Morals by Convention

The rationality of moral behaviour

Vangelis Chiotis

Ph. D. Thesis

University of York

School of Politics, Economics and Philosophy

September 2012

Abstract

The account of rational morality presented in *Morals by Agreement* is based, to a large extent, on the concept of constrained maximisation. Rational agents are assumed to have reasons to constrain their maximisation provided they interact with other similarly disposed agents. On this account, rational agents will internalise a disposition to behave as constrained maximisers. The assertion of constrained maximisation is problematic and unrealistic mainly because it does not explain how the process of internalisation occurs. I propose an amended version of constrained maximisation that is based on a conventional understanding of social behaviour and the social contract. Repeated interactions between rational agents lead to the creation of social conventions, which in turn serve as supportive mechanisms for behaviours that reinforce their stability. In addition, established social conventions facilitate and ensure information sharing, thus making it possible for conventional agents to know others' dispositions. The development and establishment of social conventions are best described and explained through an evolutionary account of social structures. The evolutionary account offers a more powerful and more realistic method of discussing cultural evolution, since it considers large populations over long periods of time and the interdependence between social structures and individual behaviour. In this context, information availability ensures that the most efficient conventions take over and maximising strategies become dominant. While for Gauthier moral behaviour depends on constrained maximisation, in the conventional account of morality it comes about as a result of repeated interactions between rational agents within the bounds of social conventions.

Table of Contents

Acknowledgements....5
1. Introduction....6
1.1 The social contract tradition....7
1.2 Game theory as a tool for the moral philosopher....9
1.3 Naturalising the social contract....12
1.4 Thesis overview....14
2. The Gauthier Project and its critics....17
2.1 A morally free zone....19
2.2 Minimax Relative Concession....22
2.3 The Lockean Proviso....26
2.4 Constrained Maximisation....28
2.5 The Archimedean Point....32
2.6 Overview of the "Gauthier Project"....34
2.6.1 Weak Morality....35
2.6.2 Moral Principles....39
2.7 Conclusion....43
3. Evolutionary Theory in Moral Contractarianism....45
3.1 Methodological Aspects....46
3.2 Spontaneous order....50
3.3 The Stag Hunt....52
3.3.1 Location....55
3.3.2 Communication....57
3.3.3 Association....58
3.3.4 Coevolution....60
3.4 Game Theory and the Social Contract....62
3.5 Criticism....67
3.6 Conclusion....73
4. Rationality and Evolutionary Theory....75
4.1 Why evolutionary theory....76
4.1.1 Evolution....76
4.1.2 Evolutionary game theory....78
4.1.3 Conclusion....79
4.2 Rationality....80
4.3 Functionalism and conventional behaviour....86
4.3.1 Functional analysis....86
4.3.2 Rational conventions....87
4.3.3 Conclusion....90
4.4 Evolutionary game theory and constrained maximisation....91
4.4.1 Evolutionary game theory....92
4.4.2 Rationality in evolutionary context....94
4.4.3 Conclusion....98
4.5 Constrained maximisation as conventional rationality....99
4.6 Conclusion....102
5. Evolution and the Social Contract....104
5.1 Dynamic Contractarianism....107
5.1.1 The social contract dynamics....107
5.1.2 Conventional change....109
5.1.3 Rationality in the social contract....111
5.1.4 Bargaining....112

5.1.5 Dynamic contractarianism: Conclusion....114
5.2 Evolutionary Contractarianism....115
5.3 Equilibria Selection and Justice....121
5.4 Conclusion....125
6. Conventional rationality and collective benefit....127
The problem of free-riding....128
6.1 Free-riding and collective action failure....129
6.1.1 Free-riding....130
6.1.2 Collective action failure....131
6.1.3 The rationale of free-riding....133
6.1.4 Conclusion....135
6.2 The prisoner's dilemma and the tragedy of the commons....136
6.2.1 Collective action games....137
6.2.2 Solutions for collective action failure....139
6.2.3 Conclusion....141
6.3 Free-riding within rational conventions....141
6.4 A response to the Foole....146
6.5 Conclusion....148
7. Social conventions and Justice....151
7.1 Rationality and Justice....151
7.1.1 Justice....152
7.1.2 Justice as mutual advantage....154
7.1.3 Conclusion....156
7.2 Equilibrium selection....156
7.2.1 Justice in conventions....157
7.2.2 The equilibrium selection problem....159
7.2.3 Conclusion....162
7.3 Justice and the vulnerable....163
7.3.1 The vulnerable....163
7.3.2 Justice....167
7.3.3 Inter-conventional Justice....172
7.3.4 Conclusion....174
7.4 Conclusion....174
8. Rational morality and social conventions in the real world....176
8.1 The realism of conventional rationality....178
8.1.1 Bounded rationality....178
8.1.2 Rationality in conventions....181
8.2 Social conventions in the real world....183
8.2.1 Rational conventions....183
8.2.2 Real social conventions....184
8.3 Information availability....188
8.3.1 Information availability and rationality....188
8.3.2 Information spreading in contemporary societies....190
8.4 The evolutionary time-frame....191
8.5 The individual and the social contract....193
8.6 Conclusion....197
9. Conclusion....199
Future developments....201
Bibliography....205
Alphabetical Index....212

Acknowledgements

First and foremost, Matt Matravers has been the person who has helped me the most to complete this thesis, particularly during the writing process when he has been extremely helpful and insightful. His comments and our discussions proved invaluable both for my writing and more importantly for my understanding of moral philosophy. Also, he suggested the title of this thesis and his supervision style, a practical application of libertarian ideals – or so it seemed – has allowed me to pursue my own interests and enabled me to decide the direction of the thesis. Finally, his manners and his patience facilitated my thinking and encouraged me to express my ideas and take my time with writing, which was essential for me in order to establish a coherent argument.

Moreover, Matt Matravers's support during the preparation and presentation of my paper at the "Contractarian Moral Theory: the 25th Anniversary of Morals by Agreement" conference, at York University in Toronto, was overwhelming and way beyond his responsibilities. My attendance to the conference was only made possible thanks to support of the school of Politics, Economics and Philosophy and the Department of Politics of the University of York, for which I am grateful. The conference was a truly invaluable experience that helped me develop my thinking about moral contractarianism and offered me significant insights into the field. My conference presentation has been used in this thesis, although, I would like to think in a significantly improved form.

John Bone has been very critical of my argument for all the right reasons, asking all the right questions and highlighting the issues with some aspects of the analysis. Thus, his feedback has been extremely valuable as it helped me understand some of the deeper issues and problems of the thesis. His feedback has been instrumental in forcing me to think harder about the finer methodological points.

I am deeply grateful to Mozaffar Qizilbash, for offering me a place to the programme to begin with, and making me feel very welcome from the very beginning at the School of P.E.P. and the University of York. Our discussions at the first stages of my writing helped me develop my ideas and structure my thesis.

Finally, the school of PEP, the department of Politics and the Morrell Centre for Toleration Workshop have provided an ideal environment for intellectual stimulation. This made it possible for me to explore new ideas in political theory and thus develop my own understanding further. VOX magazine of the School of P.E.P. has given me the opportunity to publish two articles, which served as the basis for Chapters Four and Seven.

Author's Declaration

This thesis is the result of my own work, except where explicit reference is made to the contribution of others, and has not been submitted for any other degree at the University of York or any other institution.

1. Introduction

Morals by Agreement (1986) proposes a theory of morals grounded on rational premises. For Gauthier, morality is the outcome of agreement between rational agents given certain conditions of rational agency. The argument in this thesis reinforces the theory of *Morals by Agreement* and suggests a closely linked alternative account of the rationality of moral behaviour grounded on the repeated interactions of rational actors.

Morality is viewed as the result of agreement between rational agents, just as in Gauthier's theory, but in addition, social conventions serve as a supportive and enforcing mechanism of the agreement. A social contract, consisting of social conventions, provides a social environment in which reaching an agreement, and complying with it, is rational. The conventional account of morality deals with some of the problems of the theory of *Morals by Agreement* and especially the rationality of "constrained maximisation" (Gauthier, 1986: 167). It will be shown that within social conventions it is rational for one to adopt a joint strategy, without the need to call upon – or to 'smuggle in' – considerations that are not rational. Unlike Gauthier's theory, the conventional account does not need quasi-rational concepts to support an argument for moral behaviour and eventually a theory of justice.

One of the main problems with theories of justice in which morality is based on rationality is that they are not considered broad enough to include all cases where justice and moral behaviour are required. For critics of a rational morality, the requirements of rational and moral behaviour are in conflict and, as a result, principles of justice cannot flow from the premises of rational agency. In contrast, a theory of moral rationality presupposes that rational and moral behaviour are based on common principles and, moreover, it answers the most fundamental question of social interaction: namely why one should care for others.

Viewing social interactions in the context of social conventions that are the result of interactions between rational individuals, can offer an alternative solution to the problem of morality. Rational agents behave morally provided there are appropriate social structures that support and bound their actions; at the same time, repeated interactions between rational agents lead to the formation of social conventions. In addition, established social conventions facilitate social welfare maximisation by ensuring long term maximisation for agents who behave conventionally. When these conventions are seen as the components of the social contract, they can facilitate and

accommodate moral behaviour by rational actors, not just in local interactions within small groups but also at the level of society. Thus, conventions of rational morality are viewed in the context of moral contractarianism.

Contractarianism makes up the basis, and provides the framework, for the theory of rational morality presented in *Morals by Agreement.* Therefore, in this thesis social conventions will be examined in the context of social contract theory and in relation to evolutionary accounts of social structure. Conceptually, the thesis can be divided into four sections; the first two chapters examine the related literature in contractarianism, rational choice theory, and evolutionary game theory. The second section consists of Chapters Four and Five, which put the literature in the context of conventional rationality in a framework of the evolutionary dynamics of social structures. The third section, made up of Chapters Six and Seven, deals with morality and justice respectively. Specifically, Chapter Seven shows that conventional rationality is the basis for moral behaviour and justice. Finally, in the fourth section, Chapters Eight and Nine discuss the application of the analysis in the real world and conclude the thesis.

In the following paragraphs contractarianism, rationality and justice, and a naturalised version of the social contract, will be examined in that order so as to provide a general overview of the main arguments of the thesis.

1.1 The social contract tradition

The contemporary social contract paradigm is primarily a method of understanding society as a hypothetical contract between the people and the government when we talk about political contractarianism or a contract dictating moral obligations in the context of moral contractarianism. Although the discussion of why one should behave morally was started by Plato (2006), there have been significant contributions to it much more recently. Hobbes (1976) in the 17th century and Hume (2008) and Rousseau (2008) in the 18th, reformulated the problem and attempted to offer plausible accounts for moral behaviour. The Hobbesian tradition is especially relevant since it is based on the assertion that humans are self-regarding rational agents, whose main aim is to maximise their benefit. The work of these thinkers has set the framework in which we think of rationality and morality today within the social contract tradition; a tradition which can be usefully thought of under two headings: political and moral contractarianism.

Political contractarianism, in its contemporary form, was first introduced by Hobbes in *Leviathan* (Hobbes, 1976). In Hobbes's argument, the rules of the social

contract include the responsibilities and rights of both the people and the government. Moral contractarianism, which is closely linked to, and to some degree embedded in, political contractarian arguments, is a method of deriving moral obligations towards other people without the need to refer to established political institutions. “The contractarian enterprise...seeks answers to questions about the moral obligations we owe to one another, about the legitimate functions of government and the nature of our obligations to it, and about justice in the distribution of income and wealth” (Gauthier & Sugden, 1993: 1). Thus, moral contractarianism addresses, and to a great extent solves, the issue of what one ought to do by referring to individual reason and rationality; one ought to behave in a certain way because it is in one's best interest to do so.

Gauthier's theory builds on the Hobbesian contractarian tradition to construct a primarily moral contractarian theory. He argues, more convincingly and plausibly than those before him, for a morality that is exclusively based on practical rationality. For Gauthier, there is no need to import moral constraints into the contract or to assume that humans are pre-disposed to act morally. The purpose of *Morals by Agreement* is to “develop a theory of morals as part of the theory of rational choice” (Gauthier, 1986: 2). His aim of the project is to show why one ought to be moral. In terms of social interactions, theories of moral contractarianism need to justify why individuals who are “mutually unconcerned (take no interest in each other's interests)” (Vallentyne, 1989: 187), ought not exploit those weaker than themselves. The social contract theory makes possible the reconciliation of moral and rational behaviour within a common framework. Even if the type of rational morality that Gauthier proposes is not convincing, and his critics correct, moral contractarianism remains the only viable theoretical paradigm that allows morality to be viewed as a consequence of rational action.

In this thesis, the social contract will be shown to be a dynamic process consisting of social conventions. Conventions, viewed as the outcome of repeated interactions between rational agents, evolve as individuals' strategies and behaviour change. The dynamics of social structures will be examined through a game theoretical analysis of repeated interactions. In this context, it is possible also to claim that the status-quo and the agreement point are part of the dynamic process characterising the social contract. Social conventions are affected by pre-existing social contracts, which define the established status-quo. In conclusion, individual strategies, social conventions and the social contract are interdependent, with each influencing the other.

The social contract will be examined in the context of repeated interactions between rational actors and therefore game theory provides the necessary tools to describe and to understand these interactions, just as in *Morals by Agreement*. The next section will discuss the importance of traditional and evolutionary game theory for analytical and moral philosophy and the advantages a theoretical paradigm can gain from using games to analyse social interactions.

1.2 Game theory as a tool for the moral philosopher

Hobbes (1976) was the first to use an analysis that later would be described as informally game theoretical. Hobbes's approach, especially in terms of rational agency and the conflicting interests of individuals, comes very close to the assumptions of modern economics and game theory. Moreover, Hume (2008) implicitly uses an informal game theoretical analysis, although in a repeated games framework. In contemporary contractarian theory, Gauthier was the first political philosopher explicitly to use game theory and to draw from a formal analysis in order to construct a theory of justice. To a large extent Gauthier's use of game theory was made possible by a 1954 inaugural lecture given by Braithwaite. In his lecture entitled *Theory of Games as a Tool for the Moral Philosopher* (2009), Braithwaite suggested a new way of thinking about issues of moral philosophy.

The great advantage of using game theory to describe human behaviour and social interactions is that game theory analyses the possibility of cooperation between people with different aims and conflicting interests. Game theory proposes a method of examining interactions between self-interested individuals who can benefit from collective action. In that respect, game theory introduces an aspect of realism into moral philosophy; individuals in the real world have conflicting interests more often than not, and asserting that they will agree on common ends and actions is idealistic. Given that moral philosophy is usually seen as disassociated from the real world, game theory can introduce some much needed plausibility and realism in the field, without altering philosophers' main function, "to think about thinking about ethics" (Braithwaite, 2009: 3).

In addition, game theory has the advantage of being ethically neutral. Being a quasi-mathematical theory, it ensures that the outcome of interactions is not biased by pre-existing views about what is good or morally desirable. It is only concerned with what is feasible in a given social context and with how individuals can maximise their

utility through their interactions. In economic analysis, where game theory has been primarily successful, considerations about the morality of an outcome are irrelevant. The same does not usually apply in moral philosophy, but the introduction of game theory ensures that moral philosophy becomes amoral and reaches moral principles that can be objectively justified. Despite the fact that game theory does not include normative suggestions, the subsequent "recommendations themselves will constitute what may be called second-order moral principles" (Braithwaite, 2009: 6). Therefore, the use of game theory in moral philosophy offers two advantages of great significance: realism and moral neutrality.

For Braithwaite – as well as for Gauthier – game theory can be used to solve some of the analytical problems of moral philosophy; namely reconciling the moral priority of the individual with justice. More generally, this translates into dealing with the conflict of individual maximisation and collective benefit, which can be done most effectively with a game theory analysis. In that respect, game theory is vital for any theory of justice that assumes rational agents. Despite the risk of transforming moral philosophy into a formal model of human behaviour, the framework and basis of all the thinking about these problems is philosophical. The premises of game theory must be grounded in moral philosophy in order for game theoretical analysis to be useful when examining morality and justice, otherwise we could end up with a mathematical model of morality which would be limited and misleading since human behaviour cannot be completely described by mathematics.

Grounding game theory in philosophical thinking bypasses the main problem with formal modelling in the social sciences and the humanities. Game theoretical assumptions are based on oversimplified assumptions about the structure of interactions and the capacities of the actors. Therefore, the philosophical background is essential if we are to avoid falling into the trap of taking game theoretical conclusions at face value (as is the case in neoclassical economics). A critical view of game theoretical models, gained using the lenses of moral philosophy, can help us understand the limitations and the true value of game theory in examining human behaviour.

Gauthier's analysis does just that; it is based on a game theoretical analysis and accepts individual rationality. However, in *Morals by Agreement* the formal economic model of behaviour is limited by the analysis of the arguments that surround individual behaviour and the concept of rationality. For Gauthier, "the rational principles for making choices...include some that constrain" (Gauthier, 1986: 3); this is unacceptable

in traditional game theory where actors have been dehumanised in order to be modelled more effectively. At the same time, it would be accurate to say that Gauthier's theory utilises game theory through rational choice theory; put differently his approach is more cautious than the classic economics approach to formal models. Just as Rawls (2005) before him, Gauthier aims at developing a theory of justice within the limits of rational choice theory (Gauthier, 1986).

In this respect, a further methodological step is needed that relates to the social structures that bound interactions between rational agents. In the final paragraphs of section 1.1 we talked about the dynamic nature of the social contract. This can also be examined through game theory. However, traditional game theoretical paradigms are not sufficient to represent complex dynamic interactions within societies. For that reason, it is essential that we use a different version of the theory of game: evolutionary game theory.

Evolutionary game theory will be shown to be compatible with assumptions of individual rationality, provided we assume boundedly rational agents. Relaxing the strict assumptions of economic rationality gives us a more realistic account of human behaviour, which takes place in a dynamic world of repeated interactions. Therefore, rationality is bounded by social conventions, that facilitate and at the same time limit individual strategies. Bounded rationality is a more realistic account of individual behaviour while being compatible with concepts of cultural evolution that describe a dynamic social contract. It is more realistic because it does not rely on agents having complete information, unlimited memory and extraordinary logical capacities, in order to be rational. Bounded rationality describes more adequately humans' behaviour, as opposed to the simplifying models of traditional rational choice theory. In this sense, bounded rationality is a more appropriate approach to human behaviour, without being an exact description of it. To sum up, there are two central concepts in the argument: the dynamic nature of the social contract and the bounded character of individual rationality.

The first five chapters of the thesis will focus on showing how the combination of the dynamic nature of the social contract and the bounded character of individual rationality can create a plausible description of social behaviour. This discussion provides the basis for a more explicit examination of morality and justice, which will take place especially in Chapter Seven. The account of rational morality that is based on local social conventions can only be topical; morality depends on conventional rules,

which in turn stem from the history of interactions within a given convention. The following section will deal with the evolutionary account of the social contract and the understanding of morality that flows from it.

1.3 Naturalising the social contract

Morals by Agreement is "an attempt to write a moral theory for adults, for persons who live consciously in a post-anthropomorphic, post-theocentric, post-technocratic world" (Gauthier, 1988: 385). When we remove all these possible grounds of a moral theory, what is left is a naturalised moral theory; naturalised not in the sense of natural law (Hobbes, 1976), but in that it does not need a reference outside the natural world to be successful. Humans are defined by their natural capabilities, one of which is the capacity for rational deliberation, and subsequently thus, a social contract that is not metaphysical can only be natural.

The Economics of Rights, Cooperation and Welfare (Sugden, 2004), proposes an evolutionary account of contractarianism, which will be examined in §3.2. Sugden bases his analysis on the fact that social conventions make up the social contract and that these conventions do not need a hypothetical – or even an actual – agreement in order to be binding. The core of his argument is that self-interested individuals who interact without outside influences will reach a state of "spontaneous order" (Sugden, 2004: 1). Sugden assumes that individuals are reasonable, but not hyper-rational, utility maximisers. Interactions among agents of this type, in a state of anarchy, will lead to the creation of social norms and as a result "people will come to believe that their behaviour ought to be regulated by convention" (Sugden, 2004: vii).

Unlike Gauthier, Sugden is more influenced by the Humean, rather than the Hobbesian, tradition of understanding human interactions and rationality. In that respect, his analysis is closer to holistic explanations of social behaviour than to methodological individualism. This is important because in the following chapters there will be an attempt to reconcile the Hobbesian and the Humean traditions and by extension, holistic and individualistic methods of social explanation. In addition, Sugden uses evolutionary accounts of social conventions to explain their formation and sustainability without abolishing the assumption of individual rationality. Therefore, or so I will argue, an evolutionary account of social conventions similar to the one proposed by Sugden can be used to explain the rationality of constrained maximisation as described by Gauthier.

Sugden uses evolutionary game theory and repeated games to give an account of the establishment of social conventions. However, although he analyses a variety of repeated games, he does not go to great lengths to discuss the relative dynamics of social conventions. *The Stag Hunt and the Evolution of the Social Contract* (Skyrms, 2004, discussed below in §3.3), provides a more detailed and formal explanation of the evolution of the social contract and the coevolution of social structures and individual strategies. Skyrms's analysis, although clearly explanatory, is more analytical in terms of the dynamics of social structures as it examines how they are affected by individual strategies and local equilibria. The coevolution of social structures and individual behaviour is central to the argument presented in Chapters Four and Five; it shows how individuals influence social equilibria and are influenced by established conventions, thus making it possible for us to argue that individual rationality can fit in an account of dynamic social structures.

Binmore (1998; discussed below in §3.4), continues the work of Sugden and combines it with aspects of the Skyrms's evolutionary account of social interactions. His main argument is that our moral norms are linked to our biological evolutionary history and consequently that our cultural evolution and social structures depend on our biological history. Thus for Binmore, morality is not exclusively an artificial virtue, but our ideas of what we ought to do are heavily influenced by our biological past. In that respect, *Game Theory and the Social Contract* (Binmore 1998), explicitly suggests the naturalisation of the social contract.

The naturalisation of the social contract tradition, as understood in this thesis, does not mean that human and animal social contracts are bound by the same limitations; nor that the rational and moral character of humans are determined by their biology. Rather, we will follow Sugden and Skyrms in arguing that the social contract paradigm can benefit from the inclusion of evolutionary theory, in conjunction with game theory, in order to promote a plausible alternative method of understanding social structures and human behaviour within society. Normative recommendations of behaviour follow from that understanding of the dynamics of social interactions, but are not determined or limited by our biological nature.

Established social conventions and the social contract follow evolutionary processes; they are dynamic and their equilibria depend on population dynamics. Therefore, established moral norms are influenced by a given society's cultural evolution. Rational morality, within the context of social conventions, is the outcome of

repeated interactions between boundedly rational agents; these interactions are best described by an evolutionary account and lead to local equilibria that define the local moral norms. Put differently, a social contract, consisting of social conventions, defines equilibria of justice; since different societies have different cultural evolutionary pasts, their equilibria of justice will differ. The conventional account allows for as many understandings of morality and justice as there are possible evolutionary paths to a social equilibrium.

The discussion in this section attempted to give a general description of the main works in contractarianism and evolutionary game theory and their relation to the argument in this thesis. The evolutionary account of the social contract will be further analysed in Chapters Three and Four. Moreover, in Chapters Four and Five evolutionary game theory will be shown to be a valuable tool for moral contractarianism, which allows us to claim that moral norms are the outcome of repeated interactions between rational actors.

In sum, this thesis lies in the field of moral contractarianism and the argument presented for a rational morality is based on the evolutionary process of social conventions. Established moral norms vary according to specific evolutionary processes that set the bounds for rational strategies. Thus, the interdependence of social structures and conventional rationality may give rise to very different moral norms that are equally justifiable. The following, final, section of this introduction provides a brief description of the chapters that follow, thus giving a fuller indication of the main arguments that will be used.

1.4 Thesis overview

Chapters two and three consist of a critical review of the literature mentioned above. In chapter one, the main points of what I call “the Gauthier project” are discussed followed by a review of the main criticisms of that project. *Morals by Agreement* aims at introducing a theory of moral behaviour and justice that is based on rational premises. However, in order for it to be successful, the definition of rational agency has to be relaxed. The principal of minimax relative concession, the Lockean Proviso, and constrained maximisation are the main loci of criticism of Gauthier's theory. Critics argue that the Gauthier project has failed since moral principles are smuggled into what Gauthier claims to be a merely rational baseline. In addition, another line of criticism deals with the fact that the theory of morals produced is not general and broad enough to

qualify as an adequate theory of justice.

Gauthier's theory may not have a plausible answer to every point of criticism. However, this is hardly surprising and the Gauthier project should be seen as part of a greater process within the contractarian tradition; a process that aspires to create a moral theory that is relevant to modern individuals and contemporary social life without the need for metaphysics. Moreover, the theory of *Morals by Agreement* can be strengthened by looking at rational behaviour in the context of social conventions and evolutionary dynamics. The literature in these fields is discussed analytically in Chapter Two. The works of Sugden, Skyrms and Binmore will be examined critically as a prelude to Chapter Four, which will look at how it is possible to combine the Gauthier project with an account of social behaviour that is based on conventions and evolutionary theory.

Chapter Four, “Rationality and Evolutionary Theory”, as its title suggests, attempts to show that it is reasonable and realistic to use evolutionary principles in conjunction with rational choice theory premises. This will make it possible to argue that the dynamics of social structures and individual rationality are interdependent and that, despite the determinism of traditional evolutionary theory, the evolutionary account of social conventions is based on individual action and rationality.

Chapter Five continues the discussion of dynamic contractarianism and focuses on the evolutionary nature of social conventions in general and on the social contract in particular. Chapters Four and Five are linked as they deal with closely related arguments over individual rationality in an evolutionary framework and the role of evolutionary dynamics in the formation of stable social conventions and the social contract. Thus, they present the main argument of how individual rationality can be reconciled with collective action and give rise to evolutionarily stable social conventions.

Chapter Six looks at the possibilities of free-riding within the context of evolutionary conventions and repeated interactions. Free-riding will be shown to be irrational if we assume non-random repeated interactions within a social structure of conventions that facilitate and accelerate information sharing.

Social conventions are the main topic in Chapter Seven. They will be shown to be essential for supporting and enforcing the moral behaviour of rational individuals, thus leading to a social contract that is just. Justice as mutual advantage will be shown to be in accordance with the previously described version of evolutionary conventions.

Finally, Chapter Eight examines under what conditions the arguments presented

in the previous chapters can make sense in the real world. Relatively recent technological and social developments make it plausible to claim that information is readily available and that rational agents within social conventions are equally rational.

The overall argument admits that there is no one single version of moral behaviour. Rational actions depend on one's environment and on available information. Depending on the established convention and the local history of interactions, a conventional rule of turning the other cheek is as rational and as moral as following a rule requiring an eye for an eye; in contemporary terms, a society that allows the death penalty is as just as a society that punishes only with short prison terms, irrespective of the crime. A stable conventional equilibrium is the topical moral norm irrespective of possible non-conventional moral imperatives. Each society reaches its own social equilibrium that dictates moral behaviour and as such, morality is relative to the local history of interactions. Hence, the understanding of rationality and morality presented here depends on social structures and cultural history and as such has no room for universal moral rules. Nevertheless, there is room for optimism that we may reach, at some point, a common account of morality.

2. The Gauthier Project and its critics

Grounding moral behaviour on reason has been a philosophical problem at least since Plato's Republic, but it was reinvented more recently by Thomas Hobbes (1976) and David Hume (2008). David Gauthier continued on the same path, providing a more plausible normative account of rational morality. *Morals by Agreement* (1986) introduces a theory of morals based on rational choice, whose principal project is to reconcile rational choice theory with morality using a contractarian methodology. The aim of this chapter is to analyse the components of rational morality as presented by David Gauthier and to discuss its strengths and weaknesses. These are seen through three main concepts of the "Gauthier project": the concept of rationality, its relation to morality and the definition of moral behaviour. A critical analysis of Gauthier's work will allow an assessment of its effectiveness and provide a framework for the following chapters that will attempt to link *Morals by Agreement* (*MbA*) to a theory of social conventions. Thus, conceptions of morality and rationality are crucial for the "Gauthier project" and have to be examined closely in order to obtain a deeper understanding of *MbA*.

Rationality is central to the "Gauthier project". According to the "maximising conception of rationality" (Gauthier, 1986: 7), it is rational for different rational agents to want to maximise different interests. Thus, *MbA* follows the concept of rationality used in neoclassical economics and game theory. In the economics account of rationality, rational behaviour is related to individual preferences over a set of choices and an individual is characterised as rational when she acts in a way that she thinks will maximise her utility, within the given circumstances. Therefore, rational behaviour is subjective since it depends on subjective individual preferences. Although rational behaviour is utility maximising behaviour, the subjectivity of preferences means individual behaviour varies, as each person maximises in a different way. In addition, the concept of rationality in *MbA* lies on individuals realising that it is rational to constrain their maximisation. Interactions between constrained maximisers are mutually beneficial and yield optimality as opposed to straightforward maximisation that is a strictly rational behaviour; constrained maximisation is a joint strategy, which maximises the constrained maximisers' utility in the long term whereas straightforward maximisation yields immediate maximisation for a single agent. Morality is then seen as the rational constraints on individual maximisation. For Gauthier, it is rational for one to

constrain one's maximisation in order to benefit from interactions with other similarly disposed agents.

MbA is based on five main components. The first one is a morally free zone, which refers to an ideal economic market. In this perfect market, interactions among rational self-regarding individuals are mutually advantageous and hence there is no need for non-rational constraints. In the morally free zone, constraints on individual actions can only be imposed as the result of a mutually beneficial bargaining procedure. The second component of the theory is minimax relative concession which describes the bargaining procedure, which is "a principle that governs both the process and the content of rational agreement" (Gauthier, 1986: 14). Constrained maximisation, the third component in Gauthier's theory, calls for rational agents constraining their maximisation when interacting with others similarly disposed. The Lockean Proviso is the fourth component of the theory presented in *MbA*, which is used as a mechanism that regulates the original position of the bargaining. Finally, the Archimedean Point is Gauthier's version of the Rawlsian veil of ignorance. A rational agent at the Archimedean Point is one who is able to make impartial decisions. These five elements constitute the basis of the theory of morality that is introduced in *MbA*. The theory is based on the idea that there can be moral constraints deriving from rational premises.

A contractarian framework is essential in order to derive morality from rational choice theory premises. Rational, self-regarding individuals will constrain their maximisation if they are bound by the terms of a rational agreement. If the terms of the contract are mutually beneficial, then compliance with the contract has to be rational. This is an improvement on Hobbes's contractarianism. Despite *Leviathan's* influence being apparent throughout the book, coercion is not part of *Morals by Agreement*. In *Leviathan*, the state is introduced as a method of forcing rational maximisers to constrain their behaviour; in the absence of a strong, coercive government individual rationality would lead to a "war of all against all" (Hobbes 1976, vii). In addition, the theory of state in *Leviathan* requires individuals to abandon their rights for the commonwealth to work. Gauthier replaces Hobbes's sovereign with a voluntary rational agreement to comply with the contract. In that respect, Hobbes's crude conception of rationality as self-interest creates grounds for criticism for Gauthier's theory; Hobbes would not accept that constrained maximisation is rational nor that it is possible for rational agents to comply with their agreements unless there were a government to force them to comply. In both *Leviathan* and *MbA*, complying with a rational agreement is

maximising only when the act of compliance is rational in itself. However, Hobbes ensures compliance through government coercion whereas for Gauthier compliance is based solely on individual rationality. The introduction of a third party, a mutually accepted enforcer, makes Hobbes's account more plausible given the assumption of individual rationality.

