory, is closely associated with a quartet of 20th century thinkers (Strauss, Schmitt, Oakeshott and Hayek) who in Andersen's view now shape "a large part of the mental world of end-of-the-century Western politics". The lineage is, of course, much longer and, as Anderson notes, Hayek (1962) himself traces the battle lines in the dispute back to the beginning of *Enlightenment* thinking.

> Hayek distinguished two intellectual lines of thought about freedom, of radically opposite upshot. The first was an empiricist, essentially British tradition descending from Hume, Smith and Ferguson, and seconded by Burke and Tucker, which understood political development as an involuntary process of gradual institutional improvement, comparable to the workings of a market economy or the evolution of common law. The second was a rationalist, typically French lineage descending from Descartes through Condorcet to Comte, with a horde of modern successors, which saw social institutions as fit for premeditated construction, in the spirit of polytechnic engineering. The former line led to real liberty; the latter inevitably destroyed it. (Anderson, 1992)

Box 5.14

A FUNDAMENTAL CHANGE IN MAN?

A further difference within the European Enlightenment traditions regarding the State centres around the feedback one can reasonably expect between (a) the formation of the State, its laws and institutions, and (b) the formation of the individual's character. Hobbes thought of the former as independent of the latter. The State simply serves the need of the 'exogenous' individual for peace, law and order. In contrast, J.-J.Rousseau's *The Social Contract* (1762), is founded on the belief that the *process* which brings citizens together (through direct political activity, dialogue, democratic process) is one which simultaneously shapes the rational State *and* the rational individual. As persons get together, argue, reach agreements, hold elections, change their minds etc. they evolve into citizens; they *become* rational. In the end, according to Rousseau, the creation of the State and the shaping of the rational citizen are two symbiotic processes.

Some of the specific arguments of the *Intransigent Right* have turned on the difficulties associated with political decision making and State action. For instance, there are problems of inadequate knowledge such that even the best-intentioned and efficiently-executed political decision generates unintended and undesirable consequences. This has always been an important theme in Austrian economics, featuring strongly in the 1920s debate over the possibility of socialist planning as well as contemporary doubts over the wisdom of more minor forms of State intervention.

Likewise, there are problems of 'political failure' (as opposed to 'market failure') that subvert the ideal of democratic decision-making and which can match the 'market failures' that the State is attempting to rectify. For example Buchanan and Wagner (1977) and Tullock (1965) argue that special interests are bound to skew 'democratic decisions' towards excessively large bureaucracies and high government expenditures. Furthermore there are difficulties, especially after the *Arrow Impossibility Theorem*, with making sense of the very idea of something like the 'will of the people' in whose name the State might be acting (see Riker, 1982, Hayek, 1962 and Buchanan, 1954).

These, so to speak, are a shorthand list of the negative arguments against 'political rationalism' or 'social constructivism'; that is, the idea that you can turn social outcomes into matters of social choice through the intervention of a collective action agency like the State. The point is simple: *Reason should know its limits!* There is, in addition, a positive argument against 'political rationalism'. It turns, as the quote above suggests, on the idea that these interventions are *not* even necessary. The failure to intervene in prisoners' dilemma/ free rider interactions (and others requiring some form of co-ordination) does not spell sub-optimality of one kind or another. Instead, an efficient order can be thrown up 'spontaneously' once people interact repeatedly (as they do in settled communities).

Hayek's own argument regarding the prospects for such *spontaneous order* depends on an appeal to evolutionary pressures. We shall examine such arguments in the next chapter. For now, the point to note is that the formal game theoretical result, showing how co-operation can be sustained in an indefinitely repeated prisoners' dilemma/free rider interaction, supplies powerful weight to the libertarian side of the argument within Liberalism. Co-operation can be sustained, apparently, without recourse to the State (or some collective agency) and the acceptance of diminished freedom.

The point can seem blunted, however, since the difference between a Hobbesian State which enforces collective agreements and the generalised *Tit-for-Tat* arrangement is not altogether clear; and so in proving one we are hardly undermining the other. After all, it might be argued that the State merely codifies and implements the policies of 'punishment' on behalf of others in a very public way (with the rituals of police stations,

courts and the like). And is this any different from the golf club which excludes a member from the greens when the dues have not been paid? Or the Pygmies' behaviour towards Cephu? Or the gang which excludes people who have not contributed 'booty' to the common fund? Or is it really very different if you pay the State in the form of taxes or the Mafia in the form of tribute?

The answer here, however, might plausibly be 'yes'. It has been argued, for instance, that there is a big difference between paying tribute to the Mafia and taxes to the State and this is revealed in the contrasting transformations of the post-Soviet Union and Easter European economies. Paying taxes according to this argument is more efficient precisely because the system is transparent in a way that Mafia-like arrangements are not. Likewise, while the inhabitants of Beirut somehow managed to maintain services that were prone to free rider problems during the long civil war without any grand design, most of its citizens prayed for one.

The case for 'political rationalism' is further strengthened by the *Folk Theorem*. For if the emergence and survival of efficient 'spontaneous orders' is only one out of many possible Nash equilibria, the theory offers nothing resembling a guarantee that they will emerge and survive (let alone flourish). Against this, the libertarians often argue that evolution will favour practices which generate the desirable co-operative outcome since societies that achieve co-operation in these games will prosper as compared with those which are locked in mutual defection. This is another cue for a discussion of evolutionary game theory which comes in the next chapter. So we leave the debate now, noting that: (a) it has been sharpened by game theory, and (b) the formal properties of the *Tit-for-Tat* (i.e. that it requires a sufficiently high probability of the game being repeated) may supply a useful guide as to when *spontaneous orders* are feasible alternatives to *collective action*.

5.6.3 The limits of the Prisoner's Dilemma

Stinchcomb (1975) provocatively asks: "Is the Prisoners' Dilemma all of sociology?" Of course, it is not, he answers. Nevertheless, it has fascinated social scientists and proved a pedagogically powerful illustration of the pervasive tension between public virtues and private vices. That it is neither "all of sociology" nor "all of social science" is revealed both in the empirical evidence on the play of one-shot games and in the theoretical analysis when the game is indefinitely repeated. The one suggests that people are more complexly motivated than game theory allows. The other shows that, while co-operation is possible, it is not guaranteed and this poses a problem of equilibrium selection. Social science ought to have something to say about both.

To take the case of arguments around the State, it is not merely a question of whether the State intervenes but, importantly, of how it does so. Should the public good be given a helping hand by direct taxation and State provision? Or should the State privatise it?[22] Should the State regulate an oligopolisitic industry (without being a player in it) or should it act directly through a State-owned production unit? And what about public goods which are demeaned and belittled the moment they are commodified (see Section 5.4.3) Should the State provide them as free, public goods? If so, who should pay for them? More generally, whenever the State intervenes to foster efficiency, whose interests exactly is the State promoting? There will always be more than one way of overcoming a *Prisoner's*

Dilemma or *Free Rider* problem and the distribution of the benefits of collective action will vary accordingly.

All of the above suggest aspects of the State's role that go beyond a simple response to a *Prisoner's Dilemma*. There is, however, a deeper limitation. Both strands of Liberalism discussed above share a common assumption with game theory that people's pay-offs can be identified *prior* to social interaction. Social interaction is, as a result, always a form of exchange, whereas many argue that individual identities are in important respects socially constituted and preferences often cannot be identified independently of social interaction. Box 5.14 on Rousseau is one version of such an argument and Chapter 7 explores others that have been developed in direct response to some of the difficulties game theory encounters in settings that are perceived as prisoners' dilemmas.

PROBLEMS

Problems 5.1 and 5.2 below refer to the (symmetrical) two-person Prisoner's Dilemma in which you play against an opponent with the following pay-offs:

	Other co-operates	*Other defects*
You co-operate	3,3	*k*,1
You defect	1,*k*	2,2

where k≥3

Assume that the game is repeated one more time with probability p independently of how many times it has been played already.

5.1 If k=4, under what conditions is the *Tit-for-Tat* strategy (τ) consistent with a Nash equilibrium?

5.2 Show that, as long as k is large enough, the alternating strategy (a) defined below is a better reply to strategy τ than τ is to itself. Is strategy a a Nash equilibrium in this case? If not, what is?

Definition of strategy *a*: "Co-operate in round 1. From then onwards, (A) every time mutual co-operation occurs defect in the next round but then immediately co-operate in the following one; (B) whenever your opponent co-operated (in a round you defected), in the next round co-operate; and (C) whenever your opponent defected after a round in which you co-operated, defect in the following round."

Problems 5.3,5.4&5.5 below refer to the Prisoner's Dilemma in which you play against an opponent with the following pay-offs:

	Other co-operates	*Other defects*
You co-operate	3,3	1,4
You defect	4,1	2,2

5.3 Suppose that players know in advance (and this is common knowledge) that the game will be repeated three times only. What is the game's SPNE (subgame perfect Nash equilibrium - see Section 3.3.2 of Chapter 3)?

5.4 Two persons, A and B, play the *Prisoner's Dilemma* above three times (say, at t=1,2,3). The difference with Problem 5.3 is that CKR is relaxed. To be precise, B begins the game with some doubt in his mind about A's instrumental rationality; that is, she thinks that there is a probability p (>0) that A co-operates blindly as long as B does so too. In contrast, A is perfectly certain that B is fully instrumentally rational (a case of one-sided asymmetrical information). Draw the extensive form of this finitely repeated *Prisoner's Dilemma* and, by means of *Nash backward induction*, compute the game's *sequential equilibrium* (see Sections 3.3.3 and 3.5.3 of Chapter 3).

5.5 Outline the critique of SPNE and of sequential equilbrium in the context of problems 5.3 and 5.4 above (see Section 3.5 of Chapter 3).

NOTES

[1] If the other is to confess, failing to confess yourself means an additional 2 years in gaol. If he denies, you are guaranteed to walk whatever you do. However, if you confess not only do you get to walk free but you get the coveted betting licence as well.

[2] Akerlof (1980) is the main reference here.

[3] See Bowles (1985).

[4] Such as shorter working hours, or a change in property rights over the means of production

[5] 'Corruption' here might range from serious 'kickbacks' to the favouring of departmental policies which benefit the minister's local constituents when alternatives would secure greater advantage for the party nationally.

[6] In part through adding a further random pay-off to each person's winnings at the end of the experiment so that no one would know how any particular person had responded to them.

[7] See interview with Varoufakis, in Kottaridi and Siourounis (2002). An English version is available from Varoufakis on request.

[8] One of course should be careful with such moral pronouncements. Kantian Mafiosi will fail to fall out with one another (opting to co-operate because co-operation is best for all of them) with an end result as far removed from an ethical equilibrium as possible. Honour among thieves is a double edged sword as far as the Kantian rationality-morality nexus is concerned.

[9] See also Sugden (1982).

[10] Three subjects A, B and C participated in a lottery which would award each DM10 with probability 2/3. Subjects where asked *ex ante* to State how much of their winnings they were prepared to share with the other subjects in their team of three who won nothing. Subject A was invited to declare the sum she would donate to B (or C) **if** A were to win DM10 **and** B (or C) was the only loser in the trio. Let us call this sum X. Then A was asked to select her donation to both B and C if neither B nor C were to win any money. Let this sum equal Y and assume that 'losers' B and C split Y between them. 52% of the subjects chose X≅Y (up to a rounding error), a finding which the authors label *fixed total sacrifice* (FTS) and show to be inconsistent with standard utilitarian altruism.

[11] For instance, in the Selten and Ockenfels experiment, symmetry means that, in A's eyes, *ceteris paribus* the loss of one expected currency unit (e.g. DM1) by a 'losing' subject B yields the *same* disutility for subject A as the loss of DM1 by a winning C who nevertheless donates DM1 to some other 'loser'.

[12] E.g. suppose that Jack defected at T but then repented by co-operating at T+1. If she is following τ, Jill will defect at T+1 (punishing Jack for his defection at T) but co-operate at T+2 (in response to Jack's co-operation at T+1). In contrast, were Jill to adopt τ', she will react to his T-period defection by defecting twice in a row (in rounds T+1 *and* T+2) even if Jack co-operated at T+1.

[13] For if Jack defects as a result of some 'tremble', Jill will defect twice in a row, so will Jack and, thus, a string of mutually co-operative moves will ensue.

[14] If Jill expects Jack to play $\tau^{\bullet}$ her expected returns from adopting the same strategy equal $ER(\tau^{\bullet},\tau^{\bullet}) = 1 + 3p + 3p^2 + \ldots = -2 + 3/(1-p)$. Meanwhile, if she responds with permanent defection (d), she can expect $ER(d,\tau^{\bullet}) = 1+4p + p^2 + p^3 + \ldots = 3p + 1/(1-p)$. Clearly, as long as p exceeds 1/3, the former exceeds the latter, thus rendering $(\tau^{\bullet},\tau^{\bullet})$ a Nash equilibrium.

[15] The pay-offs featured in Game 6.2 have been calculated as follows:

Row 1: If you play d, your expected returns depend on your expectation regarding your opponent's choice. Letting $ER(x,y)$ denote your expected return from strategy x given that you believe that your opponent will play y, we have

$ER(d,d) = 1 + p + p^2 + p^3 + \ldots = 1/(1-p)$.
$ER(d,\tau) = 4 + p + p^2 + p^3 + \ldots = 3 + 1/(1-p)$
$ER(d,\tau') = 4 + p + p^2 + p^3 + \ldots = 3 + 1/(1-p)$
$ER(d,\tau^{\bullet}) = 1+4p + p^2 + p^3 + \ldots = 3p + 1/(1-p)$

Row 2: If you choose τ, your expected returns (depending on your beliefs regarding your opponent) are:

$ER(\tau,d) = 0 + p + p^2 + p^3 + \ldots = -1 + 1/(1-p)$
$ER(\tau,\tau) = 3 + 3p + 3p^2 + 3p^3 + \ldots = 3/(1-p)$
$ER(\tau,\tau') = 3 + 3p + 3p^2 + 3p^3 + \ldots = 3/(1-p)$
$ER(\tau,\tau^{\bullet}) = 0 + 4p + 0p^2 + 4p^3 + \ldots = 4p/(1-p^2)$

Row 3: If you choose τ', your expected returns (depending on your beliefs regarding your opponent) are:

$ER(\tau', d) = 0 + p + p^2 + p^3 + ... = -1 + 1/(1-p)$

$ER(\tau', \tau) = 3 + 3p + 3p^2 + 3p^3 + ... = 3/(1-p)$

$ER(\tau', \tau') = 3 + 3p + 3p^2 + 3p^3 + ... = 3/(1-p)$

$ER(\tau', \tau^{\bullet}) = 0 + 4p + 0p^2 + 4p^3 + ... = 4p/(1-p^2)$

Row 4: Finally, your expected returns from $\tau^{\bullet}$ (depending on your beliefs regarding your opponent) are:

$ER(\tau^{\bullet}, d) = 1 + 0p + p^2 + p^3 + ... = -p + 1/(1-p)$

$ER(\tau^{\bullet}, \tau) = 4 + 0p + 4p^2 + 0p^3 + ... = 4(1-p) + 4p/(1-p)$

$ER(\tau^{\bullet}, \tau') = 4 + 0p + 4p^2 + 0p^3 + ... = 4(1-p) + 4p/(1-p)$

$ER(\tau^{\bullet}, \tau^{\bullet}) = 1 + 3p + 3p^2 + 3p^3 + ... = -2 + 3/(1-p)$

[16] As argued in Box 2.5, in games of the Stag-Hunt variety (Game 6.2 being one of them) two commonly suggested principles for equilibrium selection, *efficiency* and *security*, can pull in opposite directions. In Game 6.2, strategies τ and τ' seem attractive because the Nash equilibria based on them offer the prospect of pay-off 6, when mutual defection promises a miserly 2. (This is the *efficiency principle* working in favour of the co-operative *Tit-for-Tat* type of Nash equilibrium). On the other hand, strategy *d* guarantees you a minimum expected pay-off of 2 if your opponent defects also and a much higher pay-off otherwise, whereas strategies τ and τ' can easily live you with expected pay-off 1. (The *security principle* favouring the mutual defection Nash equilibrium.)

Van Huyck et al (1990) designed an experiment to test which principle was the most powerful. In this experiment, subjects played a discrete version of the game described in Box 2.5. Players had to choose a whole number between 1 and 7 with individual pay-offs determined by the simple formula: $a\times$MIN - $b\times$OWN, where MIN was the smallest number chosen in the group of subjects and OWN was an individual's own choice. Clearly, there are seven Nash equilibria here: (everyone chooses 1), (everyone chooses 2),…,(everyone chooses 7). Efficiency would point to the selection of the (everyone chooses 7) equilibrium. However, if security was associated with the choice of the *maximin* action (see Chapter 2), then the (everyone chooses 1) equilibrium would be selected because the action which maximises the minimum outcome is the choice of '1' for each agent.

After each choice, the minimum number was announced and the subjects calculated their earnings. The choice was then repeated and so on. Subjects were also sometimes asked to predict the choices of the group playing the game. Interestingly (see the discussion on CAB in Chapter 2), the predictions for the first play were widely dispersed and would appear to be inconsistent with the *Aumann-Harsanyi doctrine* that players facing the same information will form the same prior probability assessment of how others will play. The experiment was repeated with several groups, some with slight variations to test particular hypotheses. In the first play of the game, neither principle seems to explain most people's actions, although *efficiency* did much better than *security* with, across all the groups, 31% choosing '7' and only 2% choosing '1'.

Although no group achieved perfect coordination on any integer in the first play, the striking result of repetition is that after 10 plays almost all the subjects in the seven versions of the experiment converged on the (everyone plays 1) equilibrium. Thus whereas *security* did not seem to be important in the initial play of the game, it became very important later on. There was, however, one version of the experiment where the number of subjects was reduced to two where efficiency held sway with quick convergence on the (both choose 7) equilibrium. So the choice of principle may be both sensitive to repetition and group size.

[17] The frequencies of the efficient equilibrium (e.g. three fishers on each island) were consistent with the hypothesis that fishers adopted the unique symmetrical mixed strategy Nash equilibrium (NEMS) in this game whereby each supplies each market with probability ½. However, if the failure to achieve perfect co-ordination in all plays is to be explained in this way, then the deviations in the price from its Walrasian level in each market should be serially uncorrelated. But this was not the case. In fact, price deviations were positively serially correlated and this is consistent with the way people switched markets much less than one would expect from people following NEMS. Thus it seems that the understanding of how people make strategic choices in such co-ordination games will have to go beyond an appeal to Nash and this could help explain the origins of what economists refer to as an *endogenous cycle in market prices*.

These results are similar to those found by Ochs (1990) in a related experiment. Like Meyer et al, Ochs also tested for whether the subjects used the history of play in the game as some kind of precedent for future play. In particular, once co-ordination is achieved, do the subjects 'stay put' in the current locations

and so achieve co-ordination in all future plays of the game? The answer is: No! Some do and some don't. Thus it seems, contrary to what one might expect from the *Harsanyi-Aumann* doctrine, that not all subjects draw the same inference about how to play the game in future from the same information regarding how it has been played in the past.

[18] Myerson expressed this view in an interview with one of the authors (Varoufakis) which is published in Greek (but which is also available on Myerson's website and from Varoufakis upon request). See Kottaridi and Siourounis (2002)

[19] See Binmore (1987/8), Pettit and Sugden (1989) and Kreps (1990) for dissenters.

[20] As opposed to studying reciprocal behaviour *after* the players' utility ordering is established.

[21] Or to put this slightly differently, playing *Tit-for-Tat* allows your opponent to develop a reputation for co-operation simply by playing co-operatively in the previous play of the game.

[22] There is, for instance, a famous neoliberal argument that to prevent the degradation of forests or lakes, the thing to do is to privatise them. In doing so, the free rider problem is not so much solved as annulled as the *N*-person public good or free rider game is replaced by the dictatorship of one person/firm who has an interest in either looking after the 'asset' or charging those who do care about it (e.g. local residents or environmentally conscious citizens) for its upkeep.

CHAPTER 6

EVOLUTIONARY GAMES

Evolution, Games and Social Theory

6.1 INTRODUCTION

- 6.1.1 The origins of Evolutionary Game Theory
- 6.1.2 Evolutionary stability and equilibrium: an overview.

6.2 SYMMETRICAL EVOLUTION IN HOMOGENEOUS POPULATIONS

- 6.2.1 Static games
- 6.2.1 Dynamic games

6.3 EVOLUTION IN HETEROGENEOUS POPULATIONS

- 6.3.1 Asymmetrical (or two-dimensional) evolution and the demise of Nash equilibria in mixed strategies (NEMS)
- 6.3.2 Does Evolutionary Game Theory apply to humans as well as it does to birds, ants, etc.? An experiment with 2-dimensional evolution in the *Hawk-Dove* game
- 6.3.3 Multi-dimensional evolution and the conflict of conventions
- 6.3.4 The origin of conventions and the challenge to methodological individualism
- 6.3.5 The politics of mutations: Conventions, inequality and revolt
- 6.3.6 Discriminatory Conventions: A brief synopsis

6.4 SOCIAL EVOLUTION: POWER, MORALITY AND HISTORY

- 6.4.1 Social versus natural selection
- 6.4.2 Conventions as covert social power
- 6.4.3 The evolution of predictions into moral beliefs: Hume on morality
- 6.4.4 Gender, class and functionalism
- 6.4.5 The evolution of predictions into ideology: Marx against morality

6.5 CONCLUSION

6
EVOLUTIONARY GAMES
Evolution, Games and Social Theory

6.1 INTRODUCTION

This chapter sets out the elements of Evolutionary Game Theory (EvGT hereafter). In particular, we focus on how EvGT may contribute the problem of *Indeterminacy* (i.e. of how an equilibrium is selected in the presence of multiple equilibria) and to debates in social science. One such debate concerns the role of the State and we have spent considerable time in the earlier chapters bringing out what game theory can contribute to this debate. In this chapter, we shift our attention to the discussion of power, history and functional explanation in social science.

6.1.1 The origins of Evolutionary Game Theory

Evolutionary ideas have a long history in economics (see Hodgson, 1989). For example, Darwin(1859), in his introduction to *Origin of the Species*, acknowledged the influence of classical political economy by referring explicitly to Malthus's theory of population growth. The struggle for existence was no more than an extension of Malthus' economic theory to "…the whole animal and vegetable kingdoms" (pp. 4-5).[1] Despite this and the encouragement given by Alfred Marshall in his *Principles of Economics* that "The Mecca of the economist lies in economic biology rather than mechanical economic dynamics…" (1890,1961; xiv), evolutionary ideas remained on the margins of the discipline (see Box 6.1) until they were imported into game theory in the 1980s and 1990s.

Box 6.1

EVOLUTIONARY ECONOMIC THINKING IN 20TH CENTURY ECONOMICS

Joseph Schumpeter (1883-1950) is perhaps the best-known economist whose work turned on the use of evolutionary ideas. He still inspires scholars and politicians who believe in the power of capitalism to rejuvenate itself and forge progress, not in spite of the large corporations' monopoly power, but *because* of it. In Schumpeter's words: "The essential point to grasp is that in dealing with capitalism we are dealing with an evolutionary process." He dismissed the view that capitalism is best based on small family-owned firms competing feverishly against one another. Instead, Schumpeter argued on the basis of a distinctly evolutionary narrative (eloquently summed up by his phrase "the perennial gales of creative destruction") of how certain companies grow larger and domineering on the back of some innovation but then die out (becoming 'extinct'); victims of some newfangled, smaller competitor who gained the evolutionary upper hand. In this *Struggle for Existence*, monopoly power spawns creativity because of the 'transience' caused by inexorable evolutionary pressures (see Schumpeter, 1946). Turning to the leftwing critics of late capitalism, Paul Sweezy (who was Schumpeter's student) applied the evolutionary method to the opposite effect; in order, that is, to demonstrate how capitalism's tendency to generate monopolies intensified its propensity toward stagnation and crises (see Sweezy, 1942). In Sweezy (1972) we find a fully-fledged evolutionary explanation of the political economy of US suburbia.[2]

Evolutionary arguments have also periodically occurred in otherwise mainstream theory. For example Milton Friedman (1953) famously defended the assumption of profit maximisation through an appeal to the process of natural selection which must have 'weeded out' of markets those firms which failed to maximise profit. Even Nash seems to have foreshadowed the need for a genuine evolutionary approach; at least as a means of verifying the plausibility of his equilibrium concept. In his PhD thesis Nash included a brief but influential note suggesting a so-called 'population-statistical interpretation' of Nash equilibrium. The idea was that we posit players (drawn from a large population) who interact repeatedly (against a different opponent each time) without assuming that they "…have full knowledge of the total structure of the game, or the ability and inclination to go through any complex reasoning process" (Nash 1950b, p.21). And if this analysis shows that these less-than-rational players converge to Nash equilibrium, the latter's predictive value will have been verified. Lastly, in more recent years, the evolutionary approach to industrial organisation has been championed by Nelson and Winter, 1974).

Game theory took this evolutionary turn as a direct response to its encounter with *Indeterminacy* that proved crucial in earlier chapters. In a recent survey of game theory, from an evolutionary perspective, Larry Samuelson (2002) expressed the predicament in the following way.

> Much of the difficulty in interpreting the contending equilibrium refinements of the 1980s appeared because the models were divorced from their context of application in an attempt to rely on nothing other than rationality. The resulting models contain insufficient information about the underlying strategic interaction to answer the relevant questions, at least if game theory is intended to model real interactions rather than to ponder philosophical points. (p.57)

So how can evolutionary biology help game theory out of its predicament? A good place to start is Darwin's (1859) own description of his central idea: "... it struck me that favourable variations would tend to be preserved and unfavourable ones to be destroyed." In a game theoretical context it is natural to think of 'variations' as strategies (varieties of behaviour) with some surviving and some disappearing depending on their 'success'. The idea, then, is that this process of weaning strategies might help explain how an equilibrium is selected. This marks two, related, differences in approaching strategic interaction.

First, survival and extinction happen in historical, or real, time. By contrast, Nash conceived of his equilibrium as a static notion addressing the question of consistency between actions and beliefs (recall Chapter 2, although see Box 6.1 for the hint of an evolutionary approach in his PhD thesis). Later, his epigones created an equilibrium concept for dynamic interactions, but along identical lines. As we saw in Chapter 3, a dynamic game's SPNE is no more than a series of actions that are consistent with one another *and* with the beliefs that support them, *once backward induction is combined with common knowledge of rationality* (CKR). The passage of actual time does not help agents decide how to behave in these dynamic games in any deep sense; instead it merely complicates the problem that rational agents are set at the start of their interaction. A plan of action is decided on at the beginning (which can involve mixed strategies) and this covers all future moves. This is quite unlike how we understand real historical time where the passage of time can make fundamental differences in how we perceive problems and how we act.

Secondly, the sense of rational individual agency is weakened. People are no longer assumed to be able to think through the logical entailments of CKR and CAB. Rather in so far as they are rational, they adopt strategies on the basis of trial and error, adapting their behaviour on the basis of its 'success' with the result that they gravitate towards the relatively most successful type of behaviour. One interpretation here is that EvGT adopts a bounded sense of rationality which fits directly with the historical, evolutionary approach to dynamic interactions as 'learning' occurs with the experience of a new stage in a dynamic interaction (see Box 6.2). Another interpretation is that EvGT posits no theory of rationality at all. Indeed, EvGT is not in the least concerned with the evolution of human reasoning (unlike conventional game theory which strives to model the latter explicitly). Instead, it focuses exclusively on the evolution of *phenotypes* (or behaviour) by positing either that 'parents' transmit their strategy to their offspring (and those with successful strategies tend to have more children resulting in the proliferation of the successful strategy within the population) or that agents simply mimic behaviours which are relatively successful.

Box 6.2

KARL POPPER ON EVOLUTIONARY KNOWLEDGE

"The theory of knowledge which I wish to propose is a largely Darwinian theory of the growth of knowledge. From the amoeba to Einstein, the growth of knowledge is always the same: we try to solve our problems, and to obtain, by a process of elimination, something approaching adequacy in our tentative solutions… the growth of our knowledge is the result of a process closely resembling what Darwin called 'natural selection'; that is, the natural selection of hypotheses…" (p. 261) "I thus submit a variation of Darwinism in which behavioural monsters play a decisive part… (p.283). Our aim must be to make our successive mistakes as quickly as possible. To speed up evolution." [Popper, 1979]

6.1.2 Evolutionary stability and equilibrium: An introduction

To introduce this approach in a little more detail we shall consider how it is applied to the play of Game 2.14, the *Hawk-Dove* game, reproduced here for convenience. Chapter 2 analysed this game as a static interaction and concluded that there are three Nash equilibria: two pure strategy Nash equilibria, *hd*, *dh*, and one Nash equilibrium in mixed strategies (NEMS) [according to which players choose strategy *h* with probability $p = \Pr(h) = 1/3$].

	h	*d*
h	-2,-2	$^{+}2,0^{-}$
d	$^{+}0,2^{-}$	1,1

Game 2.14 – *Hawk-Dove*

If this game is turned into a dynamic one (that is, it is repeated indefinitely between the same two rational players) all sorts of other Nash equilibria emerge (provided the probability of repetition is high enough – recall the *Folk Theorem* in Chapter 5).

In an *evolutionary* version, this game is not played repeatedly by the same two players. The game is played repeatedly but the players are drawn from a large population and play (anonymously) against fresh opponents each time. In this manner, even agents who are smart enough to think strategically do not worry about their reputation and do not try to influence their next opponent's behaviour through current moves. Thus it is like a series of one-shot plays of the game between players who are, somehow, *programmed* to choose a strategy at any point in time such that, and this is the rub, the likelihood of choosing a strategy depends on its pay-offs *relative to average pay-offs in the population* (recall Darwin's central idea).

This is because the relative size of a strategy's pay-offs is linked to the probability of survival within this population, as sketched above, through the mechanism of either individual learning or the rate of individual reproduction. So we make no specific assumption about what goes on in people's minds save that whatever is going on, it has the effect of encouraging the proliferation of relatively successful strategies.

In particular let us assume that *Hawk-Dove* above the whole population is initially programmed to play strategy *d* in the game. In each interaction all players retreat (i.e. play dovish strategy *d*) and each receives pay-off 1 every time. Suppose now that, for some unspecified reason, one player switches to strategy *h*. This 'switch' could be a 'tremble' (i.e. an 'error' in that player's programming, similar to the trembles studied in Chapter 3) or (in the language of biologists) a 'mutation' (that is, the rare birth of a 'hawk' to a dovish parent). Alternatively we may think of it as an experiment performed by an inquisitive 'dove'.

Whatever the reason, a lone player selecting strategy *h* in a population of 'doves' will collect pay-off 2 in each interaction (with her opponent collecting 0). If this relative 'success' translates into relatively more offspring to our 'mutant hawk', or if other doves mimic the mutant's relatively successful strategy and turn into hawks, the proportion *p* of *h*-playing agents in the population will grow. In this sense, evolutionary biologists tell us, an homogenous population of *d*-players is susceptible to an *invasion* of *h*-playing mutants. Outcome *dd* is, consequently, evolutionarily *un*stable.

The same applies to outcome *hh*. For if all players are initially programmed to play *h*, the cumulative pay-offs of a *d*-playing 'mutant' will be higher than the norm. Thus, generalised *h*-playing (*d*-playing) cannot survive evolutionary pressure in an homogeneous population as proportion *p* falls (rises) following the birth of a mutant dove (hawk). In short, if *p* is too high, evolution will force it (via mutations and copying of relatively successful strategies) to fall while if *p* is too low it will tend to rise. When will it stabilise? The answer is when *p* equals exactly 1/3,

which coincides with the Nash equilibrium in mixed strategies (NEMS): that is, in any interaction there is a probability of 1/3 that each person will play *h*.[3]

The above result is remarkable and helps explain the excitement caused by EvGT. Recall how in Chapters 2 (Section 2.6) and 3 (Section 3.2.2) we struggled to find a convincing justification for NEMS. Yet in an evolutionary context NEMS emerges naturally as an *evolutionary equilibrium* (EE) of the game. Not only this, but NEMS becomes the only plausible equilibrium in an homogenous population. Thus in one short step, the problem of rationalising NEMS is solved and so is the problem of equilibrium selection (as two of the three Nash equilibria were culled)!

Box 6.3

POPULATION HOMOGENEITY ERADICATES THE PURE NASH EQUILIBRIA OF HAWK-DOVE GAMES

A population is homogeneous if players cannot distinguish one opponent from another. In this context, each player acts independently of who she is playing against (since they all 'look' the same) and, therefore, it is not possible to sustain an equilibrium whereby Jill conditions her behaviour by playing, for example, *h* against Jack and *d* against Judy. Jill is unable to tell whether she is playing against Judy or Jack and therefore plays either one pure strategy always (*h* or *d*) or a mixed strategy *consistently* (i.e. chooses *h* with probability *p* and *d* with 1-*p*). In these circumstances, evolution turns on boosting or diminishing the proportion *p* of players who opt for *h*. Note how this rules out Nash equilibria *hd* and *dh*: Since, in an evolutionary equilibrium, proportion of players *p* will be playing *h* without being able to condition their behaviour on some distinguishing mark of their opponent, it is impossible for pairs of players to co-ordinate their actions so that one plays *h* and the other *d* with certainty – it is *as if* each will play *h* with probability *p*.

Having established that population homogeneity means that each will play *h* as if with probability *p*, it is easy to show that the evolutionary process will rest only when *p* stabilises at the NEMS value of 1/3. For if $p<1/3$, the expected returns from *h* exceed those from *d* (and thus the proportion of *h*-players *p* will rise) and vice versa (see note 8). Evolutionary equilibrium is thus achieved at $p=1/3$; the repercussion being that, on average, outcomes *hh*, *dd*, *hd*, and *dh* will be observed with (NEMS) probabilities 1/9,4/9, 2/9 and 2/9 respectively. [None of the above applies in heterogeneous populations. As we shall see, the possibility of conditioning behaviour on observed characteristics undermines NEMS and renders it evolutionarily unstable.]

The study of other games confirms the capacity of EvGT usefully to narrow down the set of Nash equilibria, even if this does not mean that there is always a unique *evolutionary equilibrium* (EE). For instance, consider Game 2.16, the *Stag-Hunt* game (reproduced below for convenience)

	h	*s*
h	$^{+}1,1^{-}$	2,0
s	0,2	$^{+}3,3^{-}$

Game 2.16 – *The Stag Hunt*

Applying the same logic as in the case of *Hawk-Dove*, a different result ensues: NEMS is evolutionarily *un*stable and thus *not* an EE. By contrast, both of the pure strategy Nash equilibria emerge as potential EE. To see this, let us consider an evolutionary version of the *Stag-Hunt*. In each round two players are paired off with a fresh colleague and set off for the hunt. Each can either chase hare (strategy *h*) or hunt the stag (*s*) [revisit Box 2.5 for Rousseau's original narrative]. Suppose now that the population is programmed according to NEMS: proportion $p=½$ hunt hare with the rest opting for the stag. At this stage, the pay-offs from the two strategies are (by definition of NEMS) identical.[4]

However, if there is a mutation, and one individual unexpectedly switches from hunting the stag to hunting hare, then *p* rises slightly but sufficiently to destabilise NEMS: the moment *p* exceeds ½, the average pay-offs from *h* exceed those from *s*. Given the assumption that a strategy's relative success translates into it being copied, more and more players will opt for *h*. And as this bandwagon gathers pace, switching from *s* to *h* benefits *s*-playing agents increasingly. The only limit to these rises is $p=1$ (i.e. Rousseau's nightmare in which the collective stag hunt fails). Of course, not all is necessarily lost. For if the initial mutation goes the other way (i.e. a player

Box 6.4

NATURAL SELECTION DOES NOT MEAN THE SURVIVAL OF THE FITTEST

It is a common fallacy to think that Darwinian evolution favours the 'fittest' and discriminates against inefficient practices. In 1930 R. Fischer put forward the so-called *Fundamental Theorem of Natural Selection* according to which (under special conditions) evolution boosts a population's average fitness. However, by the 1960s (see Moran 1964) it had become clear that most evolutionary processes violate Fisher's theorem. Darwin had, in fact, renounced the term 'survival of the fittest' (a term not coined by him in the first place) and evolutionists nowadays reject the term outright.

In the context of the *Stag-Hunt* above, we saw that there are two evolutionary equilibria: An efficient EE (in which all players hunt stags) and an inefficient EE (in which players chase after hare). Evolution may give rise to either one of the two depending on the initial conditions and the type of mutations that occur. There is no evolutionary reason why the efficient social outcome will dominate.[5] Stephen J. Gould has popularised this point in the context of biological processes, arguing against simplistic Panglossian adaptationism. For example, he warns against the tendency to presume that we humans have held on to our eyebrows (after having lost most of our body hair through the aeons) because they "must have been good things". Gould (1980,1981) has shown that observed traits (like our eyebrows) are the result of complex processes which have little to do with their current function and discernible advantages. The next box gives some similar examples from economics.

Box 6.5

QWERTY AND OTHER COORDINATION FAILURES

Have you ever wondered how the QWERTY arrangement of keys emerged as the standard for keyboards (and typewriters)? (It is called QWERTY because these are the first six letters on the top row.) One might think that it represents some optimal arrangement for typing fast and efficiently; otherwise it would not have stood the test of time. However, the QWERTY system is regularly outperformed in speed and time trials by another arrangement, the Dvorak and Dealey keyboard.