Moreover, and perhaps more importantly, the account of rationality assumed in *MbA* has been the object of much of the criticism. Gauthier has been criticised for the inclusion of assumptions that can be seen as not purely rational, such as minimax relative concession and constrained maximisation. To conclude, the theory has been criticised on two accounts: firstly, because it is not plausible to assert that rational agreement in the Gauthier framework described above will lead to rational compliance. Secondly, because its basic concepts are not purely rational as Gauthier claims. The criticism of Gauthier's account of rationality will be discussed more analytically at the end of the chapter.

The purpose of this chapter is to provide an overview of the “Gauthier Project”. Before looking at the criticism, the theory's main components will have to be examined and assessed in terms of plausibility and effectiveness in showing that rational behaviour can be moral. Each component will be discussed separately, following the order of the book and starting from the ideal market as a morally free zone.

2.1 A morally free zone

The concept of a morally free zone is used in *MbA* to describe a situation where there is no need for principles of moral behaviour. A perfectly competitive market, as described in *MbA*, is a morally free zone where moral principles are redundant. Because an ideal market combines optimality with equilibrium as is shown by economic theory, a rational agent maximises her utility through interactions without making anyone worse off. The market as a morally free zone is a state where a Nash equilibrium coincides with Pareto optimality. Thus, the perfectly competitive market is one in which no moral constraints are required in order for the market to reach optimal equilibrium. In that respect, the concept of a perfectly competitive market is a loan from theoretical economics. The free market is an ideal environment in which to study rational interaction. However, Gauthier's use of the ideal market is in the context of moral behaviour which sheds a different light on the paradigm. In order for the ideal market to work as part of a moral theory its main characteristics have to be qualified.

By definition a free market must encompass the rights of private ownership and private consumption. In addition, an ideal market requires the absence of positive or negative externalities and the existence of market certainty. These four characteristics, private ownership and consumption, absence of externalities and certainty, are essential and have further implications for the workings of a completely competitive market. When all of the four conditions are met, there is an ideal market which can be used as a model for moral anarchy. The above mentioned conditions of the perfect market and their implications will have to be discussed more analytically in order to show how they relate to the absence of a need for moral principles.

Private ownership is linked to and depends on individual factor endowments and free market activity. Individual factor endowments define each agent in the market by describing her capabilities and preferences. She can use these to profit in the market and improve her situation. The owner of a good, a product or a factor of production, is free to use it as she pleases in order to make the most of it. In other words, the assumption of individual factor endowments refers to the fact that each person in the ideal market has a set of capacities and owns a number of goods. The assumption of free market activity refers to the freedom of each individual in the market to act in any way she thinks will maximise her utility.

Private consumption comes as a result of private ownership. If there are goods that are privately owned and are exchanged freely, then these goods must be consumed. The right of private ownership of a product or a factor of production includes the right to their consumption. In essence this condition means that all products have to be private, that is, each product can only belong to only one person. This idealised account of the market presupposes that there are not any goods that are public, shared or free. A product or factor of production can only belong to a single individual. Therefore, in the ideal market all goods are private. In addition, since all consumption must be individual, the benefit from its consumption is also individual. A rational agent consuming a good can only be concerned with her own maximisation. Hence, similarly to private ownership, private consumption can be seen as consisting of two components; private goods and mutual unconcern.

In summary, the first two, conceptually linked, conditions for a perfect market are private ownership and private consumption. Private ownership requires individual factor endowments and free market activity and private consumption needs private goods and mutual unconcern. The two following conditions are the absence of

externalities and uncertainty.

The absence of externalities means that the actions of an individual only have an impact on those with whom he interacts. An exchange between two individuals changes their utilities without affecting anybody else's. For instance, my building a road to my house can create a positive externality for my neighbour who will also be able to use it. This example shows how the presence of externalities can also contradict the previous condition of the existence of exclusively private goods. The moment that my neighbour can use the road without having participated in its construction, the road becomes a shared good. Thus, all the above conditions for an ideal market, private ownership and consumption and absence of externalities have to apply simultaneously and are interdependent; the failure of one inhibits the possibility of a perfectly competitive market.

Finally, certainty about production and prices is assumed as well. Agents have full knowledge of future changes in their utility functions and the production functions of society. In an ideal market rational individuals know the prices for their products and the supply and demand functions. Rationality demands that each agent will share all available information and therefore everybody will be able to maximise their utility, given their original factor endowment, market activity, consumption and absence of externalities. Perfect competition requires equal information which is assured by honest communication between rational agents.

In a society where the above conditions for an ideal market apply, rational interaction leads to an optimal state. They are typical idealised assumptions used in economic theory models. Economic theory shows that in a perfect market in equilibrium no one can improve her situation without worsening somebody else’s; in equilibrium each agent maximises. By maximising individual utility, social welfare is simultaneously maximised. This social welfare is the optimal outcome of market interaction – the social optimum. In a perfectly competitive market the equilibrium is a point of optimality. Thus, perfect competition, as defined by the conditions mentioned above, leads to a socially optimal equilibrium. Given perfect competition, “each individual, intending only her own gain, promotes the interest of society” (Gauthier, 1986: 89). Under these circumstances, a morally free zone where both optimality and equilibrium are present, is plausible. However, Gauthier has to show that ideal market conditions ensure that moral constraints on rational behaviour are redundant. In order to achieve this, he explains how the ideal market is also impartial and there is no conflict

between moral and rational requirements, since “market interaction is rational and morality is not opposed to rationality” (Gauthier, 1986: 95).

Equilibrium and optimality coincide in the perfect market and therefore rational interactions lead to an impartial society. Just as Robinson Crusoe does not need morality to justify or motivate his actions, the rational agent in the ideal market has no use for a moral code. Free market activity is similar to the activity of Crusoe. Just as Crusoe, each agent in the ideal market makes the most out of the situation without any compulsion or partiality. Crusoe can only blame himself for sub-optimal outcomes; in the market where there is no coercion, each rational agent cannot blame others for an interaction outcome he sees as impartial. Moreover, the absence of externalities ensures the absence of free-riders and hence establishes an impartial interaction. Finally, rational interactions in an ideal market lead to an optimal equilibrium, which means that no one can become better off without worsening another's position. Ideal market conditions ensure impartiality and hence ensure that the market is a morally free zone. Therefore, both the agents and the market are morally free. Given free market activity, absence of externalities and rational interactions that lead to optimality, the market is shown to be a morally free zone. Should these conditions hold, there is no need for moral constraints on rational behaviour.

However, the ideal market and the subsequent morally free zone are hypothetical constructions and Gauthier states this very clearly when he says that his intention is to: “show that there would be a morally free zone in ideal interaction, not to argue for its presence in most of our daily activities” (Gauthier, 1986: 93). Despite his presuppositions being unrealistic, the thought experiment he uses shows it is possible to have an impartial society, without enforced constraints and coercion. This moral anarchy presupposes and requires that there are no market failures. However this is not usually the case as the idealised conditions for a perfect market are not met. Market failures require the discussion of interactions outside the realm of the ideal market and the introduction of constraints on rational behaviour. In *Morals by Agreement* cooperation and bargaining examine the possibility of inequalities in factor endowments and discuss the extent to which these lead to impartial outcomes. These ideas are discussed in the following section.

2.2 Minimax Relative Concession

Outside the ideal market, rational interactions take place in the context of cooperation

and bargaining. Cooperation is necessary when resources are scarce and there is competition over their use. Moreover, cooperation creates additional resources by increasing the total output; the collective output is greater than the sum of the individual inputs. Cooperative behaviour creates a surplus that would not be available otherwise as individual factor endowments are multiplied through cooperation. The creation of this surplus creates a need for its distribution. In order for this surplus to be divided between the contributors, a bargaining process has occurred. Bargaining between rational agents leads to a division of the cooperative surplus that is mutually accepted as impartial. Therefore, cooperation and bargaining arise from market failures and the need to maximise given these failures. The following will be a discussion of the details of the bargaining procedure presented in *MbA* and the way it may lead to fair, mutually beneficial outcomes.

The central components of the bargaining procedure are the initial bargaining position, the cooperative surplus and the bargainers' claims. The initial bargaining position refers to the factor endowments each actor has at the beginning of the interaction. A rational agent expects to leave the bargaining table with more than she had originally. The cooperative surplus should create a good that after its distribution is greater than the initial individual endowments. Otherwise, it is irrational for one to engage in bargaining in the first place. During bargaining, each actor makes a claim for her share of the surplus. After the original claims, bargainers have to make concessions from their original claims in order for an agreement to be reached. The optimal outcome of the original claims is unattainable since at the strictly optimal point both agents would receive their maximum claim which is more than the additional cooperative product. Bargaining in *Morals by Agreement*, partly based on Zeuthen's bargaining theory, includes two stages: making a claim and making a concession. A claim asks for a concession from the other party or parties in the bargain. In this procedure, cooperation is a given; everyone wants to share the cooperative surplus and reach an agreement. And everyone would be worse off should an agreement not be reached. Therefore, all bargainers want to avoid endangering the agreement by making claims that are too great to be accepted by others. At the same time, bargainers have to make concessions in order for the process to continue; they have to be willing to accept less than their original claim. Rational agents want to maximise their claim and minimise their concession, in order to maximise their share of the cooperative surplus. An agreement is reached once all bargaining parties are satisfied that their share of the cooperative

surplus is the maximum they could have gained, considering their contribution by comparison to the other party's contribution. In conclusion, bargaining involves an initial position, claims and concessions. These elements of the bargaining process are used in *MbA* as a basis for introducing and examining the relative concessions and the subsequent minimax relative concession.

More specifically, relative concession is introduced as a meaningful comparison of the relative concessions of the bargainers. Relative concession is defined as the proportion of absolute concession by comparison to the absolute concession at the starting point of the bargaining. In this way, relative utilities can be compared while interpersonal utility comparison is avoided. Since a rational agent would try to minimise her concession, the optimal outcome would be one where the magnitude of the concession required by each bargainer is a minimum. Each agent compares the concession she makes at the bargaining table with the concessions others make. He only accepts the agreement if he feels the difference between these concessions is not too great. If the relative concession is small, then it is more likely that the eventual agreement will be seen as fair and hence more likely to be accepted. Hence, the idea of relative concession is critical for the bargaining process in *MbA.* Furthermore, minimax relative concession is introduced to describe the relative concessions that rational agents may accept.

According to minimax relative concession, the maximum concession each person makes is the minimum from all possible outcomes. In other words, the maximum concession one is prepared to make is the minimum of all alternative concessions. In a sense, this is a tautological claim given rational agents; rationality calls for utility maximisation and conceding the least possible amount is essential in maximising. In Gauthier's words: "in any cooperative interaction, the rational joint strategy is determined by a bargain among the cooperators in which each advances his maximal claim and then offers a concession no greater in relative magnitude than the minimax concession" (Gauthier, 1986: 145).

More formally, there are four conditions of bargaining theory in *Morals by Agreement*: a rational claim, a concession point, a willingness to concede and the limits of concession. The two latter conditions are derived from the former ones. The willingness to concede refers to the fact that all bargainers are rational and expect others to behave rationally as well; thus, they do not expect others to concede more than they would. This expresses the "equal rationality" (Gauthier, 1986: 143) of the bargainers.

Similarly, rational agents have limits to the amounts they are willing to concede; the limits of concession cannot be so large so as to cause a decline in utility for any party.

These steps of bargaining make up the internal rationalisation of cooperation. They are a rational procedure for bargaining and cooperating and lead to an optimal bargaining outcome from which no rational individual has an incentive to defect. Since this outcome is optimal every other option will be inferior in utility for the parties in the bargain. Moreover, agents have accepted each step of the process as rational and in the absence of coercion; they will accept an outcome that is based on rationality. The optimality and rationality of the outcome ensure the stability of the agreement. Rational agents will have no reasons not to comply with the optimal outcome.

Thus, minimax relative concession is central to the theory of bargaining. First, it requires agents to maximise their utility at the bargaining table. Then, it defines the bargaining procedure which is meant to leave them better off than they were in their original position. Finally, mimimax relative concession moves the centre of bargaining from individual rationality to maximising the bargaining outcome making bargaining the vehicle in which rational utility maximisation coincides with impartiality. In other words, minimax relative concession is a constraint, on rational behaviour. And since it maximises social welfare it is a moral constraint exclusively derived from rational premises; “the principle of minimax relative concession serves … as the ground of an impartial constraint” (Gauthier, 1986: 150). Minimax relative concession links the concepts of cooperation and bargaining with that of justice. Rational individuals will engage in cooperation and bargaining in order to maximise their utility. The constraints on maximisation that are agreed at the bargaining table are constraints on individual maximisation. In other words, the constraints decided during a rational interaction are impartial and thus serve as a principle of justice.

The ideal market model shows how morality may be redundant in a perfectly rational world; then rational bargaining exhibits how even in the case of market failures rationality does not need explicitly moral constraints to reach an optimal equilibrium. Bargaining and its subsequent fair outcome have focused on making and accepting claims and concessions according to the principle of minimax relative concession. They have ignored the original endowments as a possible means of instantiating impartiality. However, given this process, an injustice at the beginning will be transferred to the bargaining outcome. In the limitations of the above discussion of bargaining, the unjust outcome will still have to be characterised as impartial and optimal. Therefore, it

becomes clear that an additional construct is needed to ensure that the original bargaining position is also impartial. For Gauthier, this construct is the Lockean Proviso, which will be discussed next. In combination with rational compliance with the terms of the agreement that will follow it, they make up Gauthier's theory of justice.

2.3 The Lockean Proviso

The discussion of bargaining and justice has ignored the effect the original position can have on the agreement point. A coercive or extremely unequal starting bargaining position will lead to a contract that is similarly coercive or unequal, given the above mentioned bargaining process. In order to address this issue, *MbA* needs a new component that ensures that the original position does not have an unjust impact on the bargaining outcome. The Lockean Proviso ensures that bargaining starts from a non-coercive, impartial position that is rationally acceptable. Then, following the bargaining procedure described earlier, rational interactions reach an optimal equilibrium. The following paragraphs will examine the need for the inclusion of an interpretation of the Lockean Proviso in Gauthier's theory and will consider whether it is a rational and not strictly a moral principle.

The proviso is used at the starting point of bargaining, where there are no moral constraints on individual behaviour. The initial bargaining position is central in the bargaining procedure of *MbA*. Having accepted the priority of individual freedom and rationality over moral principles, Gauthier needs to address the issue of bargaining between rational agents characterised by great factor endowments inequalities. Utility maximisers cannot accept a contract that does not allow them to maximise, or that leaves them worse off. Therefore, the bargaining process must not threaten their endowment for the benefit of the weaker bargainers. Within this context of rational agency, the role of the initial bargaining position becomes even more important as it directly affects the bargaining outcome. That is not to say that the initial factor endowment distribution has to be roughly equal but that it has to be considered impartial. If one agent in the agreement feels that the endowment distribution is not impartial, then the contract cannot be stable as described in Gauthier's slave society example (Gauthier, 1986).

If the contract is to be impartial then the original factor endowment distribution has to be impartial and the initial bargaining position has to be non-coercive. Only then will compliance by rational agents lead to a stable mutually beneficial contract. A

pre-contract state of nature where there is coercion, such as a slave and master society, cannot lead to a stable, mutually accepted agreement. If one is being taken advantage of originally, then one has no reason to comply with any agreement made. In the masters and slaves society story in *Morals by Agreement* (Gauthier, 1986: 190), the masters promote a new contract for the abolition of slavery the terms of which the slaves will not have a rationally compelling incentive to comply with once they are free. During the original social contract, where the bargaining takes place, the slaves were being taken advantage of. Hence, the bargaining process in a coercive contract could not have been fair or mutually beneficial, which nullifies the concept of rational bargaining. Therefore, if they are rational they will not accept the contract. The slaves, being in a position of coercion will have reasons to accept any deal that improves their situation. However, once their situation has been improved, even incrementally, they have no rational incentive to keep the agreement if its terms are still partial in favour of the slave owners. Therefore, it is essential to start from a state where none of the bargaining parties is impartially advantaged; in a different situation the eventual contract would be unstable

A stable contract between rational agents must be based on an impartial initial position. Minimax relative concession by itself does not justify cooperation between the ex-slaves and the ex-masters since a cooperative agreement is impossible without a commonly accepted original position. Cooperation is rational if, and only if, the original bargaining position is non-coercive. Therefore, coercion has to be removed from the initial position, for the contract to be optimal. In addition, removing coercion will not be enough in Gauthier’s theory. The possibility of taking advantage by worsening others' situation has to be removed from the original position. From both a moral and a rational perspective, the original bargaining position has to be impartial to all parties if the agreement is to be kept by rational agents. A form of the Lockean Proviso is inserted in the original position to enforce an impartial starting point of interactions. According to Gauthier's restatement, the Lockean Proviso demands than no one will benefit by worsening another's situation; or put differently the Lockean Proviso, “prohibits worsening the situation of another person, except to avoid worsening one's own through interactions with that person” (Gauthier, 1986: 205). In essence, the use of the Lockean provision in *MbA* ensures that there will be no exploitation when there is no subsequent benefit.

The Lockean Proviso is imported as a moral constraint and as a mechanism to prevent coercion that serves as an impartial constraint on interaction. However, rational

agents will accept this moral constraint on their behaviour in order to maximise. Without an impartial starting point, an impartial and stable contract becomes impossible. Since, all bargaining parties are benefited by an agreement, they would all accept an original position characterised by the Lockean Proviso. The Proviso is essential for a fair original bargaining position and a fair eventual agreement. It is made clear that the proviso "merely constrains natural interaction…" (Gauthier, 1986: 208) and thus leads to a fair initial bargaining position and in turn a fair agreement. It is also made clear that it is rational to accept the Proviso. It enables the stability of the agreement and makes the fair division of the cooperative surplus possible. Masters will not have to bear the cost of having slaves should they accept a constraint in their maximisation. In return they will share the cooperative surplus that derives from a stable agreement. Therefore, it is both utility maximising and moral to accept the Lockean Proviso as a constraint; the moral constraint is also rational.

The proviso is used to ensure that the agreement point is impartial. In the context of rational, mutually indifferent agents who are free to act as they will, the original bargaining position is crucial in determining the terms of the contract. In order for the contract to be accepted by rational agents who find it in their best interest to keep their agreement, its terms have to be impartial and mutually accepted. In turn, an impartial contract needs an impartial factor endowment distribution which is assured by the Lockean Proviso. The discussion so far has centred around rational bargaining and impartiality. The ideal market model shows that is plausible to assert that rational interactions without constraints lead to an optimum equilibrium. The principles of minimax relative concession and the Lockean Proviso ensure that the final bargaining outcome will be fair and accepted by all. Compliance with the terms of the rational agreement is thus assured by the rationality of its terms. However, this is only true to an extent. Rational agents act in a strategic environment where their maximisation depends on others' strategies and behaviour. Therefore, compliance with the terms of the rational agreement is not always realised without the introduction of a further theoretical component. Constrained maximisation shows how it is rational for an individual to only cooperate with those similarly disposed and will be discussed next.

2.4 Constrained Maximisation

Minimax relative concession and the Lockean Proviso are necessary but not sufficient conditions for rational cooperation. A rational agent still has reasons to break an

agreement that is based on rational grounds. A further component needs to be added to *MbA* in order to ensure that it is rational to comply with agreements that meet the above mentioned conditions. Constrained maximisation yields optimal outcomes while maximising individual utility within a given environment. Constraining one's maximisation is rational given that others also constrain their maximisation. An interaction between constrained maximisers yields higher utility for the individuals by comparison to straightforward maximisation, while at the same time leading to a socially optimal outcome. In conjunction with minimax relative concession and the Lockean Proviso they constitute a set of rational conditions that lead to moral outcomes. Hobbes's Leviathan forces individuals to keep their agreements. On the contrary, Gauthier speaks of voluntary rational compliance; the contractors should keep their agreement because rationality dictates it without the need for external enforcement. Gauthier's argument on compliance is based on the idea of constrained maximisation.

A constrained maximiser will choose a strategy that does not strictly maximise her utility if she knows others will behave similarly. This enables her to participate in future interactions with constrained maximisers, which increases the overall benefit (Gauthier, 1986). A constrained maximiser cooperates as long as she expects others to constrain their maximisation as well and as long she expects that her behaviour will create a higher cooperative surplus. In other words, constrained maximisation is a strategy of conditional cooperation. In addition, constrained maximisation is rational behaviour. According to Gauthier, someone who constrains her maximisation is still rational but "reasons in a different way" (Gauthier, 1986: 170). Put differently, a constrained maximiser is willing to compromise with less than full utility maximisation in a given instance, provided that the rational agents she interacts with are willing to do the same. Moreover, a constrained maximiser will only accept a small decrease in her immediate utility, aiming at long term gains from interactions. Constrained maximisers as opposed to straightforward ones will comply with a contract that requires them not to maximise their utility, provided they think that the other members in the contract will adopt the same strategy. Therefore, constrained maximisation rests on the assumption that rational agents adopt joint strategies in order to maximise. The rational behaviour of one individual depends on the strategy employed by the other. Thus, it is important to be able to predict others' strategies.

The rationale of constrained maximisation separates the act from its disposition. Constrained maximisation relies on rational agents being able to identify

dispositions to act and adjust their own behaviour. Constrained maximisers choose to interact with other agents similarly disposed. Gauthier follows Hobbes in moving “the question from whether it be against reason … to keep one's agreement … to whether it be against reason to be disposed to keep one's agreement” (Gauthier, 1986: 162). Thus, the rationality of the disposition to comply with an agreement becomes central in arguing that constrained maximisation is rational.

In terms of a prisoner's dilemma game, constrained maximisation is cooperative behaviour. Cooperation in repeated PD games is maximising, provided that both players cooperate. Cooperators are more likely to be accepted in an agreement, as all parties in the agreement will prefer to interact with agents who are disposed to cooperate. Since cooperation leads to optimal equilibria, it is rational for all parties to be disposed to enter a cooperative agreement. Therefore, a constrained maximising strategy gives the advantage of being accepted in a cooperative agreement and hence optimising joint strategies. Provided the agents are fully informed, the constrained maximiser does better. Thus, constrained maximisation behaviour yields a greater utility over a series of interactions. Hence, disposition translucency is vital for constrained maximisation to work as a rational strategy. If constrained maximisers cannot identify each other, they will be taken advantage of by straightforward maximisers. It is imperative that other agents' strategies are known. More specifically constrained maximisation is rational only when others' dispositions are known, or can be guessed at a high probability.

The presupposition that the agents’ disposition to cooperate or defect is known poses problems to the plausibility of the argument that constrained maximisation is rational. In order to have a realistic analysis, dispositions can only be known as a probability in a society which includes both constrained and straightforward maximisers. In this case, there can be four possible outcomes from bargaining: non-cooperation, cooperation, defection and exploitation. The ability to be fully informed becomes critical. If cooperators can identify each other, then they will commit to agreement among themselves. In this case, straightforward maximisers will not be able to maximise their utility through entering agreements.

Communication among constrained maximisers thus makes constrained maximisation a rational strategy. However, the presence of communication and information do not solve the problem of compliance. An agent may pretend to be disposed to cooperate or defect after cooperating. Disposition translucency is used by Gauthier to face this problem; people are neither opaque nor transparent. Translucency

is more plausible than assuming transparency and opaqueness. In reality most people cannot completely hide their dispositions and be opaque, nor be transparent and expect others to trust them blindly and at the same time a transparent CM in a mixed populations risks being exploited. Given that dispositions are translucent, constrained maximisation becomes rational as straightforward maximisers can only limit their interactions to those similarly disposed. An interaction between constrained maximisers leads to an optimal equilibrium and is always preferred to interactions between straightforward maximisers, both in terms of individual utility and social welfare.

Constrained maximisation is rational provided the other party is also disposed to constrain his maximisation. This means that it is essential that agents who are disposed to constrain their maximisation can identify others' dispositions so as to interact with constrained maximisers and avoid straightforward maximisers when this would lead to being exploited. Therefore, as long as disposition translucency is possible, constrained maximisation is a rational strategy. In addition, the number of constrained maximisers also plays a role; the more the constrained maximisers the more likely it is they will be able to interact. The more interactions among themselves, the more likely that cooperative outcomes will be utility maximising. In a population with a majority of constrained maximisers, it is more likely that constrained maximisation will become the norm and hence utility maximising.

In conclusion, constrained maximisation is a rational strategy when others are similarly disposed to constrain their maximisation, given disposition translucency. Constrained maximisation, like the Lockean Proviso, is a moral constraint that is grounded on rational premises. As such it is in accordance with the principles of rational morality that *MbA* puts forward. However, it is made clear that this does not include all moral behaviour or moral institutions. Constrained maximisation is not always rational. Constrained maximisers have to recognise and decline interactions with straightforward maximisers in order to achieve utility maximisation.

Constrained maximisation, in combination with the Lockean Proviso and minimax relative concession make up the core of the argument in *MbA.* Based on these concepts we see how it is possible for a rational utility maximiser to behave morally. Constrained maximisation and minimax relative concession are constraints on rational behaviour; however they are constraints of rational morality and not explicitly moral. They are derived from rational premises and hence rational agents can adopt them without loss of rationality. The same applies for the proviso which is a condition to be

accepted by rational individuals. It is based on rationality but in essence it is a moral constraint on rational maximisation. These concepts of rational morality are supplemented with a fifth one; the Archimedean Point, which will be examined in the following section.

2.5 The Archimedean Point

The Archimedean Point is introduced in *Morals by Agreement* as a position from which a rational individual can be impartial about the social arrangements that will be the result of the social contract. It is similar to Rawls's veil of ignorance (Rawls, 2005), but unlike Rawls's original position, the person at the Archimedean Point is a rational actor with full knowledge of possible capabilities and preferences; freed not from her individuality but from the content of any particular kind of individuality (Gauthier, 1986: 233). However, she is not aware of her position in society, which makes her impartial and unbiased. She has no way of knowing how her choice of an ideal social structure will affect her personal situation and this puts her in a position to decide impartially about the optimal social arrangement. The introduction of the Archimedean Point is used to show that an impartial observer will make choices that are in accordance with those made by rational agents who adopt minimax relative concession and constrained maximisation and accept the proviso. In other words, it is a mechanism that confirms that morals by agreement is a working theory of morality. Hence, "the Archimedean Point reveals the coherence of morals by agreement" (Gauthier, 1986: 17). The next paragraphs will describe the point of impartial choice through the characteristics of the rational agent occupying it.

The conditions of rationality are still present at the Archimedean Point. The ideal observer is an objective utility maximiser. However, she knows which choices will maximise social welfare. In a sense, she is only partially rational. Her rationality is limited to her understanding of the type of society her choices will construct. She is ignorant about her own utility function once she becomes a member of that society. This premise assures "equal concern" (Gauthier, 1986: 236) for all members of society. By being equally concerned, this ideal person is making certain that she – or anyone else in her position – will not have a disadvantaged position in the future society. Therefore, the main positive characteristic of the person at the Archimedean Point is equal concern, which ensures the fairness of social structures.

The Archimedean Point is more plausible than alternative accounts of

impartiality in that it is occupied by a person who has the same attributes as any other individual; “there is no other conception of the person involved” (Gauthier, 1986: 256). The ideal actor is unlike other ideal observer paradigms in moral philosophy but it is rather close to the popular understanding of an impartial point of view (Sidgwick, 1907). She is not completely disassociated from society, nor is she an ideal sympathiser. On one hand, her interest in others’ well-being exists only as long as she is unaware of her future position. On the other hand, the typical ideal observer is neither apathetic nor empathetic; he is not taking part in social interactions and he will not be a member of society eventually. In that respect, to accept him as an impartial observer would mean accepting the presence of an external force that decides fairness criteria by which it is not bound. Ideal observers of the type of the Rawlsian veil of ignorance imply the existence of impartiality outside the social realm disconnected from individual rationality. Therefore, these alternative impartial observers have to be rejected as their account of fairness is not as effective and plausible as the one provided by the Archimedean Point. The choices of the actor at the Archimedean Point will affect her utility directly and she is fully rational – with the exception of her impartiality. Therefore, she is bound, by reason and not metaphysics, to make impartial choices that construct a fair society. In other words, the agent at the Archimedean point is not restricted by idealistic assumptions to the same extent as in other models of ideal observers such as those proposed by Sidgwick (1907) and Rawls (2005). This means that the overall argument that is based on the choices made at the Archimedean point can be integrated more easily in a normative theory of social behaviour.

The analysis of the Archimedean Point does not include personal projects. This, in turn, shows how from the Archimedean Point a rational actor can choose morally without abandoning rationality. A rational ideal observer would make choices that are maximising for all. Thus we are able to characterise the ideal observer's choices. A rational actor will choose freedom and mutual benefit as was shown in the ideal market model, mutually beneficial rational interactions and interactions among similarly disposed agents. Therefore, from the impartial position she will choose a society where there is a cooperative surplus that is distributed impartially. Moreover, she will choose a society where she has the freedom to maximise her utility without exploitation or coercion or social structures that do not allow bettering one’s position by worsening someone else’s. And her choice will involve every contributor to the social outcome without exclusions. If she did not, society would be divided in conflicting groups which

would result in a sub-optimal equilibrium.

The above choices are the same rational choices that lead to justice; they include the conditions of the ideal market, minimax relative concession, the proviso and constrained maximisation. The rational impartial observer will select a society defined by a completely competitive market, which also produces an optimal outcome and a fair distribution of resources. Thus, the rational choice at the Archimedean Point leads to an ideal market society. The market is a mechanism that allows the society to maximise its cooperative surplus without coercion. The fair division of the cooperative surplus is then ensured by employing the principle of minimax relative concession. Gauthier's analysis of the Archimedean Point exhibits how an impartial observer will make the same choices as rational agents who pursue their self-interest free of coercion.

The above principles of choosing a social structure comprise an application of the theory of *Morals by Agreement*. The ideal observer at the Archimedean Point will choose a society defined by a free market, the proviso and minimax relative concession. The Archimedean Point confirms the validity of Gauthier's theory. From the impartial point the ideal observer will choose the principles of minimax relative concession, the Lockean Proviso, and constrained maximisation. Since the ideal impartial observer will adopt these principles as fair, anyone in the society will as well. The Archimedean Point concludes the theory of the *MbA*, by confirming that rational utility maximisers make the same choices as impartial observers.

2.6 Overview of the "Gauthier Project"

It seems that the most crucial and radical element of *Morals by Agreement* is constrained maximisation. The other elements of the theory had been used previously by other thinkers and Gauthier merely adjusts them to his theory. This means that they are more tested against criticism as opposed to the claim that a rational agent should be disposed to be a constrained maximiser. In addition, there are several variations of the bargaining theory that can be used to replace Zethen's model (Gauthier, 1986: 74) and the principle of minimax relative concession. Hence, out of the above discussed five elements of Gauthier's theory, we can argue that the four play a supportive role to constrained maximisation. Constrained maximisation refers to individuals adopting maximising joint strategies and so its basis lies on individual behaviour more than the proviso or minimax relative concession. Since both rationality and moral constraints are aimed at the individual, it is reasonable to assert that the core of *MbA* lies in constrained

maximisation.

Overall, *MbA* makes a convincing case for rational constraints on maximising behaviour that lead to moral outcomes. All the elements of the theory appeal to individual rationality but at the same time they are the basis for moral behaviour. However, the extent to which Gauthier succeeds in grounding moral behaviour on rationality has been disputed. The criticism of *MbA* can be put in two categories: first, critics claim that Gauthier has not been able to provide a plausible account of rational morality, but merely a rationalised version of morality. A second category of criticism focuses on the fact that Gauthier's rational constraints are moralised. These two categories are closely linked in that they attack morals by agreement from different sides, but in essence target the possibility of a rational morality. On the one hand, minimax relative concession and constrained maximisation have been criticised as moral concepts. For Gauthier, rational agents accept constraints on their maximising behaviour, but from the perspective of standard economic theory it is possible that reasoning differently (as Gauthier puts it) is non-rational. Constraints such as these have to be moral in that they require that rational agents should not maximise; therefore, they cannot be justified by traditional ways of understanding rational choice. On the other hand, moralists claim that these constraints are indeed moral, but not strong enough to lead to a fair outcome.

Understanding the argument of the critics will help understanding Gauthier’s contractarian theory better and its main weaknesses. The criticism exposes the implications of Gauthier’s theory in new ways. Therefore, criticising *Morals by Agreement* is a very effective method to show its limitations and the possibility for improvements. The following section will examine the arguments of the two categories of critics and, through that, attempt a deeper overall understanding of the theory.

2.6.1 Weak Morality

The main criticism of the theory presented in *MbA* is that it fails to generate moral requirements on the basis of rational agency. Critics accuse Gauthier of merely creating a theory of alternative morality, not an original theory of morality. The elements of his theory that he claims are rational, are in reality moral constraints on rational maximisation. As such, Gauthier has failed in producing a theory of rational morality, in which moral behaviour will be explained exclusively by rational premises. The criticism of this stream focuses on the fact that Gauthier's moral constraints in their attempt to

seem rational are too weak to lead to a just outcome. The charge refers to all of the elements of the theory, the proviso, minimax relative concession and constrained maximisation.

The Lockean Proviso as described by Gauthier is unjust as it allows for great inequalities and immoral behaviour. The example of the starving woman who is not even allowed to take the crumbs out of the rich person's table is indicative. The wealthy man is not required by the Lockean Proviso to help someone who starves, even if it would have a very small negative impact on his utility (Gauthier, 1986). And this behaviour is not moral in the general sense of the word. Gauthier does not defend it as moral element either and thus it is difficult to accept the Lockean Proviso as a necessary condition for a moral state. The implication is that this situation is the outcome of a violation of the proviso or minimax relative concession in the past (Hausman, 1989). But even if this is the case, the theory in *MbA* does not have with any method to correct the situation in the present, but relies on the fact that there have not been violations of its premises in the past. In this sense the Lockean Proviso is too weak a moral constraint to sustain a moral state of affairs.

Moreover, the Lockean Proviso does not specify the utility amounts by which one can improve his situation. The implication is that, according to the proviso, it is acceptable for one to become marginally better off while another's utility increases greatly. Another criticism of the proviso is that it allows one to exploit another provided that if one did not, someone else would (Hubin and Lambeth, in Vallentyne 1991). This is made explicit through the slave owners' example: in the absence of a horrible slave owner all his slaves would be owned by an even more horrible slave owner. Therefore, the slaves are better off since the first slave owner exists. Finally, Gauthier's use of the proviso has been criticised because it is seen as an attempt to reduce morality to utility maximisation. Although this is not necessarily true in Gauthier's case, proponents of pure moral theory (one that is based on individuals' motives being moral, not rational) are sceptical of the inclusion of utility in describing the outcomes of the proviso. In sum, Gauthier's use of the proviso has been criticised for not providing a strong enough moral claim; it is arbitrary and allows immoral situations thus making it an inappropriate condition for justice.

Gauthier's bargaining theory components are weak in terms of their rational justification; both minimax relative concession and constrained maximisation are introduced as rational elements, but they contradict the principles of rational choice

theory in the sense that they are based on the assumption that “rational” agents will forego immediate benefit in order to benefit from future maximisation. The criticism against the minimax relative concession principle and constrained maximisation is in the same vein as they are not considered strong enough requirements always to ensure an outcome that will be mutually beneficial and accepted by those worse off. The bargaining process that Gauthier suggests is problematic and therefore it is likely to lead to an unjust outcome. Minimax relative concession compares the concessions the bargainers make in order to agree on a contract. However, the concept of concessions contradicts the definition of rational, self-regarding utility maximisers; the bargaining process entails accepting less than the maximum desirable gain and in this sense can be seen as problematic from a rational choice theory perspective. Gauthier defends this because he argues that MRC is the best one can do compatible with others having sufficient incentive to agree to the distribution, but this is not at all clear and indeed in his later work Gauthier adopts a bargaining principle that more closely resembles Nash’s solution (Gauthier & Sugden 1993). Therefore, according to this line of criticism, a more powerful moral constraint is needed; one that will set more specific moral limitations on agents’ behaviour.

Constrained maximisation is also problematic from a traditional rational choice theory perspective. It is imported as a moral constraint but is not strong enough to enforce moral behaviour on self-regarding utility maximisers. In a sense it is a moral constraint on rational behaviour but it is not clear whether it applies primarily to rational or moral agents. In other words, it is not clear how constrained maximisation as a principle is transferred from rational choice theory to moral philosophy. An interpretation is left open that it is a moral constraint that is presented as a rational one. Thus, by the use of constrained maximisation, Gauthier reduces a moral rule of behaviour to a rational one without adequate explanation. He sees constrained maximisation as a “core element in the agent's overarching life plan” (Gauthier in Gauthier & Sugden, 1993: 188), but this contradicts rational agency as accepted by economists and game theoreticians. Moreover, instrumental rationality is at odds with the principle of constrained maximisation (Hollis 1993: 40). Gauthier's strict assumption of instrumental rationality conflicts with the concept of rational morality; in the economic account of rationality, rational behaviour leads to utility maximisation. On the contrary, in Gauthier's account constrained maximisation is rational because it promotes utility. Once a constrained maximiser finds himself in an agreement with a similarly

disposed agent, it is rational for him to turn into a straightforward maximiser. The only way to avoid this gap in Gauthier's theory is to use expressive instead of instrumental rationality. Using instrumentally rational maximisers as the actors in *MbA* does not provide a convincing argument for rational constraints on maximising behaviour. Put differently, someone who is disposed to cooperate does not have to cooperate all the way through a series of interactions. He does not have to keep acting according to his disposition once he benefits from others' constrained maximisation behaviour. Practical rationality cannot provide a bargaining solution as disposition to behave morally depends on how one expects others to behave. However, wishing that others cooperate does not make any difference. Thus, it is unlikely that rational dispositions will lead to moral behaviour (Nida-Rumelin, in Gauthier and Sugden 1993). In conclusion, rational compliance in the "Gauthier project" is seen as problematic in that it is not in accordance with the assumptions of practical rationality and rational agency as these are perceived by traditional economic theory.

Similarly, Hampton (1986) exposes the problem of compliance in the contractarian approach; the difficulty in keeping one's promises without an external enforcer. She argues that it is not important whether individuals behave based on rationality or passions since the outcome in both cases will be defection from the agreement. If humans are motivated by passions such as fear, they have no reason to behave differently after the contract. In this case, they will not comply with the agreement, even if they have agreed to it. In the case where individuals are rational actors living in a non-cooperative state of nature, there is no incentive for them to cooperate after they have benefited from a contract. Individuals can neither be forced to change their behaviour by Hobbes's Leviathan, nor convinced by Gauthier's argument for rational morality.

The above paragraphs have looked at how proponents of purely moral constraints on rational behaviour have criticised Gauthier's attempt to produce a theory of rational morality. Critics have focused on the concepts of constrained maximisation, minimax relative concession and the proviso to show that the rational constraints in *MbA* are in reality moral. According to the critics, Gauthier's attempt to derive moral principles from rational premises fails in that his principles are only pseudo-rational. The following section will examine the second stream of criticism for the "Gauthier project"; that includes rational choice theorists who claim that Gauthier has "polluted" rational constraints with moral principles and thus his project fails.

2.6.2 Moral Principles

The second category of critics include those who claim that the "Gauthier project" is moralised by the inclusion of clearly moral constraints on rational behaviour. The Lockean Proviso, minimax relative concession principle and constrained maximisation are not adequately justified as rational constraints. Therefore, Gauthier's theory is not a purely rational choice theory of moral behaviour. These constraints will have either to be altered so as to fit rational choice theory premises or removed.

The theory of *Morals by Agreement* does not allow coercion. However, this is not justified as rational agents have no reason to avoid coercing others, if by doing so they maximise their utility. On one interpretation, Gauthier's claim is that coercion is ruled out not for (hidden) moral reasons, but because coerced contracts will turn out to be unstable. This is because those who are coerced realise that they could have done better. However, it is always true that parties to a contract could have done better than they did if their initial bargaining position had been stronger. What motivates the coerced must be that in some sense their initial position was illegitimate; it is because they think they have been treated unfairly that they wish to renegotiate. That, though, is a moral claim.

The Lockean Proviso is therefore not effective as a rational constraint. The same applies for the other constraints in Gauthier's theory, the principle of minimax relative concession and constrained maximisation. Gauthier does not offer a convincing justification why rational agents should adhere to them. The fact that rational agents will benefit in the long term by accepting constraints on their behaviour might be a reasonable assertion, but it is a very weak argument in terms of the theory of rational choice, which demands immediate utility maximisation. Therefore, the moral constraints in *Morals by Agreement* are not supported efficiently and remain unjustifiable. They could be an integral part of the theory, should there be a valid justification of how they help us move from rational to moral behaviour. In sum, this line of criticism takes on Gauthier's theory on the grounds that his rational premises are not indeed purely rational; they are moralised as the constraints on behaviour are moral and not exclusively rational. Gauthier needs to do more in order to explain how the conditions he uses are explicitly rational.

The principle of minimax relative concession is Gauthier's bargaining solution. However, it has come under fierce criticism as it probably it is the weakest component of the theory from a rational choice theory perspective. Minimax relative concession is

unjustified and very difficult to defend; especially from a game theoretical side. Gauthier himself acknowledges that “at most it may have a heuristic value in presenting the idea of minimax relative concession as uniting rationality and morality” (Gauthier, 1993: 178). The Nash bargaining solution can be as efficient as minimax relative concession – if not a more efficient one. Thus, Gauthier’s theory would have been better served by the Nash bargaining solution. The criticism for minimax relative concession arises from the fact that it is weak in game theory terms. Moreover, Binmore (1993) proves how Gauthier has no reason to import his own bargaining solution and ignore the Nash solution. In *Morals by Agreement* there is no convincing argument that minimax relative concession is a more appropriate solution than Nash’s. The fact that Gauthier does not provide a formal proof of minimax relative concession makes it easier for game theorists to attack his bargaining theory on mathematical grounds. Nash’s solution is bound to be more attractive from a mathematical point of view.

Moreover, Gauthier’s model for minimax relative concession principle assumes a two person interaction. There is no argument for how it would work on a group level. Gauthier’s theory refers to how to fairly distribute a social surplus. Therefore, he should have provided a bargaining solution that addresses interaction among more than two agents. His minimax relative concession solution could be generalised to N persons, but he does not make any attempt to do so. Neither does he mention the possibility of applications to interactions among more than two agents. A formal approach to the bargaining problem would be more likely to discuss distribution in groups. Both Nash and Kalai-Smorodinsky offer bargaining solutions that are not limited to two person games (Gaertner and Klemish-Ahlert, in Vallentyne 1991).

In short, minimax relative concession as a bargaining solution is problematic. The lack of an axiomatic analysis means there are possibilities that are not addressed. An axiomatic analysis would not necessarily make minimax relative concession a sounder solution, but it would provide a stronger argument (Gauthier & Sugden, 1993). It is a fact though that the Nash formal bargaining solution is eventually more convincing for Gauthier as well (Gauthier & Sugden 1993). Much of the criticism from the side of rational choice theory has focused on minimax relative concession. However, a fair amount of criticism also concerns other elements of the theory such as constrained maximisation.

An additional relevant criticism is that constrained maximisation asks too much. It is a moral requirement that is very difficult to incorporate in a theory of

rational choice. Gauthier assumes that agents are fully rational utility maximisers. However, paradoxically he expects them to constrain their maximisation. If we accept the premises of rational choice theory, how are we supposed to accept that in some situations, rational actors will not maximise? Gauthier talks about the disposition to constrain one's maximisation, but this disposition has to be converted to action in order for his model to work. He does not provide a convincing enough reason for which a rational agent will have to act according to his disposition. The fact that a disposition is rational does not necessarily mean that the action, even if undertaken, will be rational as well. Holly Smith (in Vallentyne 1991) argues that even in the case that all agents are disposed to cooperate it does not follow that they will indeed cooperate. It actually pays one more to be disposed to cooperate but do not actually cooperate. Moreover, two constrained maximisers are not certain to always cooperate: it is impossible for each of them to know that the disposition of the other will be translated into cooperative action. Therefore, interaction between constrained maximisers does not always lead to the cooperative surplus. A rational agreement does not always ensure that cooperation will be the utility maximising response. It is "quite implausible to assume that any intention of mine inevitably causes my subsequent carrying out of that intention" (Smith in Vallentyne, 1991: 236).

Dispositions to act do not always lead to acting: "Covenants are but words… having no force to oblige, contain, constrain or protect…" (Hobbes 1962: 146). There is a weak link between being disposed to act morally and actually acting morally. Kavka (1983) presented this criticism as the toxin puzzle; an eccentric billionaire pays one million dollars to anyone who will promise to drink a toxin that will make him sick for a day, but will have no long term effects. It is rational then to form an intention to drink the toxin in order to collect the money. However, once the money has been paid, a rational agent would decide against an action that would make him ill. In other words, a rational disposition to act is not always followed by the corresponding rational action. Kavka's paradox exhibits how Gauthier's account of the relationship of a disposition to action and an actual action, can be problematic. In addition to that criticism, the conditions of constrained maximisation are problematic for rational choice theory.

Transparency and translucency are essential conditions for cooperation in Gauthier's argument. However, they are not plausible as it is very difficult, if not impossible, for an agent to detect others' dispositions. This does not apply only to empirical claims about the possibility of translucency, but to its theoretical implications

as well. Assuming that rational agents have capacities that are super-human weakens the theoretical argument. Of course dispositional translucency is not an exclusively idealistic characteristic of human nature. People within societies can guess what others will do, especially if they know them. However, this is not always the case and it is not the case that we only interact with people we do not know. A stranger, just like a friend, can deceive us if there is a strong enough incentive to do so. Sayre-McCord (in Vallentyne, 1991) argues that the translucency assumption makes the applicability of the model impossible. Fully informed individuals are assumed to engage in mutually beneficial interaction. In the real world though, it is virtually impossible to detect a disposition for cooperation. Straightforward maximisers may pretend to be disposed to cooperate and exploit constrained maximisers. Therefore, transparency "may generate a defence for being moral" (Sayre-McCord in Vallentyne, 1991: 188) but it is limited to that. Although of course using certain idealised assumptions is not fatal in an argument in moral philosophy, it is certainly preferable to use realistic assumptions instead when it is possible to do so. Translucency is more realistic but still not plausible and the same criticism applies for both transparency and translucency. People can be deceived about others' dispositions and rational actors have reasons to pretend. And being effective in deceiving may prove to be more beneficial than cooperating.

Being opaque should be a burden in order for Gauthier's argument to work. However, as Sayre-McCord argues, opacity is not always an obvious characteristic. Opaque agents may very well appear to be transparent so as to attract cooperation. The discussion revolves around issues of collective action then. A society can make deception very costly, so as that agents will have to appear in their true colours before the bargain. Societies however have failed to create institutions that make free-riding or parasitic behaviour too costly. On the contrary it usually pays one to act as a cooperator, while in reality being a defector. In conclusion, Sayre-McCord's main line of criticism for Gauthier's theory is the realism of the translucency and transparency of rational agents. Since cooperation is vital for *Morals by Agreement*, showing that it is unlikely for rational agents to be disposed to cooperate reveals a major disadvantage in Gauthier's theory.

The criticism for morals by agreement coming from rational choice theory focuses on the premises of the theory that are not exclusively rational, and especially minimax relative concession and constrained maximisation. However, the spirit of the criticism is the same and applies to the "Gauthier project" as a whole; Gauthier

introduces moral claims within rational premises.

2.7 Conclusion

The “Gauthier project” aims at creating a theory of morals based on rationality. If successful, the incentive of moral behaviour will be rational and the question of why one ought to be good becomes irrelevant. Gauthier starts from showing how an ideal market would be a morally free zone. In an ideal market all persons are affected by interaction in the same way and hence the market is impartial. And for Gauthier impartiality coincides with morality. However, there are no ideal markets. Markets fail by creating externalities and we need cooperative interaction to correct these failures. In order for rational cooperation to be optimal and fair, there is a need for moral constraints on behaviour. The Lockean Proviso ensures justice, by not allowing more powerful actors to take advantage of weaker ones. Minimax relative concession sets the rules for a fair bargaining procedure which is essential for the construction of a fair society. Gauthier’s contract theory describes how it would be possible to reach an agreement about social interactions. It starts from a natural state and through bargaining reaches a fair social contract. The social contract is based on moral norms arising from rational bargaining. Moral norms are the product of rational bargaining and compliance with them depends on individual rationality. In that respect, Gauthier shows that compliance with the terms of a rational agreement is beneficial and leads to optimal outcomes. *Morals by Agreement* has come under a lot of criticism for failing to derive a moral theory from rationality. The moral constraints Gauthier introduces are unacceptable for rational choice theorists. Thus, his theory has to be purified in a sense so that it will be a theory of moral philosophy based on rational choice constraints. Binmore, Sugden and Skyrms have contributed in this field. Their starting point though is Economics and Social Science and therefore their methodology is different. In addition, they do not use moral constraints; at least not the way Gauthier does. In this sense, they are more successful in following the conventions of rationality and using them to create a theory of rational morality. Their analysis uses evolutionary game theory and, as opposed to Gauthier, who sees social contract as the result of an agreement, they see it as having developed through social interaction. Their evolutionary account of moral norms can be used to reinforce Gauthier's morals by agreement by replacing the moralised aspects of his theory. Their work will be examined in Chapter Three. Moreover, in the following chapters an evolutionary

account of social conventions will be used to replace the need for bargaining and constrained maximisation, without loss of rationality or moral outcomes. That account will be shown to be more realistic as it refers to societies as well as individuals. Gauthier's theory is based on assumptions and analysis of individual behaviour used in order to examine social interactions. In that sense, it fails to take into account social dynamics that can only be analysed when we look at a social group as a whole and not exclusively at the behaviour of its members. Additionally, moral norms as the result of evolutionary processes of social structures and conventions can be a more efficient type of moral rationality than the one Gauthier proposes. Morality in this context is not as important as optimality in equilibrium. A society that has reached an optimal equilibrium will be characterised by the moral behaviour of its members (Gauthier & Sugden, 1993: 150).

3. Evolutionary Theory in Moral Contractarianism

The previous chapter analysed the Gauthier project discussing its strengths and weaknesses. This chapter will focus on literature including the use of evolutionary and game theoretical models in moral philosophy. More specifically, it will discuss the literature of game theoretical models regarding social and cultural evolution. Examining the literature in moral philosophy that utilises evolutionary game theory should make clearer that it is possible to view moral behaviour as the outcome of rational deliberation within dynamic social structures that are described by cultural evolutionary models.

Evolutionary game theory can explain and justify an account of rational constrained utility maximisation taking into account evolutionary factors, thus making the concept of constrained maximisation impervious to the criticisms related to Gauthier's use of moral constraints that were examined in the previous chapter. Evolutionary accounts of social explanation propose the use of a dynamic process of repeated interactions in the place of constraints on rational maximisation. Repeated and evolving strategic interactions can show that rationality does not have to be contaminated with moral concepts in order to lead to a moral outcome. Therefore, evolutionary theory can be used to advance the Gauthier project by reinforcing his argument. The main representatives of this “purist” thinking include Brian Skyrms, Robert Sugden and Kenneth Binmore.

Skyrms (2004) and Sugden (2004) attempt to show how evolutionary accounts of the creation of social structures and conventions account for human behaviour. Binmore's (1998) more ambitious attempt is to combine contractarian moral philosophy with formal mathematical models and evolutionary game theory, in order to explain theories of justice and provide rational incentives for cooperative behaviour. By comparison to the Gauthier project they are all more focused on explaining moral behaviour and social institutions of justice, rather than offering a normative account. However, explaining justice with rationality would also provide incentives for just behaviour. In other words, they try to show that rational agents have incentives to be moral. Therefore under the assumption that humans are rational, there is no need for externally enforced cooperation. Cooperation would be sustained by rational individuals.

Gauthier follows the philosophical tradition started by Hobbes's *Leviathan*, asserting that individuals are mutually unconcerned utility maximisers. However, his influences also include Hume in that he argues that moral responsibility is based on rationality and Rousseau in the sense that they are both moral contractarians. Evolutionary game theory explanations in moral philosophy follow the philosophical tradition of Hume and Rousseau by acknowledging the importance of reciprocal altruism. Evolutionary accounts of human behaviour imply that we can understand the whole by understanding its components – more specifically understand social behaviour by explaining individual behaviour.

Combining explanations on group and individual level creates tension which arises from the fact that social and individual theoretical paradigms are often in conflict. In the context of the present analysis the most significant area of conflict has to do with the definition of individual rationality. Evolutionary theory assumes agents of limited rationality whereas traditional rational choice theory asserts hyper-rationality. The following sections will deal with the methodological aspects of combining evolutionary game theory with individual rationality, before moving on to discussing the theory of social conventions which will be used to link individual with social behaviour.

3.1 Methodological Aspects

Using evolutionary accounts of behaviour in conjunction with assumptions of individual rationality causes methodological problems. This section will attempt to show how these two branches of social explanation can be reconciled through a focus on how these seemingly opposing paradigms have been used in the literature. In order to achieve this, the first point of analysis will be our understanding of individual rationality (Gintis, 2009).

Rational choice theory is defined and limited by the premises of methodological individualism. In turn, methodological individualism is based on the idea that the whole has to be explained in terms of its components (Hollis, 1996). This means that in behavioural sciences, society has to be analysed in terms of individual actions. Therefore, the model of homo economicus becomes prominent. The assertion that there are radical differences between methodological individualism and holism is not as generally accepted as it used to be (Gintis, 2009; Young, 2001). However, it needs to be examined here in the context of the literature for two main reasons. First, holistic theories of social evolution incorporate individual action; however, they fail to provide

convincing solutions to problems arising from individually rational behaviour within social groups. Secondly, the similarities and differences between social and biological evolutionary models have to be examined in order to see to what extent biological evolution can be used as a pattern for cultural evolution. These three factors will be examined shortly in the following paragraphs.

First, paradigms of evolutionary game theory such as Sugden's and Skyrms's make reference to social evolution which means that social action is viewed from a holistic perspective as well. More specifically, Sugden's use of conventions, examines how individuals in societies behave differently than they would outside society. Agents do follow the social conventions even when by doing so, they do not strictly maximise their utility. This account bypasses a line of criticism similar to the one used against Gauthier's claim that constrained maximisation is not a moral principle, since agents are not assumed to be perfectly rational (according to the economic paradigm), and their behaviour is assessed at the end of a series of interactions. A conventional agent is rational in that she participates in a convention that maximises her utility in the long term as opposed to expecting a maximising outcome out of every single interaction. Individual maximisation comes as a result of social norms being created and observed. Therefore, for Sugden, individual maximisation and conventional behaviour are interdependent and one feeds off the other.

Similarly, Skyrms discusses the evolution of social structures and group selection. Behaviour is imitated at the level of the individual but evolves within social groups. Arguably, the account presented in *The Stag Hunt and the Evolution of the Social Contract* (Skyrms, 2004) relies more on collective explanation of social evolution and leaves little room for individual rationality. However, an integral part of Skyrms's theory is that individuals imitate the most efficient strategies and adapt their behaviour to their environment. Although not explicitly stated, this can be viewed as an inclusion of a concept of limited rationality.

Finally, Binmore blurs the lines between holism and individualism. His claim that genes affect cultural evolution makes it difficult to see where biological evolution concepts stop and the rational actor model begins. His argument includes rather bold assertions about genetic predispositions for cooperative behaviour, given certain conditions such as the size of the social group. However, the core of his argument is that individual action leads to the evolution and creation of social norms that in turn create the bounds for the behaviour of rational agents. In that respect, his theory can be seen as

complementary to those of Sugden and Skyrms.

The second reason why the combination of holistic and individualistic theories of social explanation can be problematic has to do with their inability to deal with collective action failures. In essence this is related to assumptions of individual rationality within group behaviour. A group of rational individuals does not necessarily have to behave rationally. In other words, rational behaviour is different when we look at an agent who interacts with another agent than when behaviour is examined on a collective level; on the former case strict individual maximisation is rational whereas in the latter, rational behaviour has to lead to maximising group welfare. In the prisoner's dilemma game for instance, maximising behaviour changes depending on whether we are talking about a two person or an n person game and whether we view the game as a model for individual maximisation or collective action. In the typical PD game, defect is always the rational strategy. However, when the interaction is examined as a collective action problem, cooperation yields higher utility.

Finally, the third reason why incorporating the rational choice theory model into evolutionary theory is controversial relates to the discrepancies between biological and cultural evolution. In biological evolution there is no room for the concept of individual actions that may change the course of evolution. In cultural evolution a claim that individual actions have no impact over the final outcome would lead to a type of social determinism that is not in accordance with the literature examined here. Biological evolution does not select the fittest individual but the fittest gene (Dawkins, 2006). In order to be able to transfer biological evolution into behavioural sciences, we have to account for the absence of gene selection, or even individual selection, in cultural evolution. Cultural evolution discusses how specific social conventions or structures evolve to become stable over competing social models. In cultural evolution we should take into account the concept of memes which are selected, as opposed to individual selection (Binmore, 1998). The theories reviewed here, use evolutionary theory to examine strategy selection and hence biological gene selection mechanisms are not needed and do not cause problems for the theory.

The problems arising from using evolutionary theory to describe social behaviour are common throughout the literature and constitute the main similarities among the above mentioned theories. Despite the important similarities, there are also differences in the methodology of the work of Sugden, Skyrms and Binmore. All three scholars accept the premises of experimental economics and build their theories on

laboratory results and computer simulations. Sugden's work is probably the one closest to moral and political philosophy. In a nutshell his argument is that the social contract is comprised of social conventions that in turn develop from repeated interactions among rational individuals. In Sugden's work justice comes about when rational agents freely pursue their self-interest. Along with Gauthier's, Sugden's work will be the most important in forming the next chapters and providing a moral theory framework in which Skyrms and Binmore can be understood.

Skyrms's work is the most innovative one from a game theoretical perspective. He argues that in evolutionary game theory both strategies and game structures evolve and affect the bargaining procedure and the final contract point. The coevolution of strategies and structures makes his analysis more realistic in that it approaches the dynamic changes of real societies. In addition, *The Stag Hunt and the Evolutions of the Social Contract* (2004) uses computer simulations to examine the behaviour of basic biological organisms, such as colonies of bacteria, to study the dynamics of cooperation. Although there is no claim that bacteria and societies evolve in the same way, this analysis offers new insights into individual and group selection in theories of evolution.

Just like Sugden, Skyrms is forced to discuss more than one game. Although the stag hunt can describe social interaction better than the prisoner's dilemma, it is still inadequate to account for all kinds of social interaction. Therefore, the prisoner's dilemma is also used as a special case, as well as variations of the stag hunt. Different games have different basins of attraction which means that the dominant strategy in a population changes as the number following that strategy changes; for instance stag hunting has a basin of attraction of 75 percent which means that "[i]f more that 76 percent of the population hunts stag, then stag hunters will take over" (Skyrms, 2004: 11). Therefore, games with different basins of attraction can evolve into distinct social structures with different equilibria.

Binmore's theory is the most difficult to classify methodologically as it very extensive and exceptionally broad. In short, he proposes an evolutionary account of moral norms. Biological evolution plays a role in that we are all bound by our genes. Furthermore and similarly to Sugden, Binmore also suggests that social norms are subject to evolutionary forces. Evolution plays a two-fold role: it affects human behaviour biologically and also affects how human moral norms evolve. Finally, the social contract is viewed as an equilibrium in repeated games and rational agents will choose the best equilibrium for their social contract because "fairness is evolution's

solution to the equilibrium selection problem" (Binmore 2005: 14). Binmore, without abandoning the assumption of rational actors, constructs an evolutionary theory of justice.

The above paragraphs have been an attempt to discuss the literature of the general field of evolutionary game theory and moral philosophy. The work of Sugden, Skyrms and Binmore has a lot in common and in practice belongs in the same field, which is both innovative and difficult to classify. Sugden is probably the first to have written in this field combining the idea of justice with game theory and evolution. His work has clear normative implications in addition to its explanatory power. Skyrms on the contrary focuses on describing the dynamics of interactions that can lead to cooperative equilibria but makes no explicit normative claims. Finally, Binmore offers a more ambitious account that aims to explain cultural evolution with a more direct albeit complex normative account. It is clear however that all the above, and especially Sugden and Binmore, follow the philosophical tradition of Hume. In contrast with Gauthier's clear affiliation with the Hobbesian tradition, the following chapters will use key concepts from both the Hobbesian and the Humean traditions.

The next sections will discuss more analytically the work of Sugden, Skyrms and Binmore. First, Sugden's work will be examined through an analysis of his concept of social conventions in the next paragraphs.

3.2 Spontaneous order

Spontaneous order, a phrase first used by Hayek (Sugden 2004), describes very adequately the main aim of *The Economics of Rights, Cooperation and Welfare* (Sugden, 2004): to show that social interactions can lead to equilibria of moral behaviour without third party enforcement. The core of the argument is that societies, just like ideal markets, reach efficient and optimal equilibria should they be left to operate freely. In this framework, a convention is a "stable equilibrium in a game that has two or more stable equilibria" (Sugden, 2004: 32).