In some sense, it is hardly surprising that the QWERTY arrangement is not very efficient because the basic four-row configuration was designed to overcome some early technical problems afflicting mechanical typewriters. Moreover, the location of the letters across these rows was devised in part as a sales gimmick to promote the sales of a Remington machine called a Type Writer (notice how the brand's name could be typed quickly by the sales representative using only the top row of letters, thus allowing for an effortless demonstration). What is surprising is that people continue to use it today, despite its failings, to operate state of the art computers. So why do people stick with it? Well, plausibly because it remains the best course of action so long as everyone is still producing and buying QWERTY machines and acquiring the skills to use those machines. After all, it pays employers to buy these machines when everyone else uses them because the employer thereby gains access to a large pool of workers with the skills to use these machines; and likewise it pays workers to acquire the skills to operate these machines because this secures access to most job opportunities when employers have bought these machines. Thus, the selection of a key configuration has all the hallmarks of an evolutionary equilibrium like *hh* in the stag-hunt game: as long as this behavioural code has evolved, it pays to follow it, even if there is a better behavioural code for all to adopt.

The occurrence of Keynesian unemployment is also often interpreted as an inefficient equilibrium in some type of large-scale stag-hunt (or co-ordination) problem - see also Box 2.5, Cooper and Johns, 1988, and Hargreaves Heap, 1992. It arises when nominal prices (and wages) fail to adjust immediately to nominal demand shocks (that is, shocks, like changes in the money supply, which, in principle, require changes in all nominal prices to preserve the original equilibrium position). It can be regarded as an evolutionary equilibrium similar to *hh* in Game 2.18 because the selection of a price is like playing a stag-hunt game. Under certain circumstances, no change is the best response when all other prices remain constant, while a change is the best response when all other prices are changed. Thus there are (at least) two evolutionary-Nash equilibria and one is better than the other. The 'all change' outcome (equivalent to *ss* in Game 2.18) dominates 'no change' (equivalent to *hh*) when there is a deflationary nominal demand shock as the latter generates an increase in unemployment, yet it seems that economic agents often fail to select them. EvGT can explain this as a case of a mutually disadvantageous behaviour which is, nonetheless, in an evolutionary equilibrium.

Another illustration of how recessions and booms in production can be interpreted in the context of EvGT comes from Diamond's (1982) trading game. In a simple version, people produce either a high or a low amount of the commodity in which they specialise and they take it to a market to exchange with goods produced by other people. If the opportunities are restricted because everyone else has chosen low production levels, then trading opportunities are poor (making low production the best option). Conversely, if others choose high production levels the trading opportunities are good (making high production the best strategy). Thus each will tend to produce high levels when everyone else does likewise, and vice versa. In other words, we have two evolutionary-Nash equilibria, one associated with a boom and the other with a recession. Which of the two will emerge depends on the evolutionary process.

programmed to play *h* gives an unexpected birth to an *s*-playing offspring), the bandwagon will roll toward the second EE: p=0 (i.e. everyone hunting the stag).

Although the evolutionary approach yields two EE in the *Stag-Hunt*, as opposed to the unique EE in the *Hawk-Dove*, it is still comforting that at least one Nash equilibrium (i.e. NEMS) has been ruled out. It is also gratifying to encounter the theory's capacity to discriminate between cases. Rather than making blanket pronouncements on the plausibility of, say, NEMS, EvGT takes a different view depending on context. In the *Hawk-Dove*, it deemed that NEMS is not only a legitimate equilibrium but that it is the unique EE as well. On the other hand, in the *Stag-Hunt* it is neither.

In both these cases the EEs are Nash equilibria. This is always the case. The reason is simple. For behaviour (or phenotype) *b* to be consistent with an EE, it must be immune to an invasion by some mutant behavioural code, say, β. If β enters the population (via some mutation) and is a better response to *b* than *b* is to itself, then *b* is doomed. Thus, to be 'invasion-proof' (and thus evolutionarily stable) incumbent *b* must be *at least* as good a reply to itself as mutant β is to *b*. Of course, if *b* is strictly a best reply to itself, then no mutant can dethrone it. In summary, a necessary condition for behaviour *b* to be evolutionarily stable (and thus an EE) is that it is no worse a reply to itself than any potential invader. Notice that this is no more than the standard definition of Nash's equilibrium.

The ability of the evolutionary approach to weed out some Nash equilibria stems from the fact that evolutionary stability demands *more* of behavioural codes than Nash does. Behaviour *b* is in a Nash equilibrium as long as there is no other behaviour that does better against it. This requirement, though a necessary condition, is insufficient to qualify *b* as an EE. The additional requirement is that *b* is either a *best* reply to itself or that it is a better reply to any mutant β than β is to itself. To see why evolutionary biologists introduced this second condition, recall that *b* can only be in EE if it can be relied upon to resist mutant invasions. Suppose that β enters and *b* performs just as well against β as β does against itself. Granted that the mutant code β fails to gain an evolutionary upper hand when pitted against *b*, there is no reason to presume that it will die out. For this to happen (and for *b* to eclipse β) it must also be true that, following a β-invasion, the incumbent code (*b*) must gain the evolutionary advantage when pitted against the mutant β; that is, *b* must be a better reply to β than the latter is to itself – see Case 2, Condition B in the definition below.

Evolutionary Stability/Equilibrium (Definition)

Behavioral code *b* is an *evolutionarily stable strategy* (ESS) and thus consistent with an *evolutionary equilibrium* (EE) in either of the two cases below:

Case 1: *b* is strictly a best reply to *b* (i.e. *b* is in a strict Nash equilibrium),[6] i.e. $ER(b,b)>ER(\beta,b)$ for all potential mutant strategies β

Case 2: *Condition A – b* is as good a reply to *b* as any potential mutant β is to *b* (i.e. *b* is in a non-strict Nash equilibrium), i.e. $ER(b,b)=ER(\beta,b)$

AND

Condition B – b is a better reply to any mutant β than β is to itself or $ER(b,\beta)>ER(\beta,\beta)$

[Note that in homogeneous populations a strict Nash equilibrium is also an EE only in doubly symmetrical games.[7]]

Let us revisit the above examples in the context of these formal definitions of ESS and EE. We saw that in the *Stag-Hunt* (Game 2.16), both Nash equilibria *ss* and *hh* are EE. The reason is that the pure strategies supporting them are both best replies to each other and, in equilibrium, they defeat mutants. Starting at *ss* (i.e. a situation in which all hunters hunt the stag), it is clear that a mutation

turning a player into hare-hunter (or *h*-player) will not benefit the latter. Strategy *s* remains a best reply to itself [*Case 1*] as long as everyone else is playing it. Suppose now that a mutation does occur and a non-mutant's opponent is drawn out of a population that includes a few hare-hunting mutants. Will our non-mutants benefit in terms of average pay-offs (or evolutionary fitness points) by mutating also into hare-hunters?

As long as the number of mutants is small (i.e. *p* remains close to zero), the answer is negative and the mutation will die out since it will fail to spread in the population.[8] Thus, *Condition B* (see *Case 2*) applies since, not only is strategy *s* a best reply to itself, but (given that proportion of mutants *p* is close to zero) *s* is a best reply to the mutant *h* as well. In this manner, we argue that the mutant *h* strategy has no future (it will become extinct in the biologists' language) in a population of stag-hunters. By the same logic, a stag-hunting mutant will die out in a population of hare-hunters. When all play *h*, an invading stag-hunting mutant will net 0 utils every time it interacts with non-mutants compared to the 1 util she would have received otherwise.

In the *Stag-Hunt*, it is also clear that ESS and EE do eliminate the mixed strategy Nash equilibrium (NEMS). To see this, suppose that we begin with a population perched on the NEMS: pairs of hunters are repeatedly drawn from the population and, in each pair, proportion $p = ½$ of the individuals (making up each pair) chase hare while 1-*p* go for the stag.[9] It is easy to establish that *Case 1* does *not* hold. Under NEMS, $p = ½$ and there is no benefit from following any strategy as opposed to any other. The best we can say is that, though NEMS is not a better reply to NEMS than any mutant, it is at least no worse. This means that, since *Case 1* does not apply, NEMS can only be an EE in an evolutionary version of the *Stag-Hunt* if *Case 2* does. Does it?

If a mutant appears and confronts NEMS, a hare-hunter turns stag-hunter and, therefore, proportion *p* changes. Even though this change is tiny (as long as the population is large), it triggers the following evolutionary process that leads NEMS to its 'extinction': As *p* drops slightly below ½, the mutant stag-hunter gains more on average than the non-mutants. By the Darwinian hypothesis underpinning evolutionary stability analysis, this causes more mutations of hare-hunters into stag-hunters, it pushes *p* further down and, in turn, this increases further the evolutionary advantage of being a mutant stag-hunter. The resting point of this process is one in which everyone hunts stags ($p=0$). Of course, had the initial mutation been one that turned a stag-hunter into a hare-hunter, NEMS would again be destabilised only this time evolution would lead all hunters to chase hares.

6.2 SYMMETRICAL EVOLUTION IN HOMOGENEOUS POPULATIONS

6.2.1 Static games

In this section we subject some of the classic games that we first encountered in Chapter 2 to the evolutionary 'treatment'. These are: *The Battle of the Sexes* (Game 2.13), *Hawk-Dove* (Game 2.14), *Pure Co-ordination* (Game 2.15), *The Stag-Hunt* (Game 2.16), *Hide-and-Seek* (Game 2.17), and the *Prisoner's Dilemma* (Game 2.18). Of these, we have already analysed the *Hawk-Dove* and the *Stag-Hunt* interactions (see Section 6.1.2) under the assumption of an homogeneous population. We extend this analysis to all six classic games here. Before proceeding, it is important to state the conditions under which the present analysis holds more precisely:

(A) Even though the population (*N*) is assumed to be very large (so as to ensure that no player tries to influence the rest), our analysis applies only when each meeting involves precisely two players.
(B) The games are symmetrical.
(C) The population is continuous, as opposed to discrete.[10]

(D) Only one mutation is allowed to occur at any one moment. When simultaneous mutations are possible, the ensuing analysis no longer holds.

(E) An homogenous population

Of the five conditions, the first three will be assumed to hold *throughout the chapter*. Condition (E) will be relaxed in Section 6.3 and the repercussions of relaxing Condition (A) will be discussed in Section 6.4.1.

With these conditions in place, we can now define the two mechanisms making up any evolutionary process.

- A *mutation mechanism* which generates variety and thus ensures that, at any point in time, even the most successful behavioural code (or strategy) will be challenged by an invader
- A *selection mechanism* which determines which variety will gain the upper hand at every point in time

The notions of *evolutionary stability* (ESS) and *evolutionary equilibrium* (EE) defined in Section 6.1.2 relate to the *mutation mechanism*. It was the idea of mutations which led us to the conclusion that (as long as the five conditions above hold) NEMS is the unique EE in the *Hawk-Dove* but that it is not an EE in the *Stag-Hunt*. Indeed, we could use the same logic to arrive at the EE of all of the 2X2 games under investigation (Games 2.13 to 2.18).[11]

Dawkins (1976) coined the term *replicator* to signify entities (e.g. genes, strategies) which have the capacity to copy themselves and therefore spread within a population. Schuster and Sigmund (1983) later introduced the term *replicator dynamics* to refer to the *selection mechanism* determining which of the potential *replicators* will grow at the expense of its 'competitors'. In our analysis, the *replicators* are the game's pure strategies which are assumed to be copied (from parent to child or from originator to imitator) without error. The rate of growth of a replicator-strategy is assumed to be proportional to the difference between (a) the replicator-strategy's current pay-off, and (b) the current average pay-off in the population. That is, a strategy that does better than average spreads in the population at a speed that is proportional to its current relative success.

In the five 2X2 games examined below, the *replicator dynamics* can be modelled straightforwardly. Consider a general 2X2 symmetrical game of the form:

	C1	*C2*
R1	a_1, a_1	$a_2, a_3,$
R2	a_3, a_2	$a_4, a_4,$

Game 6.1 – *General Symmetrical 2X2 Game*

Suppose p is the probability that row player R will select pure strategy R1. Since the population is assumed homogeneous (that is, no player bears any distinguishing mark) and the pay-offs are symmetrical, it ought to make no difference whether a player chooses between the row or between the columns. Thus, we assume that, if p=Pr(R1) and q=Pr(C1), then $p=q$ (i.e. it is as if all players choose between their first and their second strategies with probabilities p and 1-p respectively). Players who choose their first strategy here will collect average pay-offs equal to ER(R1) [=ER(C1)=] = $pa_1 + (1-p)a_2$. Players who choose their second strategy will collect, on average, pay-offs ER(R2) [=ER(C1)=]= $pa_3 + (1-p)a_4$. The *replicator dynamics* favour the relatively more successful strategy and so we compute the net gain from selecting one's first, as opposed to one's second, strategy.

$$d = \text{ER(R1)} - \text{ER(R2)} = pa_1 + (1-p)a_2 - [pa_3 + (1-p)a_4] = A + Bp \text{ ---------- (6.1)}$$

$$\text{where } A = (a_2 - a_4) \text{ and } B = (a_1 + a_4) - (a_2 + a_3)$$

Note that d is positive when the first strategy is relatively more successful than the second. Thus a simple replicator dynamic can be based on the idea that, whenever $d>0$, the first strategy spreads at the expense of the second and, therefore, that p (which is the probability/frequency of the first strategy) rises. In summary, we have our first replicator dynamic:[12]

Replicator Dynamic:
$d>0$ means that p increases (or p rises whenever $p>-A/B$)
$d<0$ means that p declines (or p falls whenever $p<-A/B$)
$d=0$ means that p is stationary (or p stays still whenever $p=-A/B$)

From (6.1) (the definition of d) we can discern the interesting feedback effects between a strategy's probability (e.g. p) and its relative success (e.g. d). Suppose that the population is behaving in a manner that renders the two strategies equally successful (i.e. $d=0$). Is this an evolutionary equilibrium (EE)? It depends. In games with a positive (negative) parameter B, an increase (decrease) in the frequency of the first strategy (p) will boost (diminish) its relative success (d). This will, in turn, cause the strategy to spread (shrink) as p rises (falls), thus enhancing further its relative success (failure) – due to the further rise (fall) of d.

Game	***A***	***B***	***NEMS***	***EE***
Battle-of-Sexes (Game 2.13)	3	-4	$p=3/4$	$p=3/4$
Hawk-Dove (Game 2.14)	1	-3	$p=1/3$	$p=1/3$
Pure Co-ordination (Game 2.15)	-3	4	$p=3/4$	$p=1$ and 0
Stag-Hunt (Game 2.16)	-1	2	$p=1/2$	$p=1$ and 0

Figure 6.1

To illustrate, consider two antagonistic games: *Battle-of-the-Sexes* and *Hawk-Dove*. In both games, parameters A and B are positive and negative respectively. A graph for these two games (see Figure 6.2) of function d (plotted against probability p) thus begins in the positive quarter, declines toward the horizontal axis, crosses it at value $p=p^*$ and from then onwards enters the area in which d is negative.

Note that the threshold value $p=p^*$ (beyond which d turns from positive to negative) is, by definition, the game's NEMS (Nash equilibrium in mixed strategies).[13] If we begin at any value of $p<p^*$ the value of function d is positive; i.e. the first strategy is relatively more successful than the second. Thus it spreads and p rises. (Notice the arrows on the p-axis pointing rightwards.) And if $p>p^*$ then d is negative which means that the first strategy is relatively less successful than the second. In that case, the first strategy recedes and p declines (note the leftward arrows). Clearly, evolution points to the NEMS value of p^* at which one third of the population plays their first strategy at any point in time. We make a note of the coincidence of EE with NEMS in Figure 6.1

Things look quite different in the next two games. The *Pure Co-ordination* (Game 2.15) and the *Stag-Hunt* (Game 2.16). Here, see Figure 6.2, function d begins from negative d-values (as intercept parameter A is negative, see Figure 6.1), then rises (as slope parameter B is positive), cuts the p-axis from below at the NEMS value $p=p^*$ and continues to rise with p. As $d<0$ for values of p below p^*, the first strategy will become extinct (sooner or later) if, initially, the proportion of players who are programmed to play it lies below p^*. And vice versa. If at the outset a proportion of players greater than p^* are programmed to play their first strategy, then the first strategy will dominate fully. It is for this reason that the last column of Figure 6.1 reports that, in Games 2.15 and 2.16, the EE differ from NEMS and lie at the extremities of the probability distribution: either all players choose one strategy ($p=0$) or the other ($p=1$). It all hinges on the initial conditions and on whether the initial p was less than, or greater to, its NEMS value.

An Evolutionary Equilibrium's Catchment Area (Definition)

Consider probability interval $\pi=(p',p'')$ such that when the frequency of p falls within π, then p moves monotonically (that is, without turning back) toward the value of $p=p^{\varepsilon}$ which is consistent with evolutionary equilibrium ε. We call π the *catchment area* of evolutionary equilibrium ε. In effect, any game which begins from within the *catchment area* of some EE will converge to that EE.

Antagonistic Games 2.13 and 2.14 feature a unique EE and, therefore, the whole range of possible p-values (ranging from 0 o 1) is that EE's *catchment area*. We can see this clearly in the relevant diagrams in Figure 6.2 since all arrows (whichever p we start from) point to the unique EE/NEMS. A unique EE means, in other words, a single, all-encompassing catchment area. In contrast, co-ordination games feature two EE. Although NEMS is not an EE in these games, it still plays an analytical role. It demarcates the two EEs catchment areas: the catchment area of EE p=0 lies on the left hand side of NEMS and the catchment area of EE p=1 lies on its right.

Another game in which there exists a single equilibrium is the *Prisoner's Dilemma* (i.e. Game 2.18). Figure 6.2 confirms the irresistible power of defection from the mutually advantageous (co-operative) outcome. In terms of the analysis here, function d is constantly positive and independent of p. The fact that the net gains from defection (i.e. function d) never fluctuates is a reflection of defection's strict dominance: It is the best reply whatever the frequency of co-operative moves in the population. This means that, even a population oblivious to the fact that co-operation is a strictly dominated strategy, will sooner or later be drawn to the superior net returns of defection and abandon co-operation. As this happens, p (the proportion who defect) tends to 1. Again, there is a unique EE and the whole p-zone is its catchment area.

Game 2.13 *Battle of the Sexes*

ER(1st strategy) = 3(1-*p*); ER(2nd strategy) = *p*

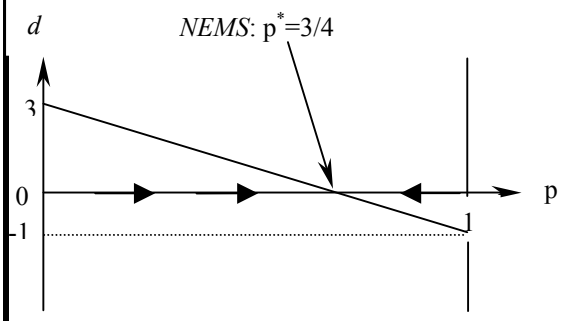
d = ER(1st strategy) - ER(2nd strategy) = 3-4*p*

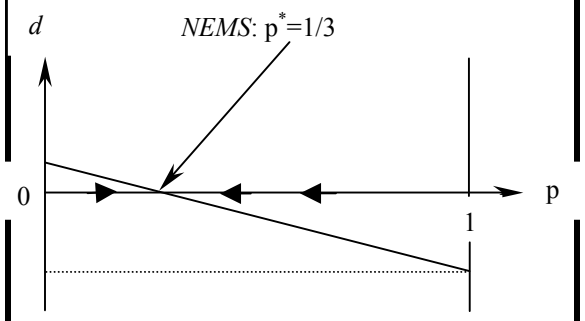
Game 2.14 *Hawk-Dove or Chicken*

ER(1st strategy) = -2*p*; ER(2nd strategy) = 2(1-*p*)

d = ER(1st strategy) - ER(2nd strategy) = 1-3*p*

In both games, when *d=0, p* equals its NEMS value (*p**) which is also the unique EE (see arrows)





Game 2.15 *Pure Coordination*

ER(1st strategy) = *p*; ER(2nd strategy) = 3(1-*p*)

d = ER(1st strategy) - ER(2nd strategy) = -3+4*p*

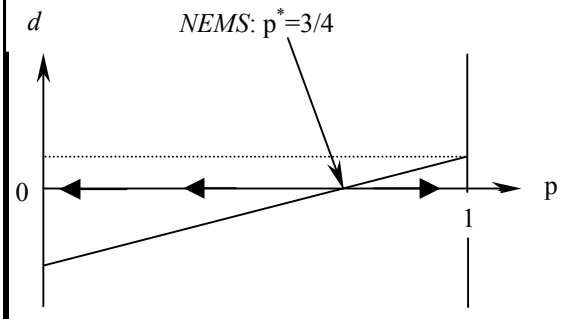
Game 2.16 *Stag Hunt/Common Assurance*

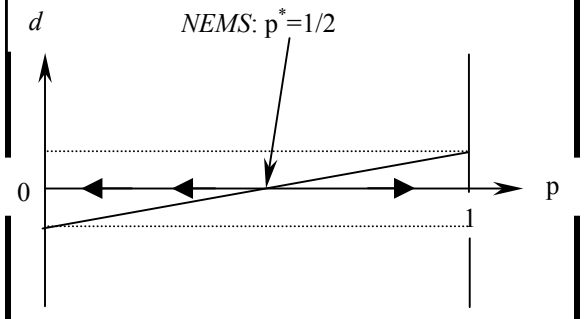
ER(1st strategy) = p+2(1-*p*)

ER(2nd strategy) = 3(1-*p*)

d = ER(1st strategy) - ER(2nd strategy) = -1+2*p*

In both games, when *d=0, p* equals its NEMS value (*p**) which ***is not*** an EE; there are two EE here, corresponding to the pure strategy Nash equilibria at *p*=0 and *p*=1 (see arrows)





Game 2.17 *Hide and Seek*

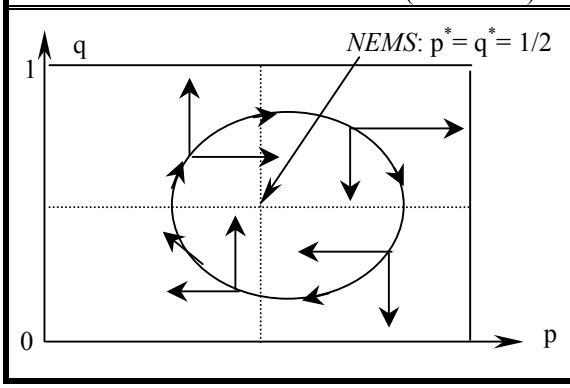
ER(C1) = (1-*p*) and ER(R1) = *q*

ER(C2) = *p* and ER(R2) = (1-*q*)

d^C = ER(C1)-ER(C2) = 1-2*p*

and d^R = -1 + 2*q*

When $d^C_{12} = d^R_{12} = 0$, *p* and *q* equal ½, their NEMS value. There exists no EE (see arrows)



Game 2.18 *Prisoner's Dilemma*

ER(1st strategy) = *p*+4(1-*p*)

ER(2nd strategy) = 3(1-*p*)

d = ER(1st strategy) - ER(2nd strategy) = 1

d>0 for all *p*. The EE coincides with the unique pure strategy Nash equilibrium: *p*=1

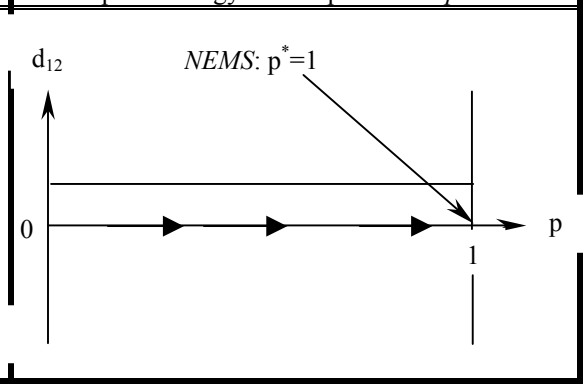


Figure 6.2 – *Replicator dynamics, evolutionary equilibria and NEMS in the classic games of Chapter 2*

6.2.1 Dynamic games

How should we envisage the evolutionary play of a dynamic game such as the repeated prisoner's dilemma? In the case of the static version, it was quite straightforward. Players were matched in each round randomly (and anonymously), played the prisoners dilemma, and then were re-matched with a fresh opponent. In the previous section we saw how the unique evolutionary equilibrium (EE) of the static prisoner's dilemma coincided with the (non-co-operative) Nash equilibrium of the conventional static game. The repeated version requires a slightly richer imagination.

In the evolutionary play of the repeated prisoner's dilemma a player is matched against an opponent, plays the game a number of times (i.e. the game is repeated once more with probability p – see Section 5.5, Chapter 5) and then, when these rounds come to an end, she is re-matched with another opponent against whom she plays another repeated prisoner's dilemma. Similarly with all dynamic (and repeated) games. In the case of, say, the *Centipede* (see Game 3.4, Chapter 3), a player plays the complete sequence of the game against one opponent before being matched against another, and then another *ad infinitum*.

To give a flavour of how the evolutionary perspective differs from that of the conventional theory of dynamic games, we shall subject the indefinitely repeated prisoner's dilemma to the evolutionary treatment. Recall how in Section 5.5 co-operation had emerged as a potential Nash equilibrium (one of many, of course). This window of opportunity opened for co-operation because the *tit-for-tat* strategy (τ)[14] co-operates conditionally on the opponent's continued co-operation and such co-operative behaviour is a best reply to itself (provided the probability that the game will not end immediately is low). Interestingly, it is easy to show that strategy τ is *not* evolutionarily stable!

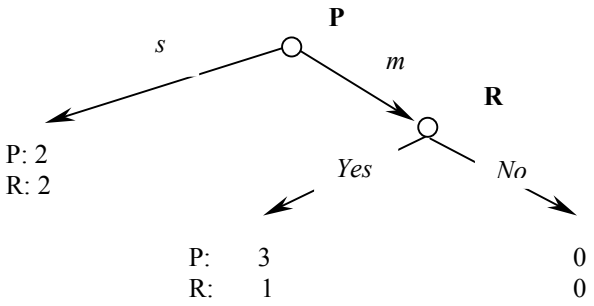
This may sound strange at first. Strategy τ should be immune to invasion from defectors (that is, mutants who defect in each round) since, as we saw in Chapter 5, it is a best reply to itself (provided the game is repeated with sufficient frequency). True enough, in a population in which everyone is programmed to play τ, co-operation rules and mutations giving rise to a small number of defectors in each round eventually die out. The reason why defectors cannot successfully invade a τ–playing population is that, on average, τ nets larger average pay-offs than defection against a τ–playing opponent; that is, τ is in a Nash equilibrium. However, τ is *not* immune to invasion from *unconditional co-operators*; that is, players who play co-operatively irrespectively of their opponents behaviour. If players also have a proclivity (however slight) toward simpler (as opposed to more complex) strategies, then unconditional co-operation will grow (within a τ–playing population) at the expense of τ. But unconditional co-operation, once it gets established, is susceptible to an invasion from defectors.[15]

In short, a τ–playing population has the tendency to evolve into one of unconditional co-operators which, in turn, mutate into ruthless defectors. Thus, for τ to be evolutionarily stable it must be rendered safe from unconditional co-operation. For this, we need to complicate the conditional strategy a little. For example, we need something like the *Tit-for-Two-Tats* strategy already discussed in Section 5.3.5 of Chapter 5. Just as this more sophisticated conditional strategy proved robust in the presence of conventional theory's trembles, it can once more be shown better equipped than τ to withstand evolutionary pressure – see also Sugden (1986). Unfortunately, it is not the only EE as defection also remains a best reply to itself in the modified world where trembles occur. Thus, rather like the analysis of previous chapters, EvGT substitutes one puzzle, regarding co-operation, for another. It can occur spontaneously but we need an analysis of equilibrium selection to explain why it happens.

The travails of strategy τ illustrates again the difference between Nash and evolutionary equilibria. Strategy τ fulfils the conditions of a Nash equilibrium but not of an evolutionary equilibrium because the latter requires more than the former: it requires that the strategy under consideration is not only a best reply to itself but, additionally, that it is a better reply to strategies

that may infect the population of τ–players (see Samuelson, 2002). Conditional co-operation (τ) may be a best reply to itself but it is susceptible to unconditional co-operation which is in turn susceptible to consistent defection.

Similarly intriguing results obtain in other dynamic games. Consider for instance the simple (two-stage) ultimatum game from Chapter 4. Conventional theory predicts that the Proposer (P) will offer the Responder (R) a tiny offer during Stage 1 that the latter has no alternative but to accept in Stage 2. Thus the unique SPNE predicts that P will keep all the money except a tiny amount that goes to the hapless R. However, experimental evidence consistently defies this prediction – see the discussion in Chapter 3. One possible explanation comes from EvGT – see Binmore, Gale and Samuelson (1995). Game 6.2 is an even simpler version of the ultimatum game. The Proposer (P) chooses between sharing equally \$4 with the Responder (R) [strategy *s*] or offering R \$1 and demanding \$3 for herself [strategy *m*]. In the latter case, R chooses between agreeing to the unequal distribution (and thus getting \$1 only) [strategy *Yes*] or turning it down [strategy *No*] in which case neither players wins anything.



Game 6.2 – *A simplified ultimatum game*

Via *Nash backward induction*, the unique SPNE is: P chooses *s* and R says *Yes*. Suppose, however, that the game is played in an evolutionary setting with Ps and Rs being matched repeatedly against fresh opponents. In addition, suppose that, at the outset, a significant proportion of Rs are programmed to say *No* to Ps who refuse to share the \$4 equally. The sharing strategy *s* will then give Ps higher pay-offs on average than the greedier strategy *m*. The replicator dynamics will, as a consequence, exert pressure on strategy *m* as a diminishing proportion of Ps will be choosing it. Consequently, the pay-off losses to Rs from saying *No* to greedy Ps diminish and so their 'recalcitrance' is favoured by the evolutionary dynamics. The population may, therefore, converge on a situation in which Ps offer to share the \$4 and a significant proportion of Rs would be programmed to say *No* to the unequal distribution.

It is, of course, true that, for the equal distribution (strategy *s*) to remain evolutionarily stable, the Rs' resolve to say *No* must not be tested too frequently. On the other hand, if the greedy strategy *m* does well in the population, it will tend to boost the frequency of *No*. Samuelson (2002) adds that the mutations which infuse strategy *m* into the population have "…the potential to introduce a counter-pressure by continually injecting *No* into the population. Either force may win, raising the prospect of weakly dominated strategies surviving". This conclusion is reinforced by Mailath (1998) who shows that almost anything is possible if we allow Rs and Ps to learn at different speeds, thus cautioning against any general presumption that EvGT will always help with the equilibrium selection problem.

6.3 EVOLUTION IN HETEROGENEOUS POPULATIONS

6.3.1 Asymmetrical (or two-dimensional) evolution and the demise of Nash equilibria in mixed strategies (NEMS)

There is one game from the list of classic 2×2 games in Chapter 2 that we have, so far, deliberately neglected: *Hide-and-Seek* (Game 2.19). The reason is its asymmetrical nature (violating Condition E) and the complications caused by it. The asymmetry in pay-offs between the two players (note that they cannot be represented in terms of Game 6.1) means that we are no longer at liberty to assume that the probability with which the row player (R) chooses her first strategy must be the same as the probability with which the column player (C) chooses his. We are forced instead to examine the behaviour of R and C players separately.

The population is no longer homogeneous as there exist two types of player: R-players and C-players. One might have thought that this is an arbitrary distinction which should make little difference to the analysis, but it is not. The moment we introduce observable differences between players, a great deal changes. To pre-empt our analytical conclusion, this game ends up *without* an evolutionary equilibrium.

To see why, we begin by defining another probability, in addition to p (which is the probability with which R-players select R1, their first strategy). Let q be the probability with which C-players choose their first strategy (C1). So, p=Pr(R1) and q=Pr(C1) and since we can no longer assume that $p=q$, we also need to define the d functions separately for the two types of player (d^R for R-players and d^C for C-players).

$d^R = ER(R1) - ER(R2) = q1 + (1-q)0 - [q0 + (1-q)1] = -1+2q$ ---------- (6.2)
$d^C = ER(C1) - ER(C2) = p0 + (1-p)1 - [p1 + (1-p)0] = 1-2p$ ---------- (6.3)

Expression (6.2) is R's net gain from adopting her first strategy (relative to her second) and is an increasing function of q. This makes perfect sense: As the probability that C will choose his first strategy rises, it makes increasing sense for R to play her first strategy too (recall the pay-off structure of Game 2.17 – see Section 2.6.1). And vice versa. Meanwhile, the opposite applies for C. As the probability that R will choose her first strategy rises, C has more reasons for scorning his first strategy. [Note that this is consistent with naming Game 2.17 *Hide-and-Seek*: Think of C as the player who tries to 'hide' and R as the one who is looking for him. If they go to the same place (i.e. follow the same strategy), R will catch C and win. If they do not (follow different strategies) then C is not caught and wins.]

Before proceeding further, the replicator dynamic needs to be updated as there are now two selection mechanisms at work: One for R-players and one for C-players.

Replicator Dynamics for Game 2.19

Dynamic for p=Pr(R1)	**Dynamic for q=Pr(C1)**
$d^R>0$ means that p increases	$d^C>0$ means that q increases
$d^R<0$ means that p declines	$d^C<0$ means that q declines
$d^R=0$ means that p is stationary	$d^C=0$ means that q is stationary

When a player-type makes positive net gains from its first strategy, that player-type will be increasingly drawn to its first strategy. And vice versa. From expressions (6.2) and (6.3) we know that:

(a) When $d^R= d^C=0$, $p=q=$ ½ and we are at the game's NEMS
(b) $d^R>0$ ($d^R<0$) when q exceeds (is less than) its NEMS value ½

(c) $d^C>0$ ($d^C<0$) when p lies below (exceeds) ½

In (a) we note that NEMS implies that both players will randomise with equal probability between their two strategies. We mark this NEMS in the graph of Figure 6.2. From (b) and (c) we derive the dynamics which are depicted on the same graph. Suppose that the population finds itself initially in the top-right quadrant where both p and q exceed ½. From (b) and (c) above it transpires that R-players have more to gain from their first strategy than from their second. C-players on the other hand gain more when programmed to play their second strategy. Thus, a population that finds itself in the top-right quadrant of the relevant graph in Figure 6.2 will tend to evolve in a manner that suppresses the first strategy of C-players but favours the first strategy of R-players. In short, p will tend to grow and q will shrink. If we perform the same analysis for all four quadrants, we find the following evolutionary dynamics.

Quadrant	***Net gain from using their first strategies***	***Evolutionary Tendency***
$p>½$ and $q>½$	$d^R>0$, $d^C<0$	$p\uparrow,q\downarrow$
$p>½$ and $q<½$	$d^R<0$, $d^C<0$	$p\downarrow,q\downarrow$
$p<½$ and $q<½$	$d^R<0$, $d^C>0$	$p\downarrow,q\uparrow$
$p<½$ and $q>½$	$d^R>0$, $d^C>0$	$p\uparrow,q\uparrow$

Figure 6.3 – *Evolutionary dynamics in Game 2.19*

Superimposing the arrows of the last column of Figure 6.3 above on the graph of Game 2.17 in Figure 6.2, we discover that evolution places this population into some never ending cycle around NEMS (that is, around values $p=q=½$). Of course there is an infinity of possible cycles that may come about and we have only drawn one in order to illustrate the lack of a resting point (i.e. the absence of an evolutionary equilibrium).

What if the population *begins* at NEMS (that is, the initial behavioural code is precisely $p=q=½$)? Since both d functions are exactly zero, does the replicator dynamic of this population not suggest that behaviour will remain unchanged? That NEMS is still an evolutionary equilibrium? Although it is true that the replicator dynamic settles down when $d^R=d^C=0$, we must not forget that an evolutionary process consists not only of a *replicator dynamic* but, crucially, of a *mutation mechanism* as well; a mechanism which ensures that every now and then mutants will test the evolutionary stability of any behavioural code. When we begin with NEMS behaviour $p=q=½$, any mutation will push the population's aggregate behaviour into one of the quadrant of Figure 6.3. Thus the population's initially NEMS-like behaviour will be destabilised and will enter into a cycle depending on which quadrant it was pushed into.

This is analytically interesting. Recall (from Chapter 2) how John Nash famously showed that all (finite) games feature at least one (Nash) equilibrium. Evolutionary Game Theory (EvGT) offers no such solace. There are countless games in which no evolutionary equilibrium (EE) exists. Game 2.17 is one of those. It has no Nash equilibrium in pure strategies, one NEMS ($p=q=½$) and no EE. All that we can predict is a perpetual cycle around NEMS.

So far, we discovered that asymmetries in simple 2×2 games (e.g. Game 2.17) engender population heterogeneity (e.g. a distinction between R-players and C-players) which, in turn, can undermine NEMS. In effect, EvGT predicts a kind of dynamic indeterminacy for games of this type. More generally, population heterogeneity is inimical to NEMS even in symmetrical games. Take for instance the *Battle-of-the-Sexes* or the *Hawk-Dove* interactions (Games 2.13 and 2.14). The relevant graphs in Figure 6.2 confirm the coincidence of a unique EE with the games' NEMS when the population is homogenous. Suppose, however, that the population is heterogeneous. For instance, if there are two types of player distinguished by some arbitrary feature (one type has brown eyes, the other blue; or one is left-handed the other right-handed etc.), the NEMS ceases to be an EE!

To see why heterogeneity has this effect, we shall focus on *Hawk-Dove* (Game 2.14). To keep the analysis simple, we shall assume that nothing changes in the evolutionary treatment of *Hawk-Dove* (Section 6.2.1) except one thing. The population now comprises two equally sized groups: one group consists of 'blue' players and the other of 'red' players. We assume that a player's colour is utterly arbitrary and is not correlated with any other character trait, talent etc. According to conventional game theory, the introduction of colours should make no difference to the game as it is unrelated to the pay-off structure and its observation conveys no meaningful observation about one's opponent.

Be that as it may, the arbitrary assignment of colours does make a difference from the perspective of EvGT. Suddenly, players can condition their behaviour on some feature of their opponent. The potential thus exists for heterogeneous behavioural codes. To the protests of conventional game theory, that there is no rational motive to condition one's behaviour to one's opponent's meaningless colour, EvGT retorts that successful strategies need have no good reason behind them other than their relative success. The question for EvGT is: Will conditional strategies (that is, strategies which instruct players to play differently against opponents of different colour) gain an evolutionary upper hand over unconditional ones?

Box 6.6

HOW BIOLOGISTS DISCOVERED THE IMPORTANCE OF ARBITRARY FEATURES

Long before game theorists, biologists studied carefully the endogenous evolution of new phenotypes (or, equivalently, new ways of playing a game). Having observed that populations of birds, ants etc. had a penchant for subdividing into groups which behaved quite differently to one another, evolutionary biologists were alerted to the possibility that behavioural differences between groups could be the result of nothing more than arbitrary differences in appearance. For instance, birds with a red feather on their head behaving more or less aggressively than birds with a blue feather. The puzzle was, since the colour of feather is uncorrelated with strength or fighting skills, why does it influence behaviour?

With this query as a starting point, biologists came to the thought that one way in which new strategies may enter a population (without altering the structure of the game) is through conditioning behaviour on an extraneous feature of one's opponent (that is, a feature extraneous to the structure of pay-offs). For instance, the extraneous feature of the interaction might be the colour of feather on a bird's head, the colours of one's eyes, of one's skin etc. The significance of the availability of such observable features is that players can now condition their behaviour on them. The new behavioural rule would then take the form of: *If your opponent has blue eyes, then play strategy x; and if your opponent has brown eyes, then play strategy y*. Such a game is now said to be played *asymmetrically* in recognition of the fact that the players have learnt to differentiate themselves and to condition their strategies on observed characteristics of their opponents (however arbitrary and seemingly empty of information they might be).

Once this happens, behavioural adaptation also becomes more nuanced as it is now feature-specific. Indeed, as we shall see below, population heterogeneity often gives rise to behavioural conventions which, in antagonistic games (like *Hawk-Dove* and *Battle of the Sexes*), are asymmetrical, iniquitous and lead to the demise of NEMS-like evolutionary equilibria.

We suppose that, in *Hawk-Dove*, p is the proportion of blue players who play h (strategy 'hawk'), i.e. p = Pr(blue plays h), while q is the proportion of red players who play h; q = Pr(red plays h). In meetings between a blue and a red player, a two-dimensional replicator dynamic emerges conceptually similar to that in the case of Game 2.17 above. When a red player acts aggressively (strategy h), her average pay-offs *might* be different depending on whether her opponent is a blue or a red player. To be precise, as long as red and blue players behave differently, a red player's expected returns from h will differ depending on her opponent's colour:

$$ER^r(h \text{ against a red opponent}) = = -2q + 2(1-q) \text{ and } ER^r(d \text{ against a red opponent}) = (1-q)$$

Similarly for blue players. Their average pay-offs from aggressive strategy h are:

$$ER^b(h \text{ against a blue opponent}) = -2p + 2(1-p) \text{ and } ER^b(d \text{ against a blue opponent}) = -(1-p)$$

Thus, the net gains from hawkish behaviour of the red and blue players respectively are:

d^r = ER^r(h against a red) – ER^r(d against a red) = -2q + 2(1-q) - (1-q) = 1-3q---------- (6.4)
d^b = ER^b(h against a blue) – ER^b(d against a blue) = -2p + 2(1-p) - (1-p) = 1-3p ----------(6.5)

Replicator Dynamics for Hawk-Dove (Game 2.14) under 2-dimensional evolution	
Dynamic for p=Pr(red plays h)	**Dynamic for q=Pr(blue plays h)**
d^r>0 means that $p\uparrow$ when q<1/3	d^b>0 means that $q\uparrow$ when p<1/3
d^r<0 means that $p\downarrow$ when q>1/3	d^b<0 means that $q\downarrow$ when p>1/3

Figure 6.4 plots the above dynamic and the phase diagram reveals that this game's NEMS (p=q=1/3) is evolutionarily unstable. Starting from p=q=1/3, any mutation that alters either p or q ever so slightly will push the population's behaviour into one of two areas: Either above the 45° line or below it. We can clearly see that above the 45° line all arrows (or trajectories as mathematicians call them) lead in the direction of the Nash equilibrium in pure strategies (p=0, q=1).[16] The meaning of this is that *any mutation that causes blue players to become slightly more hawkish than red players* will start an evolutionary bandwagon which will lead to an evolutionary equilibrium (EE) such that in meetings between blue and red players, *all blue players act aggressively* (i.e. play h) *and all red players acquiesce* (i.e. play d). Similarly, if the original mutation initially has red players acting relatively more aggressively than blue players, the reds will come to dominate the blues whenever they meet them *irrespectively of how slight the initial difference was*.[17]

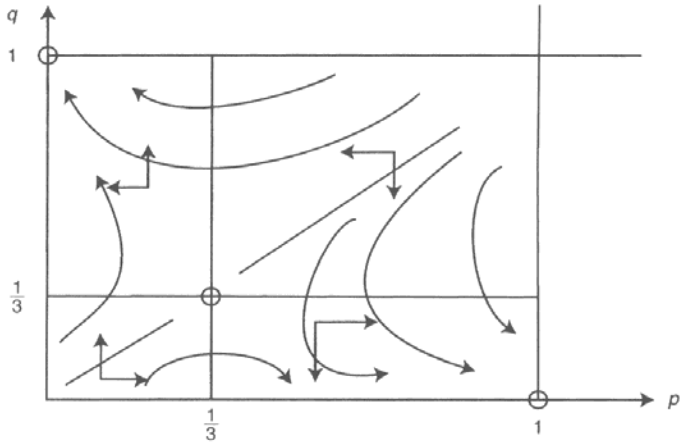


Figure 6.4 – *Two-dimensional evolution destabilises NEMS in the Hawk-Dove*

In summary, two-dimensional evolution gives rise to two distinct evolutionary equilibria:

EE1: p=0 and q=1
EE2: p=1 and q=1

Which of the two will prevail will depend on initial conditions or mutations. If the population's behaviour starts off with p>q, *EE1* will obtain. And if initially p<q, *EE2* will be favoured by evolution. Finally, if at the outset p=q, the bandwagon's ultimate destination (toward *EE1* or *EE2*) will depend on whether the first mutation will cause red or blue players to be slightly more aggressive. (For a more detailed analysis see Weibull, 1995, and Friedman, 1996).

As Problem 6.2 demonstrates, the above result applies also to other antagonistic games, e.g. the *Battle-of-the-Sexes*. From the perspective of the problem that has been following us in this book from one chapter to the next, namely *indeterminacy*, it means that there has been a setback. EvGT was hailed because it seemed to narrow down the range of potential equilibria. In *Hawk-Dove*, for

instance, we were impressed with how it discarded the pure strategy Nash equilibria, pointing to NEMS as the unique evolutionary equilibrium. However, that result holds only as long as the population is strictly homogeneous. The moment the slightest distinction is allowed to slip in (either because of pay-off asymmetries between row and column players, as in Game 2.17, or because of some seemingly meaningless characteristic of individual players), evolution occurs in more than one dimensions at once (one per type of player), NEMS collapses as the unique EE, and two new EEs emerge. We are thus left, again, with multiple equilibria (as in the original *Hawk-Dove* and *Battle-of-the-Sexes*). Couple this with the earlier thought that evolution in heterogeneous populations need not have any equilibria (e.g. Game 2.17) and the evolutionary turn is beginning to look distinctly less promising.

Population heterogeneity has other effects too. It may, for instance, cause evolution to favour resting points (or evolutionary) equilibria that are evidently wasteful. Consider, for instance, Game 6.3 below:

	C1	C2
R1	$^{+}2,2^{-}$	$^{+}1,2^{-}$
R2	$^{+}2,1^{-}$	0,0

Game 6.3 – *Example of 2-dimensional evolution favouring the inefficient Nash equilibria*

There are three Nash equilibria in pure strategies. Nevertheless, one of them (R1,C1) is mutually advantageous compared to the other two. Moreover, the same (superior) equilibrium is the only one that does not rely on weakly dominated strategies R2 and C2. One would have thought that evolutionary pressures would discard the inefficient Nash equilibria in favour of (R1,C1). In homogenous populations (and one-dimensional evolution) this is, indeed, what happens. However, the moment players can distinguish between two types of player (e.g. those with the blue and those with the red colour assignments), evolution only comes to (R1,C1) by accident!

Indeed, it can be proven that a 2-dimensional evolutionary process will lead to the following strange equilibrium: While one of the two subpopulations will always be playing their first (efficient) strategy, the proportion of the other subpopulation who do the same may be anywhere between 0 and 1 – see Problem 6.3 for a proof. Not only is this suggestive of the return of *Indeterminacy* caused by population heterogeneity but it also rings alarm bells regarding the capacity of arbitrary personal characteristics to cause social inefficiencies. Heterogeneity, in summary, spawns behaviour asymmetries which, in turn, yield indeterminacy. Once more, 'anything goes'.

6.3.2 Does Evolutionary Game Theory apply to humans as well as it does to birds, ants, etc.? An experiment with 2-dimensional evolution in the *Hawk-Dove* game

Evolutionary game theory essentially copied the method of evolutionary biology where the subject matter was the evolution of phenotypes among bees, ants and birds. A question can, therefore, be legitimately put as to whether the results hold for humans too. After all it was evolution that equipped us with the capacity rationally to *assess* our behaviour (rather than simply to follow some programmed phenotype thoughtlessly).

One way to answer this question is by means of experiments. In a recent experiment (see Hargreaves-Heap and Varoufakis, 2002), we set out to establish whether the last section's theoretical predictions would be upheld in the laboratory. As is common practice with evolutionary games (see Friedman, 1996), the *Hawk-Dove* (Game 2.14) was played anonymously amongst 640 players (separated in 32 sessions, or groups, of between 20 and 26 subjects each) who knew that in the next round (there were 32 rounds in total) they would come across a fresh opponent. Pay-offs were exactly as in our matrix (Game 2.14) with the outcomes 2,1 and –2 taking the form of

Australian dollars. Per round pay-offs were added up (per player) and were collected in cash at the end of the session.

In 8 of the 32 sessions (or groups), the game was played under conditions consistent with one-dimensional evolution; that is, the population of players could be thought of as homogeneous in the sense that, although they were all different people, complete anonymity and constant re-assignment of opponents ensured that one opponent seemed utterly indistinguishable from the next. When the population is homogenous in this way, EvGT predicts a unique EE which coincides with NEMS (recall Section 6.1.2 and, mainly, Section 6.2.1). Players ought to choose their aggressive strategy (h) with probability p=1/3.

In evolutionary terms, the prediction here is that the population's aggregate behaviour would gradually approach an equilibrium in which conflict (i.e. outcome *hh*) occurs with frequency ($p \times p$ =1/9) 11%, mutual acquiescence (i.e. outcome *dd*) is observed in [(1-p)×(1-p)=4/9] = 44% of the interactions, and the pure strategy Nash equilibrium (in which one player plays h and the other d) is observed with frequency [2×p×(1-p)=2× (2/9)] = 44%. In our experiment, the hypothesis that the population would settle down to some stable frequencies was indeed confirmed, albeit the frequency of conflict was considerably greater than predicted. Outcomes *hh*, *hd* and *dd* were observed with frequencies 29%, 39.8% and 31.2% respectively.

The above was neither terribly surprising nor particularly insightful. The real question was whether discriminatory behaviour, like that predicted in the previous Section (see Figure 6.4), would evolve when we introduce some arbitrary distinction between players. According to EvGT, when players come from an heterogeneous population and can condition their strategies on some feature of their opponents (however meaningless that feature might be), then we should observe a tendency for players of one type to become aggressive (play h with high frequency) and players of the other type to turn acquiescent (play d). Does this apply to humans as well as to insects and birds?

To find out, we ran 16 experimental sessions which differed from the homogenous population (or control) sessions. At the beginning of these sessions, each player received either a blue or a red label, with equal probability.[18] In each round, players were told the colour of their opponent. Our intention was to test the prospects of two-dimensional evolution on the basis of these colour assignments.

It did. After approximately the first 16 rounds (of 32) one of the two colours started to dominate the other quite clearly.[19] In 9 sessions the dominant colour label turned out to be the blue one (i.e. blue players played significantly more often strategy h against red opponents), while in 6 sessions it was the red. In conclusion, there is little doubt that, at least in the laboratory, discriminatory behaviour among humans evolves (just as it does, according to evolutionary biologists, in the animal world) when players can condition strategies on extraneous information, e.g. colour labels (like the ones we used).

But why does this happen? Shouldn't rational human beings, who know that the colour labels are randomly distributed, ignore them? Well, they did not. In effect, within 20 or 30 minutes, highly discriminatory conventions ('blue dominate red players' or vice versa) became established and, in meetings between players of different colour, determined whether a subject would get the $2 or receive nothing *on the basis of his/her arbitrary colour* (as opposed to something meaningful like their personal characteristics, e.g. intelligence, aggression). Our interpretation is this:

In interactions (like *Hawk-Dove*) in which there is *rational indeterminacy* in the conventional game theoretic sense, people try to condition their behaviour on *any* information that comes to hand, even meaningless information (e.g. on the relative aggression of 'blue' players). It is a simple psychological response to uncertainty. Indeed there is much evidence to support the idea that people look for 'extraneous' reasons to 'explain' what are in fact purely random types of behaviour (see Box 6.7 on winning streaks). Of course, once they do so, an initial, random difference in the behaviour of the 'reds' and the 'blues' gets a bandwagon rolling (see Figure 6.4),

leading to stable discrimination that succeeds in minimising costly conflict despite being non-rational (why should the 'blues' be dominant over the 'reds' or vice versa?). So *ex post*, it will seem to make sense as each player does take the best action given that chosen by the other, even though there is no reason *ex ante* for this selection of actions. Figure 6.5 shows how conflict dropped with the use of the convention.

Discriminatory conventions (Definition)

The evolutionary equilibria in cross-colour *Hawk-Dove* interactions result in systematic *hd* outcomes where one type of player (e.g. the red or the blue) plays *h* consistently, with the other type plays *d*. These equilibria, the products of two-dimensional evolution, can be interpreted as conventions (see Lewis, 1969). Indeed they constitute a form of discriminatory convention in the sense that they assign each person, on the basis of his or her colour, to either the hawkish or dove-like role and this results in people of one colour enjoying much higher pay-offs than those of the other for reasons which have nothing to do with superior rationality, information, talent or innate aggression.

Type of experiment\ Outcomes	**hh**	**hd**	**dd**
*Observed frequencies **without** colours*	29%	39.8%	31.2%
*Observed frequencies **with** colour assignments*	19%	52%	28%

Figure 6.5 – *How colour assignments reduced conflict*

To put these observations rather less blandly, since rationality in these games does not tell us what is our best course of action (due to multiple equilibria), rational agents have no reason to resist the general impulse engendered by evolution towards mimicking 'successful' behaviour. As a result, the history of the game depends in part on what are the idiosyncratic and unpredictable (non-rational, one might say, as opposed to irrational) features of individual beliefs and learning. Thus, the replicator dynamics of Figure 6.5 might be even more valid in the case of humans (as the reported experiment suggests) than within animal or insect populations.

Box 6.7

WINNING AND LOSING STREAKS?

People often refer to winning or losing streaks to describe a process whereby one win (loss) leads to another because confidence grows (falls), making another win (loss) even more likely and so on. However, there are very few examples of so-called winning streaks which are not what would be expected statistically from a team which has a constant chance of winning each game (Joe DiMaggio's famous hitting run in baseball appears to be one such exception). The attribution of a causal process (the growth or decline of confidence which affects performance) to explain what is a purely random phenomena seems to be part of a general human cognitive tendency which imposes or seeks explanations for events which are in fact purely driven by chance (see Kahneman, Slovic and Tversky, 1982). If, as a species, we like to impose a deterministic order on events to render them intelligible, even when none exist (and the best description comes from the operation of chance), then it seems likely that the symmetrical version of these games (where we play and expect to encounter others playing strategies probabilistically) will yield to an asymmetrical deviation because this affords a deterministic 'explanation' of our behaviour.

Another example of this kind of attribution surfaces with 'moral luck'. Driving at speed through a town is liable to produce accidents with some frequency for the simple reason that there are always people who inadvertently step onto the road. It is a matter of luck whether this happens to one speedster rather than another (because such inadvertent moves occur randomly) and yet we typically hold the person involved in an accident morally culpable in a way that a mere speedster is not. It is as if we believe that the person involved in the accident has done something more than speed to make the accident happen, but this is not the case.

While this experiment suggests that EvGT offers important insights into human behaviour, there is also evidence that all is not well with EvGT. Consider Game 6.3 below. The original *Hawk-Dove*

(Game 2.14) has been amended through the addition of a third 'co-operative strategy' for each player. This third strategy is not part of any equilibrium. Even though its use by both players would yield a superior result to any of the available pure strategy Nash equilibria, it will not be played according to conventional game theory[20] and it will disappear according to EvGT.[21]

		B		
		h	d	c
A	h	-2,-2	2,0	4,-1
	d	0,2	1,1	0,0
	c	-1,4	0,0	3,3

Game 6.4 – *The HDC (Hawk-Dove-Co-operate) game*

In our experiment, Game 6.4 was included for two reasons. First, we were interested to see if strategy *c* would 'survive' under evolutionary conditions in the laboratory, despite game theoretical (conventional and evolutionary) predictions to the contrary. Secondly, we wanted to investigate the possibility that patterns of co-operation might differ depending on the location of players in relation to the evolved convention. So, after the 32 rounds of *Hawk-Dove* play (reported above), our experimental subjects played another 32 rounds of Game 6.4. The results were quite striking.

The first observation is that, contrary to EvGT's prediction,[22] the co-operative strategy did not become extinct. Although it did decline in frequency, it stabilised at a significant frequency in excess of 30%. The second observation prelates to the different propensities of players to co-operate depending on whether they, and their opponent. were advantaged or disadvantaged by the colour assigned to them at the beginning. Recall that in 9 sessions the blue players gained the upper hand while the red players were similarly advantaged in 5 sessions. The data in Figure 6.6 tabulates the aggregate outcome frequencies in meetings between the advantaged players (A-players), between the disadvantaged players (D-players) and between one A-player and one D-player.

While co-operation did disappear from interactions between A-players, it hit a remarkable frequency of just under 90% in meetings involving players of the disadvantaged colour. So, it seems that in addition to its conflict-minimising impact, discrimination on the basis of arbitrary colour assignments had another important effect. It caused wildly different behavioural patterns when a co-operative action became available. Neither conventional game theory nor evolutionary game theory can explain this.

Outcome frequencies in the last 11 rounds of Game 6.4 (HDC) in the sessions in which players of one colour (A) gained an advantage at the expense of players of the other colour (D)	**hh**	**hd**	**dd**	**cc**
Meetings between two A players	**51.2%**	8.9%	1.81%	4%
Meetings between two D players	2.1%	3.5%	0.5%	**89.9%**
Meetings between an A and a D player	8.2%	**81%**	0.4%	0.5%

Figure 6.6 – *Data from the last 11 rounds of Game 6.4*

One interpretation of this failure of EvGT is that it may be insufficiently evolutionary. While EvGT models nicely the way in which behaviour adapts in response to a game's *given* pay-off structure, the latter may be evolving too (at least in the creative imagination of human beings). We strongly suspect that, once discrimination takes hold, the disadvantaged players begin to value outcomes differently depending on who they are playing with. If, for instance, they are playing against an opponent of the same colour as them (and are thus equally disadvantaged), they may value the co-operative outcome disproportionately to the actual money pay-off. In contrast, when

matched against a player advantaged by the evolved discriminatory convention, they may not value co-operation in the same way. Unless EvGT takes into account the agents' political or moral psychology, it may not succeed in capturing some of the most significant aspects of social evolution.

6.3.3 Multi-dimensional evolution and the conflict of conventions

In the last two sections, it was assumed that agents bear a single, potentially discriminatory feature on which behaviour can be conditioned (we used the example of blue or red colours). In reality, of course, individuals have a number of observable characteristics and each one of them could 'evolve' into the feature on which behaviour is conditioned. Which feature comes to dominate is hugely important, at least to the individual agent (see Box 6.8).

Box 6.8

DISCRIMINATORY CONVENTIONS, INEQUALITY AND PROPERTY RIGHTS

In our experiment, the players who happened to have the colour that, eventually, dominated, ended up with much larger aggregate pay-offs than the players with the 'disadvantaged' colour. Suppose there was a second potential 'discriminant'; players being assigned not only red or blue colours but also numbers (1 or 2). EvGT cannot tell us whether discrimination would evolve between red and blue players or between players with numbers 1 and 2. All that EvGT predicts is that *some* form of discrimination *will* emerge. Of course, which discriminatory convention arises (colour or number based) matters deeply to players themselves. In effect in a *Hawk-Dove* game over contested property, what happens in the course of moving to one EE (e.g. reds get $2 when playing against blues irrespective of their number; or players with number 1 get the $2 irrespective of their and their opponents' colours) is the establishment of a form of *property rights*.

This is interesting not only because it contains the kernel of a possible explanation of property rights (on which we shall say more later) but also because the probability of having the dominant role is unlikely to be distributed uniformly over the population. Indeed, this distribution will depend on whatever is the source of the distinction used to assign people to roles. Thus, for instance, the distribution of property is likely to be very different in a society where the assignment to roles depends on sex and age as compared with, say, height. In the one either the tall or the short people will be respectively advantaged and disadvantaged. Whereas in the other, it could be old females who are marginalised while the young males rule the roost; or some other hierarchical combination of these age and sex differences.

To see why this is likely, consider a situation where there are two competing sources of differentiation (i.e. observable features) which generate two types of conventions (π and φ). Suppose, for instance, that players come as either 'young' or 'old' and, *at the same time*, as 'tall' or 'short'. When players had a single distinguishing feature (blue or red – see the previous two sections), the discriminatory convention latched on to that unique feature and (behaviourally) segregated the population accordingly. When there are two distinguishing features (age and height), which of the two will discrimination turn on? This is of great importance because some individuals who might be favoured by discrimination against the short (or tall) may be disadvantaged by discrimination turning on age (and vice versa).

There are, in other words, two potential conventions pitted against each other. Let us say one (π) distinguishes players according to age and instructs the young to concede to the old, while the other convention (φ) distinguishes according to height and instructs the short to concede to the tall. In technical terms, this is a case of three-dimensional evolution. One dimension concerns the evolution of behaviour among adherents of convention π (π-players henceforth), the second concerns the evolution of behaviour among φ-players, and the third dimension concerns the evolution of behaviour in interactions between a π-player and a φ-player.

The basic intuition is easy to grasp. One convention will dominate and its selection depends critically on the initial number of people who subscribe to each convention. The reason is simple. We are dealing with conventions which, by their very nature, work and become stronger the larger

the number of adherents. Thus once the balance tips in the direction of one convention, it quickly develops into a bandwagon. But the rub is: What does the tipping?

To make this clear, consider how the pay-off to the use of a particular convention will depend on the numbers adhering to it. The convention will tell you what is actually the best action to take provided you come across someone who also adheres to your convention (for instance, in *Hawk-Dove* if you are old and you come across a young person who subscribes to the age convention, the best action is to play *h*). The convention, however, will lead you to take an inferior action when you come across someone who subscribes to a different convention and that convention indicates a different course of action. Of course, another convention will not always do this.

For instance, in our example some young people are also taller than some old people and so the two conventions will sometimes point to the same pattern of concession for your opposing player. Nevertheless for any given overlap between conventions of this sort, the probability of coming across someone who is going to play the game in a contrary manner (that is, play *h* to your *h*), and who thus turns your action into an inferior one, will depend on the number of people who subscribe to the contrary convention. In other words, as the numbers using your convention rise so it becomes increasingly likely that it will guide you to the best action. If people switch between conventions based on average returns, then eventually one convention will emerge as the dominant one.

This conclusion reinforces the earlier result that the course of history depends in part on what seem from the instrumental account of rational behaviour to be non-rational (and perhaps idiosyncratic) and, therefore, on features of human beliefs and action which are difficult to predict mechanically. One can interpret this in the spirit of methodological individualism at the expense of conceding that individuals are, in this regard, importantly unpredictable. On the one hand, this does not look good for the explanatory claims of the theory. On the other hand, to render the individuals predictable, it seems that they must be given a shared history and this will only raise the methodological concern again of whether we can account for this sharing satisfactorily without a changed ontology. In summary, if individuals are afforded a shared history, then social context is 'behind' no one and 'in' everyone and then the question is whether it is a good idea to analyse behaviour by assuming (as methodological individualists do) the 'separability' of context and action.[23]

There is a further wrinkle to this analysis which addresses actual social settings. The return from the use of a convention for a particular individual will depend not only on the proportions of the population subscribing to it, but also on the frequency with which it is assigned the dominant role. Thus the general population movement towards the emerging convention is liable to be taking place against a backdrop of cross movements which take, to use the earlier example, the old to the age convention and the short to the height convention. In fact, these cross movements could be very influential in establishing which convention becomes more popular.

To see the point a little more sharply, suppose the two conventions have an equal number of adherents. The expected return from the use of each convention is the same when every person has a 50% chance of being dominant under each convention. Now suppose one convention actually allocates the advantage of being dominant more unequally than the other. This will encourage some from the equal convention to the unequal one (namely, those who would benefit under the unequal convention more than 50% of the time). At the same time, those who lose out under the unequal convention will be attracted to the equal one. The relative movement of population will be determined initially by the movements which are sparked by the differing characters of each convention with respect to the distribution of the advantage of dominance.

Expressions (6.6) and (6.7) illustrate this point. We assume that in a *Hawk-Dove* context (Game 2.14) there are two conventions (π and φ). Below we define the proportions of the

population that adhere to the two conventions (π-persons and φ-persons) as well as the frequency of interactions between players adhering un-observably to different conventions:

p = proportion of persons following convention π; i.e. p=Pr(an opponent is a π-person)
q = proportion of persons following convention φ; i.e. q=Pr(an opponent is a φ-person)
k = the proportion of all inter-convention interactions in which the players are instructed by their (different) conventions to play in the same way.
r = the probability that a π-person will be instructed by π to play d.
s= the probability that a φ-person will be instructed by φ to play d.

In Figure 6.7 we have the tree diagram which enumerates all possibilities for a π-person who, in a *Hawk-Dove* context, comes across a stranger who must subscribe to one of the two conventions π or φ. The following expression for π-person's expected returns follows from that tree diagram:

$$\begin{aligned} \text{ER}(\pi\text{-person}) = \quad & rpr\times 1 + rp(1-r)\times 0 + r(1-p)k\times 1 + r(1-p)(1-k)\times 0 + \\ & (1-r)pr\times 2 + (1-r)p(1-r)\times(-2) + (1-r)(1-p)k\times(-2) + \\ & (1-r)(1-p)(1-k)\times 2 \\ = \quad & r[pr+(1-p)k] + 2(1-r)[pr+(1-p)(1-k)] - 2(1-r)[p(1-r) + (1-p)k] \qquad (6.6) \\ \text{ER}(\varphi\text{-person}) = \quad & s[qs+(1-q)k] + 2(1-s)[qs + (1-q)(1-k)] - 2(1-s)[q(1-s) + (1-q)k] \qquad (6.7) \end{aligned}$$

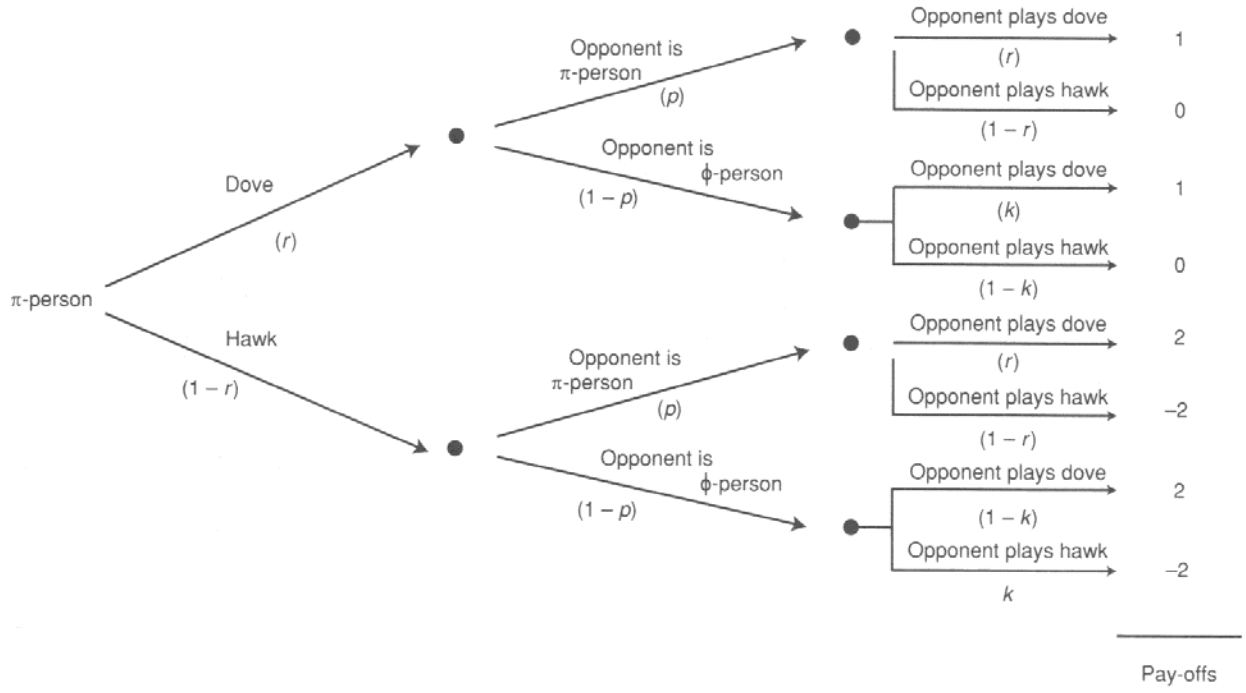
Close inspection of these expressions reveals that there is a wide range of k, r and s values for which the expected returns for players following either convention are both increasing functions of p and q respectively. This confirms the earlier observation that population movements may create a bandwagon effect in favour of one convention once it emerges as the one offering the superior expected returns. Why? The reason is that, under random pairings of players, and provided ER(π-person) increases when p increases, the higher the value of p (i.e. the proportion of π-persons) the greater the expected returns for π-persons. And the higher the expected returns, the more people will be drawn to convention π. Hence the bandwagon effect. For this to happen, however [that is, for ER(π-person) to be an increasing function of p], it can be shown that the following condition must hold:[24]

$$k > [(3r^2 - 8r + 4)/(4 - 5r)] \qquad (6.8)$$

Of course a similar condition applies for convention φ, namely that for ER(φ-person) to be increasing with q,

$$k > [(3s^2 - 8s + 4)/(4 - 5s)] \qquad (6.9)$$

Inequalities (6.8) and (6.9) tell an interesting story. Consider for instance what happens when r=2/3, or s=2/3; that is, when the conventions recommend that a player acquiesces (i.e. plays d) with the NEMS probability favoured by one-dimensional (or symmetrical) evolution within an utterly homogeneous population (see Section 6.2.1). Then the right hand side of the inequalities is zero, and the expected returns from each convention will be increasing functions of the number of people following it, provided of course k>0; that is, provided that there is at least a tiny possibility that an opponent following a different convention to yours will play the same strategy as you when the two of you meet.



(Probabilities in brackets under arrows)

Figure 6.7 – *The conflict of two conventions in the context of three-dimensional evolution*

The point of interest here is that, as long as individuals have more than one distinguishing feature (however arbitrary that might be), evolution will give rise not to one but to two conventions. Even if a single convention evolves initially, it will have a tendency to divide into two (π and φ). Suppose, for instance, that at the 'beginning' there is some symmetrical equilibrium when players cannot distinguish between one another. As they become more familiar with their opponents (or observant of their features) they may learn to observe *one* distinguishing feature (e.g. male/female). At that point, a discriminating convention emerges, instructing individuals to behave differently depending on their feature (e.g. gender). However, when individuals learn to observe some other distinctive feature as well (e.g. black/white skin colour), a second discriminating characteristic may come into play and, as it gathers more adherents, those who already adhere to it will benefit (provided $k>0$).

Which convention will do better for its adherents? We cannot tell in the abstract. What we can say is that the adherents of one convention will do better while those of the other will do worse. The reason is that when there are only two conventions, $p = 1\text{-}q$ (that is, when a person switches towards one convention he or she automatically abandons the other convention), and thus when some people start switching to π (for example) those who follow π will do better while φ-followers will suffer. Of course, it is simple to show that if players can later on distinguish more distinctive feature of their opponents (e.g. accent, size), then we have multidimensional evolution and all sorts of patterns of discrimination may emerge, co-exist and become evolutionarily stable.

An interesting corollary of the above is that, with two conventions only, a convention which can skew its followers' interactions towards fellow users of the convention will be better able to survive than one that does not. In this sense, discriminating conventions come about because of the antagonistic nature of the games people play (e.g. *Hawk-Dove*). Once they do, they divide and multiply and, in their struggle for survival, conventions end up not only segregating people behaviourally (that is, instructing people with different characteristics to behave differently) but also socially (in the above sense that conventions which manage to keep their adherents isolated from adherents of other conventions achieve greater success).

To demonstrate another point simply, let us assume that pairings are random; i.e. $p = 1\text{-}q = ½$. The expected returns for an individual with a $1\text{-}r = 1\text{-}s$ chance of the dominant role under each convention are now identical, placing the group of people as a whole on a knife-edge ready for the bandwagon to roll toward one or the other convention. Now imagine how the knife-edge is disturbed when one convention does not give every individual following it the same chance of being assigned the dominant role (that is, when r and s are not the same for all π-persons and φ-persons respectively).

Of course, for the population as a whole there will always be a $1\text{-}r = 1\text{-}s$ chance of being given the dominant role under each convention at any one time, as the per capita expected return to the followers of each convention is still identical. Nonetheless the distribution of that return can vary across the followers because particular individuals may be assigned to the dominant role more or less often than the group-wide $1\text{-}r = 1\text{-}s$ figure. For instance, under the putative height convention, the shortest person among the population is always assigned to the dominant role while the tallest person is always given the subordinate role. This is captured above through the possibility that an individual's r or s probabilities [see expressions (6.6) and (6.7)] need not be the same as the average group figure.

Thus even though the per capita expected returns have been assumed equal, individuals will be encouraged to switch to the convention offering them personally the higher expected probability of playing the dominant role (for example, the tallest person may adopt the age-based convention). Since the subjective calculation of one's personal r and s values is made difficult by the fact that it depends on who else switches with you, it would be pure serendipity if these rough and ready estimates yielded flows which balanced exactly. The population sits so precariously on the knife-edge that the bandwagon is bound to roll.

6.3.4 The origin of conventions and the challenge to methodological individualism

Social scientists might find it useful to re-interpret the equilibria spawned by multi-dimensional evolution as conventions in the sense of Lewis (1969). What sustains the practice of, say, red players conceding in *Hawk-Dove* (see Section 6.3.2), while blue players take the lot (e.g. p=0, q=1), is simply the players' forecast that this is what will happen. Such predictions become self-fulfilling because, once they are shared, no individual can profit by acting in a manner that contradicts them.

Of course, the opposite prediction is equally self-sustaining (i.e. all players expecting that the reds will dominate the blues; i.e. p=1, q=0) *provided the population held this alternative set of predictions*. Thus the behaviour at each of these (potential) evolutionary equilibria are conventionally determined and, to repeat the earlier point, we can plot the emergence of a particular convention with the use of phase diagrams such as Figure 6.4. In the case of ants and bees, adaptive behaviour is all evolutionary biologists require in explaining the evolutionary dynamics and no talk of convention is necessary. However, when the players are humans, the evolution of behaviour is underpinned by (and gives rise to) an evolving belief system.

Which behavioural (evolutionary) equilibrium will evolve will thus depend both on the presumption that agents *learn* from experience (the rational component of the explanation) and, crucially, on the particular *idiosyncratic* (and non-rational) *features of initial beliefs* and precise learning rules. This is probably EvGT's greatest departure from conventional game theory vis-à-vis the perennial problem with indeterminacy. Rather than proposing solutions through rationality assumptions of increasing complexity (and, some would add, absurdity), EvGT explains why one equilibrium rather than another is selected on the basis of idiosyncratic phenomena in the early stages of a population's evolution; a form of game theoretical Freudianism whereby much of contemporary life depends on a population's early 'childhood'.

When we note that equilibrium selection makes a huge difference to a society's structure and composition (viz. property rights, gender relations etc.), these idiosyncrasies explain why societies might differ so much from one another even if the character of the individuals were, at least initially, identical. The natural question to ponder next concerns what can be said say about the idiosyncratic processes that determine at the early stages of social evolution in which direction (or evolutionary equilibrium) society will turn. Some evolutionary game theorists have appealed to the idea of *prominence* or *salience* to explain the initial direction of evolutionary processes (see Schelling, 1963).

Certain aspects of the social situation just seem to stand out and these become the 'focal points' around which individuals co-ordinate their decisions (see Box 6.9 for some evidence of our surprising capacity to co-ordinate around focal points). Adding a further evolutionary twist, Sugden (1986,1989) argues that conventions spread from one realm (or game) to another *by analogy*. 'Possession' for instance is prominent, or salient, in property games like *Hawk-Dove* with the result that it seems 'natural' to play aggressively (strategy *h*) in some disputed property game now when you seem to 'possess' the property, while non-possession naturally leads to dovish behaviour.

In fact evolutionary biologists lend some support to this particular idea because they find that a prior relationship (rather than size or strength) seems to count in disputes between males over females in the animal world (see Maynard Smith, 1982, and Wilson, 1975). But, they also draw attention to the apparent 'salience' of sex in the natural world as a source of differentiation; so it seems unlikely that a single characteristic can, on its own, explain the emergence of these crucial conventions.