Stable equilibria are evolutionarily stable, which means that they are the result of iterated interactions among rational actors and cannot be destabilised by the adoption of alternative strategies. Society reaches an evolutionarily stable equilibrium when all its members, or almost all of them, follow their maximising strategy. Thus, an overwhelming majority of a population has to adopt the conventional behaviour in order for the convention to become established. The greater the number of individuals that

follow a convention the more likely it is that this convention will expand, until it becomes a social convention that is generally followed. The implication here is that conventions arise and become stable randomly and not so much because of individual rational deliberation. What matters is the establishment of a convention so as to regulate social interactions and avoid conflict and not the selection of a specific convention. In conclusion, the type of convention and the equilibrium point are not important. What matters in this analysis is their becoming established and stable. There are different types of conventions with various structures and equilibrium points.

Sugden distinguishes three categories of conventions: conventions of coordination, conventions of property and conventions of reciprocity (Sugden, 2004). They are all seen as equilibria of repeated games whose purpose is social peace by generating an understanding of justice. The break of conventional behaviour is viewed as unjust by those who follow it, as it is the convention in the first place that has created a sense of what is just. The breaking of a convention for whatever reason, either by mistake or weaknesses of will or because it is deemed irrational, creates a feeling of injustice to others, as established conventions serve as social behaviour regulators. Since it is rational to keep conventions as long as others keep them, it follows that is also rational for one to want others to keep the convention (Sugden, 2004).

Conventions are characterised as moral and rational: "...conventions are normally maintained by both interest and morality..." (Sugden, 2004: 155). They come about as the result of rational interaction, but rationality alone cannot sustain them. In a sense, in Sugden's work our sense of morality is being informed by established conventions, which are also the outcome of rational interactions. Moreover, since there is no equilibrium selection mechanism provided, an established convention while maximising for its members may very well be random, not moral. In other words, rationality leads to conventions of justice through an arbitrary evolutionary path. In this respect, Sugden is close to Gauthier. They both justify a constrained rational behaviour that is not always maximising, based on interactions with other rational agents who are disposed to behave in the same way in *Morals by Agreement*, or have been behaving in the same way in *The Economics of Rights, Cooperation and Welfare* (2004). Sugden uses game theory more formally and extensively than Gauthier. This has an impact on his use of the idea of morality which is evidently more difficult to incorporate in formal analysis. Justice for Sugden is a side effect of the interactions between rational agents in an evolutionary context. Thus moral behaviour is the outcome of rationality and does

serve as its constraint. Morality is the result of spontaneous order that arises in the form of natural, and not designed, conventions. Sugden manages to show that it is plausible to assert that social order does not depend on an external enforcer and that conventions of justice are self-enforced once they become established.

Although very plausible, the theory of conventions proposed by Sugden does not offer an account of how conventions become established in the first place. However, Skyrms's *The Stag Hunt and the Evolution of the Social Contract* (2004) focuses on the characteristics of the games and their dynamic structure in attempting to explain how evolutionary equilibria develop. The following section will discuss Skyrms's work by looking at his version of the stag hunt game and the subsequent importance of location, communication, association and coevolution.

3.3 The Stag Hunt

Although *The Stag Hunt and the Evolution of the Social Structure* (Skyrms 2004) is not directly linked to *Morals by Agreement* or moral contractarianism in general, it can be used to advance Gauthier's theory. As we have seen, constrained maximisation is rational provided interactions occur between similarly disposed agents. In other words it is rational for one to be disposed to behave as a constrained maximiser if one is in a group of constrained maximisers. Skyrms's analysis describes a similar concept of equilibria that depend on the behaviour of a critical mass of agents and utility maximisation as a function of his neighbours' strategies.

In *The Stag Hunt and the Evolution of the Social Contract* (2004) the concepts of morality and justice that are central in *MbA* are replaced by cooperation and cooperative equilibria. Skyrms's main aim is not to show how morality can arise from rational premises but to argue that social cooperative equilibria incorporate individual maximisation. The two theories have similar objectives, as Gauthier's notion of morality is closely related with cooperation. However, the cooperative social structure in Skyrms's case is the result of evolutionary processes. For Skyrms rationality "is not necessary for solving the social contract" (Skyrms 2004: xii), but on the contrary cooperative equilibria are the result of the correlated dynamics between social structures and individual strategies. Despite having similar aims to Gauthier, Skyrms adopts naturalistic premises about individual rationality, social interactions and the social contract.

Social contracts are present in the animal world as well. Skyrms, just like

Aristotle and Hobbes before him, accepts that animal societies have solved the problem of peaceful coexistence by adhering to natural social contracts, that is social contracts that are defined and bound by the natural characteristics of each species. On the contrary, human social contracts are artificial and in conflict with human nature since humans are naturally rational and outside family and social circles mutually unconcerned. As a result, they realise that their self-interest is threatened by the collective benefit and as Hobbes (1976) argued human social contracts fail because of rationality. In this respect, rationality threatens the stability of the social contract, which is sustained by the dynamics of social structures. Therefore, the use of evolutionary theory to describe human interactions and the social contract can be problematic, especially in context of rational choice theory. In addition, the prisoner's dilemma game which is commonly used to describe social interactions, is not as effective in describing dynamic interactions within social groups as the stag hunt game.

Unlike most contractarians, Skyrms uses the stag hunt game in his analysis instead of the prisoner's dilemma. The stag hunt, originally an allegory used by Rousseau, is "a prototype of the social contract" (Skyrms, 2004: 1). A tribe has to decide whether they should hunt stag or hare. Stag hunting requires everybody to participate and is therefore a joint decision. If the tribe decides to hunt stag collectively, they will need all the available hunters to participate, but collective action is not necessary for hare hunting; on the contrary hare hunting implies collective action failure. Unlike the typical PD analysis, the stag hunt discusses a group of people thus making it easier to examine issues regarding collective and individual maximisation.

The stag hunt example shows how individual rationality contradicts social welfare and is more directly linked to issues of free-riding. The cooperative social contract is sustained if the tribe collectively hunts for stags. A stag hunt equilibrium can be destabilised by a large enough number of hare hunters. In a more realistic stag hunt one or very few defectors do not affect the success of the stag hunt, or in other words the stability of the social contract. The critical number of stag hunters needed depends on the formal representation of the game. However, the implication is that in order for a cooperative social structure to be sustained, it is essential that a majority agrees and behaves according to the agreement. In this respect the stag hunt is a more appropriate type of game to be used in social explanation as it takes into account populations and not just individuals. However, their differences are not as great as they seem at first. Repeated interactions between two individuals can be described by the PD game

structure. Although, strictly speaking the PD game is not iterated, its pay-off structure can be used to examine repeated interactions. In this case benefit from future interactions becomes important.

The term “shadow of the future” (Skyrms, 2004: 4) refers to the importance attributed to future interactions. Agents who believe it is important to maximise pay-offs in future interactions, are more likely to cooperate in the present. Those who perceive interactions as repeated will see the obvious advantage of participating in cooperative interactions. Also known as “future discount factor”, the shadow of the future is a central idea in repeated games. The smaller the future discount factor, the greater the importance of future interactions for the players. In the prisoners’ dilemma for instance, if the two prisoners expect to be arrested and to be offered the same deal again, it is more likely that they will keep silent. When taking into account the shadow of the future, cooperation seems a more rational strategy and the games of the stag hunt and the prisoner's dilemma are similar. However, “[t]he shadow of the future has not solved the problem of cooperation in the prisoner’s dilemma; it has transformed it into the problem of cooperation in the stag hunt” (Skyrms, 2004: 6).

Drawing connecting lines between the two games is important because it shows firstly that cooperation in a two person game can be examined in a group context and secondly that the premises underlying game theoretical models are more important than their formal description. By extension, individual behaviour can be examined in conjunction with collective behaviour with the help of a theory of cultural evolution. Evolutionary game theory describes strategic interaction over generations. When individuals of the original generation cooperate, they will produce cooperators who will in turn cooperate eventually reaching a group or society where cooperation is the norm. Put differently, social contract games are played over generations. The game played by the original generation of a population affects the eventual equilibrium point reached after several generations. Whether this will be a cooperative or a non-cooperative equilibrium depends on whether the founding generation cooperated or defected. Therefore, the stability of the social contract depends on interactions that took place several generations earlier.

Skyrms analyses how depending on factors that will be discussed later, both stag and hare hunting equilibria are possible and will be stable once established. Despite the stag hunt equilibrium being optimal, individuals can opt to hunt hare which is also an equilibrium. In an evolutionary context either equilibrium is acceptable since they

are both stable. Gauthier offers reasons for which constrained maximisation is rational but Skyrms's limits his argument to explaining how equilibria evolve. Although they are very similar in that they distinguish between two types of behaviour whose rationality depends on one's neighbours, the evolutionary account focuses on explaining social structures. Moreover, Skyrms's explanatory model is based on the coevolution of strategies and social structures which in turn depend on the relative location of cooperators and defectors, the possibility of communication and the subsequent association. These concepts will be discussed in the following sections.

3.3.1 Location

The importance of "spatial structure, location and local interaction" (Skyrms, 2004: 15) is paramount for the theory presented in *The Stag Hunt and the Evolution of the Social Structure* (Skyrms, 2004). The relative location of stag or hare hunters in the stag hunt, changes the probability of reaching a stag hunt or a hare hunt equilibrium. A stag hunter whose neighbours are hare hunters will be forced to change behaviour and similarly a hare hunter will have to convert to stag hunting in an environment dominated by stag hunters. Individual maximising behaviour depends on one's environment and varies according to one's neighbours.

In that respect we can distinguish between two types of interactions: interactions with neighbours and interactions with strangers. In the former case an individual adjusts her strategies by imitating successful strategies in her neighbourhood. In the latter case, interactions occur between individuals selected randomly from a large population. In both cases there is a convergence of strategies towards those that are maximising; those who maximised in the first iteration will follow the same strategy in the future, whereas those who did not will change their strategy.

In the divide the dollar game that was used to examine the importance of location in equilibria selection, both types of interaction eventually lead to a division that is near the fifty-fifty mark. The game consists of two players who have to divide a dollar, but they will only get a share if they agree on how much each should get. Being rational, any alternative would be preferable to disagreement. Rational agents will cooperate, as long as they prefer the cooperation outcome to the current status. The current status is the point of disagreement. Therefore, an agreement will be reached if the bargainers' utility is greater after the agreement than before. Interactions with neighbours converge faster whereas interactions with strangers depend on the original

spatial arrangements. When interacting with strangers almost 60 percent of simulations lead to a fifty-fifty division while “[f]air division becomes the unique answer in bargaining with strangers if we change the question to that of stochastic stability in the ultra long run (Skyrms, 2004: 28). The divide the dollar game shows how the relative position of individuals provides significant insights into the evolution of equilibria.

Similar principles apply for the stag hunt game. The evolution of a stable social equilibrium depends on the spatial dynamics of stag and hare hunters. If the majority in the neighbourhood hunts hare then the few stag hunters will be converted, whereas in a neighbourhood or a population of stag hunters, a single hare hunter will change the equilibrium from all hunt stag to all hunt hare. Hare hunting is a risk free strategy as it does not require cooperation and thus it pays off irrespective of the behaviour of others. A hare hunter surrounded by stag hunters will change his behaviour next time they bargain. His close neighbours though, will also change to hare hunting. Therefore, after several interactions, hare hunting will be the equilibrium. In terms of interactions between strangers hare hunting has a replication effect on the whole population. “Hare hunting is contagious” (Skyrms, 2004: 36) and eventually a single hare hunter may change the social equilibrium. The evolutionary process in this case leads to a suboptimal social contract, which is however a sustainable equilibrium.

The unambiguous conclusion is that “local interaction makes a difference” (Skyrms, 2004: 40) in equilibrium selection. In addition, equilibrium selection is affected by whether strategies are being imitated within a neighbourhood or replicated in a population. The cornerstone of the above spatial analysis is that social structure and repeated interactions matter more than individual rationality. Individuals imitate the best available strategy not because of rational deliberation but because of the evolutionary dynamics in a given population. In any case, there is individual utility maximisation even through an evolutionary analysis. In that respect, location can be seen in relation to Gauthier's disposition translucency. For Gauthier it is rational for one to constrain her maximisation provided others are disposed to do the same; the likelihood that a constrained maximiser will interact with others similarly disposed increases in a social group that consists mostly of constrained maximisers. The similarities between disposition translucency and the rationality of constrained maximisation and the evolutionary dynamics in Skyrms's analysis should become clearer in the following section discussing communication.

3.3.2 Communication

Communication before the interactions is vital for the structure and the equilibrium of the game. Provided that pre-play communication was possible, a group of cooperators could invade and eventually take over a population of defectors, since communication in this context ensures that future behaviour can be agreed upon within a group of co-operators. Communication does not solve the problem of compliance in a PD game, but in a stag hunt game – as described by Skyrms (2004) – it does make a cooperative equilibrium a more likely outcome. Since cooperation is maximising in the long term and individuals can know that others cooperate, it is in everybody's best interest to cooperate. Therefore, costless pre-play communication is important in evolutionary games. However, costless communication can only have an effect if the cooperators can signal each other using "a secret handshake" (Skyrms, 2004: 66) in order to avoid misinformation. In other words, defectors should not be able to pretend they are cooperators. In this context, non-cooperative equilibria can shift to cooperative ones. Thus, the secret handshake is a type of behaviour the precedes interactions and shows that an agent is disposed to cooperate and also that she has cooperated in the past and as a result is aware of the secret handshake. By comparison to the account of disposition translucency presented in *MbA*, the secret handshake does not require an ability to predict the future or others' dispositions; in this respect it is a stronger argument for the possibility of identifying cooperators.

A non-cooperator can pretend to be willing to cooperate so as to take advantage of honest cooperators. There is no efficient way to detect liars and punishment can only take place after the fact. Cheap talk can therefore be advantageous for those who are able to use it effectively to misrepresent their intentions. Especially, when communication is costless, their work is made easier. Therefore, the effectiveness of the secret handshake becomes essential for the establishment of cooperative equilibria. The original claim is then used as an indication of one's intentions or, put differently, "disposition". When players have the choice between demanding ½ or ⅔, they will choose ½, as this will increase their chances of interacting. Demanding ½ means that they will be more attractive than those demanding ⅔ in an interaction. In a way, cooperators advertise their intention to cooperate, which serves as the secret handshake mechanism (Skyrms, 2004). This is based on the same rationale as the MRC in *Morals by Agreement*, but the emphasis here lies on the importance of communication between the interacting agents.

The possibility of communication before the game, just like the relative location of the players, makes a great difference in the outcome. Skyrms (2004) shows how low or no cost communication can be central in sustaining a cooperative equilibrium provided there is an effective secret handshake. Therefore, communication is as important as location in achieving a maximising equilibrium. A stag hunter surrounded by hare hunters will not attempt to cooperate, provided there is efficient communication. If both the provisions of location and communication are met, it is very likely that there will be an increasing number of stag hunters. This in turn will lead to a cooperative equilibrium.

Again communication can be seen as an alternative to Gauthier's disposition translucency. In *Morals by Agreement* it is asserted that rational agents are able to see others' dispositions. Skyrms replaces the need for translucency with special signalling between agents, offering a more realistic solution to misinformation and dishonesty. The difference between the two analyses is that Skyrms takes into account social and group dynamics whereas in *Morals by Agreement* the focus is on the individual and two person interactions. In this respect it seems more plausible to argue that agents within a group have developed a subtle code of communication that is known only to members of a group, rather than that they have the ability to know strangers' dispositions.

In Gauthier's analysis the assumption of disposition translucency becomes more realistic if we take into account that constrained maximisers will rationally choose to interact repeatedly with other constrained maximisers. In other words they will form groups where constrained maximisation is generally adopted. In Skyrms this notion is viewed through the concept of association which will be described in the following paragraphs.

3.3.3 Association

Association refers to network formation. Stag hunters who communicate and are located close to each other will create networks of stag hunting equilibria. Therefore, association is based on location and communication which are essential requirements for successful network formation. In addition, association requires agents with learning abilities; successful interactions are repeated. Skyrms examines association using an example of ten strangers in a new location visiting each other (Skyrms, 2004). The greater the pleasure derived from each interaction, the more likely it is that the visit will be repeated. This leads to certain individuals exchanging visits and making friends and

enemies depending on the pleasure they derive from each visit. A visit resulting in higher pleasure – or utility – is more likely to be repeated. The contrary applies for a visit that is not pleasant.

Moreover, a pleasant visit reinforces both individuals and makes it even more probable that the interaction will be repeated. Adding the concept of memories makes the example more realistic and plausible as it brings the ideal agent of the model closer to the characteristics of real persons and thus makes it easier for the model to be applied in, and become relevant to, real life situations. When an agent has a good memory of past interactions, she will be able to choose who she visits based on previous success. Even in the case where memory fades relatively quickly or interactions occurred a long time ago, we assume that individuals can have an account of their history of interactions and whether it has been mostly pleasant or not. Although not directly acknowledged, the example with strangers exchanging visits is reminiscent of the Robinson Crusoe allegory discussed by Gauthier, which shows how the rationale behind each analysis is to an extent similar. With it, Skyrms shows how individuals with no history and free from outside constraints, form bonds as a result of repeated interactions. The stag hunt game also embodies a form of reinforcement of behaviour; stag and hare hunters' behaviour is reinforced when they interact with individuals following respective strategies.

Each agent has the opportunity to “look around and if another strategy is getting a better pay-off than his, then he imitates that one” (Skyrms, 2004: 106). This sentence encompasses the central point of the concept of association. The player has the option to imitate more successful strategies and strategies developed in different groups of interaction. The main argument here is that different strategies will lead to different equilibria and eventually different social contracts. All interactions start from similar initial positions, but the final social contract depends on the agents’ choice of strategies, their relative position and the ability to identify each other's intentions accurately. Location, communication and association all play a central role in the evolution of the social contract as presented by Skyrms. A final element of the theory is their coevolution; each one develops in parallel with the other and affects it; as a result evolutionarily stable strategies also change in time. With each new iteration, or generation in evolutionary context, the parameters have changed and therefore the maximising strategies have to adapt. Coevolution will be analysed in the next, last section on Skyrms's theory.

3.3.4 Coevolution

Individuals who choose their strategy and the agents with whom they will interact, also affect the structure of the game. Therefore, there is a simultaneous evolutionary process by which maximising strategies develop in parallel with structural changes in social groups. Association is paramount for the theory developed in *The Stag Hunt and the Evolution of the Social Structure* (Skyrms, 2004) because it deals directly with strategy revisions and learning during the game. Coevolution of structure and strategy adds to association the fact that now agents are able to imitate behaviour that is observed in their social group, even if they have not encountered it; "a player looks around, and if another strategy is getting a better payoff that his is, he imitates that one" (Skyrms, 2006: 106).

Coevolution encompasses the previous elements of the theory making it a plausible model for actual social interactions; individuals do change their behaviour when they see someone else doing better or realise that interacting with new individuals will maximise their utility. Humans make friends and enemies based on the pleasure of their interactions as described in the simulations; the more the interactions continue being pleasant, the more the bonds between the interacting agents strengthen, creating social structures that support a type of behaviour. Location, communication and association provide a solid description of human behaviour that is put in a realistic dynamic context with the concept of their coevolution.

The dynamic analysis of the social contract offered by Skyrms is more realistic than the static analyses found in moral contractarianism. Societies are dynamic in that individuals change behaviour and social structures shift accordingly. Evolutionary theory advances the theoretical models of contractarianism since its description of social structures approximates real life societies more accurately than normative theories. Coevolution exhibits how the equilibrium in a stag hunt game depends on a variety of factors that have to be examined in conjunction in order to be meaningful.

The stag hunt story is a close analogy of reality although it still is an oversimplification of actual social contracts. In reality, one hare hunter (or defectors in the PD game) in a society of stag hunters (or cooperators), will not cause the breakdown of the social structure. However, a critical number of defectors in the population can affect the stability of the social contract. If a stag hunting social contract has been established as an evolutionarily stable equilibrium it cannot be affected by a minority of defectors. However, contagion or free-riding is a problem. When all, or most hunt stag,

it pays to free ride by not hunting but participating in the distribution of the successful hunt.

Free-riding behaviour in the long run is “fatal” (Skyrms, 2004: 121). Eventually it can cause the collapse of cooperative equilibria and lead to an all-hunt-hare situation. In a model such as Skyrms's, where social structure and its evolution are central for the argument, free-riding is a maximising strategy if the secret handshake safety mechanism can be bypassed. However, the same applies when the analysis rests exclusively on individual rationality. It is rational for one to benefit from the social output without participating if he can get away with it. Thus, free-riding is central to any discussion of individual rationality in the context of social interactions and poses a significant challenge to the possibility of a rational morality, as it has been exhibited by Hobbes, Hume and Gauthier. Gauthier's constrained maximisation, should it be effective, can provide a more convincing answer to the free-riding problem as it calls for an internalisation of cooperative behaviour (Gauthier, 1986). However, as was discussed in the previous chapter, the concept of internalisation of constrained maximisation is problematic. Free-riding and the possible rational incentives against it will have to be analysed in Chapter Six where there will be an attempt to show how it is not rational in a repeated interactions framework.

The theory presented in *The Stag Hunt and the Evolution of the Social Structure* (Skyrms, 2004) advances the social contract tradition by enriching it with evolutionary concepts. It is essential that its components, location, communication and association, are examined in conjunction so that it becomes clear how evolutionary game theory can lead to cooperative social structures. The main conclusion is that there can be a society at a stag hunt evolutionary equilibrium, given enough time and repeated interactions.

Overall, there are two significant methodological contributions of Skyrms's work in the contractarian tradition: first, by examining the stag hunt and the PD game he shows how games' structures and pay-offs are not as distinct as they seem at first and game theoretical premises can be used to explain social behaviour without the need to refer to a specific game. Secondly, his theory shows how it is possible to use evolutionary game theory to describe efficient social contracts. Typically, contractarianism is based on individual rationality but in *The Stag Hunt and the Social Contract* (2004) it is shown that an evolutionary account can be used to complement assumptions of rationality without necessarily negating them. Implicitly, agents are

assumed to be boundedly rational since they can learn and imitate more successful strategies while participating in utility maximising interactions. Bounded rationality will be discussed more analytically in the following section and especially Chapters Four and Five as a tool of reconciling rational choice theory with evolutionary game theory in a contractarian framework.

Despite being an innovative step in evolutionary contractarianism, Skyrms's work is limited in that although it is an effective description of social structures, it does not make normative suggestions. By comparison to Sugden's analysis of social conventions which was discussed previously, it is a more detailed account of social behaviour but at the same time more limited in scope. An intellectual synthesis of these analyses could provide a deeper understanding of the social contract while at the same time offering normative suggestions about individual and collective behaviour. Binmore's work, which attempted to achieve just that in *Game Theory and the Social Contract* (1998), will be examined in the following section.

3.4 Game Theory and the Social Contract

Binmore advances the work of Sugden and Skyrms by examining social behaviour through evolutionary theory. However, his work is much broader than Sugden's and more complex that Skyrms's. He is not limited in a naturalistic account of fairness conventions like Sugden, or simulation models like Skyrms's, but attempts to create an evolutionary account of the social contract theory, following mostly the tradition reinvented by Rawls (Rawls, 2005). As a result Binmore's is a naturalised normative theory of justice and not simply an explanatory model of cultural evolution and social conventions.

Despite the fact that he concedes that "…justice is an exclusively human phenomenon" (Binmore 2005: 11), many of the arguments are grounded on biological and not only cultural evolution. For Binmore, human genetic identity forces us to behave in certain ways which give rise to conventions of behaviour that bound our social behaviour. *Game Theory and the Social Contract* (1998) is an attempt to create an evolutionary theory of justice with cultural evolution as part of biological evolution. Naturalists, like Binmore, argue that natural sciences can be a valuable tool in social sciences and moral philosophy by offering significant insights into how social structures evolve. However, naturalism is the starting point as the final aim is the creation of a contractarian theory of justice.

Binmore distinguishes between two kinds of games: the game of life which “is beyond our power to alter its rules at will” (Binmore, 1998: 4) and the game of morals which refers to the artificial laws of morality. The game of life is modelled as an infinitely repeated interaction between two players, which makes it likely that reciprocity will emerge. The rational agents interacting in the game of life, in a state of nature similar to the one described in *Leviathan* (Hobbes, 1976), are indifferent between the game's multiple equilibria. The game of morals includes the moral principles that provide the mechanism to distinguish among the many equilibria of the game of life based on principles of fairness. Therefore, the game of morals is an extension of the game of life whose players have the ability to choose from additional strategy sets (Binmore, 2005). The moral rules are decided behind a Rawlsian veil of ignorance which ensures their fairness and these rules cannot contradict the natural laws of the game of life. In summary Binmore's games of life and morals include the following steps: two hypothetical rational agents interact in a state of nature in the game of life that has infinite iterations; the two agents bargain about the moral rules of their ideal society without knowing their identities in that society; the rules they will agree on are fair. Using these steps, Binmore introduces evolutionary game theory in a Rawlsian framework that he argues ensures a fair social contract.

The original position is used as a mechanism to select a fair equilibrium. It is a useful tool to analyse bargaining but only because it is the result of combined biological and social evolution (Binmore, 1998). Therefore, although Binmore accepts the original position as the centrepiece of contractarian theory, he does not see it from a Kantian perspective in contrast to Rawls who also bases his theory on a similar concept and Harsanyi's utilitarian approach. For Rawls and Harsanyi individuals at the original position are behind a veil of ignorance, although their respective assumptions about the thickness of the veil are different. Since they do not know their position in the future society, they will agree on a fair social contract (Rawls, 2005; Binmore, 1989). However, for Binmore this means that they import necessary metaphysical claims in the theory. For him the perceived fairness of original position mechanism lies on our common evolutionary history. Moreover, in Binmore’s original position the veil of ignorance is thinner. Players do not know their identities but they do know their preferences. In this sense, Binmore uses a contractarian mechanism to derive rules of fairness that he has naturalised and justified as a mechanism of biological evolution.

The starting point for Binmore’s theory is a state of nature that is different from

Hobbes's war of all against all or Locke's more benign original state. We cannot and should not choose an ideal starting point when constructing a social contract and we have to be realistic when considering the current status: "Like it or not, we are what history has made of us" (Binmore 2005: 25). Binmore's original position is the current state of society which makes his theory less controversial and more easily applicable to contemporary societies since it bypasses the criticism linked to the idea of an original position in contractarian theories.

Game Theory and the Social Contract (1998) describes a fair social contract based on a theory of rational bargaining. Thus, game theory and bargaining broadly perceived are central for Binmore's argument. Defining "[a] game is any situation in which people or animals interact." (Binmore, 1998: 3) shows how game theory can be applied to almost every aspect of social interaction irrespectively of assumptions of rationality or evolutionary stability. Again, this broad definition of game theory offers a more realistic and plausible account of human interactions as opposed to models of homo economicus or strictly evolutionary theory.

Game theory and bargaining are based on the Nash programme: the Nash bargaining problem and the Nash solution. These can be illustrated with the use of the divide the dollar game which was discussed earlier. In this game, the disagreement point also defines the possible bargaining outcomes. Put differently, if the utilities of two players A and B are respectively $U_A (X_A, Y_A)$ and $U_B (X_B, Y_B)$ at the disagreement point, then the bargaining should yield an outcome where their utilities will be $U_{A*}(X_{A*}, Y_{A*})$ and $U_B (X_{B*}, Y_{B*})$, where $U_A > U_{A*}$, $U_B > U_{B*}$. When player A's demand is the best response to player B's, we have a Nash equilibrium point. Furthermore, rational agents will always reach an agreement that is Pareto efficient. A Pareto optimal agreement means that it is impossible to reach a new agreement without anyone's utility being reduced. In a stable contract the agreement point is a Nash equilibrium that is also Pareto efficient.

Another essential feature of the Nash game is that both players make their demands simultaneously. This means that they will have to make sure that their demands lie in the bargaining set defined by U_{A*} and U_{b*}. Failure to do so will result in the game breaking down and each player having to compromise with the status-quo or look for outside options, that is utility points that are outside the bargaining set and therefore suboptimal. In this light, commitment is central to the Nash bargaining programme. Players are assumed to select the best response and commit to it because it

is utility maximising. However, simultaneous demands imply that it is not possible to make counter offers. Binmore argues that an alternating offers game variation of the Nash programme is more realistic where players make their offers consecutively as times passes until an agreement is reached. The agreement point is again Nash equilibrium. Obviously, bargaining with consecutive offers occurs over time; after each offer, there is a time lapse before it is accepted or rejected. Therefore, in the Rubinstein model that has just been described, time and bargaining skills have a role. Patience, or in other words a small time discount factor, and efficient bargaining skills will yield better long term results for a player. The weighted Nash equilibrium is thus identical to the equilibrium of the alternating offers game. The Nash programmes deals with interactions between fully rational agents that reach a mutually beneficial equilibrium. However, in repeated games the most significant problem is the selection of one optimal Nash equilibrium; in the literature this is solved with the folk theorem (Binmore, 1998).

The folk theorem shows that cooperation is sustainable in repeated games; “all interesting outcomes in the cooperative payoff region of a one-shot game are also available as equilibria in an indefinitely repeated version of the game” (Binmore, 1998: 265). Self-interested agents will cooperate in order to maximise their utility. Therefore, in repeated games reciprocity can substitute a central enforcing authority and sustain a stable social contract through individuals' rationality. Hence, reciprocal altruism can be sustained in large groups in the long term. In this context, reciprocal behaviour can spread to the whole population when it has been established as an equilibrium in a group within the population. When the social contract is seen as a super-game, its equilibrium is affected by the sub-game equilibria by means of transmission of efficient strategies. A meme is any behaviour that is imitated and spreads from a group to the population. It is introduced by Binmore in order to explain how cooperation can become the social norm. Memes replicate themselves just like genes. In *The Selfish Gene* (2006), Richard Dawkins argued that evolution of species is based on the survival and evolution of the genes. The analogy to social science is that social contracts evolve based on replicating cultural memes, that is types of behaviours that are more efficient.

Summing up, for Binmore biological evolutionary forces set the framework for cultural evolution. Humans are biologically programmed to live in small groups. This is why moral behaviour is a rational equilibrium when we are faced with small scale coordination problems such as when we wait for an old lady to pass first from a short passage (Binmore, 2005). Game theory can show that the same applies on a large scale.

Biological and cultural evolution are related to the social conventions we follow and the social contracts we use. Biological evolution made it possible for humans to create societies and social contracts that manage social interaction. The rules of these contracts in primitive societies were very closely linked to the rules humans adhered to because of their biological needs. In time, more complex social contracts evolved to deal with the complexities of bigger societies. The two kinds of evolution are interdependent. Any kind of cultural activity has to depend on our biological capabilities and limitations. The paths and limits of cultural evolution are predefined by biological evolution. The relationship between biological and cultural evolution is examined through the games of life and morals.

The game of life is determined by our human nature. We cannot escape its rules any more that we can change our genes. The game of morals is used to help us select an equilibrium, which is also an equilibrium in the game of life. In other words, the bargaining set of the game of morals is a subset of the bargaining set of the game of life. Our genes determine the bargaining set from which we choose a social contract. Therefore in a sense, they set the bargaining limits. Within those limits we select a social contract that does not have to be fair. However, a fair social contract will always be in the bargaining set defined by the game of life. Within these bargaining sets, we use the bargaining procedure to move from one Nash equilibrium point to the next. Every consecutive point is Pareto superior to the previous one. Therefore, we move from the status-quo to a Nash equilibrium – Pareto optimal point. The social contract at that point should be a fair social contract.

In other words, Binmore's core argument is that we are able to solve every day small scale coordination problems using norms of fairness, because these norms are part of our genetic heritage. If it were not, every simple interaction would result in conflict or at least in costly bargaining. The issue here is how to move from one-to-one coordination to large group coordination. Evolution has not equipped the human race with a mechanism for this. And this is where rational choice theory can be most useful to moral philosophy. If it can be shown that it is rational to cooperate on a large scale the same way as we cooperate in small groups, then moral norms can be generalised. In conclusion, cultural evolution leads to a fair social contract, which is supported both by the rationality of the contractors and the evolutionary stability of the biological evolution. Binmore bases his argument on the theories of cultural evolution introduced by Sugden and Skyrms and expands by introducing a contractarian framework as it was

used by Rawls. However, his argument also incorporates the weak points of the theories he uses in addition to the inherent complexity of his work. In the following section the main points of criticism for the inclusion of evolutionary concepts in moral philosophy will be examined, attempting to show the limitations of the approach and set the scene for the next chapter that will discuss a possible alternative of a convention based evolutionary contractarianism.

3.5 Criticism

The three theories presented above can be examined as a whole since they share many common concepts. The main idea is that social contract theory can be grounded on evolutionary game theory with naturalistic elements. Moreover, there is a common underlying principle that conventions and norms evolve as a result of social interaction. Sugden, Skyrms and Binmore represent a distinct field of synthesis of moral philosophy and evolutionary theory. In this light, criticisms of each author often apply to a great extent to the work of the others as well. The following review of the criticism will be organised per author; however, it will become clear that most of the problems are common to all three theories.

Sugden has been criticised for not making explicit references to any form of government (Hamlin, 1987) as there is no external enforcer in his theory. But this line of criticism seems to be missing the point of *The Economics of Rights, Cooperation and Welfare* (Sugden, 2004) and to a certain extent of the literature on evolutionary social contract theory. Sugden's intention was to construct a theory with no need for an external enforcer, or at least a theory where government is a secondary factor. Rational conventions for Sugden sustain themselves, simply because they evolve and are followed by rational individuals. Moreover, Sugden claims that we follow conventions because we are expected to do so. We behave morally because we care about what others think of us which is what sustains conventions and turns them into norms. From a rational choice theory perspective, the argument can be problematic. There is no justification as to why we would care about what others would think if we broke a convention. The explanation given lies in psychology and its justification "is about the psychology of morals" (Sugden, 2004: 153). However, this is an apparent weakness. Sugden was the first to apply ideas of biological evolution in social theory. This has drawn criticism of the plausibility of his ground-breaking theory and the possibility of using evolutionary accounts in a normative theory. The main problem, common to all

theories examined here, is the conflict between rational agents who are responsible for their behaviour and individuals who follow conventional rules that are explained through evolutionary game theory. Sugden's theory focuses on groups following a convention and there is no explanation about how conventions of different groups converge to one social norm. The same criticism applies to Skyrms's work as well. The premises of their theories are very similar in that their focus is individual behaviour within groups and how social equilibria affect it.

Simpson (2004) provides the most comprehensive criticism of *The Stag Hunt and the Evolution of the Social Structure* (Skyrms, 2004). He argues that Skyrms does not provide a model of how cooperation was initiated that is convincing enough. For Skyrms (2004), a group of people decided to hunt stag, and once it became obvious stag hunting yields a higher pay-off, it became the norm. However, Skyrms does not examine how the group came to hunting stag in the first place; the given social group may have turned to hare hunting just as easily. The stag hunt is probably a better example than the prisoner's dilemma when we think about cooperation in society. However, in a realistic setting it is not always straightforward to decide who actually participated in the hunt. In other words, Skyrms's model does not deal with free-riding and collective action issues effectively.

Similarly, although in evolutionary models such as Skyrms's, stag hunting is contagious, the same does not have to apply in human societies where individuals have the capacity to deliberate and reflect before deciding on a strategy. Both stag hunt and hare hunt are Nash equilibria; the optimal outcome is to hunt hare if everybody hunts hare and hunt stag if everybody else hunts stag. However in reality, where individuals are at least reasonable, it is best to hunt hare when everybody else hunts stag, or put differently "the chances of a successful deer hunt go up sharply with the number of hunters" (Skyrms, 2004: 1). A hare hunter can be a recipient of the stag distribution and thus maximise his utility. This line of criticism is a significant drawback given that Skyrms's theory aspires to describe social life in general and not just small group behaviour. The models in *The Stag Hunt and Evolution of the Social Contract* (2004) are based on computer simulations of the evolution of bacteria. These are then used to derive conclusions about the evolution of the social structure in human societies, assuming that human individual behaviour is similar to the evolutionary behaviour of mindless agents. Even if this holds true in most cases, once more there is no discussion about the side effects of human rationality and egotistic behaviour. In sum, Skyrms's

evolutionary models fail to provide convincing arguments about how human individuality can be incorporated in evolutionary accounts of behaviour.

The Stag Hunt and the Evolution of the Social Structure (Skyrms, 2004) offers a robust explanation of how social structure evolves to reach a stable social contract. Moreover, it analyses how behaviour observed locally can influence the eventual equilibrium and how the strategies adopted by one individual can have an effect on the group. Skyrms's analysis, unlike Sugden's, relies heavily on laboratory experiments as opposed to Sugden's however it is clear that they lie in the same intellectual tradition. Therefore, they are open to similar criticism whose main point has to do with the assumption of rationality in evolutionary accounts of the social contract. In both theories agents of limited rationality are implied or assumed as opposed to the more traditional assumption of rationality that is adopted by Gauthier. In that respect, there has to be an examination of how a more realistic account of individual rationality can fit in the evolutionary model. Binmore has advanced the evolutionary account of the social contract tradition in that respect, but his theory has also been heavily criticised. This criticism will be reviewed in the following paragraphs, where the main lines of criticism will be listed.

The theory of *Game Theory and the Social Contract* (1998) has not been extensively reviewed. Its most analytical criticism comes from Sugden (2001). The first criticism is of the organisation of the content in *Game Theory and the Social Contract* (1998). There are many repetitions and it is especially difficult to follow Binmore’s argument. As Sugden notes “[t]he problem for the conscientious reader is to find a canonical statement of Binmore's core argument.” (Sugden, 2001: F215). Binmore addresses a complicated issue and his writing makes it even more complicated. His argument lacks “clarity, brevity, sensitivity” (Sugden, 2001: F214) and often it becomes incomprehensible.

Furthermore, “Binmore practices social science as a branch of mathematics” (Sugden, 2001: F219) as he argues as if there is one absolute truth in political theory, in the same way as there is one correct solution for each mathematical problem. However, he does not, and possibly cannot, use mathematical formulas to establish a theory of fairness that is generally accepted, without using solid philosophical arguments at the same time. Binmore is a naturalist. He therefore has to naturalise rational choice theory but he fails to do so convincingly. His arguments for showing that the outcomes of evolution and rational deliberation of individuals will be the same are at best confusing.

The introduction of cultural memes is heavily criticised by Sugden (2001) as well as Riley (2006), as he does not explain how memes replicate and how much room they leave for individual rationality. Probably, memes work similarly to social conventions but this is not discussed by Binmore. But, just like conventions, memes are products of cultural evolution, which are selected by societies of rational individuals.

As opposed to Gauthier who in *Morals by Agreement* goes to great lengths in discussing how he perceives rationality, before laying out his theory, Binmore does not explain his views on rationality and its connection with evolutionary theory in an unambiguous way. Furthermore, in Binmore's analysis it is not made clear how much each person is free to act at will and how much otherwise rational agents are constrained by memes and biology. While the theory uses rational choice premises, there is a strong naturalistic and deterministic aspect to it. We are either guided by our rational utility maximisation, or we are slaves of our genes and the memes that guide our social behaviour. It seems that the implication is that the two can be reconciled. But they are not in a clear and effective way, which adds to the confusion.

The folk theorem, which has a central role in Binmore’s game theoretical argument, requires people to have a long term horizon for their behaviour. It is only when individuals have a small future discount factor that there will be Nash equilibria in repeated games, but there is no explicit justification to be found in Binmore's argument as to why rational agents would have a small future discount factor. An implied response could be that, humans care about the survival of their genes. Therefore they care about their offspring surviving in a similar relationship they care about family members. But then this poses questions about the rational choice theory premises of the argument. Binmore argues that we solve everyday coordination problems using fairness norms. Yet, he does not provide any evidence for this. Apparently we do so in some cases, but there is no solid empirical evidence provided to support this claim (Gintis, 2000). Use of anthropological evidence for fairness norms from primitive societies, although useful cannot always be transferred to modern, more populous societies and in addition it is not always possible to show that specific topical norms are not the result of specific topical conditions. There are as many examples showing that we do not use fairness norms in minor coordination problems of everyday life. Strangers have silly arguments and rudeness in public spaces is a common phenomenon. Binmore uses examples that are anecdotal and intuitive to promote very bold arguments about the relationship of biology and rationality.

Similarly, Binmore dismisses Kantian normative philosophy because it is based on intuition. He accuses Harsanyi and Rawls of basing their theories on metaphysical claims as they do not propose a mechanism that will enforce the agreement at the original position. However, Binmore's naturalistic approach to morality is not any less contradicting. Our genes are important and determine our behaviour, but it cannot be shown to what extent this is true. In addition, even more paradoxically Binmore uses the original position and a veil of ignorance to derive principles of fairness. Therefore, it seems that Binmore criticises the Kantian tradition only to borrow some of its aspects that he sees as useful, without further explanation.

Moreover, in *Game Theory and the Social Contract* (1998) there are many appeals to anthropological evidence, specifically from primitive societies. However, the use of actual anthropological evidence is limited despite many references to primitive societies' fairness norms. According to Binmore, primitive, small scale societies can allow us to see how fairness is perceived when social interaction is basic, before the development of complex production and property relationships. But anthropology can be useful in more ways. There is a wealth of anthropological and historical data that allows us to have a very good understanding of past societies and their social contracts. In fact there is no reason why we cannot use this instead of experiments on human behaviour (Kitcher, 2011). This would allow us to see the evolution of social contracts in the very long term, in realistic historical circumstances, as opposed to in the laboratory environment used in social science experiments.

Moreover, laboratory experiments examining human behaviour in a social context cannot be trusted. Individuals in the laboratory will not behave naturally since they know they are being watched and we cannot be certain they will not alter their behaviour. Natural scientists control for the effect light has on particles, so as to have unbiased results. For social scientists it is not as straightforward to control external to the experiment factors as human are immensely more complicated; Heisenberg’s uncertainty principle (Hilgevoord & Uffink, 2012) holds in the social science laboratory as well. Unlike natural sciences, social science has not yet come up with a way to control for behaviour changes. Individuals being observed in the laboratory will not necessarily behave the same way as in society. Laboratory results do not provide a long enough time frame or a big enough population.

A typical experiment may take at most several days. However, people change over the course of their lives and their behaviour depends on their circumstances.

Experiments cannot test for behaviour changes over time and under different personal circumstances. Moreover, laboratory experiments are difficult to test how behaviour changes when group size increases since collective action problems have a central place in social science. Evolutionary game theory is not as experiment-based as other branches of game theory. If we could use historical and anthropological evidence as a laboratory for cultural evolution, we would be able to have some rough idea on how social evolution works in the long term and in the real world.

Part of the criticism for the use of game theory in moral philosophy is similar to the criticism of game theory in general. The mathematical field of game theory is still developing and as of now there is no single game to describe all types of social interaction. Sugden proposes three categories of games that include most social activity. Skyrms (2004) suggests that the stag hunt is most appropriate without dismissing the prisoners' dilemma. Binmore finally, examines a range of games. This is not necessarily a criticism. However, a simplified all-encompassing game theoretical model would be much easier to accept and work with. If more than one game is needed to describe human behaviour then it follows than more than one theory is required to cover all aspects of social interaction, which in turn can mean that the theories can be in tension. Game theory has not developed sufficient tools so that a single game will be able to represent all versions of the social contract. This is probably the greatest shortcoming of using game theory in moral philosophy. Although it is not necessarily true that the sole aim of game theory and its use in the humanities is to provide a unified model of social interactions, to a large extent a single commonly accepted game theoretical paradigm would make game theory a more prominent method of discussing social interactions. In addition, despite its restrictions and limitations, a game theoretical model can be used across all social science sub-fields and seems more promising in unifying social science than any alternative (Varoufakis, 2008; Gintis, 2009). Game theory tools are not advanced enough to describe complicated social interactions. The majority of game theoretical models discuss interactions between two agents for simplicity. However, game theory does not provide us with sufficient tools to discuss collective action and coalition building.

The evolutionary accounts of social interaction offered by Sugden, Skyrms and Binmore – despite the innovation and the advances they can offer to contractarian theories such as Gauthier's – are problematic in respect to the reconciliation of collective and individual rationality. Sugden and Skyrms allow for low rationality

agents, which is incompatible with Gauthier's theory. Binmore attempts a more ambitious synthesis with mixed results, which is however an extremely significant advance in the field and can be used as a sign post to the right direction.

3.6 Conclusion

Evolutionary contractarianism can be viewed as a new field within political and moral philosophy. So far it has not solved any problems of social organisation. However the combination of mathematics, economics and political philosophy is promising. Valuable insights can be gained by the use of other sciences concepts, such as psychology, anthropology and history. All three scholars discussed here aim at creating a theory of fairness based on interdisciplinary premises. They employ different methods, but they all accept similar premises. In Binmore's words "the science of fairness has hardly been born" (Binmore, 2005: 197). Although Binmore apparently refers to axiomatic analysis of social interaction, the science of fairness can also be taken as a method to derive morality out of rationality. In this respect the science of fairness was born with *Morals by Agreement.*

An effective synthesis of traditional contractarianism with evolutionary game theory can provide a more plausible account of social interaction. Binmore has made the boldest steps towards this direction, but not necessarily the most the successful one. His theory is not easy to follow and has no clear normative conclusion despite his claims. There are valuable innovations in rational contractarianism though that can be used in the future. Therefore, it seems that, if Gauthier's theory is enriched with the work of Sugden, Skyrms and Binmore it will be able to give a realistic explanation of social contracts. Our better understanding of this field can then lead to a normative and explanatory science of rational choice.

In the field of evolutionary moral philosophy, we can say that all games are infinitely repeated. Every single interaction can be seen as a sub-game of a greater game, which is played over the generations and in groups within the population. In this context, similarly to biological evolution that does not stop, cultural evolution can be seen as an infinitely iterated process. A Pareto efficient, Nash equilibrium does not necessarily have to be a permanent equilibrium. Put differently, cooperative equilibria can be seen as evolutionary steps towards a superior equilibrium. Equilibria may be destabilised by factors outside the game and in these cases, there is a need for selecting a new agreement point.

Within this evolutionary context describing social structures and group behaviour, individual rationality becomes paramount in selecting equilibria and the direction of social evolution. The next chapters will focus on the discrepancies between individual rationality and collective action and the means with they can be used in conjunction. To that end, heterodox economics represented by authors such as Young (2001), Gintis (2009) and Bowles (2011) offer a more realistic account of bounded rationality that can be used in evolutionary theory without loss of individuality. A conventional account of boundedly rational individuals within a framework of evolutionary dynamics can incorporate concepts from Gauthier's moral contractarianism – such as constrained maximisation – with evolutionary game theory as it has been developed by Sugden, Skyrms and Binmore.

4. Rationality and Evolutionary Theory

The aim of this chapter is to show how constrained maximisation in Gauthier's theory can be reformulated using concepts from evolutionary theory. Gauthier, in *Morals by Agreement* (1986), attempted to create a theory of moral behaviour based on rational premises. The extent to which he has been successful is debatable. Many of his critics, especially those coming from Economics, accuse him of polluting rational choice theory with moral principles in order to construct a theory that is only ostensibly a theory of rational morality. His theory could have been more coherent and consistent in terms of its rational choice premises, if it did not use any pseudo-rational elements that are in effect moral. One of the concepts of the theory that has been criticised as importing moral consideration into rational deliberation is constrained maximisation.

Constrained maximisation constitutes the cornerstone of the Gauthier programme. By introducing it he attempts to show that moral behaviour can be grounded on rationality. The rest of the theory in *Morals by Agreement* is based on the plausibility and effectiveness of constrained maximisation. Gauthier's argument is that mutual constrained maximisation is rational in that it maximises individual utility while maximising social welfare at the same time. An interaction between two constrained maximisers (CMs) is rational. However, the trouble with it is that it is not convincing to argue that it is always rational for a self-interested individual to constrain her maximisation. The rationality of the behaviour of a CM is problematic on two counts that are evident in Gauthier's definition of a constrained maximiser: "a person who is disposed to comply with mutually advantageous moral constraints, provided he expects similar compliance from others" (Gauthier 1986: 15). The first problem has to do with the plausibility of disposition translucency which is an essential requirement for constrained maximisation being rational. The second problem stems from the fact that constrained maximisation requires agents to comply with constraints that can only be characterised as moral, which attracts criticism from proponents of a theory of morals exclusively based on rational premises.

In the following section, constrained maximisation will be discussed in light of evolutionary dynamics and there will be an attempt to show how an evolutionary theory of conventional behaviour can smooth the moral edges in the theory of *Morals by Agreement.*

4.1 Why evolutionary theory

Evolutionary theory can offer significant insights and strengthen moral contractarianism by inserting a holistic approach into the traditionally individualistic approach of the paradigm. Specifically, in *Morals by Agreement*, at least implicitly, there is a strong relationship between individual rationality and social structures. This relationship becomes clearer when discussing the rationality of constrained maximisation. The rationality of constrained maximisation depends on agents employing a joint strategy and acting on their dispositions; merely forming those dispositions is not enough for constrained maximisation to be a rational strategy.

4.1.1 Evolution

The requirement for employing a joint strategy exhibits how the rationality of constrained maximisation is conditional upon the behaviour of others. It is essential that a CM "makes reasonably certain that she is among like-disposed persons before she actually constrains her direct pursuit of maximum utility" (Gauthier, 1986: 169). The rationality of constrained maximisation is affected by an agent's environment, or in other words an agent's neighbourhood (Skyrms 2004). Moreover, a CM has to form a disposition to act and actually act upon it. Perhaps more realistically, this means that rational individuals will not form dispositions unless they are certain that it will be rational for them to act accordingly. However, the real problem with Gauthier's emphasis on dispositions is related to their supposed translucency. Constrained maximisation is based on the assumption that agents' dispositions will be visible and that rational agents will only constrain their maximisation when interacting with those similarly disposed. Although not explicitly acknowledged in *Morals by Agreement*, constrained maximisation implies repeated interactions and requires the occurrence of a network or a social structure of CMs. Repeated interactions and social group dynamics are better studied by evolutionary theory.

Embedded in the concept of constrained maximisation is the idea of a small future discount factor. CMs must have a small discount factor so as that their constraining maximisation behaviour can be defensible from a traditional rational choice perspective; rational agency as described by economic theory and game theory cannot allow for a rational agent to limit her maximisation and, in this respect, Gauthier's account of constrained maximisation is not rational. Together with joint strategy, another key component of constrained maximisation, they can be used in the

context of repeated games. A rational agent will cooperate with someone who is also disposed to cooperate in order to share a cooperative surplus. The dividend of this surplus has to be greater than the outcome of non-cooperative behaviour. When two agents expect to interact again in the future, cooperation becomes more likely and future maximisation is a rational choice. In conclusion, the traditional rational choice theory paradigm calls for immediate maximisation while it does not take into account the dynamics of repeated interactions. If we allow for repeated non-random interactions, the rationality of constrained maximisation in a prisoner's dilemma game is defensible.

An evolutionary analysis of individual behaviour can replace Gauthier's moral constraints with a more plausible account of why rational agents become constrained maximisers; individuals act within certain social constraints and are influenced by them, similarly to agents restricted by evolutionary forces. Rational agents' behaviour can be shown to be bounded by social conventions which are the result of repeated interactions. The dynamics of the development and sustainability of social conventions are best explained with the use of evolutionary game theory. The analysis of conventions and social structures is central in Sugden's work (Sugden, 2004), and very important for the arguments made by Skyrms (Skyrms, 2004) and Binmore (Binmore, 1998), which were discussed in Chapter Three. Moreover, it needs to be clarified that the evolutionary convention analysis presented here serves as complementary and reinforcing of Gauthier's moral contractarianism and not as its replacement, as it might seem at first sight. Despite the use of evolutionary theory, individual rationality remains the main assumption of the analysis and the basis for the argument.

The social contract and its bargaining process can be seen in an evolutionary context of repeated interactions that generate social dynamics. In *Game Theory and the Social Contract* (1998) Binmore explains how pre-play bargaining is in essence part of the game and therefore the bargaining process and the game itself can be analysed together as a single game. In addition, the secret handshake mechanism analysed in *The Stag Hunt and the Evolution of the Social Contract* (Skyrms, 2004) exhibits how pre-play communication can have the form of actions as well as claims during bargaining. Therefore, in the evolutionary account, bargaining in the context of contractarianism occurs as agents' actions. For instance, the disposition for cooperation is known from an agent's history and her verbal commitment is irrelevant once her past interactions are known. In conclusion, evolutionary game theory can reinforce contractarian arguments such as Gauthier's and justify moral behaviour with explicitly

rational premises that derive from the dynamics of social structures.

4.1.2 Evolutionary game theory

The problem with incorporating evolutionary concepts in a rational choice framework is that evolutionary game theory assumes agents of limited rationality or organisms with no capacity for reason. Evolutionary game theory studies populations and not individuals and this makes it difficult to fit into the traditional rational choice paradigm. However, assuming that an action is considered rational when it leads to maximisation, individuals maximise – and thus are assumed to be rational – when they adopt a maximising strategy, irrespective of whether their behaviour is the result of rational deliberation or imitation. In this sense therefore, they are being rational by learning through a trial and error process and imitating more successful strategies. Moreover, the traditional expected utility model is not essential for the account of rational agents used here. Agents look back in past interactions when deciding their future behaviour, instead of trying to adopt maximising strategies for the future based on complete information and perfect rationality.

Agents are neither hyper-rational nor mindless cyphers and this does not have to mean rational choice theory or evolutionary game theory (EGT) cannot be used to examine their behaviour; “they gather information and the act fairly sensible on the basis of their information most of the time” (Young, 2001). Furthermore, rational agents reflect on the outcomes of different evolutionary processes and select the best, based on rational calculations. Finally, rational agents can select to participate in an evolutionary process, when this maximises their utility; they realise that participating in this process will eventually maximise their pay-off. Alternatively, they can abstain from social interactions or not follow what is deemed rational behaviour by most; for instance defecting in a society of cooperators. In this sense, it is possible to use the evolutionary paradigm without abolishing rationality.

An additional feature of the rational agents described here is that individuals in the first generation (or during the first iteration of a repeated interaction) analyse their strategy and try to maximise their utility by trying new strategies until a stable equilibrium is reached. Individuals who are part of a population described by evolutionary game theory remain conditionally rational. If they realise that the evolutionary strategy in which they participate is not stable or maximising, they have the rational capability to start interacting with new agents using new strategies.

Initiating new interactions means of course that the past behaviour of individuals will not be known for a time; however, the absence of information at the beginning of a series of interactions is mutual and therefore does create a disadvantage for any one party. The assumption of repeated interactions ensures that complete, or near complete information is a more plausible assertion. Individuals can have knowledge of their group's history and hence know who has cooperated in the past. In repeated interactions, especially within small groups, individuals get to know each other's behaviour and form expectations accordingly. Cooperators will choose to interact with other cooperators or not participate in interactions. Similarly, groups develop a reputation of cooperative or non-cooperative equilibria; this adds new force to Gauthier's above quoted stipulation that a CM will know that he is among other CMs.

Importing rationality in an evolutionary model is at best controversial. Sugden (Sugden, 2001) has been very critical of the approach used by both Binmore and Young who attempted to combine classical with evolutionary game theory. With these exemptions, EGT has been used formally to examine economic and social behaviour without dealing with foundational issues of its application in social behaviour. For Sugden EGT remains as theoretical as classical game theory, being unable to take into account the empirical reality of human behaviour. Sugden's central point is that the assumption of rationality has not been dropped in the evolutionary account but instead is being smuggled in. Thus, the use of EGT is being based on mistaken presuppositions. Sugden does not criticise the use of EGT but the way it is being used; evolution has been adjusted to economic theory and not vice versa.

Sugden proposes an account of evolutionary history, which is much closer to biology and therefore more accurate from an evolutionary standpoint, to help us make sense of social behaviour; "a crucial feature of evolutionary explanation [is] that historical contingencies are important" (Sugden, 2001: 10). Cultural EGT has to be empirical as well as theoretical and in order to achieve this; it has to take into consideration actual historical facts just like biology takes into consideration empirical evidence. "A genuinely evolutionary approach to economic explanation has an enormous amount to offer" (Sugden, 2001: 16), should it be used sensibly with reference to its premises and not just its formal models.

4.1.3 Conclusion

The paragraphs above attempted to explain why EGT can be used as a tool for

Gauthier's constrained maximisation and looked at some of the problems with incorporating evolutionary ideas in moral contractarianism. The following sections will expand on this analysis in order to show how evolutionary game theory and holistic explanations of human behaviour can be incorporated in a contractarian framework that is based on individual rationality. Rationality and evolutionary game theory have to be shown to be complementary (at least in some respects). The next section will discuss the concept of rationality and how relaxing some of the assumptions of economic rationality and accepting a more realistic account of bounded rationality can allow individual rationality to be examined in an evolutionary framework.

4.2 Rationality

The following paragraphs will examine an alternative concept of rationality to the traditional rationality of homo economicus (Heath 2011). The main conditions remain the same, however some of the unrealistic assumptions of neoclassical economics are relaxed. By doing this, one of the main problems of neoclassical economics – namely that it uses an unrealistic model to draw conclusions for the real world – can be bypassed. A rational individual is still one who has a consistent order of preferences over a set of alternatives and will always look to maximise her utility. The content of the preferences is not examined in assessing the rationality of one's behaviour. In other words, "[w]e do not know what [the rational man] wants...but we know his indifference curves are concave to the origin." (Hollis 1975: 75). Or in Gauthier's spirit, the behaviour of both the grasshopper and the ant are rational (Gauthier, 1986). They both maximise their utility, or else their enjoyment, by being heedless and prudent respectively. Provided that imprudence is the result of reflective thinking and its long term implications are being appreciated, then there is no reason to classify it as irrational. The present discussion will rely on this broad understanding of rationality which is more realistic and plausible than an economic rationality that assumes full information and infinite predictive power. More specifically, the following paragraphs will include a discussion of constrained maximisation as rational behaviour in a framework of repeated interactions.

In *Morals by Agreement* rational agents are expected to constrain their maximisation in order to "enjoy opportunities for cooperation which others [SMs] lack" (Gauthier, 1986: 15); put differently, rational agents have the opportunity participate in cooperative ventures that will maximise their utility through the cooperative surplus.

Thus, for Gauthier rational maximisers become constrained maximisers. Constrained maximisation is therefore a way of transferring maximisation into the future; a CM will choose to waive a smaller immediate pay-off in order to benefit from a larger future cooperative surplus. Given that the future pay-off for a CM is higher than the present one for a straightforward maximiser, constrained maximisation is still justifiable from a rational choice theory perspective. And in this sense constrained maximisation does not have to be considered a non-rational principle or a moral constraint; "[t]he constrained maximiser...reasons in a different way" (Gauthier 1986: 170). However, constrained maximisation is a weak principle from a rational choice perspective. The rational incentives for someone to comply with mutual constraints and accept a smaller immediate pay-off are not strong and if we adhere to the rigid assumption of economic rationality the behaviour of CMs can be seen as irrational.

In games with the structure of the prisoner's dilemma, cooperating when the other party defects yields a smaller pay-off than defection; it pays more to defect when the other party's disposition is not known. This is true when the game is not (infinitely) repeated. A small discount factor means that agents value future pay-offs highly and in any case not much less than present pay-offs. Therefore, given that the future pay-off is higher, they will choose to maximise at the end of a series of interactions in a repeated game. Another essential condition for the rationality of cooperation is that all games are infinitely repeated or more realistically, players perceive them as infinitely repeated. When A interacts with B, playing a game such as the prisoner's dilemma, and at the same time interacts with C, playing a game such as hawk-dove, his experience from one game is transferred to the other. Therefore A is a link between the two games or among all the games she plays at any given time and the strategies employed in these games. A's behaviour is affected by the outcome of each interaction and by the behaviour of B, C and so on. If A and B are strangers and they do not expect to meet again, their interaction history will affect their behaviour in their interaction. Hence, the game they play is affected by the games they had played before they met. Or in other words, the game played between A and B is a sub-game of all the games A and B play and their choices are affected by their history. In this context, A and B never interact in a one-off game, as they perceive every game they play as part of one single large game. All these sub-games are repeated and consequently, mutual cooperation in prisoners' dilemma type of games yields the highest pay-off. This can be best described using concepts of cultural evolution and evolutionary game theory.

The previous discussion focused on describing an alternative account of rationality that is generally in accordance with the one used in *Morals by Agreement* and whose emphasis lies on the repetitiveness of interactions. A type of constrained maximisation is rational when agents expect to interact again in the future. The following paragraphs will examine how this concept of rationality is affected by one's environment – mainly one's history of interactions.

A person who does not maximise her utility immediately, does not necessarily behave irrationally. She still behaves rationally provided that based on the knowledge and information she has, she believes her actions will maximise her utility. Her actions or set of strategies are confined by environmental parameters that cannot be influenced by her. Therefore, when evaluating an agent's behaviour, we must take these restrictions into account. But even then the predictability of her behaviour cannot be a given. The amount of information she possess and her perception of the strategy restrictions in place, makes accurate prediction of her behaviour unlikely. The homo economicus model of neoclassical economics is an approximation and simplification of human behaviour and as such it can only be used under certain conditions for the construction of theoretical models. Therefore, a realistic account of human behaviour would have to include a relaxation of the strict assumptions of economic rationality.

Predicting a rational individual's behaviour at a high probability is possible and when we know her interaction history successful prediction becomes even more likely. However, we have to bear in mind that rationality and maximisation depend primarily on the agent's perception of the available strategies and environmental limitations, which in turn depend on available information. What we can say is that we expect rational agents acting in the same environment to make the same decisions. There is no full information in this respect, but there is access to roughly equal information. Each rational agent has access to the same amount of information and therefore should adopt the same strategy in order to maximise her utility. Rational individuals with access to similar information should make similar decisions and adopt similar strategies. This does not assume that preferences will have to converge for a convention to be formed; rather it refers to a common understanding of the principles of social interactions. Individuals do not have to have similar preferences and it is likely that they will have conflicting interests. However, equal rationality should lead to a decision making process that leads to an optimal outcome. Put differently there is equal rationality within a social group or convention, as opposed to the neoclassical models which assume equal

rationality for all.