Box 6.9

PROMINENCE AND FOCAL POINTS IN SOCIAL LIFE

Thomas Schelling conducted a series of experiments on his students which reveal a surprising capacity for people to co-ordinate their decisions. As far as formal game theory is concerned the experiments pose in sharp relief the problem of equilibrium selection which we have been discussing, yet it seems people are able, in practice, to solve the problem by finding some aspect of the situation prominent in a way that formal game theory overlooks. Here is a flavour of those early experiments (see Schelling, 1963).

(1) Name 'heads' or 'tails'. If you and your partner name the same, you both win a prize.
36 people chose heads and only 6 chose tails.

(2) You are to meet somebody in New York City. You have not been instructed where to meet; you have no prior understanding with the person on where to meet; and you cannot communicate with each other. You just have to guess where to go.
The majority selected Grand Central Station.

(3) You were told the date but not the hour of the meeting in (2). At what time will you appear?
Virtually everyone selected 12.00 noon.

(4) You are to divide $100 into two piles, labelled A and B. Your partner is to divide another $100 into two piles labelled A and B. If you allot the same amounts to A and B respectively as your partner, each of you gets $100. Otherwise both of you get nothing.
36 out of 41 divided the sum into two piles of $50 each.

As Schelling suggests: "These problems are artificial, but they illustrate the point. People can often concert their intentions or expectations with others if each knows that the other is trying to do the same. Most situations…provide some clue for co-ordinating behaviour, some focal point for each person's expectation of what the other expects him to expect to be expected to do. Finding the key…may depend on imagination more than on logic; it may depend on analogy, precedent, accidental arrangement, symmetry, aesthetic or geometric configuration." (p. 57) See also Mehta et al (1994) for some more recent evidence regarding salience and focal points in pure co-ordination games.

A further and deeper problem with Sugden's concept of salience based on analogy is that the attribution of terms like 'possession' plainly begs the question by presupposing the existence of some sort of property rights in the past. In other words, people already share a convention in the past and this is being used to explain a closely related convention in the present. Thus we have not got to the bottom of the question concerning how people come to hold conventions in the first

place.[25] Indeed, the implicit assumption of prior sharing extends also to shared ways of projecting the past into the present. In this particular instance, the appeal to prior 'possession' relies on what is a probably innocuous sharing of the principle of induction. But, in general, the shared rules of projection are likely to be more complicated because the present situation rarely duplicates the past and so the sharing must involve rules of imaginative projection.

There are two ways of taking this observation. The first is to acknowledge that people actually do come to any social interaction that we might be interested in with a background variety of shared conventions (witness Box 6.9). So, of course, we cannot hope to explain how they actually achieve co-ordination without appealing to those background conventions. In this sense, it would be foolish for social scientists (and game theorists, in particular) to ignore the social context in which individuals play new games.

This, so to speak, is the weak form of acknowledging that individuals are socially located and if we leave it at that then it will sit only moderately uneasily with the ambitions of game theory, in the sense that game theory must draw on these unexplained features of social context in its own explanations. However, it could also be a source of more fundamental questioning. After all, perhaps the presence of these conventions can only be accounted for by a move towards a Wittgensteinian ontology, in which case mainstream game theory's foundations look decidedly wobbly. To prevent this drift a more robust response is required.

The alternative response is to deny that the appeal to shared prominence or salience involves either an infinite regress or an acknowledgement that individuals are necessarily ontologically social (i.e. to concede the practical point that we all come with a history, but deny that this means methodological individualism is compromised). Along these lines there are at least two ways in which, as an ideal type exercise, one might explain a shared salience in one of two other ways without conceding any ground on methodological individualism. First, salience could be biologically based (and therefore shared) in a certain bias in our perceptual apparatus. This, of course, is always a possibility. However, we doubt that biology can be the whole story because it would not account for the variety of human practices in such games (see Box 6.10).

Secondly a source of prominence could be explained if it emerges from an evolutionary competition between two or more candidate sources of distinction (see Section 6.3.3). This seems a natural route to take (and it is the one taken by Lewis, 1969). It is also of more general interest because there will be many actual settings where an appeal to a shared social context will not unambiguously point to a single source of prominence. However, it reproduces, as we saw in Section 6.3.3, an earlier problem in a different form: namely, that the initial distribution of beliefs (now regarding salience) is crucial in determining which source of salience eventually acquires the allegiance of the population as a whole.

Box 6.10

EATING DINNER

Take candlesticks, the place settings and all the other frippery out of eating dinner in company and ask: What is left? It is not implausible to imagine that what is left is something like a *Hawk-Dove* game. Humans apparently are quite unique as a species in sharing their dinners. By contrast, most species eat 'on the hoof', so to speak, with food only ever taken to another place for consumption when there are immobile young which have to be fed. The point about such dinners is that, once the food is on the table, it becomes a potentially contested resource in exactly the manner captured by the *Hawk-Dove* game. Yet rarely do we observe 'fighting' breaking out over the distribution of food. In practice, as we all know, around the dinner table property rights over the food are established by conventions: what we call 'manners' or more generally the rituals of eating. Or as Visser (1992) so nicely puts it: "…behind every ritual with respect to eating lies a simple concern of each person to be a diner and not a dish."

The emergence of 'manners' here is not unsurprising from an evolutionary perspective. It is the evolutionary equilibrium which our analysis leads us to expect. However, we have no way of predicting how the convention with respect to manners will operate. Salience might be invoked in a non-question-begging way, if it was biologically based. But in this case we should expect similar conventions to arise around the activity of dining between peoples who have very similar perceptual apparatus. Yet, as Visser (1992) marvellously demonstrates, this is not what we observe. The

rituals of eating are richly varied across time and space: from the vestiges of sacrifice, the formalism of the Oxbridge 'high table', to the Malawian perception that westerners value peanuts most highly because they are one of the few foods they hold in both hands.

6.3.5 The politics of mutations: Conventions, inequality and revolt

Evolutionary theory stems from the biologists' interest in processes which acted upon animals, plants and bacteria over millions of years. For their purposes, postulating two *independent* mechanisms, one for producing variety (the *mutation mechanism*) and a separate one for selecting among varieties (the *selection mechanism*), made much sense. When evolutionary theory is transplanted from these biological processes to the social world, and cultural processes become the subject of our study, it is easy to forget that the time frame shrinks from millennia to a few generations. And when we take our models to the laboratory in order to test them on real human participants, the time horizon reduces further to an hour or so. Moreover, unlike genes, ants and bees, humans have a capacity to think about the 'laws' that govern their behaviour. These are all good reasons for re-examining the main assumption of EvGT: that the occurrence of mutations is independent of the selection process.

In this section we concentrate on the mutation mechanism and consider what is the meaning and the nature of mutations in a social context. Is it a good idea to assume (as biologists do) that the two mechanisms (*selection* and *mutation*) are independent? To begin with, it is instructive to compare the notion of 'mutation' in this chapter with that of 'tremble' in previous chapters. Trembles, or perturbations, we introduced in Chapter 3 in order to 'shake up' games featuring multiple equilibria in the hope that some of those equilibria would 'collapse', thus reducing the problem of indeterminacy.

The explanation of these trembles was that they were meaningless, small, random errors of the type that even rational humans are capable of. When they happen, they carry no strategic information, as they are uncorrelated with anything that matters. Thus, they are to be (axiomatically) ignored because the probability that the same person will err again any time soon is infinitesimal. No wonder, the conventional analysis was left largely unchanged by the introduction of trembles except that, in their presence, it became possible for rational agents to discard some Nash equilibria.

Mutations seem, at first glance, similar.[26] A behavioural code is evolutionarily stable (i.e. an EE) if it can be shown to be invasion-proof. However, there is a crucial difference. Whereas conventional game theory's trembles are *hypothesised* deviations,[27] EvGT's mutations are *actual* deviations unfolding in real, historical time.[28] The empirical nature of mutations is what gives EvGT its edge in, for instance, rejecting an assortment of Nash equilibria as evolutionarily unstable. Whereas conventional theory's trembles simply check the resistance of Nash equilibria to behavioural 'noise' occurring in the confines of rational agents' minds, EvGT's mutations strengthen *in practice* certain strategies thus helping them conquer the rest. This is, for example, why EvGT's mutations successfully discard at least one Nash equilibrium in antagonistic Games 2.15 and 2.16 (*Battle-of-the-Sexes* and *Hawk-Dove*).[29]

In our conclusion to Chapter 3 we commented that the trouble with the so-called *Refinement* (of Nash Equilibrium) *Project* was that it offered no theory of trembles. Thus it failed to tackle game theory's nemesis: *indeterminacy*. Trembles were simply presumed to happen with vanishingly small probability. Although EvGT's mutations seem to be carrying more explanatory power than conventional theory's trembles, EvGT runs a risk similar to conventional theory's *Refinement Project* unless it can provide a theory of mutations (rather than assume simply that they occur with small, positive probability).[30] In particular, as suggested by Friedman (1996), EvGT needs to be supplemented with a theory of mutations which allows human agents forward-looking attempts to influence others' behaviour.

This suggestion is as important as it is troubling. It is important because only weak social theories trade on the assumption that their worth depends on *not* being understood, and responded to, by the very humans whose behaviour they seek to explain. Put differently, if EvGT is to mature into a significant social theory, it should abandon biological mechanism and investigate whether the behavioural processes it predicts would still come about when the human agents it models are familiar with EvGT. Of course, if humans caught up in evolutionary processes *do* understand the nature of those processes, surely they will want to behave in a manner that influences the evolutionary process in their favour. In this sense, we need a new concept of evolutionary equilibrium. One that can withstand not only random mutations but also *political mutations*; that is, rational (and thus non-random) 'experiments' at subverting the equilibrium or, equivalently, the established convention.

Of course, the very strength of the evolutionary-equilibrium-cum-convention examined in this chapter was its durability in the presence of mutations – see Box 6.10. Could we not think of mutations as experiments at subversion? Not really. Recall Conditions (C) and (D) from Section 6.2.1 under which the preceding analysis holds. Populations must be continuous and mutations must happen one at a time. In effect, our notion of evolutionary stability (and equilibrium) was built on the premise that mutations are single episodes within a huge population which occur sufficiently infrequently so as to ensure that *behavioural adaptation is faster than the rate at which mutations* (or new behaviour) *is introduced into the population*.

This assumption perhaps poses no problem for biology (given the enormous time span and the non-rationality of genes and ants). However, it is highly problematic when it comes to human communities. At the individual level, Chomsky (1957,1966) demonstrated that the mutations which drive the evolution of the one characteristic separating humans from beasts, language, are certainly not random events. Indeed young children seem to be programmed with a capacity for some linguistic errors but not for others. Meanwhile, at the social level the problem (and beauty) with most societies is that they harbour sizeable minorities which, unhappy with the established order, seek new 'behaviours'. These people are perfectly capable of mixing and matching a whole variety of (rebellious and thus correlated, as opposed to random) mutations in a short space of time. Moreover, in human societies there is no guarantee that the population at large will adapt its behaviour faster than the rate at which 'rebels' (or 'deviants' as it used to be fashionable to call them) experiment with alternative social conventions.

So it seems that the notion of evolutionary equilibrium that we inherited from biology is too brittle to capture the subtleties and richness of human culture and politics. The presumption that the mutation mechanism is apolitical is the root-cause of this brittleness. To their credit, some evolutionary game theorists have understood this well and tried to respond analytically. Foster and Young (1990), for instance, acknowledge that politics is what happens when mutations are co-ordinated into aggregate shocks which test the established conventions. Kandori, Mailath and Rob (1993) examine the impact of rational experimentation in finite and discrete populations. Bergin and Lipman (1996) demonstrate that allowing the mutation probabilities to depend on the current behavioural codes (as opposed to being random and uncorrelated to the present conventions) yields a new type of *Folk Theorem*: almost any conventional behaviour can become disestablished and any alternative may take its place if mutants co-ordinate their mutation probabilities appropriately and in response to the current behavioural conventions. This sounds like a celebration of politics as the practice of shaping a society's mutation probabilities and, eventually, the game. But it also ends all hope that evolutionary theory will be *Indeterminacy*'s deathknell in game theory.

Box 6.10

DISCRIMINATORY CONVENTIONS, INDIVIDUAL DEFIANCE AND COLLECTIVE REVOLT

In the *Hawk-Dove* (Game 2.14) one-dimensional (i.e. symmetrical) evolution yields an EE at which each player behaves aggressively (strategy *h*) with probability 1/3. Thus, each gets an average per round pay-off of 2/3. With two dimensions (i.e. two types of player) we saw how (Section 6.3.1) a discriminatory convention has one group of players always playing *h* and the rest always *d*. Thus, the same outcome occurs every time (*h*,*d*) with the result that half the population collect pay-off 2 and the other half 0 making for an average pay-off of 1. This is the rationale of discrimination in these antagonistic games: Though it condemns half the population to a consistent pay-off of zero, average pay-offs rise from 2/3 (without discrimination) to 1. [Note that this would also be the case if everyone played *d*. However, dovish behaviour is inconsistent with the evolutionary stability in games like this.] Can those disadvantaged by discrimination do anything about it? Even though they would be better off if they could wreck the discriminatory convention (thus raising *their* average pay-offs from 0 to 2/3), they cannot do so individually. If a lone disadvantaged player plays *h* all the time, this is a mutation to which the discriminating convention is immune! However, even though *individual* attempts to buck an established convention are unlikely to succeed, the same is not true when individuals take *collective action*. Indeed when a large number of individuals take common action in pursuit of a new convention then this can tip the individual calculation of what to do for the best in favour of change.

6.3.6 Discriminatory Conventions: A brief synopsis

To summarise this section, evolutionary theory predicts that the slightest asymmetries will engender a convention even though it may not suit everyone, or indeed even if it short-changes the majority. It may be discriminatory, inequitable, non-rational, indeed thoroughly disagreeable, yet some such convention is likely to arise whenever an antagonistic social interaction (like *Hawk-Dove*) is repeated. Which convention emerges will depend on the shared salience of extraneous features of the interaction, initial beliefs and the way that people learn. In more complicated cases, where there is competition between conventions, a convention's chances of success will also depend on its initial number of adherents, on how it distributes the benefits of co-ordination across its followers, and on its ability to skew interactions towards fellow users. In particular, one would not expect a convention which generated relative losers and which confined them to the interactive margins (that is, placed them in a position where they were less likely to interact with their fellow adherents) to last long. Or to put the last point even more simply, where conventions create clear winners and losers, two conventions are more likely to co-exist when communication between followers of different conventions is confined to the winners of both. Finally, to undermine discriminatory conventions, individuals' action stands no chance of success, unless it is part of collective action.

6.4 SOCIAL EVOLUTION: POWER, MORALITY AND HISTORY

6.4.1 Social versus natural selection

Social evolution is likely to be different from the Darwinian natural version in part because of what we have just noted. The 'mutation' mechanisms in the social world are unlikely to be independent of the selection mechanisms. There are also other reasons for doubting that the two are the same. For instance the social conventions and associated beliefs, which will be discussing later in this section, can be passed from one to the next through language, a route that is not available in the natural world. Nevertheless, there is much that EvGT can contribute to debates in social science. We begin by considering how it tempers what has been a popular and powerful projection of a Darwinian insight on to the social world: the idea of the 'selfish gene'. We then turn to matters of power, history and functional explanation in social science.

> Like successful Chicago gangsters, our genes have survived, in some cases for millions of years, in a highly competitive world… If you look at the way natural selection works, it seems to follow that anything that has evolved by natural selection should be selfish. (Dawkins, 1976)

There are two types of challenge to the suggestion that we might at some level be prisoners to our selfish genes. The first is philosophical and turns on the false attribution of intentions to genes (see Box 6.11). Genes do not form concepts like selfishness and so it makes no sense to describe genes as motivated by this or any other concept. A possible defence here is to argue that the idea of 'selfish genes' should be understood metaphorically. It is 'as if' genes were selfish. Against this there are a variety of arguments that *group-interest* is capable of guiding humans quite independently of private-interest. Indeed, this is precisely how anthropologists like Carr-Saunders (1922) and Edwards (1962) made their mark; they reported famously on clusters of humans avoiding overpopulation by doing the opposite of what Dawkins' selfish-gene theory would predict; i.e. *by limiting their own fertility*. In this section we briefly look at work *within* EvGT that lends theoretical support to the argument that *social selection* can be guided not only by individual-interest but also by group-interest.

Box 6.11

CAN GENES BE SELFISH?

Dawkins uses the (by now legendary) allegory of genes 'trying' to propagate themselves. Do genes *try* (to do anything)? Are they purposeful *agents*? Do they really have *interests*? Though it makes a narrator's life easer to let us presume so, philosophically speaking such presumptions are deeply troubling. Consider, for example, the narrative Dawkins would offer to explain the evolution of the giraffe's long neck. It would go like this: *Long necks help giraffes to feed themselves in tall-tree forests and are, thus, tools for indirectly spreading instructions for making more long necks through the spreading of the giraffe's DNA. The* ***point*** *of the evolution of giraffe long necks is to propagate giraffe genes.* However interesting this allegory might sound, it is merely a metaphor which ought not be taken literally. Prioritising genes, and narrating their evolution *as if* they are agents (albeit of the Chicago underworld variety), gives them a moral character which they lack.

To make this point more clearly, consider a reversed narrative which tells the same story from the viewpoint of the giraffe's neck: *Long necks help giraffes survive, propagate their giraffe-DNA and, in this manner, spread the DNA instructions for the creation of more long necks. The* ***point*** *of the evolution of giraffe genes is to propagate long necks.*[31] As Jerry Fodor (1996) has pointed out, both metaphors above are nonsense. Giraffes' necks, peacocks' beaks, genes and DNA have no point view, selfish or unselfish. "All that happens…" Fodor reminds us is that, "microscopic variations cause macroscopic effects, as an indirect consequence of which sometimes the variants proliferate and sometimes they don't. That's all there is: there is a lot of 'because' out there, but there isn't any 'for'."

The message for social theory is: Beware biologists (or social theorists uncritically influenced by them) assigning 'interests', 'characters' and 'motives' to genes, strategies and other components of evolutionary processes!

EvGT suffers from two major drawbacks of: (a) Its lack of a proper account of how mutations occur in a social context (see the previous section), and (b) its exclusive reliance on 2-person interactions.[32] A theory built on bilateral games can appear rather amateurish as a model of multilateral relations between inherently gregarious humans. Unless the evolutionary approach extends to games in which N persons interact at once, its claims to understanding social processes will be correspondingly weakened.

In recent years, evolutionary game theorists have made some progress by simulating such social games (for a survey see Bergstrom, 2002). Their better known results come out of the so-called *haystack models* (see Maynard Smith, 1964, for an early effort). Imagine individuals (e.g. mice, men) living in a finite number of 'haystacks' and suppose that each period is divided in two phases: The *reproduction phase* and the *dispersion phase*. During the reproduction phase the N_i individuals within haystack i interact with one another in the context of a repeated N_i-person game which lasts T periods; e.g. a free-rider (or N_i–person prisoner's dilemma) game. The cumulative pay-offs of each of the N_i individuals determines her reproduction rate (or evolutionary fitness); i.e. the number of one's descendants depends on one's behaviour and on the proportion of the haystack population that behaved in similar fashion.

Once the T-periods of the *reproduction phase* are over, the *dispersion phase* begins and individuals can migrate from one haystack to another. If this mass migration (or dispersal) is random, and T is short, then the evolutionary process is not affected by the fact that individuals are confined to specific haystacks for a few periods (T) at a time: Co-operative individuals will still suffer lower evolutionary fitness and natural selection favours Dawkins' selfishness. The first departure from this standard result is observed when T is large; i.e. haystacks stay isolated from one another for lengthy periods. Such confinement causes co-operation to become viable.

To see this more clearly, let us specify the interaction between members of a given haystack during the reproduction phase: Each individual belonging to haystack i must choose either to make a contribution to their haystack (at private cost c) or defect (i.e. act selfishly at no cost to the defector). Let x be the proportion of the N_i individuals in haystack i who co-operate and suppose that the (per period) benefits from co-operation equal bx for each individual, regardless of whether she co-operated or defected. Put simply, in each period a co-operator collects pay-off $bx–c$ and a defector pay-off bx. As long as $b/N<c$ and the *reproduction phase* is relatively short (i.e. T is small), defectors enjoy higher pay-offs, reproduce faster, and defection dominates co-operation in the haystack's phenotype.[33] Later, they migrate to other haystacks randomly and infect any co-operative cluster accordingly until co-operative mutants become extinct and Dawkins' *Iron Rule of Selfishness* is established.

However, when T is sufficiently large (and as long as $b>c$) this result changes. Because haystacks stay together for lengthy periods, residents of haystacks in which co-operation rules earn a serious evolutionary advantage over those inhabiting haystacks characterised by all-round defection. Of course, at the dispersion phase it takes only one defector to enter a co-operative haystack to turn it into a haystack of defectors. However, as long as the defectors are few, they shall *not* survive despite receiving higher pay-offs relative to co-operators in their path! The reason is simple (see Cohen and Eshel, 1976): By invading a co-operative haystack, defectors boost their own pay-offs in relation to their new neighbours *but bring crashing down the average fitness of their new group*. In this sense, defectors do a little better than the co-operators they infect but much, much worse than the residents of haystacks unblemished by a defecting mutant. The lengthy separation of one haystack from the next ensures that social selection of co-operation is a strong possibility.

This fascinating result turns on the co-existence of two simultaneous, and often contradictory, evolutionary contests: One *within* the invaded co-operative haystacks, and one *among* haystacks. In the first contest, the defector-mutant boosts her pay-offs vis-à-vis the co-operative folk of the haystack she invaded. However, she will only enjoy this higher pay-off for a

short while (until, that is, all other haystack residents start defecting too). So, all co-operative haystacks invaded by defectors will cease to be co-operative. The result of this is that the longer the life of haystacks (*T*) the larger the relative evolutionary fitness of un-invaded co-operative haystacks over invaded ones.

Furthermore, if migration is *not* random, co-operative haystacks do even better against haystacks populated by defectors. The simple reason is that co-operators select other co-operators as their neighbours (or family) and do better than the rest both as groups and as individuals. With non-random migration, the cost of co-operation is recouped in terms of a higher probability of landing in a neighbourhood of other co-operators. Of course this does not mean that co-operation is assured in all places and at all times. It is not hard to show that evolution gives rise to co-existence between co-operative groups and groups made up of defectors. The intuition here is that, while mutant-defectors may die out whenever they venture into areas dominated by co-operative groups, defection may remain prevalent in other areas provided a mutant-co-operator reproduces less rapidly in a group of defectors.

Additional insights are forthcoming when we locate groups geographically; that is, when we introduce a spatial dimension to the analysis. Suppose, for example, that individuals are located on a circle (e.g. they live around a lake) and repeatedly interact (in the context of a free-rider problem) with their neighbours, copying the relatively more successful strategy. Bergstrom and Stark (1993) show that, in an EvGT context, stable co-operative clusters emerge comprising more than three players. By contrast, defection comes in clusters of more than two. By changing the game's rules slightly, the authors manage to show that co-operative waves will move along the circle, ensuring that at any fixed point on the circle, periods of co-operation are succeeded by periods of defection.[34]

Eshel, Samuelson and Shaked (1998) complicate the model a little by allowing for behaviour to evolve in the presence of random mutations. Starting with widespread defection, the appearance of a single mutant-co-operator in one place alone can give rise to a small but sustainable string of altruists. Interestingly, the opposite does not hold! Starting with widespread co-operation, mutant-defectors will also engender small pockets of egotism. Nevertheless, these pockets will not survive if they are located too close together because they have a tendency to destroy one another. This means that, whereas the proportion of co-operators faces no upper bound, the number of defectors cannot grow beyond a certain limit.

In summary, evolutionary models cut both ways. They favour the notion that no one can argue with 'success' but, on the other hand, they engender a richer notion of 'success'; one that includes the benefits to a group and thus the potential evolutionary fitness of non-selfish behaviour.

6.4.2 Conventions as covert social power

In 1795 Condorcet wrote:

> …force cannot, like opinion, endure for long unless the tyrant extends his empire far enough afield to hide from the people, whom he divides and rules, the secret that real power lies not with the oppressors but with the oppressed. (1979, p.30)

Runciman (1989) is a recent work in social theory to place evolutionary processes at the heart of social analysis and we aim to give some indication of how EvGT can be used for this purpose. We do so by focusing more narrowly and briefly on the relation between evolutionary processes and the debates in social science regarding power, history and functional explanations. We begin with the concept of power; that is, the ability to secure outcomes which favour one's interests when they clash in some situation with the interests of another.

It is common in discussions of power to distinguish between the overt and the covert exercise of power. Thus, for instance, Lukes (1974) distinguishes three dimensions of power. There

is the power that is exercised in the political or the economic arena where individuals, or firms, institutions, etc., are able to secure decisions which favour their interests over others quite overtly. This is the *overt exercise of power* along the first dimension. In addition, there is the more *covert power* that comes from keeping certain items off the political agenda. Some things simply do not get discussed in the political arena and in this way the status quo persists. Yet the status quo advantages some rather than others and so this privileging of the status quo by keeping certain issues off the political agenda is the second dimension of power. Finally, there is the even more covert power that comes from being able to *mould the preferences and the beliefs of others* so that a conflict of interest is not even latently present.

The first dimension of power is quite uncontentious and we see it in operation, in fact, whenever the State intervenes. In these cases, there will be political haggling between groups and issues will get settled in favour of some groups rather than others. Power is palpable and demonstrable in a way that it is not when exercised covertly. Not unsurprisingly, the idea of the covert exercise of power is more controversial. It is interesting, however, that the analysis of 'spontaneous order' developed in this chapter suggests how the more covert form of power might be grounded. Indeed, and perhaps somewhat ironically, it is precisely because we can see that active State intervention and 'spontaneous orders' are in some respects alternative ways of generating social outcomes. Evidently, both involve the settling of (potential) conflicts of interest. In short, just as we have seen that the State does not have to intervene to create an order, because order can arise 'spontaneously', so we can see that power relations do not have to be exercised overtly because they too can arise 'spontaneously'.

To see this point in more detail, return to the *Hawk-Dove* property game. There is a variety of conventions which might emerge in the course of the evolutionary play of the game. Each of them will create an order and, as we have seen, it is quite likely that each convention will distribute the benefits which arise from clear property rights differently across the population. In this sense, there is a conflict of interest between different groups of the population which surfaces over the selection of the convention. Of course, if the State were consciously to select a convention in these circumstances, then we might observe the kind of political haggling associated with the overt exercise of power. Naturally, when a convention emerges spontaneously, we do not observe this because there is no arena for the haggling to occur, yet the emergence of a convention is no less decisive than a conscious political resolution in resolving the conflict of interest.

EvGT also helps reveal the part played by beliefs, especially the beliefs of the subordinate group, in securing the power of the dominant group (a point, for example, which is central to Gramsci's notion of hegemony and Hart's contention that the power of the law requires voluntary co-operation). In evolutionary games, it is the collectivity of beliefs, as encoded in a convention, which is crucial in sustaining the convention and with it the associated distribution of power. Nevertheless, we can see how it is that under the convention *the-advantaged-will-not-concede*, the beliefs of the 'disadvantaged' make it instrumentally rational for them to concede their claims.

The figure of Spartacus captured imaginations over the ages, not so much because of his military antics, but because he personified the possibility of liberating the slaves from the beliefs which sustained their subjugation. This is especially interesting because it connects with this analysis and offers a different metaphor for power. This is scarcely power in the sense of the power of waves, wind, hammers and the like to cause physical changes. Rather, this is the power which works through the mind and which depends for its influence on the involvement or agreement of large numbers of the population (again connecting with the earlier observation about the force of collective action).

In conclusion, beliefs (in the form of expectations about what others will do) are an essential part of a particular convention in the analysis of 'spontaneous order' and they will mobilise power along Lukes's second dimension. The role of beliefs in this regard is not the same as Lukes's third dimension. In comparison, Lukes's third dimension of power operates with respect

to the substantive character of the beliefs: that is, what people hold to be substantively in their interest (in our context this means the game's pay-offs) or what they regard as their legitimate claims and so on. At first glance the evolutionary analysis of repeated games will not seem to have much relevance for this aspect of power since the pay-offs are taken as given; but there is one which we develop next.

6.4.3 The evolution of predictions into moral beliefs: Hume on morality

Aristotle (1987) wrote in *Nicomachean Ethics* that

> moral virtue comes about as a result of habit… From this fact it is plain that none of the moral virtues arises in us by nature; for nothing that exists by nature can form a habit contrary to its nature. The stone, for instance, which by nature gravitates downwards, cannot be induced through custom to move upwards, not even when we try to train it… Neither by nature, then, nor contrary to nature do the virtues arise in us; rather we are furnished by nature with a capacity for receiving them, and are perfected in them through custom.

The idea that the virtues are not divinely conferred, but rather evolve haphazardly through custom, was too revolutionary for the *Dark Ages* that followed Aristotle's times and caused us almost to lose his writings. Only during the *Enlightenment* did this theme resurface energetically. A major part in its resurgence was played by David Hume and his famous account of justice as an artificial rather than a natural virtue. "All reasonings" he writes in connection to his theory of knowledge, "are nothing but the effects of custom." Children learn self-restraint in the same manner they learn how to walk: gradually and by mimesis.

Hume's account begins with an analysis of the *Powerlessness of Reason*; that is, of the problem of *Indeterminacy* to which we have been returning in this book. It is *because* reason is frequently impotent in providing practical guidance, claims Hume, that custom emerges as the principal determinant of behaviour. Thus, conventions owe their birth and survival to their capacity to lift the veil of uncertainty, to reduce random conflict. To the extent that they have been adopted widely, they create harmony out of otherwise chaotic circumstances and co-ordinate actions in a manner that reason cannot even dream of. (Recall how the discriminatory convention eliminated the *hh* outcome in *Hawk-Dove*; see Section 6.3.1).

At this phase of behavioural evolution, something extraordinary happens, according to Hume: *Mere conventions annex virtue to themselves and so become norms of justice.* We learn not only to *predict* that others will follow the established convention but, additionally, we *expect of them* to do so. Indeed, when they fail to do so, we are filled with moral indignity at behaviour "prejudicial to human society". Our *predictions* (or calculative beliefs) vis-à-vis others' behaviour have become *normative, or moral, expectations*. In Hume's (1740,1888) own words, at some point, the 'is' and the 'will' become a 'must' or an 'ought': "…when of a sudden I am surprised to find, that instead of the usual copulations of propositions, *is* and *is not*, I meet with no proposition that is not connected with an *ought* or an *ought not*."

Why and how does this deontology emerge? Sugden (1986,1989) takes a neo-Humean perspective in which moral beliefs emerge from playing evolutionary games because, the moment a convention turns moral, it gathers additional resistance to mutations. Put simply, a convention that makes us not only predict that we shall all adopt a certain behaviour but, also, that *we ought to*, is far less susceptible to mutations.[35] And since robust conventions minimise conflict and enhance benefits *on average*, morality is an illusion functional to the individuals' petty interests.

In contrast to Kant who thinks that "the majesty of duty has nothing to do with the enjoyment of life" (*Critique of Practical Reason*, 1855), Hume's disciples[36] see morality as the reification of conventions whose *raison d'etre* is to co-ordinate behaviours to some equilibrium devoid of waste and conflict.[37] They also see norms of justice in the same light; namely, as

conventions that imbue people with expectations of what is 'right', or 'just', and what is plainly 'wrong'. At the political level, this conversion of predictions to ethical beliefs gives rise to the notion of the 'common good' – which is, in this account, another illusion brought on by the observation that convention-following brings greater average benefits (unequally of course).

This last thought worries Sugden (1986). For he recognises the paradox in Hume's thought regarding the social utility of conventions. How can it be that the 'common good' is an illusion while, at the same time, proclaiming that conventions are good for society? Sugden, doubts along with Hayek and Nozick, that there is such a thing as 'society' (recall Margaret Thatcher's infamous line) which has 'interests' by which we can judge any convention – the 'myth of social justice', in the lingua of the *Intransigent Right*. There are only individuals pursuing their own diverse goals, doubtless informed by a variety of views of the good.

This worry deepens when we observe (recall Section 6.3) that, in heterogeneous populations, conventions do not operate in the interest of all. Sugden thus argues differently that the moral sense of 'ought' which we attach to the use of a convention comes partially from sympathy that we directly feel for those who suffer when a convention is not followed and partially because we fear that the person who breaches the convention with others may also breach it with us when we may have direct dealings at some later date. This, Sugden believes, is sufficient to explain why individuals have an interest in the observance of a convention in dealings which do not directly affect them.

There is another line of argument which is open to his position. The annexing of virtue can happen as a result of well-recognised patterns of cognition. Recall Box 6.7 on winning streaks: People, it seems, are very unhappy with events which have no obvious explanation or validation, with the result that they seek out reasons even when there are none. The prevailing pattern of property rights may be exactly a case in point. There is no obvious reason that explains why they are the way they are and since they distribute benefits in very particular ways, it would be natural to adjust moral beliefs in such a way that they can be used to 'explain' the occurrence of those property rights.

Of course, like all theories of cognitive dissonance removal,[38] this story begs the question of whether the adjustment of beliefs can do the trick once one knows that the beliefs have been adjusted for the purpose. Nevertheless, there seem to be plenty of examples of dissonance removal in this fashion, which suggest this problem is frequently overcome. Thus, whichever argument is preferred, moral beliefs become endogenous and we have an account of power in the playing of evolutionary games which encompasses Lukes's third dimension.

Box 6.12

MORAL BELIEFS IN THE LABORATORY

In our recent experiment (recall Section 6.3.2) subjects clearly developed moral expectations of one another. Once the discriminatory convention was established and players of one colour (red or blue) became advantaged, they started believing that the disadvantaged players ought to accept their 'lot'. To test this hypothesis, we programmed the computer to lie to 'advantaged' players in the last round of one of the sessions. The 'lie' constituted telling them that their disadvantaged opponent played aggressively (strategy *h*) – when they had not. Some of the advantaged players exclaimed loudly, even rising from their seats angrily, looking around for those disadvantaged players who dared challenge them (and their 'authority'). Their expression made it obvious that they were not just surprised; they were indeed indignant.

6.4.4 Gender, class and functionalism

Our final illustration of how EvGT might help sharpen our understanding of debates around power in the social sciences relates to the question of how gender and race power relations arise and persist. The persistence of these power imbalances is a puzzle to some. Becker (1976), for instance, argues that gender and racial discrimination are unlikely to persist because it is not in the interest of

profit maximising employers to undervalue the talents of women or black workers. Those who correctly appreciate the talents of these workers, so the argument goes, will profit and so drive out of business the discriminating employers. On first reading the point may seem convincing. However, the persistence of gender and race inequalities tells a different story and evolutionary game theory may provide an explanation of what is wrong with Becker's argument.

For example, suppose sex or race is used as a co-ordinating device to select an equilibrium in some game resembling *Hawk-Dove*. Groups which achieve co-ordination will be favoured as compared with those that do not and yet, as we have seen, once a sexist or racist convention is established, it will not be profitable for an individual employer to overlook the signals of sex and race in such games. Contrary to Becker's suggestion, it would actually be the non-racist and non-sexist employers who suffer in such games because they do not achieve co-ordination.

Of course, one might wonder whether sex or race seem to be plausible sources of differentiation for the conventions which emerge in the actual playing of such games. But it is not difficult to find support for the suggestion. Firstly, there are examples which seem to fit exactly this model of convention embodying power (see Box 6.13). Secondly, the biological evidence is instructive and it does suggest that sex is a frequent source of differentiation in the biological world. The point is that, since an initial differentiation has a capacity to reproduce itself over time through our shared commitment to induction, it would not be surprising to find that an early source of differentiation like sex has evolved into the gender conventions of the present. Thirdly, there is some support from the fact that gender and race inequalities also seem to have associated with them the sorts of beliefs which might be expected of them if they are conventions on Sugden/Hume's account. For example, it is not difficult to find beliefs associated with these inequalities which find 'justice' in the arrangement, usually through appeal to 'natural' differences; and in this way what starts as a difference related to sex or race is spun into the whole baggage of gender or racial differentiation.

Box 6.13

WHO GETS THE BEST JOBS IN WEST VIRGINIA?

Faludi (1992) recounts the case of the American Cyanamid Willow Island plant. This is located in West Virginia where there has traditionally been extreme competition for jobs because the area has one of the highest unemployment rates. Until the 1970s its workforce was predominantly male. Indeed the relatively high paid production lines had apparently never hired a woman until 1973 when the Federal government "put American Cyanamid on notice to open its factory doors to women or face legal action".

Men at the plant were most resistant to the idea, claiming that it was "hard work" and "no place for a woman"; and the personnel officers warned against having to "work midnights with a bunch of horny men". Nevertheless women were hired on to the production line under Federal pressure and the men complained: for instance, "Women shouldn't be in here working, taking jobs away from men." One woman worker was told: "if you were my wife, you'd be home darning my socks and making my dinner." The foreman complained that "women were a safety risk because they could get [a] teat caught in the centre feed". And so on. Faludi continues with the story: "As the women's numbers mounted, so did the reprisals. One day the women arrived at work to find this greeting stencilled into a beam over the production floor: SHOOT A WOMAN, SAVE A JOB. Another day, the women found signs tacked to their lockers, calling them whores… in two separate incidents, women fended off sexual assaults." (p. 480) "In 1976 the plant abruptly stopped hiring women. That same year back at headquarters, company executives decided to develop a foetal protection policy. American Cyanamid had never demonstrated a strong desire to protect factory workers in the past… Suddenly though management was worried about the reproductive hazards in the factory… Dr Robert Clyne quickly drafted a policy statement that would prohibit all women of child bearing age from working in production jobs that exposed them to any of twenty nine chemicals… Clyne did not consider reproductive hazards for men." (p. 482)

Two women stayed in the production department by complying with the regulations through sterilisation operations. When they returned to work they were branded: "Men in the department jeered that the women had been spayed…The management's attitude was little better: its own literature referred to the women as neutered" (p.486). Eventually they were laid-off in the early 1980s.

Finally, in so far as this analysis of gender and racial stratification does hold some water, then it would make sense of the exercises in consciousness-raising which have been associated with the Women's Movement and various Black movements. On this account of power-through-the-working-of-convention, the ideological battle aimed at persuading people not to think of themselves as subordinate is half the battle because these beliefs are part of the way that power is mobilised (recall Section 6.3.5). In other words, let us assume that consciousness-raising political activity is a reasonable response to gender/race inequality. What account of power would make such action intelligible? The account which has power working through the operation of convention is one such account and we take this as further support for the hypothesis.

The relation between class and gender/racial stratification is another issue which concerns social theorists (particularly Marxists and Feminists) and again an evolutionary analysis of this chapter offers a novel angle on the debate. Return to the *Hawk-Dove* game, and recover the interpretation of the game as a dispute over property rights. Once a convention is established in this game, a set of property relations are also established. Hence the convention could encode a set of class relations for this game because it will, in effect, indicate who owns what and some may end up owning rather a lot when others own scarcely anything.

However, as we have seen in Sections 6.3.2 and 6.3.3, a convention of this sort will only emerge once the game is played asymmetrically and this requires an appeal to some piece of extraneous information like sex or age or race, etc. In short, the creation of private property relations from the repeated play of these games depends on the use of some other asymmetry and so it is actually impossible to imagine a situation of pure class relations, as they could never emerge from an evolutionary historical process. Or to put this slightly differently: asymmetries always go in twos!

This understanding of the relation has further interesting implications. For instance, an attack on gender stratification is in part an attack on class stratification and vice versa. Likewise, however, it would be wrong to imagine that the attack on either if successful would spell the end of the other. For example, the attack on gender stratification may leave class stratification bereft of its complement, but so long as there are other asymmetries which can attach to capital then the class stratification will be capable of surviving.

Of course, these suggestions are no more than indicators of how the analysis of evolutionary games might sharpen some debates in social theory. We end with one further illustration (again in outline) of this potential contribution. It comes from the connection between this evolutionary analysis and so-called *functional explanations* (see Box 3.9). In effect, the explanation of gender and racial inequalities using this evolutionary model is an example of functional argument. The difference between men and women, or between whites and blacks, has no merit in the sense that it does not explain why the differentiation persists. The differentiation has the unintended consequence of helping the population to co-ordinate its decision-making in settings where there are benefits from co-ordination. It is this function of helping the population to select an equilibrium, in a situation which would otherwise suffer from the confusion of multiple equilibria, which explains the persistence of the differentiation.

Noticing this connection is helpful because functional explanations have been strongly criticised by Elster (1982,1986b) – see Box 3.9. In particular, he has argued that most functionalist arguments in social science (and particularly those in the Marxist tradition) fail to convince because they do not fill in how the unintended consequences of the action help promote the activity which is responsible for this set of unintended consequences. There has to be a *feedback mechanism*: that is, something akin to the principle of natural selection in biology which is capable of explaining behaviours by their 'success' and not by their 'intentions'.

The feedback mechanism, however, is present in this analysis and it arises because there is learning-through-adaptation. It is the assumption that people shift towards practices which secure better outcomes (without knowing quite why the practice works for the best) which is the feedback

mechanism responsible for selection of practices. Thus in the debate over functional explanation, the analysis of evolutionary games lends support to van Parijs's (1982) argument that 'learning' might supply the general feedback mechanism for the social sciences which will license functional explanations in exactly the same way as natural selection does in the biological sciences.

Box 6.14

EVOLVING DISCRIMINATION IN ARTIFICIAL SOCIETIES

In 1825 J.S. Mill delivered a speech at the London Union Society, parts of which sound to us today like a programmatic statement on instrumental rationality and all social theory based on it: "…inasmuch that if you know what is a man's interest to do, you can make a pretty good guess of what he will do."[39] Evolutionary models warn us against such bravado. They demonstrate how hard it is to discern individual motives from the social outcomes they bring about. And they show that knowing the individuals' intent does not help us predict the social outcome. Schellling's (1969) simulation of segregated neighbourhoods is a good example. Imagine a large chequered board with each square occupied either by a blue or a red chip. Suppose that the initial distribution is random and that there are an equal number of blue and red chips. Schelling wondered what would happen if chips had the following simple, non-'racist' motive and they moved about the board in response to it: Of its six 'neighbours', a chip would 'want' a minimum of two chips to be of the same colour as itself.

Note that Schelling's chips are in no sense 'racist' (or enthocentric); they would all be perfectly content in an integrated neighbourhood, consistent even with a positive valuation of 'diversity'. All they 'wish' for is that one third of their neighbours are like them. To find out what would happen, Schelling set a simple evolutionary process into motion such that chips who had fewer than two neighbours of the same colour as their own would move to another part of the board. At first Schelling used real chips on a real board which he moved around manually. His initial results astonished him and he took to the computer for confirmation and further research. The result was the same: The reds gravitated to their own neighbourhood and so did the blues. Soon the segregation was complete, even though the 'agents' were far from being 'racist'. (See also Schelling 1971a,1971b,1978)

Epstein and Axtell (1996) report on another famous simulation (*Sugarscape*) in which a tribe of 'hunter-gatherers' move around a fictitious landscape looking for, amassing, and consuming a single substance: Sugar. Even when they assumed identical 'agents', an unequal distribution of sugar evolved. Lastly, Ross Hammond built a remarkable simulation of social corruption (see Rauch, 2002, for an account). In his model there are two types of agents: citizens and bureaucrats. Each agent has a different proclivity to corruption and her own network of acquaintances. The interaction between a citizen and bureaucrat leads to a corrupt deal if both co-operate, to a report of corruption if only one makes a corrupt 'move', and mutual honesty if both shun corruption. Pay-offs are similar to the *Prisoner's Dilemma* (Game 2.18) or the *Stag-Hunt* (Game 2.16) variety, depending on the proclivities of the specific players.

Interestingly, Hammond assumed that individual has only a vague knowledge of how many reports of corruption it takes before she is arrested. She only knows what has happened within her network of acquaintances and, of course, to herself. A wave of arrests within her social circle is the only sign she gets of a crackdown by authorities. Initially agents are randomly distributed and corruption soon becomes the norm. However, as the simulation proceeds, a random spurt of arrests destabilises the norm and suddenly every agents becomes upstanding. A new norm of honesty is, thus, established. Does it last? No, it does not. The switch from one norm to the other proves inescapable but its timing and momentum are impossible to predict. Indeed, no two simulations produce the same patterns.

6.4.5 The evolution of predictions into ideology: Marx against morality

Karl Marx was among the first thinkers to have engaged critically, but also enthusiastically, with Darwinism's connection to social theory. Indeed, he wanted to dedicate *Das Kapital* to Darwin. Although the dedication was thwarted (by a certain lack of enthusiasm on Darwin's part), the intellectual connection seems to have persisted until the end. At Marx's graveside, Friedrich Engels clearly thought that his dear friend would be gratified by the analogy between his achievements in social science and Darwin's contribution to biology. So, what would Marx have to say about EvGT and the discussion so far in this chapter?

Marx would be fascinated by the results concerning the evolution of hierarchies among virtually identical agents. His conviction that systematic social stratification requires no systematic difference in human ability or character would resonate nicely with the models in Section 6.3. Marx

would be even more impressed by the discussion of how discriminating conventions enhance their evolutionary stability by infecting the individuals' moral beliefs. Indeed, Marx spent much of his working life explaining how people seek, or invent, 'moral principles' and 'virtues' that justify the actual social conventions they labour under on the basis that they are 'just', 'fair' or some such. 'Ideology' was the term he used to describe this melange of normative, or moral, beliefs.

In this sense, Marx would join Hume against the idealists: thinkers like Plato or Kant who like to look for the origin of morals in some realm of ideas independent of material conditions. EvGT's characterisation of virtues as *conventions reified in the process of gaining the requisite evolutionary fitness* would have met with Marx's approval. Marx, after all, led a campaign within the trades union movement *against* the adoption of the slogan 'A Fair's Day Wage For A Fair Day's Work" (see his *Wages, Prices and Profit* in Marx and Engels, 1979). The reason was a strongly held view, similar to Hume's and EvGT's, that people's perception of what is 'fair' and 'right' has little to do with some universal ideal but is tied up with the social conventions that emerge out of the specific social 'game' they play. So, if trades unionists do not like the 'capitalist game', it is hypocritical to adopt the moral codes that it has spawned.

Indeed, EvGT reveals several insights with respect to social life which sound quite like observations that Marxists might make: the importance of taking collective action if one wants to change a convention; how power can be covertly exercised; how beliefs (particularly moral beliefs) may become endogenous to the conventions we follow; how property relations might develop functionally; and so on.

So the major similarity between Marxism and the neo-Humean strand of EvGT is that both see morals as illusory beliefs which are successful only as long as they remain illusory. From there onwards, however, the two traditions diverge. Humeans think that such illusions play a positive role (in providing the 'cement' which keeps society together) in relation to the common good. So do neo-Humeans (like Sugden) who are, of course, less confident that invocation of the 'common good' is a good idea (see the Section 6.4.3) but who are still happy to see conventions (because of the order they bring) become entrenched in social life even if this is achieved with the help of a few moral 'illusions'.

In contrast, Marx insists that moral illusions are *never* a good idea (indeed he dislikes all illusions). No one can be free unless everyone is liberated from illusory moral beliefs, from what he called 'false consciousness'. Following in the footsteps of Ancient Greek philosophy, Marx identifies the Good Life, *eudaimonia* as Socrates and Aristotle called it, with the a multi-faceted flourishing of the person's historically-bred capacities. It is this notion of *eudaimonia* that a divided society poisons, with morality being its most lethal potion. To gauge the perspective from which Marx would view EvGT, we must take things from the beginning.

In humanity's beginning, there was *primitive accumulation*; namely, the daily toil of hunter-gathering. This was no romantic state-of-nature with savages roaming around, exercising their freedom from State and Law. Instead, rigid social hierarchies governed the distribution of gains and burdens in a manner that EvGT has unveiled brilliantly (see section 6.3). But as geographical and climatological conditions necessitated more co-operative patterns of primitive accumulation (e.g. nomadic or collective hunting), these hierarchical conventions spread by analogy from the realm of *Hawk-Dove*-like interactions to the way in which collective production was privately appropriated. To the extent that the community's evolutionary fitness was intimately linked with the solidity of those conventions (recall Cephu from Section 5.3), social evolution favoured developments that weakened any tendencies to 'disobey' the established conventions. With the evolution of human language, around 100,000 years ago, the emergence of concomitant ethical beliefs was made possible.

Humanity's *Great Leap Forward* came with the development of farming which put us on the path of *socialised production*, *organised armies* (for the protection and/or appropriation of stockpiled food), *bureaucracies* (for the organisation of collective effort and the distribution of the

resulting surplus), *writing* (for the purposes of book-keeping), the evolution of differential *resistance to new diseases* (leading to the genocide of those without it by those with it; e.g. the tragic fate of Native Americans and Aboriginal Australians), the technological developments that lead to greater capacities to create (e.g. *metal technology* for the manufacture of ploughs) as well as to destroy (technological advances in the development of *weaponry*) etc. However, even before we embarked collectively down that path, we came to it fully equipped with the discriminating conventions developed at the earlier, hunting-gathering stage of socio-economic organisation.

The emergence and subsequent dominance of farming is crucial to our understanding of human societies and the power structures underpinning them. As farming was foisted upon *some* tribes by geography, climate and serendipity, hunter-gathering communities were either exterminated by farming ones or survived only by adopting food production themselves (see Diamond, 1996). Either way, the norms of hunting-gathering were transformed radically, though not altogether lost. Food production could suddenly support denser populations and so the conventions had to be able to deal more efficiently with larger numbers of potential absconders in social environments where anonymity (courtesy of the greater population size) posed special threats to conformity.

In short, the new farming techniques (or *means of production* as Marx called them) demanded new social relations. The discovery of the plough meant that work had to be organised co-operatively and the community had to be tied to particular plots of land (as nomadic farming is a contradiction in terms). Due to the ubiquitous scarcity of fertile land, control of it translated directly to control of the surplus produce. By simple deduction, the norms which determined who controlled the land fed into new, analogous norms regarding control of the surplus. With a minor leap of the imagination we can visualise the conversion of relatively primitive norms for distributing stags and hares to complex norms of distributing (a) the work load in the fields, warehouses, barracks etc., and (b) the share of the agricultural production enjoyed by each.

In the pre-farming era, one's power could be gauged by the likelihood of securing one's favourite Nash equilibrium in some *Hawk-Dove*-like game (e.g. the best piece of the jointly captured meat). With socialised food production, however, the epicentre of social power shifted from appropriation-cum-consumption to control over the production process and its surpluses. New, more complex, conventions were functional to such power, spawning ideas of 'right' and 'wrong' that even went as far as to convince the disadvantaged that their rulers were deities. Metaphysical angst turned into organised religion in a manner functional to the exercise of power over the production and distribution of surplus. McPherson (1973) formalises Marx's notion of social control over the labour of others as *extractive power* as follows:

Extractive Power (Definition)

Person A exercises extractive power over person B if:

(a) A can compel B to perform tasks (i.e. work) that A does not wish to perform

(b) Surplus X results from these tasks

(c) A claims and enforces property rights over part of X usually with the help of the collective (or, in more advanced societies, the State)

(d) B would not perform these tasks if in possession of the same options as A in a social setting in which these options are just as asymmetrically distributed

(e) Ethical beliefs (or ideas) evolve which maintain steps (a) to (d).

Extractive power is thus a straightforward extension of asymmetrical conventions for distribution of non-produced goods to a community which produces assets in the context of collective manufacture. In principle, extractive power can emerge in hunter-gatherer communities too; in the sense that some group may develop, theoretically, a capacity to compel others to hunt/gather on their behalf. However, such conventions are less likely to take hold and command a significant

proportion of work effort when individuals have the opportunity to abscond and fend for themselves. The development of the technical skills to grow plants from seeds, as well as the invention of the plough, boosted the surpluses and, at the same time, made access to productive resources more restrictive (e.g. by means of fences around fertile land). The end result was the emergence of wholesale extractive power.

Marx placed great emphasis on the effects of the diminished opportunities for autonomous production on culture and society. As technology improved, these opportunities shrank further and, consequently, the extractive power of those controlling the means of production increased. History is marked, Marx thought, by the 'lumpy' transition from one type of society (or *mode of production*, to use his term) to another, with each shift coming about when some technological innovation (or other material development, e.g. climate change) rendered the previous social conventions of production obsolete. With each new phase, the ideas in men's and women's heads as to what is 'right' and 'proper' changed radically: "All fixed frozen relations, with their train of ancient and venerable prejudices and opinions, are swept away, all new-formed ones become antiquated before they can ossify. All that is solid melts into air, all that is holy is profaned…"[40]

When capitalism burst upon the scene, extractive power was propelled to its apotheosis as access to productive means for the disadvantaged vanished completely. What was it that brought on the transition to capitalism, and the maximisation of the extractive power of the advantaged over the disadvantaged? Gradual technological innovation, is Marx's answer. *Ship-building technology* enabled international trades routes to be established, thus leading to the appreciation in the exchange value of certain commodities (e.g. wool, silk, spices). Those who traded in them acquired economic power over and above their 'social station'. Landlords in Britain were thus encouraged to replace sheep for the peasants which used to work the land (producing crops of little value) and who were suddenly expelled onto the dirt roads and into the fledgling cities. The invention of the *steam engine* allowed the expelled (i.e. those who had no alternative but to sell their labour) to work in confined spaces (factories) with no independent access to prey or crops.

Just as farming had had a huge impact by shifting the centre of social life from the norms of distribution of exogenously generated assets to the norms of distributing land, labour and the resulting output, capitalist production added another crucial complication: the extraction by property owners of the generated output was shifted from the *post* to the *pre* production phase. Rather than collecting by stealth (as feudal lords and slave-drivers used to) part of the output *after* the latter was produced, capitalists paid a retainer for the workers' services *in advance*; a retainer large enough to secure their surrender of *future* time and toil but less than the *expected* value of their labour.

The transition from distributing assets contemporaneously to distributing them intertemporally, against the backdrop of highly asymmetrical extractive power, made the whole process more productive but, at the same time, more reliant on *belief*. The norms of capitalist society had to exude the complexity of its technology. Marx spent much ink describing meticulously the evolution of the *commodity* and of *capital* as phenomena which did not (and could not) exist prior to capitalism's ascendance. As commodity exchange became the exclusive means of survival, the commodity-relation replaced human relations. Capital, i.e. the manufactured means of production controlled by the few, "…was not a thing, but a social relation between persons… Property in money, means of subsistence, machinery, and the other means of production, do not yet stamp a man as a capitalist if there be wanting the correlative – the wage-worker" (*Capital Vol.1*, in Marx and Engels, 1979).

The point here is that the whole gamut of capitalist endeavour is based on particular social relations. If Marx is right and capital is but a relation-of-production (as opposed to some physical 'thing'), then its value is a matter determined by the network of conventions ruling over this relation. These conventions, in turn, reflect the *jointly* evolving technologies and relations of production. Steam engines, mechanical looms and computerised robots are, at once, the secret force

behind splendid productive capacity *and* the midwives of our ideology. Before they were invented, and while state-of-the-art technology was confined to ploughs and sickles, control over production largely remained in the hands of the labourers. It was only *after* the crop came in that the distributional conventions of slave or feudal societies would kick in.

Under capitalism, however, the temporal reversal of residual claims meant that workers lost control over the production process. For the first time in human history the residual claimants paid in advance for the privilege of exercising their extractive power. Given the inherent risks of paying for something in advance, the removal of the cognitive dissonance resulting from considerable social asymmetries was critical in reinforcing the 'evolutionary fitness' of capitalist societies. Those privileged by the new capitalist conventions could legitimise their booty based on the mythical notion of being rewarded for 'risk-taking'. More importantly, those disadvantaged by the same conventions could live with their situation more easily by a combination of normative beliefs shaped by (a) the seemingly symmetrical position of capital and labour ("we receive profit in return for laying out in advance our capital and you receive this capital in return for your labour"), and (b) the soothing impact of formal liberty for all.[41]

The deep invisibility of the social conventions of capitalist production thus played a central role in solidifying both. The resultant dominant ideology is founded on the illusion that observed inequality is not to be explained in terms of the social power of one class or group over the other but, instead, is the result of different abilities, work ethic etc. According to the dominant creed, rather than being capitalists or workers, men or women, blacks or whites, we are all entrepreneurs (even if some have nothing to sell other than their labour or even their bodies). Indeed, mainstream economics, and by association game theory, may be thought of as the highest form of this ideology in the sense that class, gender, race etc. are conspicuous by their absence in their narratives on how the social world functions.

Our world may have never before been so ruthlessly divided along the lines of extractive power between those with and those without access to productive means. And yet never before has the dominant ideology been so successful at convincing most people that there are no systematic social divisions; that the poor are mostly undeserving and that talent and application is all the weak need in order to grow socially powerful. But, even if this is true, what is *really* wrong with a world in which the dominant ideology has made most people accept (and even like) the social underpinning capitalism?

An answer along the lines of a moral judgment about the unfairness of capitalism (based on inequality and the like) was not open to Marx. For he had dismissed moral judgements as illusions functional to the current conventions. Deeply aware of this, and the concomitant need to ground his criticism on something *outside* the current belief system, Marx focussed his indignation on the inefficiency of capitalist social relations. In summary, his critique of capitalism turns on the argument that it represents a transitory phase of human history; one in which the social relations (e.g. the arrangement according to which the set of workers and of owners are, mostly, mutually exclusive) have not evolved sufficiently to take full advantage of the technology available.

As a result of this mismatch, Marx claims, we live in a society which wastes human resources (in the form of chronic and fluctuating unemployment), devalues humanity (by reducing our relations to commodity fetishism) and requires war in order to maintain some degree of compatibility between (a) what the economy can produce and (b) what consumers have the purchasing power to absorb. In short, Marx dismisses angrily the notion that capitalism is efficient but unfair, opting instead for the line that it is grossly wasteful of human capabilities because it is *one evolutionary stage behind* the productive capacity of human-made technologies. If he is right, it is easy to understand his loathing of both bourgeois *and* proletarian moralities: for they are the different sides of the same coin that prevents humanity from achieving its potential.

To the extent that the above is a fair description of Marxian thought, it is now possible to imagine Marx's verdict on EvGT: It is a fine theoretical tool for elucidating, in part, social

evolution in pre-farming societies. But it is utterly ill-equipped to deal even with simple societies (e.g. slavery or feudalism) in which assets are co-operatively manufactured and privately appropriated. EvGT has little to offer the moment the game changes from a simple *Hawk-Dove*-like interaction over given assets to a fully-fledged *N*-person game of individuals who simultaneously produce and distribute assets, as well as the social norms that govern these parallel processes. Moreover, if farming communities are an explanatory bridge too far for EvGT, capitalism is even further away from its grasp since the study of systematic extractive power cannot be elucidated by simple evolutionary models which map out the trajectory of behaviour against the background of a *given game*, with *given rules* and *given pay-offs*.

Just as Marx had the audacity to criticise capitalism for being too primitive, he might have lambasted evolutionary game theorists for being insufficiently… evolutionary, asking:

> How is it that you can explain moral beliefs in materialist terms, but you avoid a materialist explanation of beliefs about what we consider to be our in own interest? If we are capable of having illusions about the former (as you admit), surely we can have some about the latter! If morals are socially manufactured, then so is self-interest.

But is it true that there is a fundamental difference between the method of EvGT and Marx? Or is it just a technical difference that will wane as EvGT is developed further until it meets with Marx's approval? Are social classes and extractive power mere by-products of individual interactions (just as the consequences for the species in EvGT are a by-product of individual interactions)?

Fans of both EvGT and Marx[42] might argue that the former *can* become 'more evolutionary' provided it manages to model the *complete* process of self-interest feeding into moral beliefs and moral beliefs feeding back into self-interest. If EvGT can mature in this manner, could it perhaps disarm those who argue (with Marx) that social relations are primarily (though not deterministically) constitutive of the individual.[43] For them, Marx's invocation of class-determined extractive power, as well as of the analytical categories *capital* and *commodity*, is of major ontological significance and distinguishes his method to that of EvGT. But would that remain the case if EvGT 'matured' in the manner described above?

Modelling historical change as a feedback mechanism between desires and moral beliefs would be too circular for Marx. It would not explain where the process started and where it is going. By contrast, Marx's theory of history reserves a special place for the evolution of technologies as a source of non-random mutations, closely linked to human inventiveness, that help destabilise the prevailing social norms. Especially in his philosophical (as opposed to economic) works, Marx argued strongly for an evolutionary (or more precisely historical) theory of society with a model of human agency which retains human activity as a positive (creative) force at its core.[44]

Two questions remain: How useful is Marx's contribution to the debate on EvGT and, further, how relevant is the latter to those who are engaged in debates around Marxism? Our answer to the first question is that Marx seems aware of the ontological problem to which we keep returning from Chapter 2 onwards: The fact that *Indeterminacy* beckons, unless we adopt a model of human agency richer than the one offered by instrumental rationality. The answer to the second question is trickier and had caused us to disagree as authors in this book's first edition (see p.232 of that edition); it still does.

6.5 CONCLUSION

Evolutionary game theory was greeted with enthusiasm because it offered hope of addressing three concerns about mainstream game theory. Two were theoretical in origin: one related to the model of rational agency employed and the other was the problem of pointing to solutions in the absence of a clear-cut equilibrium. The third arose because game theory has some controversial insights to offer the debate on the role and function of collective agencies (such as the State). Evolutionary game theory has thrown light on all three issues and it is time now to draw up a balance sheet.

On the first two issues, we have found that evolutionary game theory helps explain how a solution comes about in the absence of an apparent unique equilibrium. However, to do so it has to allow for a more complex notion of individual agency. This is not obvious at first. Evolutionary game theory does away with all assumptions about motives and beliefs and, instead, assumes that agents blunder around on a trial and error basis. This learning model, directed as it is instrumentally by pay-offs, may be more realistic but it is not enough to lead unambiguously to some equilibrium outcome. Instead, if we are to explain actual outcomes, individuals must be socially and historically located in a way that they are not in the instrumental model. 'Social' means quite simply that *individuals have to be studied within the context of the social relations within which they live and which generate specific norms*. When this is not enough to explain their current beliefs and expectations then, of course, we have to look to the individual idiosyncrasies and eccentricities (in belief and action) if we are to explain their behaviour.

Thus evolutionary game theory, like mainstream game theory, needs a changed ontology (which will embrace some alternative or expanded model of human agency) if it is to yield explanations and predictions in many of the games which comprise the social world. We have left open the question of what changes are required. Nevertheless, it is entirely possible that the change may make a nonsense of the very way that game theory models social life. For example, suppose the shared sources of extraneous belief which need to be added to either mainstream or evolutionary game theory in one form or another come from the Wittgensteinian move, sketched in Chapter 1. Or, imagine a model in which preferences and beliefs (moral and otherwise) are simultaneous by-products of some social process rooted in the development of organised production – as in Marx's theory in the previous section.

These theoretical moves will threaten to dissolve the distinction between action and structure which lies at the heart of the game theoretical depiction of social life because it will mean that the structure begins to supply reasons for action and not just constraints upon action. On the optimistic side, this might be seen as just another example of how discussions around game theory help to dissolve some of the binary oppositions which have plagued many debates in social science – just as it helped dissolve the opposition between gender and class earlier in this chapter. However, our concern here is not to point to required changes in ontology of a particular sort. The point is that some change is necessary, and that it is likely to threaten the basic approach of game theory to social life. The following chapter takes this point further.

Turning to another dispute, that between *social constructivism* and *spontaneous order* within liberal political theory, two clarifications have occurred. The first is that there can be no presumption that a spontaneous order will deliver outcomes which make everyone better off, or even outcomes which favour most of the population. This would seem to provide ammunition for the social constructivists, but of course it depends on them believing that collective action agencies, like the State, will have sufficient information to distinguish the superior outcomes. Perhaps all that can be said on this matter is that, if you really believe that evolutionary forces will do the best that is possible, it is beyond dispute that these forces have thrown up people who are predisposed to take collective action. Thus it might be argued that our evolutionary superiority as a species derives in part precisely from the fact that we are pro-active through collective action agencies, rather than reactive as we would be under a simple evolutionary scheme.

Secondly, in antagonistic social settings, in which equilibrium selection favours the interests of some at the expense of others (i.e. when there exists no equilibrium which is better for everyone), it is not obvious that a collective action agency (like the State) is any better placed to make this decision than some decentralised process leading to a 'spontaneous order'. This may come as a surprise, since we have spent most of our energy here focusing on the failure of instrumental rationality[45] to dissolve *Indeterminacy*. But the point here is that, just as instrumental rationality cannot promise to guide agents (participating in repeated or evolutionary games) to some collectively desirable equilibrium, it cannot guarantee a successful outcome for collective action either.

To see this, one need only model collective action (or the political process) as an *N*-person bargaining game, since the 'spoils' of collective action must be distributed according to agreed principles. However, we have already demonstrated (recall Chapter 4) that, in such settings, instrumental rationality leads once more to *Indeterminacy*, just as it did in non-co-operatively repeated games (see Chapters 3&5) or evolutionary models (in the present chapter).

In other words, the very debate within liberal political theory over *social constructivism* versus *spontaneous order* is itself unable to come to a resolution precisely because its shared ontological foundations are inadequate for the task of social explanation. In short, we conclude that not only will game theory have to embrace some expanded form of individual agency, if it is to be capable of explaining many social interactions, but also that this is necessary if it is to be useful to the liberal debate over the scope of the State.

Problems

6.1 Explain the rationale behind the one-dimensional evolutionary equilibrium of Game 2.13 using the definition of evolutionary stability in Section 6.1.2.

6.2 Repeat the analysis of asymmetrical evolution in the *Hawk-Dove* game (see Section 6.3.1) when the game under investigation is the *Battle-of-the-Sexes* (Game 2.13).

6.3 Find the evolutionary equilibria in the case of Game 6.3 under (a) one-dimensional evolution (i.e. an homogeneous population) and (b) two-dimensional evolution (i.e. assuming that the population is split in two equally sized sub-populations, each with a distinctive feature).

6.4 Draw the phase diagram of the evolutionary process that corresponds to game below when the population is homogeneous:

		B		
		C1	C2	C3
A	R1	1,1	-1,-1	2,0
	R2	-1,-1	1,1	0,0
	R3	0,2	0,0	1,1

NOTES

[1] Malthus was concerned that human population grew geometrically while food production could only grow arithmetically. If so, a struggle for existence would occur as increasing numbers of people would have to starve. Darwin (1860) was clearly impressed by this. In his own words: "In the next chapter the Struggle for Existence amongst all organic beings throughout the world, which inevitably follows from the high geometrical ratio of their increase, will be treated of. This is the doctrine of Malthus applied to the whole animal and vegetable kingdoms" (pp. 4-5).

[2] The whiff of similarly evolutionary critical thinking can be detected also in Albert Einstein's argument in favour of socialism: "The profit motive, in conjunction with competition among capitalists, is responsible for an instability in the accumulation and utilization of capital which leads to increasingly severe depressions. Unlimited competition leads to a huge waste of labour, and to …[the] crippling of the social consciousness of individuals..." Einstein (1949).

[3] To see this, recall that when $p<1/3$ the expected pay-offs from h [i.e. $-2p+2p$] exceed those from d [i.e. $0p+p$]. When this happens, p grows as the proportion of hawks increases. And vice versa. Thus, the proportion of hawks stabilises when p is neither larger nor less than 1/3; i.e. when $p=1/3$.

[4] Expected returns from playing h and s equal $ER(h) = p+2(1-p)$ and $ER(s) = 3(1-p)$ respectively. Value $p=1/2$ equalises these two expected returns and coincides, by definition, with NEMS.

[5] The same applies to the *Hawk-Dove* game in which NEMS is the unique EE (as long as the population is homogenous). The average pay-offs/fitness in this game equal 2/3 (check that when $p=1/3$ the expected returns from either strategy equal 2/3). By contrast, average pay-offs fitness would be higher (equal to 1) if the whole population played d.

[6] For example, all pure strategy Nash equilibria in Games 2.14 (*Hawk-Dove*) and 2.16 (*Stag-Hunt*) are strict equilibria; in the sense that they comprise strategies that are strictly best replies to one another. By contrast, none of the NEMS are strict (by construction): each NEMS is based on the premise that players are indifferent between a number of pure strategies comprising the NEMS (and the probability mix recommended by that NEMS). NEMS are thus weaker equilibria in that they comprise strategies which are as good replies to themselves as others. Eg. In *Hawk-Dove*, NEMS recommends $p=1/3$. However, when both play h with $p=1/3$ no player has a strict preference for any of the strategies since their expected returns from each of their available strategies equal 2/3. The most we can say is that, in a NEMS, there exists no alternative strategy that does better than NEMS. Another example of a non-strict, or weak, Nash equilibrium is (R2,C2) in Game 3.1 (see chapter 3) – this is the case of a Nash equilibrium in pure strategies sustained on weakly dominated stategies.

In addition to Nash equilibria sustained by weakly dominated strategies, there are other Nash equilibria which do not qualify as EE even though they are based on strictly best replies. Recall for instance that in Game 2.14 (*Hawk-Dove*) the pure strategy Nash equilibria are not EE when the game is played within an homogeneous population. This is so in spite of the fact that *hd* and *dh* are based on strict best replies. The reason why they are not EE (as we have seen above) has nothing to do with their 'qualifications' but is merely due to the fact that it is impossible to have one player play h and the other d with certainty when the players are indistinguishable. See Box 6.2.

[7] In doubly symmetrical games (e.g. pure co-ordination or the stag-hunt) behaviour b can spread across the whole population so that each player adopts it *independently of whom she is playing against*. E.g. in *Stag-Hunt* (Game 2.18) behaviour h is not only consistent with a strict Nash equilibrium (*hh*) but it is also possible for all players to adopt it without the need to condition their behaviour on their opponent. Similarly with strict Nash equilibrium *ss*. However, this is not the case in games like *Hawk-Dove* (Game 2.14) in which the strict Nash equilibria (*hd* and *dh*) are asymmetrical in that they require that paired players choose the opposite strategies. Such Nash equilibria can only evolve if it is possible for players to recognise some trait or feature in their opponents so as to co-ordinate their actions; i.e. so that one plays h and the other d in equilibrium (see the analysis of heterogeneous evolution that follows later in this chapter). In an homogeneous population this is not possible (as players are indistinguishable from one another) and thus the asymmetrical Nash equilibria cannot 'evolve' (and therefore cannot become evolutionary equilibria).

[8] Proportion p rises to just above 0, but still $ER(h) = p+2(1-p)<ER(s) = 3(1-p)$. Thus the mutant hare-hunter gains no advantage and dies out.

[9] In NEMS, therefore, the probability that a pair will unite in hunting the stag equals ¼.

[10] By this we mean that were we to draw a diagram with N on the horizontal axis, N is so large and the players so well bunched up, that it would be possible to imagine N to be a continuous variable. This is important analytically because, as we shall see below in the context of so-called *Replicator Dynamics*, we can utilise differential equations to model the convergence toward the game's EE. If, on the other hand, the population were discrete (that is, with gaps separating one person from the next), the evolutionary process might 'fall between these cracks', so to speak, and convergence might be impeded as a result.

[11] Just as in *Hawk-Dove*, NEMS will be shown to be the unique EE in the *Battle-of-the-Sexes* (Game 2.13). And similarly to the *Stag-Hunt*, NEMS is not evolutionarily stable in *Pure Co-ordination* (in which there are, as in the *Stag-*

Hunt, two EE: the game's pure Nash equilibria). *Hide-and-Seek* (Game 2.17) will be discussed in Section 6.3 (because it violates Condition (E). Finally, in the *Prisoner's Dilemma* (Game 2.18) the unique EE coincides with the unique non-co-operative Nash equilibrium. See Problem 6.1.

[12] For readers not deterred by a little mathematics, the simplest depiction of the following dynamic is to posit the rate of change in p as a function of d: $\dot{p} = f(d)$.

[13] NEMS is defined as the probability distribution over pure strategies (i.e. the values of p and 1-p in 2X2 games) such that both players are indifferent between their pure strategies (see Chapter 2). Function d on the other hand is the net gain of the first strategy (relative to the second). When d=0 the net gain is zero and the player is indifferent. Thus, the value of p=p^* setting d=0 is, by construction, the value corresponding to NEMS.

[14] The tit-for-tat strategy begins by co-operating in the first round and then copies the opponent's previous move. Clearly, if both players adopt it, they will co-operate in each round.

[15] It is easy to show that defectors reap higher pay-offs than unconditional co-operators in a population of unconditional co-operators.

[16] Perusing Figure 6.4 we notice that there exists a trajectory which is consistent with NEMS – i.e. its arrows point toward it: It is the 45° line! Nevertheless, this is neither here nor there. For even if we are prepared to fathom the possibility that the population might tread this thin path toward NEMS, any mutation that will occur along the way will 'throw' it either slightly above the 45° line or below it. Once this happens, the replicator dynamic will push the system; either toward p=1,q=0 or toward p=0, q=1.

[17] "The intuition is that a stable mixture of hawks and doves will evolve in a single population, but with two interacting populations, one will become all hawks and the other all doves." Friedman (1996), p.7.

[18] Our choice of colours is not random. Mehta et al (1994) report on a laboratory experiment of the 'name any colour' type which shows that blue and red are, roughly, equally salient. This is important because we wanted to preclude an additional source of salience; e.g. a situation in which *at the very outset* players of one colour (i.e. the one with higher salience) are seen as more likely to play aggressively as those of the other (i.e. the less salient) colour.

[19] This pattern was observed in 15 out of 16 sessions.

[20] If one expects one's opponent to play c, one's best reply is h. This being true for both players, c is not rationally playable. Under commonly known rationality of 1st degree, no one expects an opponent to play c. Therefore, the availability of c should not make any difference (according to conventional game theory).

[21] Recall that an evolutionary equilibrium had to coincide with some Nash equilibrium (though the opposite is not true). Yet, see Camerer and Thaler (1995) who review experimental evidence on the persistence of symmetrical, non-equilibrium outcomes.

[22] EvGT predicts that the moment there are three strategies to choose from per player, evolution is two-dimensional even if the population is heterogeneous. If the population is also divided in two groups, then we have an additional (third) dimension. See Problem 6.4 for the derivation of the evolutionary equilibrium in the two-dimensional evolutionary play of *Hawk-Dove-Co-operate* (i.e. the case with an homogeneous population).

[23] Again there are many political angles here. For instance, Seyla Benhabib (1987) argues against the model of human agency found in methodological individualism by noticing that "the conception of privacy is so enlarged that…relations kinship, friendship, love and sex …come to be viewed as spheres of personal decision making", and so gender discrimination is hidden under a cloak of private preference satisfaction.

[24] To see how the following inequality is arrived at, notice that the condition for ER(π-person) to be an increasing function of p is that its first order derivative must be greater than zero. Differentiating ER(π-person) with respect to p and setting the derivative greater than zero yields inequality (6.8). Similarly for inequality (6.9).

[25] In effect, this was precisely the point that Lewis (1969) was reacting against in the work of Quine. Quine was denying that language arises by convention because *conventions are agreements* and so language could not have originated by agreement since the notation of agreement between people presupposes a shared rudimentary language. Lewis' book is an attempt to show that convention does not presuppose agreement of this type.