Furthermore, it is realistic to assert that each agent has roughly the same memory. It does not have to be full memory of every decision in the game's history, but “each individual remembers a general experience of the game but not how he fared against particular opponents” (Sugden 2004: 60), so that each player has a general understanding of how agents with whom he has interacted behave. And this understanding creates a disposition to act accordingly in the future. If the player has interacted mostly with cooperators (or he believes he has), then it is more likely he will expect his future interactions to be with cooperators.

The most important environmental parameter is other individuals with whom interactions are possible. Therefore, maximising behaviour depends on what the other player is expected to do. When the other player's strategies are not known or are only known at a probability, rational behaviour is not as straightforward a process. The longer the history of an interaction, the more likely it is that the other agent's strategies will be perceived as known. Again however, in a more realistic setting we will have to take into account that each agent participates in a number of interactions with different agents. These interactions affect her conception of others' strategies and thus her behaviour. In this context, a rational agent is expected to be able to form an opinion about her environment, that means a rational agent should be able to tell whether she is surrounded by constrained or straightforward maximisers and adapt her strategy accordingly.

Not all interactions bear the same weight in rational deliberation. Each interaction has a degree of salience; interactions with many repetitions or high pay-offs are more important than interactions that are short-lived or have low pay-offs respectively. When an agent's interaction collapses after many iterations, she will be more cautious in her new interactions. By doing so, she minimises possible loses from her interactions. Also, when the proportion and gravity of interactions that collapse is high, then a rational agent will be more cautious in her new interactions. For instance, when a short series of interactions collapses, it will not affect old series that have lasted longer. However, the collapse of many interactions, even if they have been short-lived, will have an effect on the agent's perception of her environment and thus her disposition. Repeated interactions within a population highlight the importance of optimal equilibria in relation to individual rationality. To conclude, a rational agent as described above is assumed to have access to roughly the same information as her

neighbours, a similar capacity to remember past interactions and evaluate her environment which leads to developing similar maximising strategies for a given environment. Thus, the definition of rational behaviour depends on an agent's environment and maximising strategies are bounded by environmental limitations.

The above conversation of rationality does not necessarily dismiss the premises of strategic interaction between rational utility maximisers. It is impossible to discuss strategic interaction in a game theoretical context without referring to the concept of a Nash equilibrium. Two interacting individuals behave rationally when the outcome for each one is the best she could have achieved. The Nash equilibrium point is therefore a deterministic outcome if we accept the basic conditions of rationality and others being equal. Put differently it is "as if the players' thought process has converged to an equilibrium, just as surely as a rock tumbling down a hill eventually reaches an equilibrium (a state of rest) at the foot of the hill" (Varoufakis 2008: 1259). Thus, once we accept that the interacting agents are rational within the same environment, the outcome of their interaction is known. The Nash equilibrium is a point from which neither player has a rational incentive to move away. In one-shot and finitely repeated prisoner's dilemma games, defection is the only Nash equilibrium. However, when the number of iterations is high and the future discount factor low, the sum of the outcomes of cooperative rounds can exceed the loss incurred at the last round. The benefit of cooperation times the number of iterations is possible to be greater than the cost from cooperating with a defector once. And this is more likely in a population of pairwise interactions and full information. On the other hand, when the number of iterations is finite and known, rational individuals will be able to predict that their opponent will defect at the last round. As a result, both will defect in the first round. It becomes clear therefore, that whether it is rational to cooperate depends on the game's specific parameters. If the cooperation benefits are high and the number of iterations high, it can be shown to be rational to cooperate. Hence, the rationality of cooperation even in games where the Nash equilibrium is defect, depends on the number of iterations and the related information that the players possess.

In the real world every interaction includes a cost of rational deliberation and information gathering, and our decisions are usually about long term behaviour strategies rather than one-off interactions with strangers. The decisions we have to make are not so often about eating a pear or an apple. Rather we have to decide whether planting an apple or a pear tree will be more profitable in the long run. A farmer who

has to make a rational decision about whether he should grow apples or pears will have to take into account a large amount of information such as: consumers' preferences, how these preferences may change in the future, whether the local climate favours apple or pear trees, the price of each fruit and how they are likely to evolve. It is of course impossible to have all the information required, but it is completely rational for one to try to gather as much information as possible before making a decision such as this which is a costly and time consuming process. Less important decisions can be based on quick deliberation or “just do what most others do”. For similarly important decisions it is also rational to use others experience if the information costs are deemed too high. If apple growers make a bigger profit, then it is rational for a farmer to grow apple trees even without taking into account the marginal cost of the apple market for instance. It all depends on the cost of information needed for a decision.

We can estimate how the apple market marginal cost will be in the future, but we cannot predict it. The estimation error probability and the level of the proposed investment may make it worthwhile to undertake the information cost. In this context, it is rational to just do what others do and plant apple trees when information is very expensive and the future very uncertain. Therefore, for everyday decisions where consequences are not important, it is obviously rational to follow the established norm. However, the same applies for important decisions when future is deemed very uncertain or the available information very scarce. In conclusion, rational agents have incentives to follow established norms of social behaviour as opposed to undertaking the costs of rational deliberation. This results in a sort of uniform social behaviour that may seem irrational, but it is nevertheless grounded on rational premises.

This section attempted to describe the premises of rationality in realistic settings. Individuals are reasonably rational; they want to maximise and try to make the best decisions with the information available. However, they cannot have full information or complete memory and predictive power. Just like human beings, rational actors adapt to their environment and try to make the best out of the situation. They are in this sense, boundedly rational. The point of the above analysis was to set the scene for discussing individual rationality in social groups. The following section will continue the examination of bounded rationality in relation to the dynamics that develop from social interactions.

4.3 Functionalism and conventional behaviour

Functionalism refers to the fact that an action may have implied and unintended functions for a given society (Martin & McIntyre, 1994). The following section is an examination of the extent to which functional analysis of social structures can be incorporated in the rational choice theory framework presented earlier. Functional explanation will serve as a stepping stone in reconciling the methodological individualism of game theory with the holism of evolutionary theory.

4.3.1 Functional analysis

The typical example by which functionalism is usually discussed in the literature is the rain dance. Although it does not bring rain, it serves as a mechanism of reinforcing social cohesion (Martin and McIntyre, 1994). More crudely, Hollis (1994) uses a termite colony to discuss functionalism; each termite behaves in a certain way because otherwise the colony would collapse. The focus of attention in functional analysis of social behaviour is the society as a whole. Society needs the rain dancer and the termite colony needs the termites to keep behaving as they do in order to be sustainable. Functionalism is therefore a methodologically holistic approach to the examination of human behaviour.

The actions of the individuals in a society are explained through the purpose they serve for society and not in terms of individual benefit. Thus, functionalism and rationality are found at opposite sides of the spectrum in terms of individualistic and holistic approaches to social explanation. However, given the flexible understanding of rationality presented in the previous section, individual rationality can be reconciled with functionalism. The rain dancer behaves rationally whether he believes his dance will bring rain or he realises he merely behaves as society expects him to; by fulfilling his social role, he also maximises his utility. In other words, he behaves according to an established social convention of behaviour in order to retain his role in the social structure that maximises his share of the cooperative surplus. In that respect, functional analysis can be used in conjunction with rational behaviour and game theory to analyse interactions in a dynamic environment (Martin and McIntyre, 1994).

Functional behaviour supports conventions of behaviour, in a similar way to that in which the termites support the sustainability of their colony. Individuals do not need to act intentionally for the stability of a convention but their actions are essential for its stability. Social conventions arise from repeated pairwise interactions between rational

agents and over time acquire force similar to the force of the rain dance. Hence, conventions are dynamic and depend on agents' strategies that adapt to changes in their environment. According to this understanding, a convention does not always need to satisfy a specific goal in society; it may be the result of unintended actions, serving an implied function for social cohesion.

4.3.2 Rational conventions

As in real life, people's behaviour and beliefs change. In game theory agents' strategies are subject to change. A strategy shift can occur when the amount of information a player possesses changes or when a player's history changes as the game evolves. A player's history is being enriched continuously as the players interact in a game. And thus, her strategies and objectives change accordingly. In that respect rational behaviour is not set. In a dynamic context strategies and behaviour depend on or at least are affected by environmental parameters.

In the classic example of Robinson Crusoe, his strategies change when he meets man Friday. From a game theoretical perspective, when man Friday appears the interaction analysis becomes strategic. Crusoe's strategies change as his environment changes (Hollis 1994). As the number and the disposition of the players in the game change, rational players' strategies have to adapt to the changing environment. An agent therefore has to shift her behaviour depending on whom she interacts with, the frequency, and the salience of her interactions. In a group of non-cooperators, the rational response would be non-cooperation whereas in a group of cooperators, cooperation will maximise a player's utility. Therefore, a rational response is always dependent on the expected behaviour of others. From these premises it is safe also to conclude that when environmental parameters remain stable, rational agents have no reason to alter their strategies.

Therefore, environmental parameters such as the number of interacting agents and their history are central to the present understanding of rationality. The strategies dictated by rationality depend on the game parameters. In this respect, rationality is still central to the argument and the basic premises of traditional game theory remain the same. Individuals want to maximise their utility and are able to learn and adjust to changes. Furthermore, a player's environment becomes a central concept in analysing rational behaviour. The interactions between Crusoe and man Friday are the first of their kind, given that they live on an isolated island. Their repeated interactions will give rise

to habits of behaviour. These habits will replace rational analysis of the possible choices at each decision node. Therefore, without abandoning rationality, Crusoe and Friday will stop behaving strategically in their set environment. Their behaviour will be determined by the rational decisions they made at the beginning of their interaction. They will be based on their interaction history and follow patterns of behaviour, without deliberating rationally. Social conventions, developed as a result of their repeated interactions in a stable environment, will replace their rational decision making process. Over the course of infinitely repeated interactions, they behave so as to serve a social function which is essential for the continuation of interactions. But this does not mean they lose their capacity to think rationally. Rational agents have the option to follow the – rationally established – convention or to defect. Conventions are dynamic in that they are affected by the environmental parameters. Therefore, when the number of players change, convention equilibria will have to shift.

If there is a new arrival on the island, he will also seek to maximise his utility and make the most of the situation. In essence however he has two options: follow the established convention or abstain and interact according to traditional rational choice theory. The maximising behaviour would be to take advantage of the achievements made by the two original inhabitants. In this context where the island's limited resources can be divided among a small group of individuals, if the new arrival decided not to follow the established convention but instead managed to interact following new rules dictated by his rationality, the outcome would be similar to the outcome of the interactions within the convention. Since the environment is set, rational interactions would lead to the same outcome – the establishment of similar conventions of behaviour. Maximising strategies are similar since the environmental parameters have not changed.

Newcomers on the island will follow the established conventions unless they find a new way to achieve the same maximising outcome. The conventions established by the interactions of a first generation of rational individuals will hold for the following generations. In this view, social conventions are the result of a Nash equilibrium point. Those who decide that it is in their best interest not to adhere to the existing rules will bear two different costs: First the cost of rationally analysing their best strategy. The more complicated the society in which they live, the higher this cost. Second, the cost of finding agents who would be willing to interact with them and pay the cost of being excluded from the established convention. It might seem that there are

two kinds of rationality. One pointing towards behaving conventionally and another dictating defection. However, what differs is the amount of information and the perception of the environment. Provided the environmental parameters are the same, both behaviours will lead to a Nash equilibrium. The non- behaviour however, will be more costly and time consuming.

Crusoe and Friday have not entered into a contract in the traditional way. Their common understanding of what should be done and their common aim (to maximise their utility) implicitly bind them. This agreement however does not have the form of mutual obligation but of egoistic maximisation, as one needs the other in order to survive. This informal contract then takes the form of the convention that is adhered to by new arrivals (or future generations), whose behaviour is described by their structural roles: they behave according to the role they occupy in society and by doing so, they maximise their utility without necessarily intending to do so. This however does not presuppose rational deliberation; agents behaving like that can fit in with the concept of rationality as presented above. By fulfilling her social role responsibilities she maximises her utility. Rational deliberation reinforces the behaviour that is promoted by society. Agents maximise in the long term and not in each decision node. What matters is that there is maximisation at the end of each interaction or set of interactions. In this context, homo sociologicus (Rescorla, 2011) is a rational agent taking into account environmental parameters and how they change.

Following Sugden (2004), whether we drive on the left or right is a matter of habit and a result of an evolutionary process. Rational deliberation shows that it is rational to conform with the established convention. And the convention has been established in the first place as the result of pairwise interactions between rational agents. This applies not only to dove-hawk type of games; in a prisoner's dilemma game cooperation is the only Pareto efficient solution. But the Nash equilibrium depends on what the other player is disposed to do. In a repeated prisoners' dilemma game between Crusoe and Man Friday, the Nash equilibrium will be cooperation. The same applies for small groups where free-riders are known. Therefore, in small groups where complete information and memory of the history of the game are known, cooperative conventions will be established. Even in cases when the opponent is not known, it pays to cooperate as it is more probable that the opponent will follow the convention as well. Rational individuals will choose to interact with agents within their group. When they interact with agents outside their group, they will be able to know the reputation of the group.

And if they come across a free-rider, his behaviour will become known in both groups, making the cost of free-riding even higher. Therefore, small groups do not necessarily restrict the number of interactions of the possible agents one can interact with.

The above shows how evolution of interactions and social structures lead to the creation of conventions. Therefore, we have an evolutionary story of how social conventions arise and are sustained. In this story individual rationality does not play a central role; agents follow the conventions as they have evolved through repeated interactions. Before that, social structures define strategy sets. In order to show that this evolutionary story can complement the account of rationality used by Gauthier, we will have to show how agents who accept conventions that are the result of evolutionary processes are also rational. Skyrms shows how the creation of a cooperative or non-cooperative structure depends on one's neighbours and their predispositions. In this framework, it is rational to be cooperative when you are surrounded by cooperators. A free-rider will be punished by social exclusion and his gains from defecting are much smaller than the costs of social exclusion. Non-cooperative interactions do not produce a cooperative surplus. Groups of non cooperators then do worse than groups of cooperators. The rational agent in a group of cooperators will choose to participate in cooperative interactions in order to increase her pay-off. Non-cooperative groups will thus mutate into cooperative groups. When we allow for rational deliberation, in Skyrms's story, cooperation is the rational strategy.

4.3.3 Conclusion

The above understanding of functional and conventional rationality emphasizes instrumental rationality without departing significantly from traditional definitions. Rational individuals still deliberate on how to maximise. However, they also follow the established conventions for a great proportion of their decisions. Since this maximises their utility and they have the rational capability to adhere to the predefined rules or not, they are still rational maximisers. Furthermore, rational agents have the capacity to reflect and apply backward induction on their strategies and their behaviour in relation to social conventions. They have the capacity to compare the outcome of conventional and non-conventional behaviour at the end of a set of interactions.

In a sense rational, utility maximising individuals have reasons to behave like the termites to an extent; do as they are expected by social conventions but without abandoning rational deliberation about whether their behaviour yields the best possible

outcome. Individual rational behaviour can have long term unintended implications and showing that individual rationality can coexist with explanations of collective behaviour is essential before discussing how rational behaviour can be examined within an evolutionary context. The following section will look at evolutionary game theory and its relationship with individual rationality.

4.4 Evolutionary game theory and constrained maximisation

Evolutionary game theory uses aspects of traditional game theory in combination with evolutionary theory from biology. Its use in social science and political philosophy is therefore not a straightforward step. However, the theory of evolution is not as far away from social theory as it might seem originally. Darwin formulated his concept of evolution being influenced, among others, by Malthus's theory on population (Gould 2008). We could say that in Darwin's mind the biological evolution of a species is related to the evolution of social life in human populations. Evolutionary game theory refers to the evolution of social life: changes in social behaviour, norms and the understanding of rationality. Therefore, with evolutionary game theory we use concepts of biological evolution to describe cultural evolution. Assuming that biological evolution can be applied as is to social life implies that we consider societies to behave like living organisations. And this is beyond the interests and the capabilities of political philosophy.

In addition, despite the fact that the selfishness of the gene (Dawkins, 2006) makes sense in biology, there is no need or method to use this as a model of analysis in social science. The most basic level of human behaviour analysis has to be the individual and there is no method of extending the rationale of the selfish gene to humans in a meaningful way. The selfish gene rationale entails that successful genes will replicate at the expense of others; maximising human behaviour may attract more individuals over time given individual rational deliberation. Human individuals decide on a strategy based on their personal circumstances and calculations and do not blindly accept the most successful strategy available. Despite the similarities between the selfish gene paradigm and the behaviour of human populations, their underlying principles differ. However, the great strength of evolutionary game theory is that it makes use of indisputable biological concepts that examine population dynamics to describe the dynamics of human societies and individual behaviour. And biology can, in some respects, be more appropriate than traditional social science to explain human

behaviour. Despite the fact that available strategies are genetically determined in biological organisms and animal societies (Skyrms in Danielson, 1998), in human interactions genetic limitations do not play a role (or if they do, it is beyond the scope of the present discussion). In human interactions there are still biological limitations, though not necessarily determination. These limitations are imposed by social rules, such as laws and norms. Apparently these are not as strict as genetic limits. However, they still have to be taken into account when discussing interactions in human populations. The following paragraphs will describe the main concepts of evolutionary game theory, before examining its relationship with human behaviour within social conventions.

4.4.1 Evolutionary game theory

Evolutionary game theory describes repeated games played in large populations. Therefore it is the most realistic part of game theory when it comes to describing behaviour within societies. There are three distinct, though not necessarily conflicting, mechanisms within the evolutionary game theory paradigm: (i) the more effective individuals (and strategies) are more likely to survive, (ii) agents learn by trial and error, and (iii) the best strategies are imitated (Axelrod 1986). These mechanisms work in combination or in parallel or subsequently with one another, rather than one replacing the other. For instance, the survival of an individual means that she follows an effective strategy, which she can abandon once she sees that another agent has adopted a more efficient strategy – making her a rational agent in the conventional sense.

More specifically, evolutionary game theory is defined by evolutionarily stable strategies and replicator dynamics. Evolutionarily stable strategies (ESS) describe which behaviours will survive evolutionary pressures and competition; in other words, which are the best strategies. Therefore, the approach focusing on the ESS is static as it does not address how these strategies evolve and change over time. However, ESS is useful in helping our understanding of how norms of behaviour can be stable and successful enough to attract new agents. Replicator dynamics in the contrary, focus on how strategies change over time and how they are affected by the general behaviour of the population as a whole (Weibull, 1995).

A behaviour strategy adopted by the majority of the population is more likely to evolve into being the norm for that population. A strategy that gives a higher pay off, is also more likely to be adopted by the population in time. Therefore, the replicator

dynamics approach to evolutionary game theory describes how strategies evolve taking into account factors such as population and fitness (Skyrms, 2004). Put differently, in an iterated prisoners' dilemma (PD) game the two players' strategies change as the game evolves. Provided that the players are part of a population, the number of agents adopting a strategy cause the dominant strategy to change. In essence, it is the population strategies that evolve. With that, the strategies in the two-player sub-game change as well. A strategy that takes over the population, or the majority of the population, as the fittest one is an evolutionarily stable strategy. And in this respect evolutionarily stable strategies and replicator dynamics are complementary in our understanding of evolutionary game theory. Evolutionarily stable strategies develop in populations where opponents' strategies are known. They are an equilibrium point that is in everybody's best interest to support. In different terminology, a social convention can be viewed as the outcome of ESS.

Replicator dynamics mean that successful strategies replicate themselves. A behaviour that yields higher pay-off to the agents involved will be imitated by neighbours. Hence, a population in an evolutionary game theoretical context evolves in that adopted strategies evolve to become dominant. For instance, if a population of cooperators in a PD game change to being defectors, then the dominant strategy has evolved from cooperation to defection. Evolutionary game theory is deterministic in the sense that if we know the original population disposition and the structure of the game, we can predict with accuracy what the equilibrium will be after a given number of interactions. For instance, it is a given that a population consisting of 70% defectors and 30% cooperators will converge to a non-cooperative equilibrium after a set amount of time. If environmental parameters do not change, this equilibrium will be stable.

In groups of rational agents, individuals have the capacity to reflect on the history of the game played and calculate how it could have developed differently. And then they are able to analyse how their pay-off would have evolved in the alternative scenario. If they see that despite being in a Nash equilibrium, the status-quo is not Pareto efficient, they can realise that they could have done better by abstaining from social interaction or changing their location so as to be included in a different group, where the current state of affairs is different. However, in stable equilibria this is not always feasible. It is not always practical for an individual to change his group. And therefore a rational individual is limited by the evolutionary equilibrium.

In large populations effective strategies take time to spread, when we have

assumed rational agents. A strategy in a two-player interaction will be proven to be maximising at the end of a series of interactions and not immediately. Therefore, their neighbours will only imitate it once they realise it yields higher pay-offs than their own strategy. This can only happen after a number of repetitions. In a population where defection is the dominant strategy in a PD type of game, if cooperation was established as an equilibrium in the interaction between two agents and they reached higher pay-offs than they would in defection interactions, their neighbours would pick up on that only after the convention became established. And this would only be the case in a game perceived to have infinite iterations. Thus, the utility at the end of the n-th interaction of cooperation would have to be higher than the respective defection interaction. In a repeated PD type of game, pay-off is higher for cooperators. In conclusion, cooperation will not establish itself as ESS when the first cooperating couple appears. It will take at least a generation (or iteration) to spread to neighbouring groups. Even in this case, whether cooperation becomes the dominant strategy in the population will depend on the replicator dynamics within that population.

As discussed in Chapter Three, Skyrms (2004) has shown dominant strategies within a population are influenced by location, communication and association. Evolutionary game theory can be helpful in explaining social interaction in human population as it takes into account the dynamics of social structures in repeated interactions. In this sense evolutionary game theory serves as a method to examine repeated interactions and how best strategies are affected by population and social structure dynamics. The description of evolutionary game theory presented above is typically used with the assumption of agents of low or no rationality which makes it unsuitable for use in human populations. However, on the condition that individuals are reasonably rational, that is they can learn and imitate successful strategies while deliberating about their behaviour, evolutionary game theory can be used to describe human interactions. Thus, it is possible to use EGT in conjunction with assumptions of individual rationality.

4.4.2 Rationality in evolutionary context

In human populations, rationality plays a role in interactions. In his widely cited example, Hume (Hume, 2008) describes two farmers who agree on helping each other with their corn. They do so based on rational egoism; if they help each other, each one's personal profit will be greater. Therefore, it is in their best interest to help their

neighbour. Hume informally describes how in repeated interactions, cooperation is the maximising strategy. Since they interact in a stable environment both farmers expect to interact again in the future and appreciate their future pay-off as highly, or almost as highly as their present pay-off, cooperation is an equilibrium.

The importance agents attribute to future benefits can lead to cooperative equilibria even when there are more than two interacting agents, provided there is full or equal information of the history of interactions and relatively cheap ways to punish defectors (Skyrms, 2004). When punishment takes the form of social exclusion however, cheap punishment can be achieved when there are not any other agents who are willing to cooperate. Thus, cooperative equilibria are sustainable when there is knowledge of the history of the interactions. Given that we are talking about social interactions in general, iterations are always perceived to be infinite. Also, assuming that the two farmers live in a relatively stable environment, as is most often the case with farmers, their game's environmental parameters are stable. They will be cooperating for as long as they live (or, at least for as long as they farm corn in adjacent fields). Furthermore, defection in this game would hurt the defector's reputation, thus making it impossible for him to cooperate in a new game. Hume's work is probably not the obvious case for a discussion of evolutionary game theory. However, he has examined how conventions of justice evolve and become established (Sugden 2004) and the analysis of the establishment of conventions is evolutionary. This cooperation between the two farmers, provided their environment does not change, will last for as long as they grow their corn. Moreover, their off-spring will have no reason to change strategies. Since it is beneficial for both, they will continue helping each other. Over generations or many iterations, this cooperative equilibrium will become the norm. Those farmers who follow that norm, do not necessarily do so because they have rationally analysed their pay-off in the same ways as did the original farmers. They behave conventionally because their society is based on a number of similar conventions that punish defectors.

Apparently Sugden follows the Humean rationale in arguing that “...ideas of rights, entitlements and justice may be rooted in conventions...[that] have merely evolved.” (Sugden 2004: 8). We need not assume that farmers behave rationally when they follow an established convention. However, their behaviour is rational as it maximises their utility and also sustains a social equilibrium that is essential for their interaction. In this, it is similar to the way termites' work is essential for the their colony

and vice versa. Individual termites cannot reap the benefits of their work without the collective and the farmers' cooperation on its own would not be enough for their utility maximisation since their respective utilities would be much greater if they had the chance to interact with more than one other person; the farmers would maximise in their interaction, but a farmer's utility function includes more than one interaction. Each interaction depends on a plethora of other interactions that ensure a social environment where rational interactions can be realised. These conventional rules evolve as the population evolves. When an interaction reaches a more efficient interaction equilibrium – because of environmental changes such as an increase in the population of cooperators – the new behaviour can spread and hence lead to a new convention. The pairwise repeated interactions example serves as a good example of how interactions can evolve and spread but it is a simplification as it does not examine the population dynamics and how they are affected by adopting successful strategies.

The main advantage of using evolutionary game theory is that it is dynamic and therefore more realistic. Successful strategies self-replicate given the dynamics of the population which makes it a more appropriate model to describe social interaction in changing populations. Taylor and Jonker have shown how in the evolutionary version of the PD game, defecting is the only stable strategy (Taylor and Jonker, 1978); Alexander 2000). However, this was done using the pay-off matrix of the classic prisoner's dilemma game, which although not very different for a repeated game does yield different results. Depending on the pay-offs in the original game, the outcome can be a cooperative equilibrium. The analysis used in one-off games cannot be used in repeated and evolutionary games without amending it. Agents' utility functions are different as they have to take into account how the time factor affects maximising strategies and the pay-off from repeated interactions and possible punishment. Binmore goes as far as to claim that axiomatically a game is a prisoner's dilemma game only when there is a single iteration. A multi-iteration game is simply a different game and therefore cooperation is never rational in the PD game (Binmore, 1998). However, given that the analysis here is not formal and that the premises of the game are more important than its mathematical background, repeated games with the structure of a PD game can still be referred to as PDs. The main point in discussing a PD game in an evolutionary context is that there is a coevolution of strategies and population dynamics, which affect all games irrespective of their structure.

As we have seen, Brian Skyrms's (2004) description of how social structure

evolves is primarily based on the stag hunt game which shows how equilibria evolve and become stable. For Skyrms the repeated PD game can be in essence a stag hunt game. In both games cooperation yields higher group welfare whereas defection is maximising in the one-shot game or when detection is unlikely. Therefore, the conclusion of the analysis is not affected by whether we use the stag hunt or the PD game as a basis. In an evolutionary context what is vital is whether agents are disposed to cooperate. If they are, then cooperation is the maximising strategy – similarly to constrained maximisation. His analysis examines how social structure evolves and how it is affected by the population equilibria. Similarly, Sugden (2004), whose analysis is based on a theory of conventions, discusses why social structure equilibria are followed and why they are a maximising strategy. Social conventions can be seen as the result of Skyrms' social structures evolution. Social interactions that lead to equilibria can be represented as conventions. The repeated prisoner's dilemma and the stag hunt lead to social equilibria. Individuals in societies following these conventions do not deliberate about how to maximise their utility; they just follow the convention and by doing so, they maximise their utility.

In summary, the evolution of the social structure occurs by adopting conventions. When these conventions are not maximising or evolutionarily stable, they will be replaced through a new evolutionary process by a new convention. Following an evolutionarily stable convention is a maximising strategy on the individual level. And doing so replaces the need for constrained maximisation and translucent dispositions as Gauthier uses them. A concept similar to that of constrained maximisation is still to be used in the evolutionary context presented here. However, constrained maximisation here occurs not because rational agents have internalised it, but as the outcome of repeated interactions that create dynamic social structures.

Individuals do not constrain their maximisation but follow a convention that in the short run does not maximise their utility. Agents acting within the limitations of a cooperative convention are a type of constrained maximiser while their behaviour can be analysed as an evolutionary game. They do not constrain their maximisation expecting others to do so and therefore cooperate. They follow the rules of the established convention, knowing that their utility will not be maximised outside the convention. The convention calls for constraint maximisation if we only look at one-off or short-termed interactions. A cooperative convention consists of cooperators since non cooperators have been excluded from interactions. Repeated interactions ensure that all

agents' behaviour is known, not necessarily by the entire population but by each agent's neighbours, and those with whom she interacts. Social exclusion, as punishment, maximises the cost of non- behaviour and in any case this cost is greater than the cost of conventional constrained maximisation.

Since we have asserted that agents remain rational, when acting within a cooperative convention an individual will explore the possibilities of free-riding. Rational agents will search for a method by which they will receive their share of the cooperative surplus without participating in its creation. However, the benefit from free-riding is short-term. Repeated interactions in an evolutionary context increase the likelihood of a free-rider being detected since in repeated interactions behaviour can be known. And reputation is central to an individual's interactions. The cost from having a bad reputation in repeated games is greater than the benefit of a one-off free-ride. In general, free-riding is not a maximising strategy in an evolutionary context. However, since we preserve the assumption of individual rationality free-riding will still have to be discussed more analytically in the following chapters and especially in Chapter Six. However, it is possible to assume individual rationality within the context of an evolutionary account of social interactions.

4.4.3 Conclusion

Evolutionary game theory is used as a tool to describe rational interactions within a population. Looking at interactions from an evolutionary perspective, allows us to consider population dynamics that otherwise would be ignored in assessing individual behaviour. Group behaviour can be analysed in relation to the behaviour of individuals, or more accurately, the interactions between individuals. The two individuals are still bound by the same rules that govern the evolutionary game theoretical process. However, thinking of their interaction as a two person repeated game makes it easier to relate it to an interaction between rational agents.

It is possible to compare interactions between two agents whether they behave according to the premises of rational choice theory or are seen as participants in an evolutionary process. For an outsider, who does not know the incentives for their behaviour, it would be impossible to tell whether they are being rational or evolutionary organisms, if in both cases there is utility maximisation. If this maximisation occurs immediately or at the end of the series of interactions, it does not make a difference for the characterisation of the interaction.

In conclusion, evolutionary game theory examines how populations evolve into a state of evolutionary stability. Whether these equilibrium points are cooperative utopias, or a war of all against all, does not make a difference to the evolutionary analysis. In a society where war of all against all is the norm, no individual will attempt to cooperate in a PD game. From an evolutionary perspective, a cooperator will mutate after one iteration. Thus, in a state of stable equilibrium the equilibrium bounds each agent's choices. A change in the environmental parameters can cause a destabilisation of the equilibrium. For instance, when the proportion of cooperators in the population changes because of individuals relocating, the ESS will shift. And in this case new maximising strategies will arise leading to a new convention.

Destabilisation can also be caused by contact with groups with different conventions that are more efficient, or in other words from new information becoming available. Thus, an evolutionary equilibrium is not infinitely stable as “the evolutionary dynamic never settles down completely; it is always in flux” (Young, 2001). However the associations established in each equilibrium, play an important role in the establishment of a new equilibrium. The relative location within a population and the histories of local interactions will affect the possible strategies that may be adopted. Each evolutionary equilibrium, therefore, is to be seen as an evolutionary step to the next.