[26] The analytical similarity is best explained in terms of Figure 6.2. Take any of the diagrams in it (except that pertaining to Game 2.19). The intersection of function d with the p-axis marks the game's NEMS. All NEMS are trembling hand equilibria – see Chapter 3. In other words, 'shaking up' the game by introducing the possibility of vanishingly small trembles, does *not* knock out the mixed strategy Nash equilibria. However, deviations that take the form of mutations do the trick in some games (though not in all). E.g. they knock out NEMS in Games 2.15 and 2.16 but not in Games 2.13 and 2.14. Why this difference? Are mutations and trembles not analytically indistinguishable deviations from NEMS? No, they are not. Trembles are hypothesised (see next note) in a static framework and could go either way (toward the left or toward the right of NEMS). Mutations, on the other hand, take place in real time and, once they destabilise NEMS (e.g. Game 2.16) they gather strength in *real time* and in the *same* direction, destabilising NEMS and pushing the population toward one of the two pure strategy Nash equilibria (e.g. if the initial mutation takes the population to a p-value above its NEMS value, then p tends inexorably to 1; and vice versa). The only way in which trembles could be made analytically similar to mutations is if we were willing to hypothesise *correlated trembles* in a

static framework; i.e. imagine that players contemplate various errors that they or their opponent could make in selecting their strategy. Now imagine that players assume that if they err on one side of NEMS (i.e. select, accidentally, a *p*-value above its NEMS value, so will other error-prone players). Under such correlated errors, Figure 6.2 would apply to static games with trembles just as it applies to evolutionary play with mutations. The problem with this explanation is twofold: (a) Conventional game theory makes a song and dance about trembles being free of any strategic information (and thus uncorrelated); and (b) It is hard to imagine an explanation as to why players would expect their errors to be correlated.

[27] Recall (see Section 6.1.1) the main allegorical difference between EvGT and conventional game theory: The latter presupposes players who rationally work out the complete strategy in advance, whereas the former assumes that behaviour simply evolves in real time. As we saw in previous chapters, in conventional theory trembles make a difference because rational agents work out their (trembling hand) equilibrium strategies knowing that behaviour may be affected by errors (i.e. trembles). Analytically, the importance of trembles is that it makes all strategies possible (since every player may play, by accident, *any* of her strategies with positive probability). For example, in the *Centipede* games (of Chapter 3) the hypothesis of trembles allows players to combine common knowledge of rationality with backward induction and thus work out their SPNE strategies. If you ask them: "In working out your SPNE strategy, you started at the last node, right?" "Right" they would answer, at which point we may ask again: "But why would you ever *get* to that node at the bottom of the *Centipede*'s tree if, as you claim, it is irrational to play across in any node?" They would then reply: "Because there is always a positive probability that my opponent and I may choose to play across by mistake; as a result of a tremble." This answer is central to the derivation of equilibria such as the SPNE, or sequential equilibria etc. In this sense, we argue here that that trembles are hypothetical in nature.

[28] For this reason, EvGT (as noted in Section 6.2.2) is highly sceptical of *Nash backward induction* (and SPNEs) which, as explained in the previous note, is based on the possibility of a long string of hypothesised mutations (or trembles).

[29] Recall that one-dimensional evolution (homogeneous population) only NEMS was left standing whereas 2-dimensional evolution discarded NEMS, favouring one of the two pure strategy Nash equilibria.

[30] As Mailath (1998) puts it: "Both Refinements [nb. conventional theory's attempt at reducing indeterminacy] and Evolutionary Game Theory were originally motivated by attempts to find the 'right' equilibrium. That hype was not met; and it could not have been met. What has been achieved is a description of the properties of different equilibria."

[31] One might protest that this reversal is illegitimate since genes come prior to long necks. However, on closer inspection this is not really true. Necks are part of the animal's feeding device. Feeding devices and genes have been evolving in tandem from Life's Day One to yield the giraffe's neck and current genetic code. There is no sense in which the latter preceded the former.

[32] All evolutionary models studied so far require a large population whose members repeatedly play 2-person games – recall Condition (A) in Section 6.2.1.

[33] The reader may be confused here: Does the evolutionary play of the prisoner's dilemma (PD for short) rule out co-operation? It depends. In Section 6.2.1 we examined the evolutionary play of the static (or one-sot) PD and showed that, in this setting, co-operation does not stand a chance. By contrast, in Section 6.2.2 we studied the evolutionary play of the indefinitely repeated PD where we found a plethora of potential evolutionary equilibria, some of which was involved extensive co-operation. The difference between the two, the reader will recall, is that the Section 6.2.1 pairs were re-drawn randomly *in each round* whereas in Section 6.2.2 each player would play the PD repeatedly against the *same* opponent (repeating the game against the same person with probability $p>0$) and only at the end of this sequence be re-assigned to another opponent (against whom she plays again and again with probability p every time).

[34] See also Nowak and May (1993) for computer simulations of evolution on a straight line (as opposed to a circle) resulting in chaotically changing spatial patterns in which the proportion of co-operators fluctuates continuously.

[35] Of course this is only a good thing if the conventions in question are ones which we want preserved.

[36] See also Binmore (1998).

[37] That is, to discourage the mutually disadvantageous outcomes in Games 2.14 to 2.16.

[38] Psychologists have put forward the theory of cognitive dissonance removal as follows: When we experience an inner conflict between what occurs or what we say or believe, we tend to remove the resulting dissonance by altering what we believe. Bertrand Russell (1916) had preempted them when he wrote that: "Most men, when their impulse is strong, succeed in persuading themselves, usually by a subconscious selectiveness of attention, that agreeable circumstances will follow from the indulgence of their impulse."

[39] Made available to Varoufakis by the *J.S. Mill and H. Taylor Archive*, LSE Library.

[40] Marx and Engels (1979) in their *Communist Manifesto*.

[41] Under capitalism, everyone enjoys the formal liberty to own a Rolls Royce, although few have the power to realise it.

[42] This group included, amongst others, Alan Carling, G.A. Cohen, Andrew Levine, John Roemer and Erik Olin Wright. For an interesting exchange see the Spring 1994 issue of *Science and Society*.

[43] Ellen Meiksins Wood (1989), W. Suchting (1993) and one of the authors of this book fell in this category.

[44] Marx habitually poured scorn on those (e.g. Spinoza and Feuerbach) who transplanted models from the natural sciences to the social sciences with little or no modification to allow for the fact that human beings are very different to atoms, planets and molecules. We mention this because at the heart of EvGT lies a simple Darwinian mechanism (witness that there is no analytical difference between the models in the biology of John Maynard Smith and the models in this chapter). Of course Marx himself has been accused of mechanism and, indeed, in the modern (primarily Anglo-Saxon) social theory literature he is taken to be an exemplar of 19th century mechanism. Nevertheless he would deny this, pointing to the dialectical method he borrowed from Hegel and which (he would claim) allowed him to have a scientific, yet non-mechanistic, outlook.

[45] Instrumental rationality, that is, of either of the strong CKR variety (see preceding chapters) or the weaker version implicit in evolutionary game theory.

7

PSYCHOLOGICAL GAMES

Demolishing the divide between motives and beliefs

7.1 INTRODUCTION
7.2 DIFFERENT KINDS OF 'OTHER REGARDING' MOTIVES
    7.2.1 The 'other'-regarding motives of *Homo Economicus*
    7.2.2 Beliefs as predictions and as motives
7.3 THE MOTIVATING POWER OF NORMATIVE BELIEFS
    7.3.1 Fairness equilibria
    7.3.2 Computing Fairness equilibria
    7.3.3 Assessing Rabin
    7.3.4 An alternative formulation linking entitlements to intentions
    7.3.5 Team thinking
7.4 PSYCHOLOGY AND EVOLUTION
    7.4.1 On the origins of normative beliefs: an adaptation to experience
    7.4.2 On the origins of normative beliefs: The *resentment-aversion* versus the *subversion-proclivity* hypotheses
7.5 CONCLUSION: SHARED PRAXIS, SHARED MEANINGS

7

PSYCHOLOGICAL GAMES
Demolishing the divide between motives and beliefs

7.1 INTRODUCTION

Despite game theory's many brilliant insights, its ambition to underpin some unified social science has not been fulfilled. There are two problems that we have come across in this book. The first is indeterminacy, which even after the best efforts of superb minds still remains. Two rather different tacks have been examined to combat this problem but neither breaks with the instrumental model. Either it has been taken to ever dizzier heights, as in the *Refinement Project* (recall Chapter 3), or it has been 'dumbed down', as with the evolutionary approach. Yet the second problem, which comes from the experimental evidence, concerns the failure of this model to predict action in games that, from a strategic point of view, are quite simple (e.g. the prisoners' dilemma game and dictator or ultimatum game). The trouble seems to be that people are motivated by conditional kinds of moral motivation (see Chapters 3,4 and 5). In this chapter we look at some of the theories of rational action that have been developed in a game theoretical context to try to account for this kind of motivation.

We begin in the next section with a review of how simple moral motivations work, like altruism or inequity aversion, and contrast these with a class of models where second order beliefs play a motivating role. That is, your belief about the expectations that your opponent holds with respect to your actions affect how you value those actions and so influence your choice. The distinction here is sometimes described in terms of beliefs becoming a motivating factor.

Within this class of models, one can distinguish different theories according to how they account for the influence of these second order beliefs. We focus the discussion here on those theories that make these beliefs matter because they supply the key to interpreting the intentions of one's opponent and it is these intentions which affect behaviour. So for example in Rabin's (1993) theory that we examine in Section 7.3, if you believe that your opponent has 'kind' intentions, then you derive a special pleasure from reciprocating this 'kindness'.

We conclude with a discussion of the type of breach that these models open up with the instrumental one. While we find much to commend in these theories, unfortunately they do not help with the first of our problems, indeterminacy: indeed, if anything they exacerbate this failing. This is perhaps too negative. If instrumental reason is all that one has to work with, then these models only make matters worse, but if these new models are cues for a different view of agency, then at least they supply a pointer to where to look for the origins of determinacy. This, at least, is something.

7.2 DIFFERENT TYPES OF 'OTHER REGARDING' MOTIVES

7.2.1 The 'other'-regarding motives of *Homo Economicus*

Let us begin with the chapter's title: Why psychological games? The short answer is that the title comes from the attempt to enrich the psychology of agents: people are motivated not only by material preferences but also by what are sometimes called 'psychological' pay-offs. These are preferences that are in some degree 'other regarding': they take

account of others. An immediate complaint might be that conventional game theory already allows for such a complicating psychology. Take for instance the one-shot prisoner's dilemma (Game 2.18) which we reproduce below.

	d	*c*
d	1,1	4,0
c	0,4	3,3

Game 2.18 – *The Prisoner's Dilemma*

Suppose that the pay-offs are dollars. Does this mean that rational players must defect (play *d*)? If the pay-offs were utils, the answer would be affirmative. For the assumption of game theory is that utils motivate exclusively. However, dollars do not motivate exclusively. For instance, a player may 'like' dollars but she may also like something else, say, 'fairness' or 'equity'. Game theory has no quarrel with this thought. Indeed, it insists (recall Chapter 1) that, before we study the game's strategic structure, we ought to convert dollars into utils.

In Chapter 5 we discussed one example (recall Game 6.2) of the possible psychological pay-offs associated with equity. Let us look at another similar example, as an illustration of the complex psychology that *Homo Economicus* is capable of. Suppose, for instance, that player *i* values dollars ($) but that she also dislikes unequal distributions of dollars between herself and her fellow player (*j*). Then her utility looks like this: $U_i = a\$_i - b|\$_i - \$_j|$, with ratio *b/a* reflecting *i*'s valuation of equity (or fairness) relative to own dollars. Suppose now that two players who share these preferences meet in the context of Game 2.18. Translating the dollar pay-offs into utilities yields Game 7.1 – a totally new game:

	d	*c*
d	*A*,*a*	4(*a*-*b*),-4*b*
c	-4*b*,4(*a*-*b*)	3*a*,3*a*

Game 7.1 – *The Prisoner's Dilemma when players value equity*

If the players' dissatisfaction from receiving a dollar more or less than her opponent exceeds (in absolute terms) her satisfaction from gaining 25 cents (i.e. $b/a>1/4$), then the game ceases to be a *Prisoner's Dilemma*. To be precise it becomes a *Stag-Hunt* (see Game 2.18), featuring two Nash equilibria (*cc* and *dd*).[1] What has happened here is that the players' psychological, ideological, or moral utility from equity altered the game's strategic structure, rendering *c* a best reply to *c* (as long as $b/a>1/4$).

Homo Economicus is, therefore, apparently unperturbed by accusations of having an inadequate psychology and no capacity to overcome prisoners' dilemmas co-operatively. What we *can* say, however, is that his psychological preferences are fixed before the fun and games begin; that *his psychology is independent of what others think is valuable and worthwhile (and what he thinks that they think)*. He may have preferences that are ethical, moral or whatever. However, he has no preferences that depend on what is expected of him to think or do and this is what marks the difference with the theories that we consider in this chapter.

This is an important difference not least because it can make co-operation, for instance, conditional on the expectation that the other person will co-operate and this is important for any model that is attempting to be consistent with what we know of how people actually behave in such settings. We examine this difference now.

7.2.2 Beliefs as predictions and as motives

The conversion of dollars into utility, which transformed the *Prisoner's Dilemma* (Game 2.18) into Game 7.1, can be generalised mathematically in the following simple manner: Utility is given as the sum of two sub-utility functions: $M(.)$ and $\Psi(.)$,

$$U_i(O) = M(O) + \Psi(O) \qquad (7.1)$$

where the outcome is defined in material terms by the material pay offs each player receives and $M(O)$ denotes the utility (or degree of preference-satisfaction) resulting directly from player i's *material gains* , and $\Psi(O)$ denotes what we shall call the '*psychological utility'* from this material outcome. In terms of the example above (Game 7.1), the material utility function is given as $M = a\$_i$ and the psychological utility function as $\Psi = -b|\$_i - \$_j|$, where $\$_k$ is the dollar sum awarded to player k by outcome O.

In this example, a can be thought of as the rate of increase in utility as a result of winning more dollars (or the marginal utility of dollar gains) while b is to be interpreted as the rate of increase in utility losses as one player makes more dollars than another (or the marginal psychological disutility from inequity).

In the preceding chapters we habitually assumed that the games we played (or studied) where expressed in composite, or overall, utility terms; namely, that all psychological pay-offs were already incorporated in the game's pay-off matrix or tree diagram. But what was it that made this assumption plausible? It was another, hidden, assumption: psychological utility is *only* a function of the material outcomes. Why 'only' a function of outcomes? The precise function could, of course, reflect a person's beliefs about what is 'moral' or 'right' about the outcome as well as how well it satisfies purely personal desires, but once these beliefs are known (i.e. the person's utility function is known), the only thing that causes the player's utility to change is an alteration to the material pay-offs. In this way, we can always amend the pay-offs in each cell of the game to take account of this new type of preference while leaving the basic structure of the interaction unchanged.

Is this assumption a problem? It is when the psychological pay-offs depend not only on what actually happens materially but, also, on what people expect you to do.

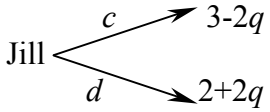
Second order (predictive) beliefs (Definition)

Suppose player A chooses between strategies $(s_1, s_2, \ldots, s_m)$ with probabilities $(p_1, p_2, \ldots, p_m)$. Suppose further that, before observing A's actual choice, B's estimates of probabilities $(p_1, p_2, \ldots, p_m)$ are given as $(p'_1, p'_2, \ldots, p'_m)$. We define A's second order beliefs $(q_1, q_2, \ldots, q_m)$ as her estimates of $(p'_1, p'_2, \ldots, p'_m)$. For example, if $p'_i = 1$ and $q_i = \frac{1}{2}$ this means that B predicts that A will choose strategy s_i with certainty but A wrongly thinks that B expects her to do so (i.e. choose strategy s_i) only with a 50% probability.

Suppose, to help illustrate the point, we begin with a simpler example than the *Prisoner's Dilemma*. Jill must decide between c and d and she plays d with probability $p = \Pr(d)$. If she *does* opt for c her material pay-off equals 2, otherwise it is 3. Meanwhile, Jack is observing Jill and, before the event, predicts that she will d with probability $p' = E^{Jack}(p)$. Jill knows Jack is 'watching' and cares deeply about his expectation. Indeed, if she thinks that he is expecting her to avoid c (i.e. Jill predicts that p' is high), she suffers psychological disutility from frustrating his expectations by playing c.

Conversely if she plays *c* when *p'* is low, she enjoys knowing that she has confirmed Jack's expectations.

Let $q = E^{Jill}(p')$ be Jill's estimate of the probability with which Jack expects her to shun *c*. In terms of the arguments in utility function (7.1) we can capture this situations by having $\Psi(c) = 1-2q$, $\Psi(d) = 2q-1$ [with $M(c) = 2$, $M(d) = 3$]. Thus when Jack's expectations are confirmed Jill enjoys an extra psychological util (i.e. when *c* and $q=0$ or when d and q=1) and when they are confounded, she loses one util (ie when *c* and $q=1$ or *d* and $q=0$). Putting the whole utility representation together, we end up with the following choice.



Game 7.2 – *Jill's dilemma*

Expectations play a radically different role here to anything we have encountered so far in this book. Hitherto, once the *outcome-specific* psychological payoffs (function Ψ) had been incorporated into the players' overall utility pay-offs, expectations were nothing more than cold-blooded information (i.e. calculative predictions). They did influence behaviour but *only* through the agents' constraints. For example, a belief that it will rain affected Jack's decision to carry an umbrella. However, it had no impact whatsoever on his preference-ordering between events: 'Walking in the rain without an umbrella', 'Walking in the sunshine with an umbrella' etc. Similarly in games such as *Hawk-Dove* (Game 2.16): The belief that one's opponent would play 'dove' might cause her to play 'hawk' herself; however, it has no bearing on how much she *desires* the outcome 'I play hawk when my opponent plays dove'. In comparison, in this simple case there are two events, Jill either chooses *c* or *d* and her ordering between these depends on what Jill believes that Jack expects of her (e.g. when $q=0$ *c* is preferred to *d*, but when $q=1$ *d* is preferred to *c*).

The immediate question is whether this difference is liable to be practically significant. In other words, do people care in this way about another's expectations and if they do, does this affect the game theoretical representation of the interaction?

To pick up on the first part of this question, while *Homo Economicus* does as he pleases and is pleased with what he does, the psychoanalysts' couches are filled by a long array of people who think too little of themselves. Other people's expectations terrify them (e.g. children trying to live up to their parents' expectations) and their wishes are often determined by what they think others expect of them. Of course not everyone queues up to make some psychoanalyst richer and so although we may recognise the point, we could equally doubt its generality.

Other examples, however, are not difficult to find. The lone marathon runner enters the stadium exhausted but is egged on by an expectant crowd, the originally wavering musician sticks it out and eventually produces a musical masterpiece in order not to disappoint some devoted sponsor. Indeed Sugden (2000b), drawing on Hume, has turned this dependence on others' expectations into a general feature of human psychology by arguing that the failure to conform with another's expectations causes a 'resentment' which, in the instrumental calculation of what to do, weighs against the action. This begs a question, though, about why the failure to conform should cause resentment.

Adam Smith is perhaps more interesting in this regard because he too made the influence of other people's expectations central to human psychology and he offered an explanation of why we were drawn to conformity. He believed that people obtained a

very special pleasure from sharing judgements regarding what was appropriate behaviour (see Box 7.1). So if Jack is expecting Jill to *c* because *c* is the worthy action and Jill shares this belief, then choosing *c* gives Jill an intense kind of pleasure; whereas if both hold no belief of this kind about the worth of *c*, then choosing *c* generates a much smaller pleasure.

Box 7.1

ADAM SMITH AND THE PARTICULAR PLEASURE OF MUTUAL SYMPATHY

Adam Smith argued that people have a capacity for sympathy. He meant by this that we experience events that directly affect others vicariously. When someone closes the car door and catches their finger, we wince with the poor unfortunate. It is, as if, our fingers had been caught too. Equally we sometimes share in the joy of others. Thus sympathy is a kind of fellow feeling, born of an imaginative projection into another's place. It does not apply to all experiences and Smith offers a range of further insights into this particular aspect of human psychology in his *Theory of Moral Sentiments*. One of these is that fellow feeling is the basis for judgements of approval and disapproval regarding an action. In so far, as we sympathise with the person when they act, then we approve of their 'passions'; and when we do not, we disapprove. These judgements of approval and disapproval are, in turn, the building blocks for morality. This is one of the key elements of his theory of 'moral sentiments'.

By itself, this insight would not seem to be very different from making *Homo Economicus* have altruistic preferences since the hallmark of the altruist is that they share in the pleasure and pain of others. However, another of Smith's insights is that we derive a particular pleasure from acts that occasion 'mutual sympathy': 'nothing pleases us more than to observe in men a fellow feeling with all the emotions of our own breast; nor are we so much shocked as by the appearance of the contrary (p.13)'. These are acts, therefore, where people know that there are experiences of fellow feeling. For example when Jack acts and Jill sympathises with him, it becomes mutual sympathy when Jack knows this and Jill knows that Jack knows this. This is very different from a mere reflective effect that can occur among altruists because mutual sympathy is always a positive pleasure. Mere reflection would reinforce the character of the initial experiences, making the pleasurable more pleasurable and the painful more painful; whereas the pleasure of mutual sympathy can be generated by initial experiences that are either painful or pleasurable.

"The sympathy, which my friends express with my joy, might, indeed, give me pleasure by enlivening that joy: but that which they express with my grief could give me none, if it served only to enliven that grief. Sympathy, however, enlivens joy and alleviates grief. It enlivens joy by presenting another source of satisfaction; and it alleviates grief by insinuating into the heart almost the only agreeable sensation which it is at that time capable of receiving" (Smith, 1759, p.14)

What gives the special pleasure to both parties in Smith's theory is that each knows that other has formed the same judgement about the experience, be it painful or pleasurable.

The origin of such shared judgements in Smith is the 'sympathy' that we feel for others and this is controversial; the observation that our judgements about what constitutes worthy action depends on the expectations of others is much less so. This is because there is a wealth of anthropological evidence that such judgements tend to be shared by members of a discernible social group. Furthermore, there are philosophical reasons for supposing that purely personal judgements of this kind are difficult to sustain. The problem here is that no set of evaluative rules can be exhaustive. They cannot cover every situation that one will come across because every new situation can always be individuated in a way that makes it different from anything that has gone before.

To have rules that covered the application of the rules in all settings would therefore be like having no rules; it would be a potentially infinite set of recommendations regarding what to do in every imaginable setting. Thus any set of evaluative rules always requires interpretation and this creates a problem: How does any individual know when they have made a mistake in applying the rules? This is a genuine

dilemma and one way of resolving it is to rely on the shared judgements of the community in which one lives.

Another way of seeing this difficulty is that actions that are worthy need psychologically to be distinguished from those that are merely self-serving. If there is scope for individual judgement over what is worthy, then such judgements will always be prone to the suspicion that they are self-serving and so undermine this distinction. Worthy actions, for this reason, depend critically on a standard that is in some sense external to the individual, and this is why we tend to rely on beliefs that are shared with others to guide our judgements. This sharing gives to these judgements the kind of external authority that permits a distinction between what is worthy and what is self-serving (see Smith on this in Box 7.2).

Box 7.2

ADAM SMITH AND THE DIFFICULTY OF FORMING INDIVIDUAL JUDGEMENTS

Box 7.1 explains Adam Smith's views on the particular pleasure that comes from mutual sympathy. Later in the *Theory of Moral Sentiments,* he considers the difficulty that any individual has in deciding whether another person is sympathising with them. Smith expresses the problem here in this way.

"When we are about to act, the eagerness of passion will seldom allow us to consider what we are doing, with the candour of an indifferent person…The passions on this account, as father Malebranche says, all justify themselves, and seem reasonable and proportioned to their objects as long as we continue to feel them…

When the action is over, indeed, and the passions which prompted it have subsided, we can enter more coolly into the sentiments of the indifferent spectator….It is seldom, however, that they are quite candid even in this case… It is so disagreeable to think ill of ourselves, that we often purposely turn away our view from those circumstances which might render that judgement unfavourable. He is a bold surgeon, they say, whose hand does not tremble when he performs an operation on his own person; and he is often equally bold who does not hesitate to pull off the mysterious veil of self-delusion, which covers from his view the deformities of his own conduct"(p.157-8).

It is in this context that Smith argues we come to rely on norms or rules of moral conduct. "Nature, however, has not left this weakness…altogether without remedy; nor has she abandoned us entirely to the delusions of self love. Our continual observations upon the conduct of others, insensibly lead us to form to ourselves certain general rules concerning what is fit and proper either to be done or to be avoided. It is thus that the general rules of morality are formed. They are ultimately founded upon the experiences of what, in particular instances, our moral faculties, our natural sense of merit and propriety, approve, or disapprove of. We do not originally approve or condemn particular actions; because upon examination, they appear to be agreeable or inconsistent with a certain general rule. The general rule, on the contrary, is formed by finding from experience, that all actions of a certain kind , or circumstanced in a certain manner, are approved or disapproved of." (p.159)

This line of argument is, of course, remarkably close to that of Wittgenstein on the difficulties with holding a 'private language'.

So, if the influence of another's expectations on the choice between actions is liable to be of some practical significance, the next part of question is: How does this affect game theory?

It may be tempting to suppose that game theory can rather easily dispose of the observation that Jill's preference for *c* and *d* depend on Jack's expectations by setting this dependence out in matrix form (i.e. turn Jack's choice into a choice about what to expect of Jill). We might also suppose that Jack's interest is merely to predict correctly what Jill will do. This would yield three potential equilibria where first and second order beliefs are confirmed by experience, that is, $p=p'=q$, and where Jill is deciding on instrumental grounds what to do. However, this is not really much of a game, since Jack's choice of belief does not have any effect on Jill's pay-offs (remember it is her expectation of his expectation, q, that matters).

Of course, there would be proper strategic interaction if we assumed that these expectations were confirmed as then $p' = q$. But, from a game theoretical point of view, this is a distinctly odd way of going about things. We have to assume an equilibrium in order to represent the interaction in a game theoretical strategic form, and once we have done this we can then use the standard game theoretical techniques to isolate the equilibria. This looks suspiciously circular.

The three psychological equilibria of 'Game' 7.2

(i) $p = p' = q = 1$ Jill chooses *d* and collects 4
(ii) $p = p' = q = 0$ Jill chooses *c* and collects 3
(iii) $p = p' = q = ¼$ Jill randomises and collects 2½

This conclusion is even clearer once we build in some strategic interaction that does *not* depend on assuming that an equilibrium is already obtained. Suppose that both Jack and Jill must choose between *c* and *d* so that there is a *Prisoners' Dilemma* in terms of the material pay-offs, as given by Game 7.3. Further, we assume that both Jack and Jill are affected by the other's expectations of what they will do in exactly the same way. So Jack's psychological pay-offs are generated in exactly the same way as Jill we let *r* stand for Jack's expectation of Jill's expectation that he will play *d* and we shall assume the same material pay-off. So when Jack plays *c* or *d*, this leads respectively to 1-2*r* or 2*r*-1 additional psychological pay-offs. Game 7.4 sets out the interaction once the psychological pay-offs are also included.

	c	*d*
c	2, 2	0, 3
d	3, 0	1, 1

Game 7.3

	c	*d*
c	3-2*q*, 3-2*r*	1-2*q*, 2+2*r*
d	2+2*q*, 1-2*r*	2*q*,2*r*

Game 7.4

It will be evident that the version of the game with psychological pay-offs cannot be analysed until *q* and *r* are specified since without this information it is not possible to determine what strategy is a best reply to one's opponents choice. For instance if $q = 0$, then Jill's best response to Jack's *c* is *c*, whereas if $q > ¼$ then *d* becomes the best response. To solve the game, it seems we must either take these expectations as given or impose some condition upon them.

One possibility here is to require that these expectations are equilibrium ones in the sense that they are confirmed (e.g. $p = p' = q$). This is in the spirit of the standard game theoretical approach, but there are at least two such equilibria for the pair (q,r): (1,1) and (0,0). So we cannot know what value to place on *q* and *r* until we know which equilibrium obtains. Thus the standard game theoretical analysis becomes entirely circular: we would have to know what the equilibrium was before we could identify the pay-off matrix in Game 7.4 and we need to do this in order to solve the game. But what would the point be of solving the game when we already know its equilibrium?

7.3 THE POWER OF NORMATIVE BELIEFS

7.3.1 Fairness equilibria

In the illustration of the transformed *Prisoners' Dilemma* above, we simply assumed that the calculative second order beliefs entered directly each agent's motivation. Jill's will was affected by what she thought Jack was *predicting* of her and vice versa. This was useful for bringing out the trouble that is caused for game theory when these second order beliefs directly motivate through influencing the pay-offs associated with any material outcome (rather than simply forming an expected constraint on action which the agent takes into account, as in the usual game theoretical account). We said nothing about the origin of this influence at that time even though we had argued earlier that the most likely explanation for this effect turned on people sharing some idea regarding the worth of an action. In this subsection, we shall examine a particular theory where this is made explicit: Rabin's 1993 *Fairness Equilibrium* concept. This is followed by a brief discussion of another theory in this vein: *Team Thinking*.

Rabin's rough idea is that people enjoy extra psychological utility when they either reciprocate 'kindness' or 'unkindness'. Such reciprocation is a *fairness norm* and hence the equilibria in games where there are these psychological pay-offs are *fairness equilibria*. The connection with second order beliefs arises because the 'kindness' or 'unkindness' of an action can only be judged once you hold an expectation of what the other player expects that you will do. To see how this works in outline, consider again the *Prisoner's Dilemma* (Game 2.18 or 7.3) and suppose that Jill expects Jack to co-operate (*c*). Her pay-offs from responding with *c* or *d* depend, claims Rabin, on the reasons for which she thinks Jack is about to play *c* and these depend in turn on her second order beliefs. Consider two different accompanying second order beliefs and their psychological effects on Jill:

(A) *that he anticipates a co-operative move* (*c*) *from her* and so is deliberately foregoing the superior material outcome that comes from *d*. He is thus being 'kind' and when Jill responds with *c* she reciprocates this kindness and so enjoys positive psychological utility. If she responds with *d* she is not reciprocating this kindness and she will suffer negative psychological utility.

(B) *that he expects her to play d and has simply made an execution mistake* and so is not intending kindness or unkindness. Since 'kindness' or 'unkindness' cannot be reciprocated in this case, there are no psychological pay-offs.

So the second order beliefs affect how each action is evaluated by Jill but only because they are crucial to Jill's interpretation of whether Jack is acting kindly and it is this that gives scope for psychological pay-offs to be enjoyed through reciprocation.

In the first case, (A), a fairness equilibrium with mutual co-operation becomes a possibility. This is because the additional psychological utility from playing *c* could outweigh the material benefits from playing *d* and so persuade Jill to choose *c*. If Jack was similarly placed, anticipating a co-operative move from Jill and weighing the psychological benefits from reciprocating kindness sufficiently highly, he too could choose *c*. Thus each would choose *c* on the basis of second order beliefs (that each expects that other expects them to co-operate) which are then confirmed. The combination of the psychological equilibrium with respect to beliefs and the Nash equilibrium with respect to actions gives a *Fairness equilibrium* with mutual co-operation.

This is a quick sketch of the idea and how it might unlock the *Prisoners' Dilemma*. We shall now set the idea out in more detail. It is, ostensibly,[2] founded on the following combination of definitions and assumptions.[3]

***Sacrifice* (Definition)**: We say that A makes a sacrifice *vis-à-vis* B when she intentionally forfeits part of her material utility in order for B to receive utility different to that which A thinks B is 'entitled' to.
***Reciprocity* (Assumption)**: When A predicts that B is about make a sacrifice on her behalf (see above definition), A experiences an urge to reciprocate (ie $\Psi>0$). If that urge remains unfulfilled, she suffers some psychological loss (i.e. $\Psi<0$).
***Symmetry* (Assumption)**: The psychological urge to reciprocate is symmetrical in that it applies equally when B is expending utility in order (a) to benefit A (i.e. increase A's utility beyond his 'entitlement'), and (b) to hurt A (i.e. reduce her utility below her 'entitlement').
***Kindness/nastiness/neutrality* (Definition):** When B sacrifices material utility in order to boost (diminish) A's material utility beyond (below) what she is entitled to, he is being *kind* (*nasty*). If his actions do not affect A's material utility, he is being neither kind nor unkind to her. He is just *neutral*.[4]
***Entitlements* (Assumption):** Consider a strategy available to A, say s_A, and examine the outcomes that are possible when playing s_A. If there are outcomes among these which happen to be dominated by some other outcome, discard them. Finally, compute the average of A's material pay-offs amongst the surviving outcomes. This average is A's entitlement when she plays s_A i.e. $e^A(s_A)$. Compute, in the same way, A's entitlement from all her strategies; and similarly for B.

These assumptions are represented mathematically by Rabin in the following way. Similarly to our equation (7.1), he supposes that overall utility is the weighted sum of the material pay-offs $M(O)$ from outcome O and the psychological pay-offs $\Psi(O)$:

$$U_i(O) = (1-v)M(O) + v\Psi(O) \qquad (7.2)$$

Clearly, the higher the value of v the more the player cares about her psychological rewards from a certain outcome, relative to her material pay-offs. Utility function (7.2) can be re-written for simplicity as (7.3) with $\mu = v/(1-v)$ capturing the relative weight of psychological to material pay-offs:[5]

$$U_i(O) = M(O) + \mu\Psi(O) \qquad (7.3)$$

Kindness/nastiness functions reflecting A's beliefs (Definition)
***A's kindness function toward B*, f_A:** Rabin defines f_A as a function which varies between –1 and +1. If it takes a value between 0 and +1, this means that A *believes* that she is being kind to B, given her estimates of what B will do and what she thinks he expects her to do. Similarly, if $f_A<0$, A thinks that she is being nasty to B. Finally, if $f_A=0$, A deems that she is being 'neutral'.
***B's kindness function toward A*, f_B:** This is a similar function, also varying between –1 and +1, depending again on A's beliefs. In this case, $f_B>0$ means that A thinks that B is being kind to her (given her estimates of what she expects him to do *and* of what she expects him to predict that she will do). Naturally, if $f_B<0$, her expectations (1st

and 2nd order) lead her to the prediction that he is being nasty to her. Equality $f_B = 0$ brings to A a feeling that B is being 'morally' neutral to her.

Rabin now defines A's psychological pay-offs in terms of these kindness functions:

$$\Psi_A(O) = f_B(O)[1+f_A(O)]\ldots\ldots\ldots\ldots\ldots\ldots(7.4)$$

This function embodies the features of reciprocation noted above. Thus when A anticipates that B is going to be 'kind' (i.e. $f_B(O)>0$), then her psychological pay-offs are positive when she reciprocates (i.e. with $f_A(O)>0$) while being nasty ($f_A<0$) turns them negative. Alternatively if A expects B to be nasty [$f_B(O)<0$], then only way of making the psychological pay-offs non-negative is to make f_A negative (i.e. to be nasty back). Finally, if she anticipates neutrality from B [$f_B(O)=0$], it makes no psychological difference to her whether she is kind, nasty or neutral to B. However, as both kindness and nastiness require (by definition) a sacrifice of material pay-offs from A, she has no reason to make it (since her psychological rewards from it are zero).

The next step in the mathematical representation is to define functions f_A and f_B. They are given below in (7.5) and (7.6) for each possible outcome. Since each outcome (of a two person game) corresponds to a combination of strategies, s_A for A and s_B for B, the kindness function is defined over these possible strategy pairs.

$$f_A(s_A, s_B) = \frac{\pi_B(s_A, s_B) - e^B(s_B)}{\pi_B^h(s_B) - \pi_B^\ell(s_B)} \qquad (7.5)$$

$$f_B(s_A, s_B) = \frac{\pi_A(s_A, s_B) - e^A(s_A)}{\pi_A^h(s_A) - \pi_A^\ell(s_A)} \qquad (7.6)$$

Kindness is given here as a ratio. The numerator in f_A is simply the difference between A's estimate of B's material pay-off and his 'entitlement' given that he plans to play s_B [that is, $e^B(s_B)$]. So this measures A's kindness to B. The role of the denominators is to keep the values of f_A and f_B within the required bound of (-1,+1) and they are the difference between player *i*'s maximum and minimum material pay-offs when he or she plays s_i: $\pi^h_i(s_i) - \pi^\ell_i(s_i)$

To give an example, take the *Hawk-Dove* (Game 2.16) and suppose that, for some reason, A expects B to play *d*. A knows that, depending on whether she chooses *d* or *h*, B's largest possible material pay-off is $\pi^h_B(s_B=d)=1$ and his minimum is $\pi^\ell_B(s_B=d)=0$). Clearly, the denominator of (7.5) equals 1. Suppose further that A plans to play *h*. In that case, $f_A(s_A=h, s_B=d) = [0-e^B(s_B=d)]/1$. Evidently, in this situation, A is either nasty ($f_A<0$) or, at best, morally neutral ($f_A=0$) toward B, depending on B's 'entitlement'. Note that, in her own mind, she is being nasty if she thinks that B was entitled to something more than pay-off 0 given that he played *d* (i.e. if $e^B(s_B=d)>0$). But if she thinks that a *d*-playing B is entitled to nothing, she believes that her *h*-choice was morally neutral (in the sense that it did not deny B of any entitlement).

Similarly, suppose that she thinks that B expects *h* of her and plans to play *d* in response. Is he being kind, nasty or neutral towards A? Let's find out the answer as reported by (7.6). If A is expected to play *h*, B knows that her highest and lowest pay-

offs are 2 and –2 respectively. Thus, the denominator equals $\pi^{h}_{A}(s_A=h)-\pi^{\ell}_{A}(s_A=h) = 2-(-2) = 4$. As for the numerator, it equals $\pi_A(s_A=h,s_B=d)-e^A(s_A=h) = 2-e^A(s_A=h)$. Suppose that A believes that, when she plans to play *h* in the expectation that B will play *d*, she is entitled to get material pay-off 2 (i.e. $e^A=2$). In that case, she thinks that B is being neutral since $f_B=0$.

Finally, to complete the model, the mathematical representation of Rabin's idea regarding entitlements is implied by its definition (see above). To illustrate their calculation, suppose that A is involved in a two-person game in which if she plays her first of three strategies there are three possible outcomes, depending on which of the three strategies available to him her opponent chooses. The following pay-offs come from Game 6.4 in the previous chapter: (-2,-2),(2,0),(4,-1). What is A entitled to, according to Rabin, when she chooses this particular strategy?

According to the above definition, $e^A(s_A)=3$. To see this, first we note that of the three outcomes associated with A's choice the first one is dominated and thus discarded. By this we mean that (-2,-2) is clearly worse for *both* players than either (2,0) or (4,-1) - and in this sense it is dominated - and therefore does not count in the computation of A's entitlement. Of the remaining two outcomes, neither dominates the other since there is no way one of them could be discarded without one player objecting. Of those two, remaining outcomes, A's possible material pay-offs equal either 2 or 4 utils. The average of these is 3 and this is, according to Rabin what A is entitled to if she plays this strategy. Of course, this being an average, it is obvious that A will either get more than she is entitled to (i.e. 4) or less (i.e. 2). But such is life. We seldom get what we deserve in life. We are either too far ahead or struggling to catch up!

With the entitlements of both players $e^A(.)$ $e^B(.)$ fully computed for each of their strategies, equations (7.5) and (7.6) can now be computed for each outcome. These values are then put back into equation (7.4) to find the psychological pay-offs per outcome before imputing these into (7.3). At that stage the game has been totally transformed and a new pay-off matrix is derived depicting the players' overall pay-offs (material and psychological). The Nash equilibria of this transformed game are the so-called *fairness equilibria*.

Fairness equilibria (Definition)

Consider games in which the players' utilities have been augmented with psychological utility (Ψ). Suppose further that these psychological utility functions (Ψ) satisfy conditions ***Sacrifice***, ***Reciprocity***, ***Symmetry,*** as well as the definition of ***Kindness and Entitlement*** above. The Nash equilibria of the resulting game are known as *fairness equilibria*. They are associated with the notion of fairness because the psychological utility underpinning them rises in response to the belief that one is acting in a manner that 'aids' ('harms') an opponent who is making a sacrifice in order to 'aid' ('hurt') one. Otherwise, the Nash idea of 'solving' a game, by focussing on some equilibrium between *beliefs* (1st and 2nd order in this case) and *actions*, remains intact.

Box 7.3

CONSISTENTLY ALIGNED BELIEFS: WHY PSYCHOLOGICAL GAMES DEMAND MORE!

In previous chapters CAB was used to generate Nash equilibria when CKR ran out of steam and could not, by itself, engender equilibrium (recall the discussion in Chapter 2 around Games 2.9,2.10,2.11,2.12 as well as the justification of NEMS in Sections 2.6 and 3.2.2). Throughout this book, we have maintained a highly critical stance toward CAB, arguing that it is impossible to attribute

it to human reason (unless we want the latter to include a form of telepathy). Psychological games appeal because of the featured link between beliefs and desires. But, as we have already seen, the problematic axiom of CAB must be given an even more important role in psychological games if the latter are to acquire a well-defined pay-off structure: Without CAB, we cannot even *describe* the game (let alone 'solve' it) since we would be lacking the necessary one-to-one correspondence between outcomes and utilities.

Recall, however, that this is not the first time we encountered games in which (*i*) a player may have more than one pay-off from a given outcome, and (*ii*) CAB comes to the rescue. For instance, in Game 3.2 (see Chapter 3) there was uncertainty about the true character of one's opponent, with different 'types' enjoying different levels of utility from each outcome. Harsanyi's masterstroke (see Section 3.2.2) was to assume that players entertained probabilistic beliefs about their opponents' potential types, before finding a mixed strategy Nash equilibrium (NEMS) reflecting these beliefs. Of course, CAB was *sine qua non* in yielding this Bayesian Nash equilibrium (for it was impossible otherwise to explain how players had common knowledge of the probabilities with which each expected the other to be of a particular type). Why is the role of CAB different here? Why can we not do as Harsanyi did in dealing with a game whose utility pay-offs are not fixed?

The answer is simple: In Section 3.2.2, and Harsanyi's derivation of the Bayesian Nash equilibrium, players' beliefs regarding their opponent's potential type or character were *exogenous*; that is, independent of the game's strategic structure. Players had complete information: (*i*) on the utility pay-offs of *each* of their potential opponent, and (*ii*) the probability that each of these potential opponents was the one facing them. This information was *prior data*, independent of any strategic consideration. Harsanyi used CAB to combine this *given* data into a set of actions that would be consistent with everyone's pre-existing probabilistic beliefs (the Bayes-Nash equilibrium).

Thus Harsanyi 'tied down' his Bayesian-Nash equilibrium on a solid foundation of *beliefs that were imported from outside the game*. Were we to enquire into the source of agents' beliefs about the potential character of their opponents, Harsanyi would say something like: "They reflect the distribution of characters or types in the broader population." His CAB thus reduces (or is explained as) common knowledge of the external beliefs regarding possible types and their respective likelihood. Though we can still dispute the reasonableness of assuming such common knowledge, Harsanyi's CAB at least boils down to common knowledge of something that lies outside the game.

In juxtaposition, the beliefs that cause A's utilities to fluctuate in the psychological games of Figure 7.1 are *internal* or *endogenous*: they cannot be attributed to anything 'objective' that comes from outside the game, like a pre-existing distribution of characters or attributes in the population at large. This endogeneity of both beliefs and pay-offs leads to an endless cycle between them. As a result, there is no sense in which players can form expectations on the basis of some mixed strategy because there are no external beliefs to tie the latter down (even if they could, ostensibly, be commonly known). In technical terms, there can be no mixed strategy fairness equilibrium!

All we can do, therefore, in the context of psychological games, is to impose CAB on the players' beliefs in order to define the games as in Figure 7.1. The stakes have been increased: If we want to end the psychological games' strategic fluidity, we are forced to assume an enhanced, and thus an even less defensible, form of CAB: Consistently aligned *endogenous* beliefs, is what we need. Of course, there is another alternative: Let psychological games retain their fluidity and abandon the ambition to 'tie them down'!

Box 7.4

COMMONLY KNOWN RATIONALITY (CKR) AS A PREREQUISITE TO MORAL CAUSALITY IN PSYCHOLOGICAL GAMES

In chapters past, CKR (of various orders) was used in order to discard non-rationalisable strategies (recall Chapter 2) and come closer to 'solving' a game. In this section, it acquires another role: that of imputing moral intent! Consider, for instance, the situation when A plans to play *h* in *Hawk-Dove*, expecting that B will also play *h because* A thinks that B expects A to pay *h*. Immediately, she brands his action 'nasty'. Now notice the use of the word 'because' in the penultimate sentence. It is the key to A's outrage. Where did this causal relation come from? From CKR, is the answer. Let's investigate: A expects B to expect *h* from her. She also thinks that he will, given this belief, play *h* himself. A knows that *h* is not B's best reply to her *h*. (In fact, it is very poor reply to it.) So, "why is B playing *h* when anticipating *h* from me?" asks a bemused A. CKR does not allow her to think that he is just careless. He *must* have an instrumentally rational reason. Given the assumptions of the model (recall the definition of *sacrifice* and *kindness/nastiness*), the only possible explanation is that B is being nasty

(i.e. B is sacrificing utility so that she loses some too). In short, without CKR there is no causal link of second order and first order beliefs, no moral indignation and, therefore, no fairness equilibrium.

7.3.2 Computing Fairness equilibria

This subsection illustrates the *fairness equilibrium* concept by applying it to two games*,* the *Hawk-Dove* and the *Prisoner's dilemma.* Figure 7.1 begins with the 'standard' material pay-off representations of the two static games. Then it notes *for each of player i's strategies* the following: (a) π^h_i *–i.e. i's* maximum pay-off possible for the given strategy, (b) π^ℓ_i – i.e. *i's* minimum pay-off possible, and (c) e^i – i.e. *i's* entitlement if indeed he/she chooses that strategy.

For instance, in the *Hawk-Dove*, the most B can expect when choosing *h* is 2 utils while the least is –2. As for his 'entitlement', we recall that Rabin demands of us that, *before* we average out B's possible pay-offs, we discard any 'dominated' outcome corresponding to B playing *h*. There are two possible outcomes when B plays *h*: (-2,-2) and (2,0). Clearly (-2,-2) is dominated by (0,2) in the sense that neither player would object if some adjudicator forced them to trade the former for the latter. Thus, we discard (-2,-2). This leaves B with only one possible pay-off (for the purposes of computing his entitlement) when he plays *h*: 2 utils. Thus, according to Rabin, if B chooses strategy *h,* he is 'entitled' to 2 utils – i.e. $e^B(h) = 2$.

Let us perform the same computation once more, only this time for when A plays *d* in the *Prisoner's Dilemma*. Since she is 'defecting', A's highest possible pay-off is 4 and her lowest 1; i.e. $\pi\pi^h_A(d) = 4$ and $\pi\pi^\ell_A(d) = 1$. Her entitlement? Just as before, we investigate whether there is a dominated outcome corresponding to A playing *d*. The two potential outcomes are: (4,0) and (1,1). Neither dominates the other, in the sense that, if the players were to be forced away from one and onto the other, one of them would protest. Thus, Rabin insists, A's entitlement $e^A(d)$ equals the average of her two potential pay-offs 4 and 1. That is, $e^A(d)$=2.5.

In this manner we compute all values of π^h, π^ℓ and *e* for both players and all their strategies. Next we input these values into expressions (7.5) and (7.6) to compute the levels of kindness/nastiness that one shows to the other for each combination of strategies (and thus for each outcome). Suppose, for example, that A plays *h* and B responds with *d* in *Hawk-Dove*. How kind/nasty is A being to B? In other words, is f_A positive or negative? From expression (7.5) we compute it as a ratio between two differences. The numerator is the difference between B's pay-off (when A plays *h* and he plays *d*) and his entitlement given that he chose *d* (the numerator thus equals 0-½). The denominator equals the difference between B's maximum and minimum potential pay-offs when choosing *d* (i.e. 1-0). In conclusion, f_A = -½ and A is deemed to be mean (or nasty) toward B.

Once f_A and f_B have been computed for all outcomes, it is straightforward to utilise expression (7.4) in order to compute both players' psychological pay-offs for each outcome. To continue with the example of the above paragraph, suppose again that in *Hawk-Dove* A plays *h* and B *d*. What are A's psychological pay-offs? According to expression (7.4), for each outcome we add 1 to the corresponding value of f_A and multiply what we find with the corresponding value of f_B. This product gives us A's psychological pay-off: When A and B play *h* and *d* respectively, f_A = -½, $1+f_A$ = ½, and f_B = 0. Thus, A's psychological pay-off equals zero [$\Psi_A = f_B(1+f_A) = 0$].

What is the meaning of this finding? Both A and B played their Nash best replies (recall that, in *Hawk-Dove*, *h* is the best reply to *d* and vice versa). However, A

was branded 'nasty' by Rabin's formulation (f_A<0). And yet she lost *no* psychological utility from this imputed 'nastiness' [$\Psi_A = 0$]. Why? The reason is that, in playing *d*, B was *not* making *any* sacrifice on A's behalf. Thus A had no moral obligation to be kind to him (that is, to sacrifice some of her material pay-offs on his behalf). Remember for Rabin, nastiness leaves a bitter aftertaste in the nasty player's mouth only if her opponent was being kind. Here, B was being neither kind nor unkind. He was simply playing his Nash best reply and A felt no obligation to make a sacrifice on his behalf. In fact she felt no psychological effects whatsoever as a result of her mild 'nastiness' toward him.

With the computation of psychological pay-offs completed, we can now add them to the original (or material) pay-offs, according to expression (7.2). The result is the psychological transformation which yields the final pay-off structures at the bottom of Figure 7.1.

Once the psychological effects of fairness have been incorporated, the games' strategic structure is transformed drastically. Of course, the extent of this change depends on how much importance players attach to their psychological 'side' relative to hard-nosed considerations of material pay-offs. One might, for example, feel bad when double-crossing a friend or foe but, at the same time, place little emphasis on this ill feeling relatively to the appreciation of the material benefits from such treachery. In this model, it is parameter v (or μ) that captures the value of psychological pay-offs (relative to the material ones). [Recall that as v tends to 1, or equivalently μ tends to infinity, psychological pay-offs tend totally to over-shadow material ones. And vice versa as v tends to zero]

Starting with *Hawk-Dove*, nothing changes as long as the relative 'weight' of the players' 'psychology' falls below a certain threshold ($v < 0.57$). Once it exceeds that threshold, the game is transformed utterly and the two original pure strategy Nash equilibria (A plays *h,* B plays *d* or B plays *h* and A plays *d*) cease to exist, giving their place to a unique (fairness) equilibrium in which both players opt for *d*. The point here is that, as long as psychological pay-offs matter sufficiently, the only possible equilibrium is one in which one believes that the other is making a sacrifice on one's behalf (in playing *d*, as opposed to reaping maximum material pay-offs) and therefore feels that, if this sacrifice is not reciprocated, the loss of psychological utility would be greater than the gain of material utility from playing *h*.

The startling aspect of the transformed *Hawk-Dove* is reinforced when players place even more importance to the psychological aspects of the game. For if $v > 0.8$, a *second* fairness equilibrium comes to light: in addition to *dd*, *hh* is an equilibrium too in the context of which each plays aggressively in the full knowledge that the other will do likewise, with the end result that both will forfeit 2 material utils. This is an equilibrium oozing mutual nastiness; a kind of eye-for-an-eye situation sustained by the urge each feels to inflict 'pain' on a player who is trying to hurt her because he predicts (correctly) that she will inflict pain on him because she thinks (correctly) that he will inflict pain on her… *ad infinitum*.

Hawk-Dove (Game 2.16)				Prisoner's Dilemma (Game 2.18)			
Original Pay-offs		**B**		*Original Payoffs*		**B**	
		h	***d***			***d***	***c***
A	***h***	-2,-2	2,0	**A**	***d***	1,1	4,0
	d	0,2	1,1		***c***	0,4	3,3

Rabin's derivation of players' maximum/minimum pay-offs and of their entitlements *per strategy*

	B:*h*	**B:*d***
π^h_B	2	1
π^ℓ_B	-2	0
e^B	2	½

	A:*h*	**A:*d***
π^h_A	2	1
π^ℓ_A	-2	0
e^A	2	½

	B:*d*	**B:*c***
π^h_B	4	3
π^ℓ_B	1	0
e^B	2.5	1.5

	A:*d*	**A:*c***
π^h_A	4	3
π^ℓ_A	1	0
e^A	2.5	1.5

Computing the kindness/nastiness functions f_A and f_B – see equations (7.5), (7.6)

f_A	**B:*h***	**B:*d***
A:*h*	-1	-½
A:*d*	0	½

f_B	**B:*h***	**B:*d***
A:*h*	-1	0
A:*d*	-½	½

f_A	**B:*d***	**B:*c***
A:*d*	-½	-½
A:*c*	½	½

f_B	**B:*d***	**B:*c***
A:*d*	-½	½
A:*c*	-½	½

Computing A's/B's psychological pay-offs: $\Psi_A = f_B(1+f_A)$ and $\Psi_B = f_A(1+f_B)$ – see equation (7.4)

$1+f_A$	**B:*h***	**B:*d***
A:*h*	0	½
A:*d*	1	1½

$\Psi_A = f_B(1+f_A)$	**B:*h***	**B:*d***
A:*h*	0	0
A:*d*	-½	¾

$1+f_B$	**B:*h***	**B:*d***
A:*h*	0	1
A:*d*	½	1½

$\Psi_B = f_A(1+f_B)$	**B:*h***	**B:*d***
A:*h*	0	-½
A:*d*	0	¾

$1+f_A$	**B:*d***	**B:*c***
A:*d*	½	½
A:*c*	1½	1½

$\Psi_A = f_B(1+f_A)$	**B:*d***	**B:*c***
A:*d*	-¼	¼
A:*c*	-½	¾

$1+f_B$	**B:*d***	**B:*c***
A:*d*	½	1½
A:*c*	½	1½

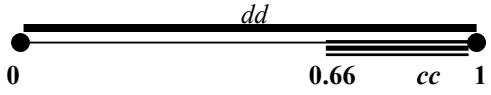
$\Psi_B = f_A(1+f_B)$	**B:*d***	**B:*c***
A:*d*	-¼	-½
A:*c*	¼	¾

The 'psychologically' transformed games *in equilibrium*

	h	*d*
h	-2,-2	2, -½μ
d	-½μ,2	1+ ¾μ, 1+ ¾μ
Hawk-Dove		

Three potential fairness equilibria depending on the relative value of psychological pay-offs [*v*; recall that $\mu = v/(1-v)$]

	d	*c*
d	1-¼μ, 1-¼μ	4+¼μ , -½μ
c	-½μ, 4+¼μ	3+¾μ, 3+¾μ
Prisoner's Dilemma		



Two potential fairness equilibria depending on the relative value of psychological pay-offs [*v*; recall that $\mu = v/(1-v)$]

Nb. The final pay-off structures above apply only in equilibrium (i.e. only if we assume consistently aligned first and second order beliefs). Out of equilibrium, it is not possible to assign given pay-offs to each outcome (or combination of strategies).

Figure 7.1 – *The derivation of Rabin's fairness equilibria using only material pay-offs*

Things are also different in the case of the *Prisoner's Dilemma*. The original mutual defection equilibrium (*dd*) survives independently of the relative weight of psychology. The idea is that, the emphasis on psychological losses makes no difference here since the decision of an opponent to defect (unlike the decision to play h in *Hawk-Dove*) involves no sacrifice and thus does not give a player an urge to reciprocate. So, mutual defection is a fairness equilibrium in the sense that one player is morally neutral to the other. What *does* change however is that, provided psychology matters sufficiently (i.e. $v>0.66$), mutual co-operation (*cc*) becomes an additional equilibrium. This is very similar to the *dd* fairness equilibrium in *Hawk-Dove*, as it is based on the mutual expectation that the other is making a sacrifice on your behalf because she expects you to make a similar sacrifice too, in the belief that you are anticipating her sacrifice… *ad infinitum*. Interestingly, as long as $v>0.66$, the *Prisoner's Dilemma* is transformed into a kind of *Stag-Hunt* game (Game 2.16) with two (fairness) equilibria: *cc* and *dd*.

7.3.3 An assessment of Rabin

Three things stand out in Rabin's 1993 theory. The first is that reciprocation matters, but the precise way in which psychological pay-offs are generated through *reciprocation* is potentially controversial. Reciprocity is a common feature of most normative theories of action in the sense that people seem more likely to be influenced by a norm when they interact with others who are similarly influenced. So if a norm dictates 'kindness' then it is indeed more likely that people will follow this dictate when they expect others to.

This is what much of the experimental evidence points to and so this is an important feature of the theory, which we return to below. Rabin however also makes it more likely that someone will behave 'nastily' when they expect 'nastiness' from others and this seems rather less plausible as a general proposition. An 'eye-for-an-eye' is, after all, but one piece of folk wisdom. 'Turning the other cheek' is another that would go directly against this kind of reciprocation of 'nastiness. Old Testament versus the New, so to speak; and it is not obvious that either could stand for the general case.

Secondly, the nature of what is being reciprocated seems rather special. 'Kindness' is plainly one aspect of behaviour that people often value, but it is not the only one. The claims of 'justice', 'goodness' and 'honour', which can all come in a variety of forms, seem just as strong.

Finally, granted that kindness is what is being reciprocated in a particular social setting, it is not obvious that it will always be identified in this precise way. The most obvious cause for concern here is Rabin's identification of how people perceive their 'entitlements'. Again, it seems more plausible, at least on the basis of the anthropological record, that people's ideas regarding 'entitlements' depend on theories of justice that in turn vary across time and space.

It is not hard, however, to see how each of these points could be met while retaining the same basic model. For instance the relation between the *f* functions in the psychological component of people's utility functions could be changed (i.e. a different mathematical form for equation 7.4). Likewise, the *f* functions could be defined in terms of the extent to which each person has chosen the act which maximises whatever social welfare function best represents the shared view of what is

just; and so on. We supply an illustration of this sort in the following subsection. All changes of this sort nevertheless beg a question of where these ideas regarding what is worthy in an action come from. The suggestion that different assumptions could be made merely highlights this point. We need, in short, a theory of norm formation. One natural place to go for this, following the discussion in the previous chapter on evolutionary processes, is history; and we pursue this thought briefly in section 7.4. We have thus arrived at a crossroads. Either we stick to game theory's often successful formula of reducing all outcomes to the initial data, or we acknowledge that good social theory demands an historical explanation of the source of agents' beliefs and, therefore, of their motivation (including their pay-offs). The book's subtitle betrays our choice.

For now, we conclude this discussion with a review of how Rabin (and other versions of psychological games) have changed game theory.

The highlight of the preceding analysis is the thought that, once psychological utilities enter the scene, the theorist needs to know the character of people's beliefs about each other's actions before pay-offs are calculated and alternative strategies assessed. To get to a unique utility assessment of an outcome, we must assume an equilibrium of beliefs. This turns what used to be a simple unidirectional system of causation in game theory, running from utilities to rational beliefs to equilibrium, into a form of circularity. This is especially disturbing when the requirement that the beliefs be in equilibrium do not typically produce a unique set of beliefs, as seems to be the case in psychological games.[6] To use the apparatus of game theory to predict what rational people will do, we need to know what beliefs *actually* obtain. But if one knows this, then the apparatus of instrumental rationality is no longer really needed to explain how people act.

It is true, of course, that action can be instrumentally rationally reconstructed once the beliefs are known, but knowing the equilibrium beliefs about action is enough to predict what actions will be taken and it seems almost simpler to say that people's actions have been guided by the prevailing norm. This is the key aspect of psychological games: pay-offs and outcomes are both norm-driven.

There is another way of appreciating what has changed in this chapter. Since indeterminacy has plagued game theory from the start of this book, it may seem that the indeterminacy of psychological games adds nothing to the argument. However, until this chapter the indeterminacy has suggested a weakness in the scope of the instrumental model of rationality, rather than a fundamental flaw. The model itself still had value once some theory of belief formation was grafted on (eg through a combination assuming a bounded form of this rationality and beliefs that are generated through an evolutionary process). So one would have to concede that something else was needed to explain action, but it still made sense to talk about people acting so as to satisfy their preferences. In this chapter, the contrast is most marked because the indeterminacy goes to the heart of the model. 'Preferences' are not given independently of beliefs and the indeterminacy of belief yields indeterminate preferences, so talk of acting on preferences becomes difficult to sustain.

7.3.4 An alternative formulation linking entitlements to intentions*

Suppose entitlements depend on intentions. In other words, if Jill intends good (bad) things to happen as a result of her actions, then Jack believes that she deserves more

* This section can be skipped without loss of continuity. The reader who intends to pass it over is advised to read only the first three paragraphs below.

(less). We consider such a possibility in this section and *the reader who has no special interest in the technical details of such an alternative definition may skip the rest of this section without loss of continuity.*

Consider some static game between A and B. Suppose that, having predicted that A will choose strategy s_A, B chooses to respond with strategy s_B. Let $E_B(s_B)$ denote our alternative definition of B's entitlement [juxtaposed against Rabin $e^B(s_B)$]. To ensure that $E_B(s_B)$ depends on the combination of B's choice (s_B) *and* his intentions (as opposed to just the former), the following must hold if $E_B(s_B)$ is to be non-zero:

(a) *B must be sacrificing utility*: i.e. $\pi_B(s_A,s_B) < \pi_B^n(s_A)$; where, $\pi_B(s_A,s_B)$ is B's payoff from choosing s_B (when he expects A to play s_A) and $\pi\pi_B^n(s_A)$ is B's payoff from choosing his *best reply* strategy in response to s_A, ***and***

(b) *A must benefit, or lose out, from B's sacrifice:* i.e. $\pi_A(s_A,s_B) > \pi_A^n(s_A)$ in the case where B is being kind A, or $\pi_A(s_A,s_B) < \pi_A^n(s_A)$ when he is nasty; where, $\pi_A(s_A,s_B)$ is A's payoff when the two players choose strategies s_A and s_B and $\pi_A^n(s_A)$ is A's payoff from choosing s_A when B plays his best reply strategy to s_A.

As long as (a) and (b) hold, A must think that B is entitled to a payoff in excess of (less than) $\pi_B^n(s_B)$ by virtue of his kindness (nastiness) to her. In other words, A must feel that B deserves to get something more (less) than what he could have expected under normal Nash-like, best-reply, play. How much more (less)? An obvious (and plausibly 'fair') answer would be that B deserves to benefit (hurt) to a degree proportional to (*i*) the benefit (loss) he has bestowed upon A, and (*ii*) to the magnitude of his own sacrifice. Finally, note that if (a) does not hold, then B deserves neither more nor less than what he will get from normal Nash-like play.

The following is one possible specification for $E_B(s_B)$ satisfying the above requirements:

$$E_B(s_B) = \frac{[\pi_A(s_A,s_B) - \pi_A^n(s_A)]R^B(s_A) + \pi_B^n(s_A)R^A(s_A)}{R^A(s_A)} \times \left| \frac{\pi_B(s_A,s_B) - \pi_B^n(s_A)}{R^A(s_A) + R^B(s_A)} \right| \qquad (7.7)$$

where $R^A(s_A)$ is the range of A's payoffs when she plays s_A ($\max\{\pi_A(s_A)\}$-$\min\{\pi_A(s_A)\}$ and $R^B(s_A)$ is B's range of payoffs when A plays s_A ($\max\{\pi_B(s_A)\}$-$\min\{\pi_B(s_A)\}$.

Note that, from A's perspective, (7.7) makes B's entitlement proportional to the absolute magnitude of *his* sacrifice (relative to the range of both players' payoffs when A plays s_A), to *her* resulting benefit or loss (relative to *her* range of payoffs when she plays s_A) and, finally, to *his* payoffs were he selfishly to stick to his best reply strategy. When B is making no sacrifice one way or another, A's normative commitment to *his* welfare vanishes; i.e. she does not think that B is *entitled* to anything.

Figure 7.2 below gives A's perception of B's entitlements corresponding to: B's choice of strategy (AbB:s_B), and A's perception of B's intention. Note that the latter perception derives from A's second order belief (AbBbA:s_A); e.g. when AbBbA:*h* and AbB:*h*, A thinks that B is making a sacrifice in order to hurt her. For why else would he be playing *h* when he expects her to play *h* too? Surely, his Nash best reply to her *h* is *d* which must mean, A concludes, that in playing h he is

deviating from his Nash best reply in order to make her suffer. So, when AbBbA:*h* and AbB:*h*, A estimates that he is entitled to utility of –2/3.

By contrast, if A thought that the reason why B is about to play *h* (AbB:*h*) is his belief that she will play *d* (AbBbA:*h*), then A no longer thinks that B is trying to hurt her. She simply interprets his (predicted) intention to play *h* as a Nash best reply (and, thus, morally neutral) action. In this case, therefore, Figure 7.2 reports that A believes that B is entitled neither to positive nor to negative material pay-offs: he is morally neutral and therefore deserves neither to be helped nor to be harmed by her. Notice that Rabin's formulation makes no such distinction (Rabin's entitlements are in brackets): According to Rabin, A thinks that B is entitled to pay-off 2 (his material pay-off from the pure strategy Nash equilibrium favouring him) regardless of her interpretation of his intentions. We believe that expression (7.7) is much better tuned into the rationale of fairness equilibria, as explained in Section 7.3.1.

Hawk-Dove				**The prisoner's Dilemma**			
		AbB:*h*	**AbB:*d***			**AbB:*d***	**AbB:*c***
E_B	**AbBbA:*h***	-2/3 *(2)*	0 *(1/2)*	E_B	**AbBbA:*d***	0 *(2½)*	2/3 *(1½)*
	AbBbA:*d*	0 *(2)*	1 *(1/2)*		**AbBbA:*c***	0 *(2½))*	4/3 *(1½)*

Figure 7.2 A's estimate of B's entitlements according to equation (7.7)
(Rabin's entitlements in brackets and italicised)

To see this better, suppose that A expects B to play *d* as a best reply to her own *h* (i.e. because AbBbA:*h*). Again A thinks that B is *not* entitled to her benevolence. *But*, if she thinks that he is playing *d* in order to help *her* (i.e. when AbBbB:*d*) she thinks that B is entitled to payoff 1; i.e. to a gain greater than payoff 0 which is proportional both to *her* gain and to *his* sacrifice.

Turning to the *Prisoner's Dilemma*, first we note that Rabin's specification of B's entitlement is counter-intuitive: B is entitled, on the grounds of fairness, to a greater payoff when he defects than when he co-operates (i.e. payoff 1 when he defects and 0 when he co-operates). This is simply unsustainable. In contrast, our specification is such that a defecting B does not deserve anything (either positive or negative), since he is not making any sacrifices either to benefit or to hurt A.[7] Indeed, whenever B is choosing a dominant (or, more generally, a Nash best reply) strategy he is being, by definition, kindness-neutral and, consequently, A 'owes' him nothing (either positive or negative). Entitlements come into play in the *Prisoner's Dilemma* only when B deviates from his dominant strategy and co-operates. In that case, his entitlement is always positive (since his co-operation always benefits A) and greatest when B is expecting A to co-operate too.[8]

We now need to define alternative functions to Rabin's (7.5) and (7.6) so as to measure the kindness/nastiness shown by one player to another given their 1st and 2nd order expectations. We specify A's kindness function to B (f_A) so that it takes a positive value when A is kind to B, a negative value when she is being nasty to him and a zero value when she is being neither kind nor nasty to him.

$$f_A(s_A,s_B)=\begin{cases} 0 \text{ when there exists an alternative strategy } s_A^* \text{ such that} \\ \pi_B(s_A^*,s_B)>\pi_B(s_A,s_B)>\pi_B^n(s_A) \text{ and } \pi_A(s_A,s_B)\le\pi_A(s_A^*,s_B)<\pi_A^n(s_A) \\ \text{OR} \\ \pi_B(s_A^*,s_B)<\pi_B(s_A,s_B)<\pi_B^n(s_A) \text{ and } \pi_A(s_A,s_B)\le\pi_A(s_A^*,s_B)<\pi_A^n(s_A) \\ \left|\dfrac{\pi_B(s_A,s_B)+R_B(s_B)}{E_B+R_B(s_B)}\right| \text{ when } \pi_B(s_A,s_B)>\pi_B^n(s_A) \text{ and } \pi_A(s_A,s_B)<\pi_A^n(s_A) \\ -\left|\dfrac{\pi_B(s_A,s_B)+R_B(s_B)}{E_B+R_B(s_B)}\right| \text{ when } \pi_B(s_A,s_B)<\pi_B^n(s_A) \text{ and } \pi_A(s_A,s_B)<\pi_A^n(s_A) \end{cases}$$

(where R_B is the range of B's payoffs when he chooses strategy s_B)

Re-definining A's Kindness/Nastiness - Expression (7.8)

Expression (7.8) replaces (7.5) and offers a more complicated 'theory' of what constitutes kindness/nastiness. The first line of expression (7.8) demands that players think of acts as kind/nasty only if they are efficient. If A intends to be nice to B by choosing non-Nash strategy s_A but, meanwhile, there exists another strategy s_A^* which would have benefited B at no extra cost to her, then A is deemed irrational rather than kind. Thus A's kindness function becomes zero. Similarly when A wants to hurt B. If her choice of strategy is 'inefficient', her nastiness function is, again, set equal to zero.[9]

The second line specifies that A's kindness to B, when B is sacrificing utility to help her, is a positive function of the proportion of B's entitlement that A's choice allows him to enjoy. Finally, the last line suggests that, when B is hurting A at a cost to himself, A's nastiness to him is a function of the extent to which A's choice inflicts on B the loss that he deserves (or that she is 'entitled' to inflict upon him).

The following table presents the values of f_A in our two games depending on A's 1st and 2nd order beliefs, as given by expression (7.8):

Hawk-Dove				The Prisoner's Dilemma		
f_A		**AbB:h**	**AbB:d**	f_A	**AbB:d**	**AbB:c**
AbBbA:h		-3/5 *(-1)*	0 *(- ½)*	**AbBbA:d**	0 *(-½)*	-2/3 *(-½)*
AbBbA:d		0 *(0)*	1 *(½)*	**AbBbA:c**	2 *(½)*	9/10 *(½)*

Figure 7.3 A's Kindness/Nastiness to B according to expression (7.8)
(Rabin's values italicised in brackets)

Note that (unlike Rabin, 1993) no unkindness is involved when A thinks that B is playing some pure Nash strategy. The reason is that playing Nash involves no sacrifice on B's part and, therefore, it cannot possibly incite (on the strength of reciprocity) any sacrifice from A. Put differently, from a psychological point of view, mutual Nash play is tantamount to kindness-neutrality. Indeed A's kindness (nastiness) surfaces, i.e. $f_A>0$ (or $f_A<0$), only when A is acting in a manner that furnishes B with a payoff greater to (less than) he would have expected under Nash-play.

Turning to our two games again, we note that in *Hawk-Dove* player A can show nastiness only to a B whom she expects is playing *h* in order to hurt her.[10] Moreover, in the *Prisoner's dilemma*, no nastiness is involved when players choose to defect (again in sharp contrast to Rabin).[11]

We are now ready to re-define the psychological pay-offs by replacing Rabin's expression (7.4) with expression (7.9) below. This is a simple yet effective way of capturing A's psychological pay-offs (Ψ_A) as the product of f_A and f_B – where the latter is computed by an expression very similar to (7.8). A's overall utility is still given by expression (7.4), only this time the psychological component of the player's utility (Ψ_A) is given by our alternative formulation above:

$$U_A(s_A, s_B) = \pi_A(s_A, s_B) + \mu\Psi_A(s_A, s_B) \quad where\ \Psi_A(s_A, s_B) = f_A(s_A, s_B) \times f_B(s_A, s_B) \quad ---(7.9)$$

and μ is, as before, the weight placed by A on her psychological utility (relative to her material utility).

As before, (7.9) confirms that, when A anticipates kindness (nastiness) from B, she loses psychological utils if she fails to reciprocate that kindness (nastiness). Note however that, in a manner reflecting Rabin's discussion better than his own formulation, the utility function above takes different values depending on A's second order beliefs.[12]

Let us now re-write in Figure 7.4 the overall pay-offs *in equilibrium* (recalling once more the crucial point that, out of equilibrium, psychological games are ill-defined) for games *Hawk-Dove* and the *Prisoner's Dilemma*.

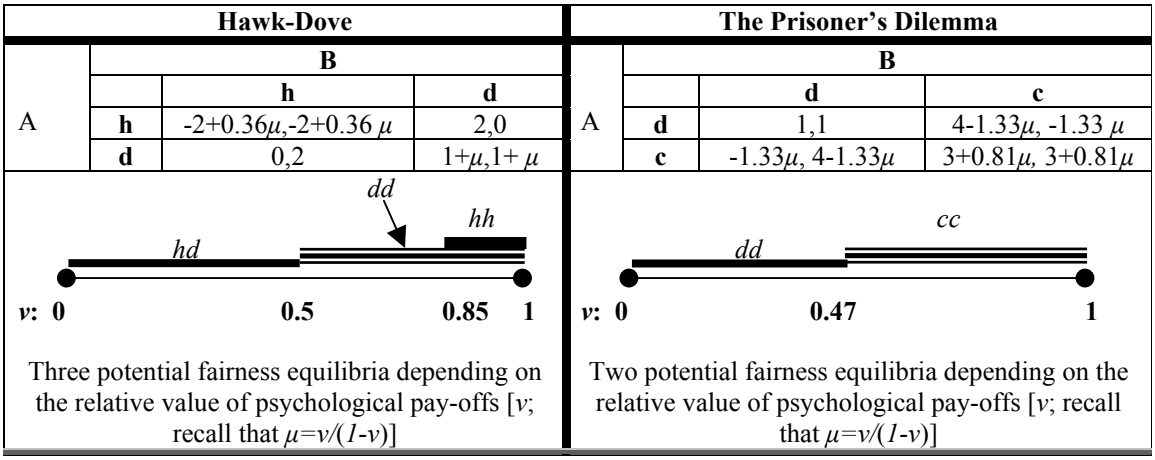


Figure 7.4 – *The transformed games under the re-defined psychological pay-offs*

The fairness equilibria that result are similar in structure to those following Rabin's transformation. However, there is one important analytical difference with Rabin's model: ***Our transformation, unlike Rabin's, is such that Nash-play is psychologically neutral (i.e. has no psychological effects).*** Moreover, they are consistent with the idea that players' perceptions of entitlements, and thus of fairness, depend on their perceptions of the motives behind their opponents' actions.

To see the point about the psychological neutrality of Nash equiliibria, consider the original pure strategy Nash equilibria of both games (*hd*&*hd* in *Hawk-Dove* and *dd* in the *Prisoner's Dilemma*): Our transformation leaves the associated pay-offs intact. The reason is that, in our case, *Nash-play is*, by definition, *psychologically neutral for players* (i.e. their psychological payoffs are zero). If *cc* in the *Prisoner's Dilemma* and *dd*/*hh* in *Hawk-Dove* become fairness equilibria, it is because they reward players with significant psychological rewards (such that the material incentive to play Nash is overcome). By contrast, Rabin's transformation assigns, wrongly we think, non-zero psychological payoffs to Nash equilibria.