In sum, evolutionary game theory provides a fluid account of social behaviour where individuals can learn by trial and error and imitate, thus affecting the equilibrium selection procedure. The coevolution of social structures and individual strategies for maximisation that offer a realistic account of human behaviour can only be studied through EGT.

The following section will emphasise the role of evolutionary theory in explaining conventional behaviour and justifying constrained maximisation as rational behaviour in an evolutionary framework.

4.5 Constrained maximisation as conventional rationality

The evolutionary approach can reinforce Gauthier's argument about constrained maximisation and provide a more realistic account of disposition translucency. In an evolutionary context, agents within a social convention will constrain their maximisation by behaving conventionally. Should evolutionary conventions of constrained maximisers yield higher utility, they will spread and replace lower utility

alternatives. Constrained maximisation as introduced in *Morals by Agreement* is problematic since it requires disposition translucency which is an unrealistic stipulation.

Constrained maximisation is based on the assumption that dispositions are translucent. When an agent knows that the agent he interacts with is going to constrain her maximisation, then he will form the disposition to constrain his maximisation as well. And in this case, constrained maximisation is rational. Ideally, dispositions would be transparent, so that constrained maximisers would only interact with agents of similar disposition and thus constrained maximisation would be a maximising strategy. It is vital then that "the straightforward maximiser and the constrained maximiser both appear in their true colours" (Gauthier 1986: 173). Gauthier rightly asserts that translucency is more realistic than transparency. It is more realistic to assume that actors can guess others' disposition at a high probability than that dispositions are just known. However, there is no argument as to how translucency comes about and how it can be achieved in the real world. The fact that translucency is more realistic than transparency does not necessarily make it realistic enough. "[T]he ability to detect dispositions of others must be well developed in a rational CM" (Gauthier, 1986: 181). It is obviously rational for one to practice detecting dispositions, but how is this ability developed? The problem of guessing others' disposition remains.

The visibility of one's disposition is not, in the evolutionary account, as much of a problem as it is in Gauthier's theory. Accepting that agents can be rational within an evolutionary equilibrium, we will have to give a plausible account of how they trust others to constrain their maximisation. Within a convention that defines the boundaries of behaviour everyone acts accordingly. Everyone behaves as if everybody else will behave conventionally. Thus, agents act conventionally and also rationally when deliberating about whether to defect. Therefore, dispositions become a secondary issue, despite this account being very close conceptually to the rationality of constrained maximisation in a group of constrained maximisers. In an evolutionary framework individuals can look at others' past interactions in order to make an educated guess about their future behaviour.

Looking at the stag hunt game can reinforce the argument. If the pay-off of hunting hare is 3 and the pay-off of cooperatively hunting stag is 4, everybody would hunt stag when they expect everybody else to do the same. This brings us back to the problem of disposition translucency. Assuming that a stag hunt was possible during the first interaction, those who decided to hunt hare will not be accepted in the next stag

hunt. After repeated hunts, the individual pay-off from hunting stag will be greater than the pay-off from hunting hare. This would result in stag hunting becoming the dominant strategy. If stag hunting were impossible at the beginning, this society would reach a hare hunting equilibrium. It is up to the individual rationality of its members then, to imitate a more successful society that has established a stag hunt. Similarly, in a repeated prisoners' dilemma, if one defects the first time the game is played, then it should be expected that the second time both will defect and the game will result in a non-cooperative equilibrium. If a cooperative equilibrium had been established instead, both would be maximising their utility in a repeated game.

In discussing constrained maximisation, Gauthier says: "The just person […] has internalized the idea of mutual benefit" (Gauthier 1986: 157). The just person though has been shown to be a rational person who complies with the principle of minimax relative concession. Constrained maximisation is therefore based on the idea that rational agents have made fairness part of their rational deliberation. Put differently, Gauthier stipulates that rational agents have made morality part of their decision making process when they interact with similarly disposed agents. Similarly, the just person conforms to conventional rules as she wants to avoid deliberation costs and exclusion from the conventions that surround her. Repeated interactions within a small population ensure free-riders are found and excluded, and thus knowing each other dispositions is not a central condition of constrained maximisation. Interactions evolve as individuals grow, learn and gain experience and this has an impact on the way individuals interact. Repeated interactions create dynamics that have an effect on the possible strategies and outcomes of a game, such as trust and social bonds. Gauthier argues that an action is rational when the final outcome of a series of interactions is maximising, irrespective of the outcome of each single action. Evolutionary game theory also moves the centre of interest from the specific action to the final outcome of the interactions that lead to an evolutionarily stable strategy. Constrained maximisation implicitly describes long term maximisation, which is more sensible as the result of an evolutionary process.

In an evolutionary game theory context, disposition visibility is not essential. The fact that cooperators or non-cooperators form groups that follow certain conventions makes it highly unlikely that these conventions will be broken. What is more important, it makes it very costly to break these conventions. Rational reflective individuals can see how breaking an equilibrium will only pay in the short term. Within their group they will be punished by being excluded from future conventions. A

population will only abandon the status-quo for a Pareto superior equilibrium point. Therefore, the behaviour inside this population can only change through imitation of a more successful interaction structure.

Rationality as understood by Gauthier, when he discusses constrained and straightforward maximisation in reference to the dispositions of the population, is very close to the understanding of rationality in evolutionary context. Gauthier implicitly takes a holistic perspective when defending constrained maximisation: “A straightforward maximiser […] must expect to be excluded from cooperative arrangements...” (Gauthier 1986: 187). It is rational and evolutionarily stable to constrain maximisation in a population of constrained maximisers. And it is also rational to be a straightforward maximiser in a population of straightforward maximisers. Rationality thus depends on the strategies employed by the whole population and not just one individual. Therefore, a cooperative society is not always the outcome of rational interactions.

The evolutionary game theory account presented in this chapter does not favour cooperation. This can be only one of the possible outcomes. A non-cooperative equilibrium is as likely. This is in agreement with Gauthier's argument that justice is only rational in a society of constrained maximisers; “[i]n a world of Fooles, it would not pay to be a constrained maximiser” (Gauthier 1986: 183). A society based on constrained maximisation achieves a higher Pareto efficiency point and therefore groups of constrained maximisers do better. Through evolutionary processes, these groups' behaviour is imitated and replicated.

4.6 Conclusion

The central argument of this chapter has been that an evolutionary account of behaviour can replace the concept of constrained maximisation as presented in *Morals by Agreement* without violating the premises of rational choice theory. Once we accept a realistic account of bounded instead of economic rationality, individual rationality does not necessarily conflict with holistic explanations of behaviour. A rational agent can preserve her rationality while behaving according to conventional rules. Her ability to reflect on past interactions makes it possible for her to learn and select future strategies and groups that maximise her utility. Given that rational agents interact within populations, the most appropriate model to examine their interactions is the one provided by evolutionary game theory. EGT is understood as a model to describe the

coevolution of maximising strategies and population dynamics when interactions are repeated.

The analysis of this chapter has to be put into the context of moral contractarianism. There is a bi-directional relationship between social contract theory and evolutionary explanation; a stable social contract stems from, and at the same time supports, the established evolutionary social conventions. The following chapter will look at how evolutionary theory can be used within a contractarian context and lead to a sustainable social contract. Contractarianism will be presented as a theory of dynamic interactions similar to the analysis of conventions. Evolutionary theory will then be used to discuss how rational agents, interacting in stable conventions, can solve the equilibria selection problem and affect the selection a fair social contract that is also evolutionarily stable.

5. Evolution and the Social Contract

The previous chapter examined the conditions under which it is possible to reconcile methodological individualism with holism. The attempt was to show that it is possible to have a conventional account of constrained maximisation, by incorporating cultural evolutionary dynamics in a rational choice framework. The argument of this chapter is closely linked and builds on the analysis of the previous chapter. The focus will shift to providing a more general account of moral contractarianism based on conventional behaviour and social conventions. The emphasis will be on showing how social conventions whose structure is described through evolutionary dynamics, can justify a type of evolutionary contractarianism. According to the proposed evolutionary contractarianism, the social contract consists of dynamic social conventions, whose structure and development is best described through evolutionary game theory.

In order to set the scene for the discussion on dynamic contractarianism, it will be useful to reiterate the main points of moral contractarianism as well as the rational framework in which it was introduced by Gauthier. This should make clearer how the social contract theory can be examined as a dynamic process.

Moral contractarianism refers to the idea that the accepted moral rules in a society are the result of mutual understanding; by participating in a society individuals accept the moral principles that govern it (Cudd, 2007). The appeal and great strength of contractarian theories is that they show that “even if we cannot agree on common ends, we can cooperate for mutual benefit.” (Sugden, 1993: 4). Contractarianism assumes that people in a society have implicitly agreed on the terms of their interactions and accepted the established method of dividing the cooperative surplus. Thus, especially moral contractarianism is based on the assumption that members of a society have consented, at least hypothetically, to the established social rules. In other words, moral contractarianism holds that social norms of morality are based on the rational agreement of all participants in the contract; moral rules are appropriate only if they are agreed by those who are expected to follow them.

Therefore, contractarianism is not a method of enforcing a given moral behaviour, but rather it is a mechanism to help rational individuals in a society decide the bounds of acceptable behaviour. Hence, its main strength is that it bypasses the discussion of what is morally and ethically desirable and focuses on what can be achieved within a given society. Contractarianism offers an incentive for self-regarding

agents to interact for mutual benefit, even when we accept the premises of individual rationality and mutual unconcern; mutual unconcern refers to the fact that rational agents do not have reasons to care about others' utility maximisation. In a sense this is a requirement for rationality; each one should care about her own maximisation irrespectively of how others fare (Morris in Vallentyne, 1991).

The Hobbesian account of the social contract theory calls for a strong government that would ensure compliance with the terms of the contract and enforce social peace (Hobbes, 1976). However, even in Hobbes's hierarchical understanding of social interactions, the terms of this peace were assumed to be a result of common agreement among the members of the commonwealth. This agreement is a mandatory prerequisite, even for a contract based on force such as Hobbes's. Similarly, Hume's account of the social contract is based on social interactions regulated by reason and the subsequent establishment of social conventions, and not on third party enforcement (Hume, 2008). In *Morals by Agreement* Gauthier follows the Hobbesian tradition but at the same time he is influenced by Hume's understanding of conventions of moral behaviour and empathy.

Traditional contractarian theories use an original position, a bargaining process and an agreement state to define the social contract. The status-quo of a given society constitutes the starting point for bargaining. Before bargaining and possible cooperation, each agent has access only to the outcome of her work. Therefore, in order for bargaining to be meaningful, each agent's share of the cooperative surplus should yield higher utility than her pre-bargaining one. Hence, the status-quo defines a set of Pareto efficient possible bargaining outcomes. Rawls in *Theory of Justice* (2005) speaks of the original position where individuals are behind a veil of ignorance which makes them unaware of their possible future position and present characteristics and thus ensures impartiality during bargaining. However, the veil of ignorance is an unrealistic limitation on rationality.

In *Morals by Agreement*, Gauthier proposes a more realistic alternative. The original bargaining position should be non-coercive in order to achieve a stable rational social contract. Gauthier argues that “it is rational to comply with a bargain...only if its initial position is non-coercive” (Gauthier, 1986: 192). For Gauthier, in the original bargaining position both parties' bargaining power should be similar enough to ensure the absence of coercion. Gauthier and Rawls have in common the fact that they both stipulate that individuals at the original position will be, in a way, moral. In Gauthier's

account rational agents will choose to constrain their maximisation based on the principle of minimax relative concession. There is no adequate explanation however as to why a rational maximising agent will choose to constrain her maximisation in accordance with the MRC and thus forfeit potential greater benefit from the bargain; in *MbA* agents aim exclusively at maximisation of individual utility not optimisation of social welfare. The Lockean Proviso that to an extent is used to justify the MRC, also relies on moral and not strictly rational grounds. Reflective rationality in repeated interactions deals with this issue. Hence, implicit moral constraints in Gauthier's theory can be replaced by rational deliberation provided we accept all interactions are repeated or perceived as infinitely repeated.

Contractarian thought in all its shades, is comprised of a two-step process: the initial bargaining procedure and the agreement point (Vallentyne, 1991). As far as the original bargaining position is concerned, there is tension as to whether it refers to an actual historical time or it is a hypothetical structure. Being a hypothetical point as in Gauthier's theory, poses problems in that a hypothetical argument cannot generate actual responsibilities (Cudd, 2007). However, a hypothetical thought experiment can be used as a method of deriving principles of moral behaviour in a rational choice framework. For Gauthier the original bargaining position is defined by the Lockean Proviso (Gauthier, 1986), which ensures the impartiality of the initial position and therefore the fairness of the subsequent contract. Hence, the eventual agreement point depends on the original position and the characterisation of the individuals which in *Morals by Agreement* are assumed to be rational, mutually-unconcerned agents.

The previous chapters have examined how the assumptions of contractarianism can be problematic through looking at the criticism aimed at *Morals by Agreement*. In the following paragraphs, evolutionary social conventions will be used to replace traditional contractarian concepts such as the original position and the bargaining process. The original position will still have a hypothetical role, which however will be shown to be dynamic. The bargaining process is not essential in an evolutionary account of interactions; infinitely repeated interactions render a bargaining process redundant. Contractarianism is to be thought of as a dynamic process where changes in the population cause shifts in the contract and successive generations are not morally bound by the agreements made without their consent. The following section will analyse dynamic contractarianism and its relation to social conventions. In order to do this, dynamic contractarianism will be examined by looking at its main characteristics,

namely its dynamic nature, the importance of conventions that evolve, the rationality of individuals and a bargaining process whose significance is downgraded by comparison to traditional theories of the social contract. The next section will examine the dynamic nature of contractarianism and argue that when the social contract paradigm is enriched is seen as a dynamic process.

5.1 Dynamic Contractarianism

The following paragraphs will examine whether an account of evolutionary contractarianism can be plausible. The discussion will revolve around the assertion that contractarianism is a dynamic process including successive generations of interacting agents selecting new equilibria that will serve as the new social contract. A description of social conventions and their relationship with the social contract, in conjunction with the dynamics in their structure should show that the evolutionary account is closer to reality than the static explanation traditionally used. Showing that selecting a social contract is a dynamic process is essential if we are to accept a description of contractarianism that is based on evolutionary game theory. Although, not explicitly stated in traditional contractarian theories, they are dynamic in that the terms of the contract are subject to change and the contractors also change as generations replace one another.

5.1.1 The social contract dynamics

After the establishment of the social contract, contractarian theory does not explicitly presuppose that this contract and its rules cannot be changed. A social contract determines the rules of social behaviour for a variable period of time with more successful social contracts lasting longer. Therefore, contracts can be changed and take effect successively, provided all contractors agree. This understanding of social contract theory can be derived from the theory's basic principles. Contractarianism is an agreement based on common understanding. Unless we assume there is a metaphysical truth or that we should aim towards a good life as defined by non-humans or humans with superior intellectual capabilities that “play God” (Gauthier & Sugden, 1993: 4), the contract is a human construct. As such it changes when individual preferences change. To argue otherwise one would have to claim that human preferences and the social environment remain unchanged irrespectively of population dynamics, technological and environmental changes. The speed of change can vary and may be incremental, but

the fact remains that once we accept the basic premises of contractarianism we also have to accept its dynamic nature.

A realistic contractarianism should be thought of as a dynamic and subsequently an evolutionary process. Just as organisms evolve to become fitter for survival, social structures evolve so as to adapt to evolutionary pressures applied by population and environmental shifts. Given that established conventions are the components of a social contract, conventions also change accordingly. Social contracts and conventions can be thought of as super-games and sub-games respectively, where the equilibria are interdependent and a shift in one causes an equilibrium change to the other. A social contract cannot evolve without its components evolving just as a super-game equilibrium cannot change without affecting its sub-game equilibria. At the same time, a hypothetical change in the social contract that is not accompanied by change in the underlying conventions, will cause the conventions to change as well, or lead to an unstable social contract.

Contractarianism is a dynamic process since the parties in the contract change as generations of individuals change while individuals' preferences change as a result of the behaviour dynamics within a population. Thus, the social contract is dynamic; as its members or their preferences change, the structure and the terms of the contract change accordingly. Changes do not need to occur in the basic principles of the contract to justify talk of dynamic contractarianism; they can be made up of small adjustments as a result of local convention changes. In addition, a single conventional shift to a higher utility equilibrium does not need to result in the social contract reaching higher social welfare. The salience and popularity of each convention is central in the effect it has on the social contract.

According to the evolutionary understanding of contractarianism, for each new generation or for new participants the original position is the status-quo when they first started interacting within the given social contract. Thus, the original position is merely another equilibrium point in the dynamic process of contractarian evolution. As discussed in Chapter Four, an evolutionarily stable equilibrium is only stable for as long as the environmental parameters allow and require its existence (Young, 2001). Similar to biological evolution that does not stop (Ridley, 1994), the evolution of the social contract is a continuous process of successive equilibria. The most successful ones last longer and historically most social contracts last longer than several human generations. Therefore, social contract change is most frequently incremental by human biological

standards and only conventional change can be perceived directly and influenced by individuals.

In a more realistic context the original position is the result of a historical and cultural process. The moral and political principles that govern our societies have developed over time to accommodate changed cultural beliefs about what is right. For instance, if enslavement of people with different skin colour is acceptable at the status-quo, the renewed social contract cannot be dramatically different; people's perceptions depend on and were formed by the status-quo. Therefore, an equilibrium change might shift perceptions to accepting that only a specific skin colour justifies enslavement before rejecting enslavement all together.

The terms and structure of successive social contracts depend on the respective equilibria that serve as original positions. A social contract leads to a new status-quo, which in turn affects the terms of the next social contract. Since a social contract is assumed to consist of social conventions, it is the dynamic structure of conventions that direct the changes on the general level of society. Social conventions and the social contract coevolve; although the contract structure depends on its social conventions similar to the ways that a super-game depends on its sub-games, a social contract also serves as the status-quo and therefore, influences the possible changes of the conventions as well. Their relationship is bidirectional with the social contract defining the bounds for the social conventions, whereas conventional change is essential for a change in the social contract.

In that respect, it is essential to look into the structure of conventions as well as the dynamics of their changes in order to obtain a better understanding of the dynamics of the social contract.

5.1.2 Conventional change

The evolution of conventions is influenced by pre-existing social contracts. A stable social contract implies the previous existence of stable social conventions. As discussed previously, repeated interactions lead to rational conventions that in turn lead to the establishment of a social contract. Thus, the stability of the conventions directly influences the stability of the social contract and the absence of – stable – conventions is equivalent to the absence of a – stable – social contract. Driving on the left, stopping at the red light, signing on a turn, giving priority to an ambulance, are all instances of conventional behaviour that make safe driving possible. If one collapses, it does not

necessarily follow that the rest will also collapse. But if all are followed by all, or almost all drivers, then the driving contract will be more stable and efficient. Moreover, although these conventions are not immediately related to conventions such as slave ownership they can be components of the same social contract. Changes in one convention will not necessarily lead to changes for very different social conventions; however, given adequate time in a cultural evolutionary framework they can lead to the destabilisation of other conventions either via a topical change or through a change in the social contract.

Successful conventional change provides incentives for rational individuals to learn new behaviour and adopt new strategies. A successful shift to a new convention is more likely to make a greater number of rational actors willing to abandon their old strategies, thus accelerating the shift. Hence, a conventional change can have an effect on the strategies employed in a different convention as long as some of the individuals participating in the first convention also participate in the second one. Conventional change is contagious and it can lead to a change in the social contract, in ways similar to the ways a hare hunter can pollute a stag hunting population (Skyrms, 2004). This account of conventional change also explains its slow pace, which will be discussed more analytically and in relation to the real world in Chapters Seven and Eight.

Conventional change is based on individuals' learning process. Individuals learn through a trial and error process and by imitating behaviour that yields higher utility. Both mechanisms require information availability and also a trial period for the newly adopted behaviour. Even in cases where information spreading is quick and accurate, adopting a new strategy within a group is costly and time-consuming. Shifting from driving on the left to driving on the right includes practical costs and a learning process for the drivers; abolishing slavery requires a new understanding and organisation of the economy (among other things). Moreover, reaching a point where deliberation about change is possible is also part of the changing process. In the driving convention the change may take centuries (Young, 2001), whereas abolishing slave ownership in a given society may happen a few years after a long process of perceptions shifting. These examples do not mean that conventional change must occur exclusively over very long periods of time, but they are supposed to show how learning and the subsequent equilibrium shift can be very time-consuming and incremental. Therefore, social contracts need to shift slowly or otherwise their stability will be questionable. In other words swift changes of the equilibrium may lead to social contracts that are not

evolutionarily stable (they can be destabilised by intruders) and are not supported by an adequate basis of stable topical social conventions.

A social contract needs all or at least an overwhelming majority of the population to behave according to its rules in order to be effective. In addition, the slow process of conventional change discussed above means that there is no effective way to enforce an abrupt change in conventional behaviour. Conventions that are the outcome of rational interaction are stable because no party has an incentive to abandon them. At the same time repeated interactions between rational individuals can only lead to stable conventions. Enforced conventions by a third party that are not in accordance with the previously established conventions and do not take into account the interaction history cannot be stable; since a social convention is the equilibrium at a series of interactions, its stability and duration are based on the fact that these interactions were among rational agents and thus cannot include coercion. Therefore, a stable convention must be the result of repeated rational interactions and its stability is ensured by the rationality of its members.

After having discussed the dynamics and the possible changes in social conventions, the focus will shift to how rational agents can affect the equilibrium point in a convention and subsequently in the social contract.

5.1.3 Rationality in the social contract

Contractarianism is a method to derive principles of justice that will govern our behaviour. These principles of justice do not have to agree with specific ideals of justice. Since they are the outcome of a contract among individuals of similar rationality, we have no reason to denounce them as non-just. A rational agent would not accept a contract if she thought it limits her maximisation. And since all contractors would do the same, the final contract will be one that maximises the utility of all contractors given the limitations of social interaction. Hence, it will be in a situation where social welfare is Pareto efficient. This understanding of justice does not have to be in agreement with any form of cultural understanding of justice. However, this does not mean that culture is irrelevant. The history and culture of a society determines the culture of individuals who draw the contract. Their rationality is, in a sense, defined by their cultural environment. The ability to deliberate and the availability of information are similar in all members of a group and therefore they are all equally rational.

In *Morals by Agreement* rationality and justice are presented in terms of a

contract. Gauthier's is a contractarian theory of rational choice and justice. Contractarianism plays a connecting role; the underlying idea of mutual agreement refers to both rationality and justice. Justice is however broadly understood here as something that agents of roughly equal status would agree on. On the contractarian account any interaction between similarly rational agents is just (Murray, 1999). Contractarianism thus provides a framework where it is possible to combine justice and rationality. Furthermore, contractarian theory is central to the understanding of evolution and rationality presented here. Contractarianism does not need to play a central role in explaining how rational interactions can be described in the context of evolution, but it is ideal in linking rationality with morality and justice. And thus it can be used to discuss the relationship between moral behaviour and evolutionary game theory.

The following subsection discusses the bargaining procedure as it is typically used in the contemporary contractarian tradition. Although the evolutionary account of the social contract bypasses the need for bargaining, the discussion in the following paragraphs is useful in examining the advantages of a conventional explanation of the social contract. In addition, a type of bargaining still takes place in the conventional account; rational agents' interactions lead to an equilibrium. These interactions can be seen as a bargaining procedure embedded in the game and with no need for a distinction between bargaining and game interactions. The next subsection will look more closely at the bargaining procedure in the account of dynamic convention that was presented earlier.

5.1.4 Bargaining

Although the dynamic account of the social contract and its component conventions presented above does not require a bargaining process like the one found in traditional contractarianism, the concept of bargaining is still present and essential in understanding the equilibrium selection process. Bargaining is assumed to take place between rational individuals who want make the most of the possibilities of cooperation and subsequently maximise their share of the cooperative surplus. In Hobbes's *Leviathan* (Hobbes, 1976) the state of nature is used as the starting point and incentive for bargaining. Similarly, in *Morals by Agreement* bargaining begins at the original position, which is the status-quo to be compared with the eventual share of the cooperative surplus. In the conventional account, the original position will be taken to

be the given status-quo which does not need to be idealised or abstract.

A basic bargaining procedure is defined by the Edgeworth box and Pareto optimality and describes a simplified model of the interaction between two rational individuals. It is valuable because, in its simplicity, it provides a concise description of the possible outcomes of an interaction. The Edgeworth box in its simplest form exhibits the interaction between two persons trading with two goods. Each player's utility is represented by her indifference curves which are tangential when both agents maximise. The contract curve is the line connecting all such points and constitutes the set of Pareto optimal trading points. Since both agents are rational, the outcome of their interaction will be found on this contract curve. At the beginning of their interaction each player has a set amount of a good which, on its own, does not maximise her utility. As trading proceeds the players move to higher utility levels until they reach a Pareto optimal trade point, where their utility is the maximum possible. There is a single optimal point where both players maximise, but there are many points where collective welfare is at a maximum. The exact point of agreement then depends on the players bargaining skills.

Bargaining is central to contractarianism and to the concept of the social contract. In essence, bargaining is an agreement on the responsibilities that derive from the contract. Put differently, it sets out the rules of the game. Two agents who interact repeatedly will either have to bargain repeatedly over the rules of their interaction or agree that their first agreement will be binding for all their subsequent interactions. However, their interactions will be continuously changing their history and therefore their maximisation strategies. It is more plausible then to assume that agreement points will be more stable when they are decided on a more frequent basis. Each agreement point can be used for a number of interactions. Then a new bargaining procedure can be initiated by either agent when he believes the existing contract is outdated. In this account bargaining is part of the interaction; the agents' repeated bargaining and interactions are part of an enlarged game consisting of periods of negotiation and longer periods of interaction (Binmore, 1998). Assuming repeated interactions means that the interacting agents have similar histories, or at least each agent's history is known. Therefore their strategy can be predicted. In game theoretical terms, repeated interactions make that the game played cooperative.

Following Binmore's discussion of the Nash bargaining solution – which eventually was accepted by Gauthier as well (Gauthier & Sugden, 1993) - the two

bargaining agents have roughly similar bargaining skills (Binmore, 2007). Their bargaining skills are included in the rationality function and since they act in similar environments their bargaining skill-set will be similar. The bargaining game therefore is symmetric as far as the players' rationality and bargaining powers are concerned. Given repeated interactions, even in the case where their bargaining skills are not strictly symmetric, they will converge to being similar enough to not have an impact on their bargaining. Their repeated interactions are a trial and error procedure, where the least skilful player has the opportunity to learn and improve. And by reflecting on past interactions, she will be able to improve her bargaining skills. Therefore, once we assume repeated, non-random interactions bargaining skills are also assumed to be similar.

The bargaining procedure changes once we accept the repetitiveness of interactions. A bargaining problem has a starting point, break-down point and agreement point. Rational agents compare their utilities under each and adopt maximising strategies. In repeated bargaining games, there cannot be a break-down point. In the case where there can be no agreement, the agent will bargain with someone new who will be more willing to accept her claims. Rational reflection on the bargaining procedure and the contract point of each interaction results in players choosing whether they will interact with the same person in the future. In this sense, the role of rationality in the bargaining is two-fold: first, during bargaining players are assumed to be rational. Secondly, when bargaining has reached a likely agreement point, when the available information about the terms of the agreement can be contrasted with other contracts. Thus, salient maximising strategies will develop. Individuals with similar maximising strategies will tend to bargain with each other giving rise to specific bargaining strategies and solutions. In conclusion, in a repeated interaction framework, bargaining strategies converge and over time rational agents adopt similar strategies.

5.1.5 Dynamic contractarianism: Conclusion

In the above paragraphs, it was argued that rational agents will look to bargain and interact with agents of similar dispositions and behaviour, as these are revealed by their interaction history. Viewing the original bargaining position as a point in historical time ensures that it is plausible to assume that agents' past interactions are roughly known within a social group. Therefore, rational agents will seek to interact with those who have adopted similar strategies and with whom there are not extreme inequalities in

power or bargaining skills.

A rational agent will not initiate bargaining with another agent when the latter is in a position to enforce his claims; if there are not any other options, any interactions will be in essence coercive. However, this is not a requirement for complete equality at the original bargaining position. The two agents can vary in their rationality and bargaining skills as long as this inequality is not so great as to allow unilateral costless coercion. In addition, coercion includes application costs which should be taken into account during bargaining. When A forces B into agreement, A's pay-off should be discounted by the cost of enforcement. This is a cost A does not have to bear when the agreement is in B's best interest as well. Columbus would have to enslave native Americans only if the cost of enslavement were smaller than the concession he would have to make during bargaining. Thus, coercion is not always the easy solution for the more powerful party, as it bears costs that last for the duration of the contract.

Compliance with the terms of the contract is a secondary issue. Each agent's history is known and the bargaining process, being part of the interaction itself, is a repeated game. A rational agent will not continue participating in a game when this does not help her maximise her utility. Therefore the issue of compliance is in essence a problem of participation in the repeated interaction. The repeated game that starts with bargaining concludes with compliance.

The following section will analyse further the understanding of contractarianism in an evolutionary framework. Having shown that contractarianism can be seen as a dynamic process, this understanding will have to be linked more explicitly to the evolutionary account of social structures.

5.2 Evolutionary Contractarianism

An account of evolutionary contractarianism can be used complementary to the above discussion of the dynamic nature of the social contract. Although evolutionary contractarianism is in many ways difficult to distinguish from dynamic contractarianism, it is essential to discuss them separately. The following is an analysis that is heavily based in the previous section on dynamic contractarianism. However, this section attempts to make more explicit the ways that evolutionary theory can work in a contractarian framework. The account of evolutionary contractarianism presented here depends on the explanation of dynamic contractarianism and they share a common understanding of conventional change and individual rationality.

Social contracts are defined by the original position, the bargaining procedure, and the agreement point reached. In dynamic contractarianism, the distinction is not clear. The original position is also the outcome of bargaining and defines the set of possible outcomes for subsequent bargaining. The contract itself – the agreement bargainers reach – is apparently linked and bounded by the procedure and its strength lies on the fact that it is the result of repeated interactions between rational agents. Therefore, dynamic contractarianism does not need to be defined by an original position as the previously reached equilibrium serves as the status-quo for bargaining; it also does not need a bargaining procedure as in a context of repeated interactions claims are replaced by actions. Finally, the evolutionarily stable equilibrium constitutes the social contract; however, unlike mainstream contractarianism, the dynamic social contract has a long-lasting but not permanent status.

Dynamic contractarianism as understood through social conventions views the social contract as a fluid succession of equilibria points where the more successful social contracts are the most long-lived ones. Similarly, the original position is just another equilibrium point defined by the previously established social contract. Furthermore, the agreement point does not have to be implicitly or hypothetically agreed upon; it is enough that rational agents continue behaving in a way that will preserve the equilibrium.