Now, this difference is not a mere technicality as it affects our interpretation of fairness equilibria. For example, consider mutual defection in the *Prisoner's Dilemma*. Both our transformation and Rabin's report that *dd* is a fairness equilibrium (independently of the players' relative valuation of material and psychological payoffs). However, Rabin is forced to insist that *dd* must necessarily be a mutual nastiness equilibrium. In our case this is not so: *Mutual defection is a mutual kindness-neutral equilibrium.*

We think this is important because our re-configuration of fairness equilibrium throws a bridge between the concept of fairness equilibria examined in this section and the idea that games are regulated by a sense of justice springing out of the expectations that people will (or ought to) comply with the current conventions for playing the game.

Conclusion

In this section, we re-defined players' entitlements so as to reflect not only their actions but also their intentions. Moreover, we postulated that players are only entitled to 'something' when they deviate from their Nash best replies either to benefit or to harm their opponents. Under these assumptions, we showed that mutual defection in the Prisoner's Dilemma *cannot* be a mutual-nastiness equilibrium. Rather, it occurs when players are locked in expectations of mutual kindness-neutrality. More

generally, in an equilibrium between 1st and 2nd order beliefs, a game's original (pure strategy) Nash equilibria come with zero psychological payoffs (unlike in Rabin, 1993). However, they may not be fairness equilibria (e.g. when μ>0.5 in *Hawk-Dove*) because other competing outcomes (e.g. mutual dovishness) may offer players a positive inner glow which the original Nash equilibria cannot match.

Box 7.5

INTENTIONS MATTER!

In 1351 English law made it illegal to "compass or imagine the death of our lord, the King." The fear among Royalist circles caused by French Revolution led to the 1794 treason trials. Those tried did not plan to kill the King, nor even depose him. Robert Watt and David Downie, two of the defendants, were accused, by the Scottish Lord Advocate, "of overt acts of imagining thoughts which had a tendency to touch the life of the sovereign". Thomas Eriskine dismissed the defendants protests that they never took a single step against the King thus:

"I protest against all appeals to speculations concerning *consequences*, when the law commands us to look only to *intentions*." (see Barrell, 2000).

Standard game theory occupies the opposite extreme to that of Mr Eriskine: only consequences determine utility and thus only outcomes matter. In Law (excepting periods of witchhunts, as in 1790s Britain or McArthite America), both intentions and consequences are important. The prosecution needs to prove intent beyond reasonable doubt. Psychological game theory comes closer to this recognition of the joint importance of consequences and intentions. It breaks down the Humean distinction between desires and beliefs and thus allows the former to be influenced by second order beliefs – just as any decent Court of Law would expect.

7.3.5 Team thinking

Some game theorists (eg Sugden, 1993 and Bacharach, 1999) have drawn on ideas from the philosophical discussion of collective intentionality (see Tuomela, 1995, and Gilbert, 1989) to reconceptualise how people sometimes act in social interactions. The basic idea is easy to appreciate.

Consider how best to describe why the central defender in a soccer match decided, upon tackling in his own penalty area and winning the ball, to pass it promptly to a colleague 6 metres away in the midfield. Why didn't he take the ball out of defence and try to beat a few players before passing or shooting at the opposition goal? Anyone who has played football will know that the satisfaction of taking the ball past an opponent is second only to scoring a goal, while the execution of a simple 6 metre pass is humdrum. One explanation, using a simple version of the rational choice model, turns on the risks associated with each type of action and if this is not enough, the model will appeal to the thought that the player is employed to defend and if he does not concentrate on this team role, then he will probably be sold on at the next opportunity. In this way, his individual preference for a bit of artistry with the ball is subordinated to the team interest which has him as a defender. The alternative explanation is that when he puts on the number 5 shirt, he stops being an individual and he becomes a member of a team; and once he does this, he thinks and decides what best to do with reference to the team's interests and not his own.

Here is another illustration from a soccer match. The striker is facing his midfielder who has the ball, he has the opposition central defender behind him and he has to decide whether to spin off to the right or the left. The midfielder meanwhile has

to decide whether to play the ball into open space to the right or the left of his striker. This is a form of coordination game where (right, right) Pareto dominates (left,left) because there is a bit more space to the right than to the left, so the chances of scoring are slightly better with the ball played to the right.

Magically at the same moment, the striker spins to the right and the ball is played into open space on the right for him to run on to. Again one explanation of this is that both players are paid to perform in the team's interest and although coordination is difficult among pure rational choice agents (as we have seen so often), they have rehearsed such moves so often that they simply 'know' that each will choose the right. The alternative is that they both take on the 'interests' of the team when playing and so they both know that playing to the right offers the best chance of achieving the team's goal. There is no complex problem of coordination now as each acting in the interests of the team takes the unique action that maximises the team's objectives.

More formally, there are two aspects to the idea of team thinking. The first is that there are 'team preferences' over outcomes. For instance in the *Prisoners' Dilemma* the team might order outcomes according to the average pay-off for each player as follows:[13]

	D	*C*
d	1,1	4,0
c	0,4	3,3

	D	*C*
d	1,1	2,2
c	2,2	3,3

How 'team thinking' may transform a *Prisoner's Dilemma* (Game 2.18) into a *Co-ordination Game* (Game 2.15) simply by allowing/encouraging players to pursue the maximisation of (the team's) average pay-offs

The second is that a person can 'team think' when they have reasonable confidence that the other players are also team thinkers. To be a team thinker means that a person considers what action each person should take in order to maximise the team's objectives and he or she then takes whatever is his or her part in this optimal plan. So in the *Prisoners' Dilemma* illustration the actions that optimise the team's objectives are (c,c) and each 'team thinker' decides to cooperate.

This account seems to raise three immediate questions. Where do team preferences come from? Possibly related to this, what determines the character of these preferences? And when do people have reasonable confidence that others are also team thinking?

The first two of these questions are sometimes answered together by positing that people agree to become members of a team and so, we can deduce that minimally a team ordering will satisfy the Pareto principle when applied to the members' individual preferences. Some teams may be built around richer types of agreement, others may not; so not much more can be said about this *a priori*. The last is trickier because it seems to reproduce a kind of coordination problem as each seems likely to team think on this account provided they expect others to. Even if this problem is overcome, since the Pareto principle will typically only help in coordination games, the first two answers may not appear to be particularly useful as far as what might happen *a priori* in a reasonably wide range of social interactions when people potentially team reason. In short, this approach may not be able to say very much.

This, though, seems a rather hasty conclusion. After all, Rabin's theory (and others in the same vein) have to specify the character of the norm that produces

psychological pay-offs and this is really no different to having to know what are a team's preferences. Indeed they seem to be two different ways of saying the same thing. Again we have seen with 'fairness equilibria' that the same basic question as to whether the norm actually influences behaviour remains as there are typically multiple fairness equilibria. So the same criticism applies to these theories.

The real issue here, however, is how to interpret these points. Are they weaknesses with these theories or are they weaknesses with an approach that tries to work only with the data of given individual preferences, instrumental rationality, common knowledge of that rationality, etc. In short, is this a problem for these theories or game theory as conventionally constructed? Our answer will be clear.

As a final argument in support of our position, Sugden (2000b) interestingly notes that, in relation to the ideas of team thinking, what needs to be assumed by this theory is really no different from what is standardly assumed by the conventional game theoretic approach. Thus the conventional approach assumes individual preferences and offers no account of their provenance; so when 'team thinking' does the same with respect to 'team preferences' there is no obvious difference. In addition, of course, 'team thinking' requires the unit of analysis to be specified: is it a case of team or individual thinking that guides an action. But this is very similar to the kind of assumption that the simple rational choice theory has to make about when two goods are the same/different for the purposes of its theory. For example, during the boycott of South African goods, one needed to know, when applying the theory of rational choice to the purchase of fruit, whether a South African orange was the same as a North American one. In short, the theory equally depended on knowing whether a certain set of beliefs influenced behaviour.

7.4 PSYCHOLOGY AND EVOLUTION

7.4.1 On the origins of normative beliefs: an adaptation to experience

Recall the laboratory experiment we reported in Section 6.3.2. We found that in *Hawk-Dove* type games norms emerge which distribute gains asymmetrically on the basis of a totally arbitrary distinction: an initial random colour assignment. Players of one colour end up in the better role consistently while players of the other colour accept their lot and play to minimise their losses. This result was explained well by evolutionary game theory. However, there was a second observation which the theory failed to explain. When the *Hawk-Dove* was augmented with a third co-operative (but non-equilibrium) strategy, the players who had, by that time, found themselves in the disadvantaged role (under the evolved discriminatory convention) co-operated with one another with remarkable alacrity. By contrast, players who were advantaged by the norms of distribution almost never co-operated with each other (recall Figure 6.6). Why?

Our decision to end this book with a chapter on psychological game theory was motivated in part by the desire to answer such questions. Why does evolutionary theory fail to explain the different attitudes to co-operation between different groups of people? Has psychological game theory shed any light on this? Are we wiser now about the reasons why people co-operate with others like them who are also the victims of evolved, arbitrary discrimination? The answer is 'yes' and 'no'.

While it is true that psychological game theory *does* explain non-Nash co-operation in games like *Hawk-Dove* and the *Prisoner's Dilemma* (see Figures 7.1,7.2 and 7.4), it cannot explain why the 'weak' are drawn to co-operation while the 'strong' are not in our experiment. One possible explanation of the experimental data in Figure 6.6 is that perceived entitlements adapt to the players' past pay-offs. In other words, to return to the argument of 7.3.3, that 'entitlements' should not be treated exogenously, one way of endogenising them is to appeal to an evolutionary process.

In particular, recall Sugden's(1986) argument that predictive beliefs became normatively charged. Echoing David Hume, he suggested that agents find it hard to accept that the convention which determines their behaviour could have been otherwise (even though it might easily have been), so people develop normative reasons to support the convention. It is not only a coordinating device, it embodies ideas of 'justice', 'fairness', etc. Interestingly, if this is how the normative beliefs of both advantaged and disadvantaged players evolve in the *Hawk-Dove* game, the data in Figure 6.6 ceases to be a puzzle. The players who are disadvantaged by the colour convention would develop humbler entitlement expectations as compared with those who are advantaged. As a result, with a sufficiently lower set of perceived entitlements, the conflictual outcome *hh* could cease to be a fairness equilibria and *dd* could become one. By contrast, advantaged players with higher perceived entitlements may not be spared *hh* and may find themselves locked into a mutual *hawkish* equilibrium with players of the same colour.[14] Why don't they 'evolve out' of these normative expectations if the latter cause them to fight each other at a great cost? The simple evolutionary answer is that such normative beliefs, despite causing much conflict between the 'strong', reward them amply in meetings with the 'weaker' players.

Box 7.6

HOW ADVOCACY CAUSES BIAS AND ALTERS PERCEPTIONS OF ENTITLEMENT

Babcock, Lowenstein and Issachoroff (1995) report on a fascinating dispute-resolution experiment. Law students were separated in two groups and given a complete dossier containing all the evidence that a judge was presented with in a real court case regarding the level of compensation of a car accident victim. They were asked then to build cases which they presented, in person, in a mock trial. Members of the first group were told to prepare their case as if they were the victim's attorney while members of the second group were asked to prepare the case for the attorney acting on behalf of the insurance company. Later (that is, after the students had prepared their speeches) they were asked to predict, on the basis of the dossier's contents, the award that the *real* judge had decided upon. The more accurate their prediction the higher their dollar pay-out. As it turned out, the first group's average estimate of the judge's compensation figure was significantly higher than the second group's, even though all students had read the same dossier. The authors conclude thus: "Even when the parties have the same information, they will come to different conclusions about what a fair settlement would be and base their predictions of judicial behaviour on their own views of what is fair."

7.4.2 On the origins of normative beliefs: The *resentment-aversion* versus the *subversion-proclivity* hypotheses

Sugden (2000a) offers a rather different account of how conventions (or mere empirical regularities) come to motivate through affecting player's pay-offs. He proposes a psychological equilibrium that he calls a *normative expectations equilibrium* (NME) which is similar to Rabin's *fairness equilibrium*, but which dispenses with any account of the character if the norms that affect behaviour. Instead, it is enough, rather like the early discussion of how second order beliefs motivate in section 7.2, that people expect someone to behave in a particular way (whatever it is) for that person to incline towards that action on psychological grounds. This psychological mechanism seems to be traced to its evolutionary role in conflict avoidance and is captured by the idea that humans are averse to the resentment of others. This is his *resentment hypothesis.*

Sugden's *Fairness* as founded on his *Resentment-Aversion Hypothesis* (Definition)

Fairness: In equilibrium, person A is *fair* towards person B as long as A does not do anything that B had not expected A to do (conventionally) and A is not hurting herself.

Resentment: If A acts *unfairly* in the sense above (that is, unpredictably), B will feel resentment toward A.

Resentment aversion: Players who cause resentment in other people's minds forfeit utility. Thus, utility maximising agents are resentment-averse.

So Sugden (2000a), in effect, defines 'fairness' as conformity with the evolved status quo. Anything that frustrates others' expectations is deemed unfair, goes against the grain of their expectations, and is the cause of negative psychological utility. People, in this view, are driven by the psychological desire to avoid the disapproval that comes from frustrating others expectations.[15]

Box 7.7

PSYCHOLOGICAL DARWINISM (PsyD)

Why do people love their children? Pinker (1997), a proponent of PsyD, answers: "People love their kids not because they want to spread their genes but because they cannot help it. Genes try to spread themselves by wiring animals' brains so the animals love their kin… and then they get out of the way." This is not uncontroversial. To argue that genes wire parents' brains to love their children is to say that loving one's kin is innate. But this is not unique to PsyD; most people (including creationists) believe

that too. PsyD requires more: that our psychological traits be *adaptations*. Some argue that "we know enough about ourselves that does not square with the theory that our minds are adaptations for spreading our genes" (see Fodor, 1998). This is an important question regarding the origin of normative beliefs (e.g. Sugden's *resentment-aversion* hypothesis). Are they to be traced to some conspiracy by our genes? Or is there something in human reason and the 'games' unique to human communities that is irreducible to our genes' interests?

One of the lessons of evolutionary game theory (see the previous chapter) is that little can be deduced *a priori* about what will happen when individuals interact in a variety of settings. This is because there are typically an abundance of evolutionary equilibria. On these grounds, it is probably unwise to assume that a process of genetic adaptation would always produce the same kind of outcome for the personality of individuals. This observation not only weighs against simple suggestions like genes being responsible for why people love their children, it equally also counts against those who explain psychological gender differences in terms of the evolutionary pressures that have moulded behaviour. In short, the fact that evolutionary processes seem not to be associated with unique outcomes ought to make it difficult to tell simple stories about how our genetic inheritance is a direct consequence of the types of problems we typically encountered in the hunter-gatherer societies that have characterised most of human history.

Granted that we all experience a certain dissonance from causing resentment in others, it is still unlikely to be the only primitive urge. It seems to us that humans equally have a *subversive tendency* (that is, our tendency to want to subvert others' expectations of us); otherwise it is likely to be difficult to explain how people ever consciously escape the status quo. Formally, we might define our proclivity to subverting others' beliefs as follows: if A expects B to expect A to perform *X* and yet A chooses some other action, *Y*, for the purposes of causing resentment in B (through frustrating B's expectations about A), then A is being purposefully subversive.

Clearly, subversion is a disequilibrium phenomenon as it implies that A's higher order beliefs about B are out of alignment. By contrast, Sugden's *resentment-aversion hypothesis* is an equilibrium notion since it relies on common knowledge that B has good reason to form the empirical expectation that A will do *X* rather than *Y*.

The *Subversion-Proclivity Hypothesis* (Definition)

Conformism: In equilibrium, person A is a *conformist* as long as A does not do anything that B had not expected A to do (conventionally) and A is not hurting herself.

Subversion: If *A* acts *contrary* to B's expectations (that is, unpredictably), B will think of her as subversive and will feel a combination of *resentment* and *admiration* toward A.

Subversion proclivity: Players who gain net utility from causing in others this combination of *resentment* and *admiration* are characterised by *subversion-proclivity*.

"The main constituents of a satisfied life appear to be… two: tranquillity and excitement." J. S. Mill

The reader will notice immediately the dependence of *subversion-proclivity* on the prior evolution of Sugden's *resentment-aversion* as well as the tension between the two. When these two tendencies are played out in historical time, and in the context of the simultaneous evolution of behaviour and motivation, the result is a never-ending cycle between periods of stability (during which some convention is established in accordance with the *resentment-aversion hypothesis* - RAH) and subsequent periods of flux (during which older conventions are being disestablished, in accordance with our *subversion-proclivity hypothesis* - SPH). It is interesting to

recall that this conflict of primitive (though not irrational) urges, was the foundation of the critique of subgame perfection in Sections 3.4 and 4.4; i.e. that games like the *Centipede* (or Rubinstein's, 1982, bargaining game), are indeterminate due to the irrepressible tension between an equilibrium and a subversive logic. [16]

To support SPH, we need two things. First, we need to link SPH to some primitive human psychological trait (as Sugden did with his RAH). Secondly, we need a plausible story as to how the proclivity to subvert and frustrate others' expectations subversive has been reinforced through the evolutionary process. With regard to the former, it seems to us that we are often torn between seeking others' approval though conforming with their expectations and wanting to impress (others as well as our selves) through uncommon behaviour; behaviour that helps us 'stick out'.

The tensions between these two urges arises because, often, the most effective way of getting noticed is to frustrate others' expectations about us and (at least initially) causing them to resent us. Indeed, causing resentment in others (at least initially) may be a prerequisite for the success of our strategy to impress and get noticed. This is a fascinating aspect of our confused, and at once majestic, nature that we often admire persons for precisely the same reasons for which we also resent them. The question now is: what is its social function and how did it come about?

In the case of the *resentment-minimisation* psychological trait, Sugden's neo-Humean (evolutionary) explanation is clear: conformity generates regularities which help populations reduce the chances of costly conflict. However, by the same token, we can explain evolutionarily the reinforcement of the subversive trait if we can show that a periodic purge of established conventions increases a community's fitness.

As we saw in Chapter 6, a well-established convention may well be 'inferior' compared to alternative ones (e.g. the inefficiency of QWERTY). Indeed, a discriminatory convention which reduced conflict effectively in the past (by arbitrarily advantaging one subpopulation over another) may have exceeded its use-by date (e.g. as a result of technological change). Therefore, communities benefit from a capacity to undermine (and thus test for the evolutionary stability of) what Sugden refers to as *normative expectations equilibria.* If that capacity is related to the subversive trait in us all, one can argue that the evolutionary process reinforces *at once* two contradictory traits of human nature: *resentment-aversion* and *subversion-proclivity*.

We mentioned above the possibility that, in stratified societies, a person belonging to a disadvantaged group can gain substantial kudos from subverting the established discriminating convention, at least within her own group. To make this point, however, we need to re-draft our SPH in terms consistent with evolution in more than one dimension.

The *One-Dimensional Subversion-Proclivity Hypothesis* (ODSPH) (Definition)
Suppose there is a population P and some convention C which has evolved earlier one-dimensionally (recall Section 6.2). By definition, in interactions of a given kind between members of P, C recommends to each person i the same action X (as opposed to Y). If some person j chooses Y, then this choice will engender a degree of resentment in the person she has interacted with and even among the rest of the population (assuming common knowledge of j's behaviour). Finally, suppose there has been a history H of continual choices in accordance with C by all members of P. Then (and this is the hypothesis) if j chooses Y, she will secure a degree of notoriety, admiration etc. proportional (a) to H and (b) to the degree of resentment caused by her

choice of Y. Thus, as long as persons within P have a taste for notoriety, admiration etc. (however small), there exists some H which will trigger subversion.

The *Two-Dimensional Subversion-Proclivity Hypothesis* (TDSPH) (Definition)
The difference with **ODSPH** above is that (the previously evolved) convention C is two-dimensional and discriminatory (recall Section 6.3). That is, it segregates (on the basis of some arbitrary feature) population P (conventionally) between two subpopulations (P_1 and P_2) and gives different instructions to $i \in P_1$ and to $j \in P_2$: It directs i to play, in a meeting with j, X and j to play Y. Suppose that i's utility is such that $U_i(i \text{ plays } X, j \text{ plays } Y) > U_j(i \text{ plays } X, j \text{ plays } Y) > U_i(i \text{ plays } X, j \text{ plays } X) = U_j(i \text{ plays } X, j \text{ plays } Y)$. Finally, if there is a history H of continuous adherence to C by members of both subpopulations, then j's choice of X (rather than Y) in violation of C will lend her some psychological utility from 'sticking out', notoriety etc. which is proportional (a) to H, and (b) to the resentment caused among members of subpopulation P_1.

So, the obvious parallel with evolutionary biology is to think of SPH as equivalent to mutations testing the stability of the established evolutionary equilibrium C. Will individual subversion succeed in undermining C? It depends on its capacity to spread by infecting others. One might speculate that in the one-dimensional case, the chances of subversion are limited. Each subversive move will be a tiny drop lost in a sea of conformity. Nevertheless, even under those circumstances, conventions are disestablished and customs change when the bandwagon of a new norm is ready to roll. The world of fashion is one area that comes to mind (see Box 7.8).

Box 7.8
SUBVERSIVE FASHION

In a private communication, Robert Sugden responded to the use of fashion as an example of one-dimensional subversive-proclivity as follows: "I think fashion would be an ideal area for creative game-theoretical analysis. But my preferred approach… would be to postulate that people have fairly constant preferences for the 'messages' transmitted by fashion goods (e.g. 'I care about appearances', 'I am up-to-date', 'I am young at heart', 'I am unconventional' etc) but the messages associated with different goods are endogenous (e.g. if only old people wear yellow, yellow is likely to signal oldness). If some of these messages are positional goods, a game in which people try to send the messages they want to send may have no equilibrium. Notice that this has elements of co-ordination (if I am buying a durable good, I want it to convey a certain message, and this may require that other people who want to convey the same message will buy it too) and of disco-ordination (e.g. if I want to signal 'I am unconventional' or 'I am at the cutting edge of fashion')."

The multi-dimensional case is, of course, far more interesting. Norms of honour among gentlemen are functional to norms of excluding women from the benefits of equality. Since convention C segregates P into subpopulations, each with its own behavioural pattern and normative/calculative expectations about the other, the success of subversive moves will clearly depend on whether j's subversion will give rise to collective acts of subversion by members of subpopulation P_2. To the extent that such 'collective spirit' is functional to the interests of subpopulation P_2, the emergence of correlated deviations from C (by members of P_2) are likely to be associated with other 'bonding' practices within P_2, e.g. greater reluctance to succumb to the norm of adhering to mutual defection in the *Prisoner's Dilemma*, or sub-population-specific lifestyle choices *vis-à-vis* music, fashion etc. To give a celebrated

example, the defiance of a sole middle aged black woman riding on a segregated bus in the American South would have gone unnoticed in the 1960s had there followed no coalition of black men and women who turned her subversive act into a campaign.

More grandly, it is tempting to claim that SPH lies behind behaviour which helps society discover not only new *ways* to play given games but of new *games* to play as well. In our conclusions to the previous chapter, we lamented evolutionary game theory's reliance on fixed payoff structures. This was the reason we found it to be insufficiently evolutionary. In this chapter, however, the idea that pay-offs are contingent on beliefs allows us to imagine a genuinely evolutionary theory of society.[17]

From this perspective, we should expect a theoretical account involving a mixture of (often opposing) social forces constantly equilibrating and subverting the evolving 'system'. As a result, we expect to find periods of continuity which are interrupted by severe discontinuities not only in the behaviour but, importantly in the structure of the social interaction (i.e. of the dominant games). At the level of beliefs, history makes itself felt in the never-ending establishment and (subsequent) subversion of normative belief equilibria. At the level of the cultural, the primitive appeal of subversion manifests itself in the best works of drama and literature (see Box 7.9).

Box 7.9

SUBVERSION AND CLASSICAL DRAMA

Theatre offers a beguiling glimpse of the timeless puzzlement caused by our predilection for subversion. The original myth of Prometheus was rather pedestrian: according to Hesiod, Prometheus stole fire from Mount Olympus, and delivered it on a whim (or, at best, in a paroxysm of thoughtless altruism) to an appreciative humanity. But when Aeschylus tells the story (in his play *Prometheus Bound*), suddenly we come across a colourful depiction of an intentional violation of the master's (Zeus) expectations regarding the behaviour of one of his own. Dramatic tension builds up as Prometheus subverts one of Zeus' expectations after the other, culminating in the final embarrassment when Zeus, surprised by and resentful of how gracefully Prometheus is taking his endless punishment, decides to end it.

Spartacus is another relevant figure who has also inspired theatre, operas, film etc. Like Prometheus, he did not gain fame through the centuries merely because of what he did (that is, for having started a war pitting slaves against the Roman legions). Rather, he gained prominence (both historically and culturally) because he succeeded in liberating the slaves from a Sugdenian *normative expectations equilibrium* which kept them contented with their slave-status. However, Spartacus's success would have been unthinkable had it not been for the incredible resentment he inspired among the Roman slave-masters.

Finally, no example is more pointed than that of Medea, the princess who caused maximum resentment within a whole community (and still does among contemporary audiences). By killing her children, she puts on display an extreme act of insubordination viz. the conventions of patriarchal society. Perhaps the best support for our *subversion-proclivity hypothesis* comes in the guise of Euripides' monologue in which Medea explains her strategy fully.

7.5 CONCLUSION: SHARED PRAXES, SHARED MEANINGS

The present chapter parted ways with conventional game theory in one important respect: it linked beliefs *directly* to desires. The result was, we wish to argue here, reminiscent of the distinction drawn in Section 1.2.3 between game theory's rules of the game, which are *regulative*, and Wittgenstein's rules of language games, which are *constitutive*.

It is indeed possible to interpret the norms of Rabin (1993) and Sugden (2000a) as akin to the rules of a Wittgensteinian language game. This interpretation seems plausible because, in this Chapter, norms are no longer simple regulative devices (as they were in previous chapters). They do a lot more than simply help satisfy pre-existing preferences (as they might be doing in a Humean or neo-Humean account). In fact they help *constitute* the players' actual preferences. Interpreting the rules is quite different to subscribing to them.

The analogy is helpful because it ties in with the change to the existence of symbolic properties associated with *action*, namely with their *meaning*. On Wittgenstein's view, the attribution of shared meaning to words in a language cannot come from some shared experience of either the external world or our inner feelings. *Shared meanings depend on shared practices*. This is a controversial claim because it depends in part on the impossibility of holding a private language. Nevertheless, it makes us social from the outset with language marking this fact, as does the existence of norms above, rather than either being a derivative from some version of exchange between pre-social individuals.

Throughout this book, we have made clear our dissatisfaction with the reduction of human *reasonableness* to the assumption of *instrumental rationality*. The conclusion here concerning the impossibility of knowing what one wants *outside a web of shared practices* is grist to that mill. One of the great gifts of game theory to the social sciences is that it has caused some thoughtful economists to question the assumption of instrumental rationality. The present chapter is based on results that emanated from such scepticism within the economics profession.

We choose to end this book by returning to the very first questions a total novice might ask of someone who just finished reading this book: *What is a game? How is it constituted?* Beyond saying that a game is a situation in which the outcome for one participant depends jointly on the actions of all, the answer must address the crucial issue of the players' motivation. Other textbooks deal with this issue concisely and without much discussion: players have pre-ordained preferences over the range of outcomes and they act in a manner that satisfies these preferences.

In contrast, our answer regarding motivation is quite different. Inspired by Wittgenstein, it comes in the form of a suggestion. Players' perception of their preferences is ill defined before the game is played. More formally, what is instrumentally rational to do is not well defined unless one appeals to the prevailing norms of behaviour.

This may seem a little strange in the context of a superficial reading of other game theory textbooks on, say, the one-shot *Prisoners' Dilemma*. In that game, game theorists proclaim, the demands of instrumental rationality seem plain for all to see: *Defect!* But, in reply, we would complain against the presumption of pay-offs which have fallen *as if* out of thin air, unvarnished by social experience. Games and humans evolved side-by-side over millennia. The norms that govern our behaviour also govern our interpretation of the events unfolding around us. A social setting requires interpretation before we know what we want and how much we value different

outcomes. But if the same norms that govern our behaviour are also implicated in those interpretations, how can we claim that motives are prior to games?

In conclusion, the study of psychological games has clarified the game theorist's dilemma: she or he may continue to pursue game theory's Holy Grail of 'closing' game theoretical explanations without 'outside' assistance. Or they may admit that *Indeterminacy* has won the day anyway, and use analyses like those offered in this chapter in order to understand the limits of methodological individualism. If we are right, game theory will keep tilting at the windmills of *Indeterminacy* until it goes out of fashion as the futility of this task becomes evident. This would be a shame. For game theory has a lot to offer, as we hope this book has demonstrated. It is a powerful tool with which to explore liberal individualism's limits and the difficulties of conjuring up satisfying social explanations. To go beyond this requires a change. Rather than 'solving' insoluble strategic interactions, or thoughtlessly applying existing 'solutions', the point is to figure out what games we play, how these came about and, perhaps, how we ought to change them.

Box 7.10

WHAT CAME FIRST? CAPITALISM OR THE PROFIT MOTIVE?

Adam Smith suggested that the division of labour in society was dependent upon man's "propensity to barter, truck and exchange one thing for another". This phrase was later to yield the concept of *Homo Economicus* whose clones populate all economics and game theory texts. Polanyi (1945) famously challenged Smith's view that there is something *natural* in people that turns them into merchants when the opportunity arises. According to Polanyi, Smith misread the past (by recognising potential merchants in the serfs, Lords and artisans of pre-capitalist societies). But, he added, "…[i]n retrospect it can be said that no misreading of the past ever proved more prophetic of the future..." (Polanyi, 1945, p.50-1). Polanyi's own view was that the newfangled motives (i.e. the propensity to barter etc. for profit) emerged at the same time, and for the first time, as genuinely new social games (that is, market societies, or capitalism) were being formed on the ruins of the feudal era:

> "The outstanding discovery of recent historical and anthropological research is that man's economy, as a rule, is submerged in his social relationships. He does not act so as to safeguard his individual interest in the possession of material goods; he acts so as to safeguard his social standing, his social claims, his social assets. He values material goods only in so far as they serve this end." (Polanyi, 1945, p.53)

The above view is consistent with this Chapter's analysis of the psychological aspects of pay-offs. The pursuit of social standing gives rise to different motivations, depending on the prevailing norms. Without knowing the norms, it is impossible to know their motivation. The two evolve, and bring new patterns to the fore, simultaneously. Neither the game nor the motivation comes first.

Karl Marx has often been disparaged for not grounding his theories of capitalism on the individual. His reasons can be seen more clearly in the light of the present discussion: For if the individual is not prior to capitalism, nor vice versa, what is the scope of any theory (e.g. methodological individualism) which takes the individual's motives as givens and only then tries to explain society against the background of these given motives? Marx's chosen solution was to deal with individuals theoretically "…only in so far as they are the personifications of economic categories, embodiments of particular class relations and class interests. My stand point, from which the evolution of the economic formation is viewed as a process of natural history, can less than any other make the individual responsible for relations whose creature he socially remains, however much he may subjectively raise himself above them." Marx, Preface to the first German Edition of *Das Kapital*.

Problems

7.1 Using Rabin's (1993) formulation, find the psychological pay-offs and the range of fairness equilibria in the case of the two games below. (Note that the first is a variant of Stag-Hunt, Game 2.14, while the second one is the *Hide and Seek* interaction, Game 2.17).

	s	*h*
s	10,10	-5,0
h	0,-5	5,5

A version of the Stag-Hunt

	Up	*Down*
Up	1,0	0,1
Down	0,1	1,0

Hide and Seek

7.2 Find the fairness equilibria in the case of *Hawk-Dove-Co-operate*, i.e. Game 6.4.

7.3 Let there be N bureaucrats and suppose $c_i \in [0,1]$ denotes bureaucrat i's chosen level of corruption; $p_i = \Pr(1-c_i)$ be the probability with which bureaucrat i will select level of honesty $1-c_i$ (or, equivalently, level of corruption c_i); $p_i' = E^{\text{public}}(p_i)$ be the public's average estimate of p_i; and $q_i = E^{\text{bureaucrat } i}(p_i')$ be B's estimate of p_i'; i.e. her second order belief regarding the probability with which she will be honest.
Let U_i be the utility function of bureaucrat $i=1,\ldots,N$ where:

$$U_i = \text{constant} + (\beta - \gamma q)c_i - \alpha(\sum c_i)/N$$

(A) Assuming that bureaucrats choose c_i once, find the psychological equilibria of this N-person game.

(B) Model the above game in terms of a one-dimensional evolutionary process.

NOTES

[1] Note that *c* is a best reply to *c* when $3a > 4(a-b)$ or $b/a > ¼$. When this inequality holds, *c* is a best reply to *c* while *d* remains a best reply to *d*. In this sense, we have two Nash equilibria in pure strategies and a strategic structure identical to that of the *Stag Hunt* (Game 2.18).

[2] We say ostensibly because, as we shall see in Section 7.3.3, the actual model diverges somewhat from the narrative in the introduction of Rabin (1993).

[3] The following classification of Rabin's (1993) assumptions, along with their labels, are not to be found in the original paper; they reflect our own interpretation of his paper.

[4] Combined with **Reciprocity**, this definition of neutrality implies the following: When Jill expects that Jack is being neutral toward her, she feels no urge to be either kind or nasty back. That is, unless neutrality is reciprocated with neutrality, psychological utility is forfeited ($\Psi<0$).

[5] Clearly, as μ rises psychological utility matters more to this person than material rewards and vice versa.

[6] For example, even though equilibrium has been assumed in Figures 7.1 no unique fairness equilibrium emerged.

[7] Note that a defecting B is making no sacrifice regardless of whether he expects A defect or to co-operate.

[8] Since when AbBbA:*c* A's relative gain is greater than when AbBbA:*d*, while B's sacrifice by co-operating is the same in both cases

[9] To give an example, consider Game 6.4, reproduced below for convenience. Suppose that A expects B to play strategy *c*. If she responds with *c* too, she is clearly being kind. The reason is that she is playing a non-Nash strategy (the co-operative *c* as opposed to her aggressive best reply *h*) in order to aid B. Suppose, however, that she replied with another non-Nash strategy: *d*. B would again benefit (albeit less) from her non-Nash behaviour *at her expense*: he would receive pay-off 0 as opposed to –1. Expression (7.8), however, determines that *d* is not really an act of kindness, even though B benefits at A's expense; it is, rather, an act of folly on A's part. If she wanted to make a sacrifice on his behalf, she should have chosen *c*. In this manner, both B and A would benefit most from reciprocated kindness. For this reason, function f_A takes the value zero when A responds to *c* with *d*: it simply rules out inefficient behaviour (i.e. a kind of foolishness) as a case of rational kindness.

		B		
		h	d	C
A	h	-2,-2	2,0	4,-1
	d	0,2	1,1	0,0
	c	-1,4	0,0	3,3

[10] This is in sharp contrast with Rabin whose rather crude specification insists that A's choice of *h* in response to B's *d* is always tantamount to an unkind act.

[11] The reason is that if B chooses *d* because he anticipated *d* from A, he is not hurting A at a cost to himself. Thus he means her no ill and deserves no nastiness from her. It is this thought that renders *dd* a mutual kindness-neutral equilibrium. On the other hand, if one co-operates with a defector one is being hugely kind (a kindness value of 2 is reported) toward a kindness-neutral person

[12] The reader who has not grasped this point yet may see it clearly by observing that the kindness functions f_A and f_B (see Figure 7.3) are determined by the players' entitlements (as in Rabin) which are in turn determined (and this is our innovation) by the player's second order beliefs.

[13] Note that this is analytically identical to the transformation we discussed in the context of Game 7.1 earlier in this chapter.

[14] It is easy to show that if in Section 7.3.2 entitlements e_i (or E_i in Section 7.3.3) are an increasingly function of the player's average pay-offs from previous interactions, then equilibrium *hh* shrinks in Figure 7.1 (and in Figure 7.4) in meetings between two disadvantaged players. The reason is that, as they learn to expect less, they increasingly feel that they deserve little and, thus, outcome *hh* ceases to be an equilibrium even for very high values of *v*. In contrast, when advantaged players meet, they bring with them great expectations of gain and, consequently, think that they deserve a lot. Even small values of *v* can then turn *hh* into a mutual-nastiness equilibrium.

[15] Notice that fairness was defined differently in the previous sections: Something more was demanded from A before we could proclaim her action 'fair': a degree of sacrifice (however small) when compared to what A could have got away with; a sacrifice that would lead to a benefit, or a loss, from someone who made a similar sacrifice to aid or pain us. In short, and unlike Sugden (2000a), to be fair one needed to pay a price.

[16] In the discussion of Section 3.4 Pettit and Sugden (1989) were among the authors who launched a remorseless critique of subgame perfection utilising a similar idea to our *subversion-proclivity hypothesis*. They showed that a mechanism which works through *frustrating* others' expectations (and thus causing resentment in their minds) renders subgame perfect equilibrium logic inconsistent. That critique of subgame perfection boiled down to a rejection of the possibility of common knowledge of the probability of attempts to subvert equilibrium beliefs (that is, the probability of a bluff).

[17] Of course, our account has left most interesting psychological categories out of the analysis. The features of shame and guilt, which often guide human behaviour, are two examples of motivation which is not subject to our control. However, this chapter does contain interesting pointers for some of the absent psychological categories. For instance, the neo-Humean attitude toward shame and guilt is not hard to imagine: deliverances of illusions bestowed upon us through social evolution. There is nothing *objectively* wrong, they might argue, with littering the streets or killing our mothers. It is just that society has become more stable and better able to reproduce itself when we all live under the fantasy that it is wrong to such things. As for shame and guilt, they are the psychological mechanisms which provide the requisite motiation at the level of our 'souls'.