As discussed in the previous section, according to the dynamic understanding of contractarianism agreement is implicit and a result of repeated small scale interactions. A common understanding of behaviour under certain circumstances leads to the establishment of conventions of behaviour. And as interactions multiply, these conventions expand to cover more aspects of social life. The result is social conventions that deal with all or almost all aspects of social interaction. These conventions that solve problems of everyday interaction provide the basis of a commonly accepted social contract. The social contract is thus an extension of established social conventions. A convention that is imported or enforced without the necessary time to evolve cannot be effective in dealing with problems of social interaction. For instance, a law by which the British should drive on the right would obviously be inefficient; it would lead to traffic accidents and road chaos. A stable convention cannot be changed without being destabilised by forces similar to those that lead to its creation. Driving on the right could become the norm over time, even if it were enforced; however, political decision making processes are also based on social conventions that have been established

through similar procedures. The point is that a social convention that has to come into being and is stable and efficient implies and needs social approval; in a sense this is tautological as it is social practices that lead to the creation of conventions in the first place. Thus, the social learning processes that occurred during the creation of the convention also have to take place during its replacement. If the social convention does not pass through stages of trial and error and testing to the specific conditions, it cannot be successful.

Binmore argues that a social contract is “the set of all the commonly understood coordinating conventions” (Binmore 1998: 5), which is a common account for the social contract of both human and animal societies. Also for Binmore, our individual understanding of moral behaviour comes from our evolutionary past. A second understanding of how evolution can be used to explain social structures and concepts of morality, also proposed by Binmore, is that human social structures follow an evolutionary pattern. Similar to natural selection, human societies have developed mechanisms to create and select social structures that are stable and maximise utility and welfare. Social and cultural evolution as discussed here refer to the latter; the conventions and norms that are the product of repeated interactions between and among individuals and groups. The evolution of social structures define societies' moral principles; and as different societies have followed different evolutionary paths, their moral social contract is different.

The “game of morals” (Binmore 1998: 12) is about the rules of behaviour determined by society. It concerns a definition of morality as understood by a society. Since it is the outcome of social interactions over extended periods of time and generations, a change in the game of morals would take long periods of time. Therefore, even if there is an individual or a group of individuals within a society with the will to change the game of morals, it will take a long process to change a behaviour pattern that was established over generations. Each game of morals consists of conventions of behaviour that can be seen as sub-games of the super-game of morals. And these conventions are usually easier to change than the game of morals as a whole. Driving on the right or left is such a convention. Helping those in need is another. The aggregation of all those conventions results in a game of morals – a social contract that defines social values.

A convention – a sub-game equilibrium that is Pareto superior to the feasible alternatives – will become established over time as a result of repeated non-random

interactions. If for instance two British drivers start driving on the right and for some reason this proves to be better for both, more drivers will follow until the convention changes. This example shows that it is not always as simple as someone figuring out a better way to do things. Many social conventions are impossible to change unless there is some environmental change that facilitates or even forces change. If there is some improbable technological development that simply makes driving on the left too costly or too inconvenient, then the driving convention will change. Conventions and convention changes cannot be imposed. A new law requiring that driving rules should be changed in Britain is not enough for drivers to get used to driving on the right and to abandon long established behaviour. Laws that oppose well-established social conventions are too costly to implement because a stable convention can only be replaced by another stable convention; and conventional stability depends on topical interaction and evolution. For instance, the alcohol prohibition in the U.S.A. from 1920 to 1933 was unsuccessful to a large extent because it did not take into account established behaviours (Thornton, 1991). Similarly, the laws against drunk driving that came into force in most western countries from the 1970's were the result of commonly accepted scientific evidence and a subsequent public dialogue with little opposition, in addition to the fact that it was based on a process that in the U.S.A. had started in 1910 (Ross, 1994).

Conventions such as the ones described in the previous paragraph can be seen as the equilibria of games played between individuals in smaller groups, just like the driving convention is an equilibrium for drivers. An equilibrium change to one of the sub-games can have an effect on other related sub-games. And changes in the equilibrium of several sub-games can cause the super-game equilibrium to shift. All these interactions and stable conventions can also complement the contractarian approach. From a contractarian point of view, a repeated game that reaches a stable equilibrium is a contract binding the interacting agents who with their past actions have agreed to the contract terms.

The contractarian approach to conventions can be analysed as an evolutionary process. A stable convention is replaced by another over several generations. Even when the new equilibrium is apparently superior, there are psychological reasons and special interests that will resist change. Given a hypothetical technological development that makes driving on the right more effective, it is easy to imagine a situation where car manufacturers oppose the change in order to avoid extra costs. Put differently,

individuals with different levels of information (and thus different maximising strategies), will have various opposing goals. Over generations, the rationality of a goal can become more obvious, and resistance to change unreasonable. Of course, this is not a one-way street; it is not a given that information about a specific issue becomes more readily available over generations.

In the case of information restrictions, provided that an alternative equilibrium is superior, change will come from imitating societies that have adopted the change and do better, unless there is a complete information blockage. Therefore, neighbouring equilibria play an important role in the evolution of conventions and the subsequent social contracts. Evolution to superior equilibria is balanced by a kind of backward evolution. A social contract can collapse if there are no successful conventions to imitate or its constitutive conventions have reached an evolutionary dead-end. In historical and anthropological terms, societies have been locked in vicious cycles such as a status-quo of a war of all against all, where the status-quo reinforces the continuity of the sub-optimal equilibria.

“Natural selection leads organisms to evolve adaptations – traits that enhance their chance of survival and reproduction” (Okasha, 2006: 11). The same is true for the evolution of social structures. The social structures that survive and become stable are the ones that develop traits that make them fitter that the competing structures. Therefore the stable social conventions that have reached an equilibrium and constitute the social contract, are the result of a process similar to natural selection. The less fit social contracts do not become extinct as a matter of natural selection though, they are being abandoned by individuals when it becomes apparent that there are social contracts that are Pareto superior. This procedure is described by Skyrms (2004). In the social structure and the social contract, Skyrms describes how the behaviour of one's neighbours affects one's rational strategy. At the level of society cooperation is Pareto efficient. In cooperative societies the cooperative surplus is higher and therefore they maximise social welfare through cooperation. In non-cooperative social contracts, the conflict costs are too high for social welfare to be comparable with that of a cooperative social contract. In natural evolution, the fittest species or organisms produce the most offspring and thus, have more chances to survive in an evolutionary time frame. In the evolutionary approach to the social contract, it is the social structure that maximises social welfare that becomes the norm, or that in any case leads to a stable equilibrium

This description of conventions and their evolution is very closely linked to the

idea of memes (Dawkins, 1976). Just like memes, conventions can imitate, replicate and evolve. More accurately, the individuals participating in a convention can change their behaviour so as to cause the conventional equilibrium to change; individuals who imitate more successful strategies in neighbouring conventions are essential for conventional evolution. Equilibria can be affected by neighbouring equilibria in a similar way that memes affect each other; information is easier and quicker spread between neighbours and hence, an individual's or a convention's neighbourhood is paramount for the local equilibrium. Thus, the concepts of biological and cultural memes, or else conventional behaviour, are very close. However, in biological memes there is no room for individual rationality and initiative whereas in cultural evolution rational reflection can explain how certain memes evolve while others disappear. A cultural meme will do better in evolutionary terms when the individuals who follow it reflect on their actions; simply put, more useful memes are more successful. Thinking of conventions as memes gives a better idea of how conventions can evolve in a manner that is similar in many ways to the evolution of memes. Just as memes affect the evolutionary path of a species, cultural memes affect the evolutionary stability of social contracts.

The fact that individuals within the evolutionary context are assumed to retain a form of rationality that allows them to deliberate about their own history and also to gather and assess information about others' strategies, reinforces the account of the evolution of the social contract presented above. Individual rationality operates as a periodic steering mechanism that keeps evolutionary dynamics under control. When evolutionary forces lead to outcomes that are unacceptable for the rationality of the majority, local rational interaction will alter the evolutionary path.

A two-person repeated interaction that maximises both agents' utilities is preferable to one that is maximising for just one party of the interaction. In repeated games, the strategy that leads to maximisation of the social welfare also leads to utility maximisation. In the repeated version of the prisoner's dilemma cooperation is maximising, just like stag hunting is the maximising strategy in a stag hunt game, and in a hawk-dove game hawkish behaviour is destructive for both. Given these two actors are surrounded by rational self-reflecting individuals, maximising behaviour will be contagious. The result will be a stable maximising social convention that regulates the behaviour of the “neighbourhood”. Similarly, this stable, maximising convention will be imitated by other rational groups.

Individuals who decide to follow a convention or enter a social contract, do so because they think it will maximise their utility. Those who think it does not, do not participate in the same conventions or social contracts. Agents with similar history of interactions living in similar environments, develop similar rationality. They will expect others to behave similarly to them and this makes it possible to have a stable social contract. Therefore agents participating in agreements, develop bonds that reinforce the stability of their agreements, without the need to abandon the assumption of rationality. They have the capacity to reflect on the terms of the bargain and its outcome and decide whether they will continue following the same rules.

This section proposed an account of evolutionary contractarianism that is based on conventions whose structure is best explained through evolutionary theory. Individualistic methodology assumes that social behaviour can be understood by analysing it to its components, that is the actions of individuals. In the evolutionary contractarian account, social behaviour is analysed first at the level of the social contract, then at the level of topical social conventions before reaching the examination of individual behaviour. The following section will turn to the examination of equilibria selection; according to the folk theorem (Binmore, 2007), repeated games have many stable equilibria. Therefore the same applies to evolutionary games that are a subcategory of repeated games. However, the evolutionary account of human behaviour can be problematic because it is indifferent between different stable equilibria.

5.3 Equilibria Selection and Justice

The discussion in the previous section and Chapter Four should have made clear that evolutionary accounts of human behaviour and the social contract do not pose a problem for methodological individualism and the assumption of individual rationality. On the contrary, the two paradigms can be used in conjunction in order to form a more plausible theory of human behaviour within societies. However, selecting among stable evolutionary equilibria is problematic when individuals are rational. In biological evolution it is indifferent which stable equilibrium will be selected. In cultural evolution there are many stable equilibria that are not Pareto efficient; rational agents would prefer a Pareto efficient status, but individual behaviour cannot always counter evolutionary forces.

Social cooperation in animal societies is essential for species survival. In human societies the same is not necessarily true. We have no reason to hope that “a species will

learn to cooperate even when the conditions are seemingly favourable" (Binmore 1998: 204). That is especially true for the human species. Being able to play a game of morals in addition to the game of life, does not mean that we are able to overcome natural restrictions and instincts. Humans, just like all species, are selfish. However, our rationality makes it possible for us to fit our selfishness in a social context. Unlike social animals like ants and bees, humans can be members of a society and at the same time be individualistic.

As was discussed in the previous section, there are many stable equilibria. A sub-optimal Pareto equilibrium can be a Nash equilibrium. In such a case, a society is trapped in a non-maximising social contract. The theory of evolution does not give an account of why a society would move from a sub-optimal equilibrium to an optimal one. In the evolutionary contractarian account there is no method by which the equilibrium can improve to a Pareto superior point. Cultural evolution, similarly to natural evolution, follows a self-determined path. Cultural evolution can lead to failed social contracts, in which case a new bargaining procedure begins. Individuals affect evolution by way of affecting the equilibrium in small scale games that lead to conventions. Therefore, individuals, and more likely groups, can destabilise an equilibrium. But given the fact that cultural evolution takes generations to evolve, individuals from just one generation cannot alter its course. The main issue is at which points cultural evolution pauses; how the equilibrium points, conventions or social contracts, are decided.

A Nash equilibrium is reached when both players have made their best possible move. Therefore repeated interactions will stop at a Nash equilibrium. Binmore (2004) argues that a second necessary condition for a stable social contract is Pareto efficiency. Efficiency is paramount and a social contract that is both stable and efficient is obviously preferable. In terms of biological evolution, equilibria that are stable but not efficient still survive. However, in cultural evolution a stable and efficient social contract will take over, given long enough time; given rational agents and communication between conventions, it is reasonable to claim that individuals within a convention will imitate the behaviour of those participating in more successful social structures. Since the time needed for a stable only equilibrium to collapse is measured in generations, the distinction between stable only and stable and efficient contracts is not clear for current generations.

Evolution will therefore lead to a selection among stable equilibria but it will

be indifferent about efficiency. Cultural evolution does not have a selection mechanism among equilibria that are both stable and efficient. Rational individuals with access to reasonable amounts of information cannot always make the distinction; they are able to compare their individual utility but not social welfare between different social states. However, the choice is not so frequently among Pareto optimal equilibria but rather individuals have to choose among stable equilibria whose effectiveness varies. A more efficient Pareto equilibrium is a fairer equilibrium given that it has been reached by a series of Pareto efficient moves; at least one individual is better off, without anybody becoming worse off. Individual rationality can direct evolutionary processes towards more efficient equilibria and thus move to a more just social contract. Concepts of justice and fairness cannot be adequately discussed with evolutionary theory and this is why it is essential that contractarianism is used in conjunction with evolution. Gauthier on the one hand, uses the Lockean Proviso to define justice and understands it mainly as impartiality. On the other hand, Binmore speaks of fairness as if it were the outcome of natural evolution and he guesses that "universal principles of justice...must be presumably written into [our] genes" (Binmore 2005: 15). That may be and even if there were evidence from biology to support it, it is not an interesting or useful assertion in the context of moral and political philosophy. The main advantage of naturalised justice is that it is free from personal and cultural prejudices. An idea of justice based on rationality can achieve the same; rationality, just like biology, can be analysed formally and as long as we accept that justice is the outcome of an interaction between rational maximisers, we can isolate and objectify it. Thus, it can be free from biased irrational characteristics.

Accepting individual rationality as a central characteristic of humans weakens the argument for a Kantian morality. The only categorical imperative for rational agents is the equilibrium that is the outcome of rational interactions. Therefore, in rational contractarianism, especially as it is understood in an evolutionary context, there cannot be any pre-existing concept of morality which is disconnected from the physical world. Rational agents will accept only what maximises their utility given the constraints of social interaction. Put differently, justice is what rational individuals of similar bargaining power accept as an agreement point. Hence, a Pareto optimal Nash equilibrium is fair and a distribution of the cooperative surplus will be fair as long as both bargaining parties have agreed to it, provided there has been no coercion. Stability, efficiency and fairness of the agreement are secured by the agents' reflective rationality.

If one party in the contract decides that the cooperative surplus is not divided in a way that will maximise her utility, she will choose to return to the original bargaining position, or in other words abandon the convention. From her new status, she has the chance to initiate a new bargaining procedure. This is a trial and error process which leads to the formation of stable, efficient and fair equilibrium points.

In *Morals by Agreement* just behaviour is seen as abiding by the principle of minimax relative concession. For Gauthier then justice is the ability to enter a cooperative agreement and constrain maximisation. The same rationally based understanding of justice is valid for the evolutionary account. However, the concept of rationality differs slightly. Reflective rationality in repeated interaction solves the problem of the rationality of constrained maximisation. Two individuals will only agree on a contract if their individual utilities are greater after the bargaining. The share of the division of the cooperative surplus has to be higher than the individual product in order to have an agreement. In conclusion, fairness is not central for a social contract, but it is a side effect of a rational bargaining procedure. Rational bargainers, who have experienced fairness – a maximising mutually advantageous equilibrium – will not accept an arrangement that is neither fair nor maximising. Nash equilibrium, Pareto optimality and fairness are all characteristics of rational bargaining. In this context, fairness is defined by a Nash equilibrium that is Pareto optimal. In other words, two rational agents who agree on an interaction and its outcome, participate in a fair interaction.

Deriving justice from rational bargaining largely depends on how justice is defined. In the literature in this area, for example Gauthier, Sugden and Binmore, justice is seen as impartiality but this is not necessarily the case. A rational agent who accepts a smaller portion of the cooperative surplus does not participate in an unfair interaction. Claiming so would mean that we assume that there is a point of rational deliberation that is superior to the agent's. However, in the naturalised account of the social contract there can be no room for a point of objectivity outside the physical world. Individual rationality as it is exhibited through repeated interactions is the only impartial point of view that is feasible as required.

Justice will be discussed more analytically in Chapter Seven, which will deal exclusively with principles of justice in an evolutionary framework and attempt to offer a convincing solution to the problem of including weaker persons in rational interactions. The above discussion assumed that extreme inequalities are not present

which makes rational bargaining plausible and effective in leading to a mutually accepted outcome. However, accepting that the original position is a point in historical time allows the possibility that it is a state of extreme inequalities. A concept of justice has to account for these inequalities or offer a method of diminish them to a point where rational interactions are possible.

5.4 Conclusion

The above discussion, continuing the analysis of the previous chapter, attempted to show how repeated interactions lead to equilibria that are to be seen as conventions of social behaviour. The aggregation of these social conventions constitutes the social contract. The selection of conventions and their respective contracts occurs through evolutionary dynamics describing social structures. Social convention, the social contract as well as individual strategies constitute the levels of selection (Okasha, 2002) in cultural evolution. Therefore, cultural evolutionary equilibria are not always fair. Repeated interactions lead to stable equilibria but in this context there is no way to ensure fairness. In this respect Sugden is right to claim that rational bargaining does not lead to a fair outcome but to norms of behaviour (Gauthier & Sugden, 1993).

However it is possible to assert that the strategies adopted by rational individuals acting in similar environments, will converge towards the same bargaining equilibria. Thus, they will lead to stable conventions and social contracts to determine the rules and limits of interaction. At the same time, reflective rational agents have the capability to compare the equilibrium they are at with neighbouring ones. Skyrms (2004) shows how more efficient equilibria will be contagious and describes how cooperative and non-cooperative equilibria evolve; a stag hunt is likely to evolve and become an equilibrium as well as a hare hunt depending on the local population dynamics. This account is reinforced by having assumed agents that are boundedly rational. Reflective rational agents with knowledge of their past and reasonable information about their surroundings will use this knowledge to further their benefit from future interactions.

Evolutionary theory cannot address free-riding because on this account free-riders cannot survive. The biological evolution of organisms and species that need social interaction has rejected parasitic behaviour. However, in cultural evolution free-riding poses a more serious problem, especially when individual rationality has been preserved. Although the account presented above does not deal explicitly with the problem of free-riding, it implies that free-riders will be known and excluded in a

society of non-random interactions. In this sense, parasitic behaviour in an evolutionary context can be exposed and punished. The evolutionary explanation of conventions offers a more realistic argument against the rationality of free-riding than Gauthier's moralised account of internalised constrained maximisation.

In the following chapter free-riders within social conventions will be shown to be irrational and parasitic behaviour unsustainable. The basis of the argument will still be the theory of *Morals by Agreement*, in conjunction with the Hobbesian tradition of the Foole and the Humean understanding of conventions. In addition, repeated interactions will be shown to be paramount in providing a convincing argument against the rationality of free-riding. Evolutionary theory allows for parasitic behaviour and in this sense the argument may seem weaker that Gauthier's. However, it also offer a more realistic incentive for cooperative social behaviour.

6. Conventional rationality and collective benefit

This chapter will attempt to show that especially when understanding social interactions as conventions of behaviour, free-riding is irrational and more easily detected and punished than in mainstream contractarian theories. In addition, the response to the problem of free-riding offered here is more realistic than the one proposed by Gauthier, who asserted that rational agents will comply with their agreements by internalising constrained maximisation. Parasitic behaviour can yield benefits but these are short-lived and smaller than the benefits of conventional behaviour, when interactions are repeated and interaction history is known.

Free-riding is not as much of problem for the conventional and evolutionary account of contractarianism and rationality that was presented in earlier chapters as it is for traditional contractarianism. Within social conventions, which are the evolutionary outcome of interactions among rational agents, free-riders can be detected and excluded from future interactions. In addition, the evolutionary framework tells a story of evolution of social structures where rational individuals' possible strategies are limited by social bounds. Thus, a society that has reached an equilibrium has done so by excluding or assimilating non-conforming behaviour, at least to a large extent. This does not require the extinction of free-riding behaviour, but it does assume that if there are individuals who free-ride, they are so few that they can neither destabilise the equilibrium nor convert cooperators. An established cooperative equilibrium implies that it has become obvious to individuals that rationality dictates cooperation and therefore there is no need for deliberating defection.

Chapters Five and Six have attempted to show that under certain realistic conditions, such as bounded rationality and dynamic interdependence of social interactions, it is possible to use holistic explanations without abolishing individuality and assumptions of individual rationality. The assumption of individual rationality leaves the analysis of conventional behaviour open to the criticism proposed by Hobbes's Foole and the possibility of free-riding being a rational strategy. The Foole suggested that it is rational to make agreements but not comply with their terms and similarly a free-rider behaves rationally if he manages to receive part of the cooperative surplus without having participated in its production.

The problem of free-riding

According to the rationale of free-riding, rational individuals will free-ride should they be presented with the opportunity. In the conventional account, agents have been assumed to retain their rationality throughout their interactions and their participation in conventions. Individual rationality is the basis for the formation of conventions that make up the social contract and rational deliberation remains a fundamental characteristic of human behaviour. Even after the establishment of the most successful social conventions that ensure that agents who behave conventionally maximise their utility, individuals have the capacity to deliberate rationally about their options and their conventions. The fact that individual rational deliberation is not essential for maximisation within a convention does not imply that conventional agents have lost the capacity for rational deliberation. Therefore, individual rationality is still present and the subsequent free-riding behaviour still poses problems for the rationality of compliance.

Moreover, there is tension between the assertion that rational agents are equally rational and the possibility of free-riding being rational. The assumption of equal rationality makes it imperative to explain why only some agents become free-riders. Free-riding arises as a possible strategy within a – mostly – cooperative society or group. Hence, for free-riding to take place there has to be an established cooperative equilibrium. A free-rider is someone who realises that his cooperating is no longer rational. In the conventional account a free-rider can also be someone who rejects the established convention and decides to maximise his personal benefit with no regard for the collective welfare. In terms of the stag hunt game, a hare hunter's behaviour is based on the same rationale as a free-rider's; a hare hunter wishes to be a recipient of the stag distribution even though he has not participated in its creation.

An alternative conception of reality and a given environment can lead rational agents to non- conventional behaviour. Free-riding can be a result of better information, for instance from a neighbouring convention that achieves higher welfare. In this case, a defector or a hare hunter, appears like a free-rider to the rest of the cooperative group but once the same information becomes more widely available, a more general strategy shift will be realised. Defection here does not have to mean defection in the PD game; any agent who takes advantage of the cooperative surplus without previously having contributed to it, can be referred to as a defector. As mentioned in the previous paragraph a free-rider and a hare hunter are motivated by similar reasoning in that they both want to take advantage of a surplus to which they have not or will not contribute.

In PD type games, a defector also attempts to maximise his benefit from interactions without contributing and so his behaviour and reasoning are similar to those of a hare hunter or a free-rider.

Hobbes described free-riding informally by introducing the Foole to argue against the rationality of compliance with one's agreements in a contractarian framework (Hobbes, 1976). Gauthier also used the Hobbesian Foole in his discussion about the rationality of constrained maximisation. In *Morals by Agreement*, the Foole argues that it is rational to free-ride; agree to comply with the terms of a contract and once the other party has fulfilled her side of the bargain defect from the cooperative project. For Hobbes, the Foole's conception of rationality requires a commonwealth to enforce the agreement, whereas for Gauthier rational individuals will comply with their side of the bargain once they have accepted the benefits of constrained maximisation. The Foole's argument is that an agent's adopting a joint strategy is only rational when cooperation yields a utility level that is “no less than what he would expect were he to violate his agreement.” (Gauthier, 1986: 165). The following discussion about free-riding will focus on Gauthier's analysis of the Foole's arguments against compliance and attempt to show how the repetitiveness of interactions and information about an individual's history make free-riding irrational.

The following paragraphs will include a discussion of rational individual behaviour in the context of evolutionary conventions. Having examined how rationality is to be understood in an evolutionary context and how conventions arise from interactions between rational agents, we will have to show that compliance with the terms of these rational conventions maximises individual utility. Free-riding is a problem that is more analytically described in economic theory but poses significant questions in moral philosophy. Therefore it seems appropriate to examine it in an economic context and in relation to collective action failure. Furthermore, discussing free-riding through collective action failure will make it easier to address free-riding as a problem of conventional behaviour within groups and not solely as a problem of commitment in repeated interactions.

6.1 Free-riding and collective action failure

Free-riding in the context of social interactions can be examined by looking at the theory of collective action failure. The following paragraphs will describe collective action failure in conjunction with free-riding and look at the circumstances under which

free-riding causes a failure of collective action. Free-riding and collective action failure are two sides of the same coin; free-riding refers to the individual level and collective action to the level of society or social group. In social groups where free-riding behaviour is not constrained or punished and free-riders are allowed to maximise their utility at the expense of contributors in the cooperative surplus, the eventual social state will be a collective action failure.

6.1.1 Free-riding

Free-riding is generally thought of as the behaviour by which individuals take advantage of the social output without having participated in its production. Although the basic principle of characterising free-riding is the same, free-riders can be seen as individuals who behave in two slightly different ways. Firstly, a free-rider is an agent who enters a cooperative convention exclusively in order to take advantage of its cooperative surplus; in this understanding a free-rider is a parasite who moves through various social groups in order to make the most of his interactions while remaining undetected. Parasitic behaviour is costly in that it requires information about the structure of social structures; in addition, parasites participate only in short-term interactions that are bound to produce less. The second category includes an understanding of a free-rider as someone who, although originally disposed to cooperate, realises that it pays more to defect. The latter case is more closely linked to the rationality of compliance as discussed by Hobbes and Gauthier and in general moral philosophy. Compliance with agreements and promise keeping is one the possible understandings of free-riding. The evolutionary and social groups approach used here makes it essential to refer to both accounts of free-riding.

In both cases, free-riding is the maximising strategy given the established behaviour of the group. So, free-riders can be seen as belonging in two different categories of rational individuals: those with a parasitic behaviour who try to take advantage of existing cooperative social structures without contributing, and those who after having participated in the production of the cooperative surplus realise that it is not in their best interest to continue participating. The essence of the free-riding behaviour is the same however; some individuals' rational deliberation dictates that they can benefit from a social good without participating in its creation. In this respect free-riding conflicts with the assumption of equal rationality; if agents are equally rational, they should adopt similar strategies to maximise, and free-riding is rational only when there

is a sufficient part of a society cooperating.

Hence, the fundamental characteristic of the behaviour of free-riders is that they benefit from conventions that have reached cooperative equilibria. In a cooperative convention one maximises, at least in the short-term, by reaping the cooperative benefits without contributing, whereas in non-cooperative conventions free-riding is not a meaningful option since there is no common resource available. Thus, free-riding is individually rational within a group of agents who are disposed to cooperate and have established a cooperative equilibrium. Moreover, free-riding is possible only when the majority or at least a critical number of individuals are not free-riders.

This short description of free-riding shows that it is a social welfare problem as well as an issue related to the concept and understanding of individual rationality. On the individual level, free-riders take advantage of others' contributions and expect that others' will not behave in the same way. On the social welfare level, free-riding leads to the lack of a cooperative surplus and in that respect it coincides with a failure of collective action; the free-riding of individuals is the cause of social collapse as collective action failure.

6.1.2 Collective action failure

Collective action failure describes a situation in which there is no social output produced as a result of generalised free-riding or alternatively society is found at a sub-optimal equilibrium. It takes place when most or a large number of the members of a group try to benefit from a cooperative product without participating in its production. If a sufficiently large number of individuals free-ride, there will be no cooperative product and everybody will be worse off. Free-riding and the subsequent collective action failure are the essence of the conflict between individual maximisation and collective benefit. At the same time, failure of collective action inevitably leads to the impossibility of free-riding, since there will be no cooperative surplus.

Rational behaviour on an individual level results in collective action failure; or in other words “...a collective action problem exists when rational individual action can lead to a strictly Pareto inferior outcome” (Taylor, 1987: 19). The classic example by which collective action failure is examined is the tragedy of the commons (Hardin, 1968), which exhibits how a group of rational individuals will prefer short term maximisation over longer term sustainability of a common resource. When agents are assumed to be rational according to the economic assumptions of rationality, all social

groups where rational agents are not constrained end up with collective action failure. The tragedy of the commons (Hardin, 1968), which will be discussed more analytically later, originally referred to the over-exploitation of an environmental resource but it also applies on the exploitation of any collective output. Its importance lies in the fact that it shows how individual rationality is in direct conflict with collective benefit in a static environment of limited and constant resources and that under these conditions free-riding is the only rational strategy.

Free-riding is a problem every time collective behaviour is examined and individual rationality is assumed. When each individual's behaviour cannot be monitored efficiently, either by other members of the group or an external policing mechanism, there will be a rational incentive to defect from a cooperative equilibrium. In the discussion in the previous chapters it was assumed that free-riding is not rational mainly for two reasons: first because behaviour is or can be known and the cost of being found out is too high, and second because it was asserted that rational agents have a small future discount factor so that they value long term benefits highly enough to accept a smaller immediate pay-off.

More realistically, detection is possible only in two person interactions and interactions within small groups or in an idealised model of the world. In real world, n-person interactions, there cannot always be complete information. In addition, the assumption of rationality is not bound to encompass a small future discount factor. The objectivity of individual preferences implies an objectivity in the chosen maximising strategies. Therefore, an agent who chooses to maximise immediately as opposed to waiting for a bigger benefit in the future cannot be characterised as irrational as long as she has taken into account all possible strategies.

Furthermore, a cooperative equilibrium even when it is Pareto optimal for society presents an incentive for individual rational agents to defect; as a matter of fact, the existence of a Pareto optimal state makes it rational for one to free-ride in the first place. Individuals' defection does not necessarily have an effect on the social cooperative outcome; social welfare will be unaffected provided that the number of defectors is smaller than a critical threshold, which varies depending on the size and the dynamics of any given social group. One can expect that his defecting will not affect the social equilibrium, at least in the short-run. And hence, he can still be a recipient of a share of the distribution of the cooperative surplus.

6.1.3 The rationale of free-riding

Free-riders are not necessarily less or more rational than cooperators, but apparently their rational deliberation leads to different outcomes. However, provided they interact in the same environment, we should expect them to maximise in a similar way. A rational agent will imitate free-riders when she sees they are doing better than her, in similar ways as maximising strategies are replicated in biological evolutionary models (Skyrms, 2004). Similarly, a free-rider will change his behaviour when cooperative behaviour in his environment is more beneficial. The critically important assumption is that both cooperators and defectors are utility maximisers and as such they will change their behaviour when there is information about a different available strategy that yields higher utility. And the maximising strategy in each case depends on an agent's environment and the history of established equilibria.

In economic theory the typical example of free-riding deals with avoiding paying for a collective good (Mueller, 2003). Assuming that the public good will be provided eventually, it is in everybody's best interest to avoid paying their share. Therefore, public good provision cannot rely upon voluntary contribution. An institution needs to exist that will detect and punish tax evaders. Collective action in a group of rational, utility maximising agents will result in failure, unless there are incentives to cooperate. When detection and punishment occur at a high probability, free-riding becomes irrational. Especially in the case of large groups, where detection of free-riders is impossible or very difficult, individual utility maximisation is an obstacle to social welfare maximisation. In most real world societies, if it were not for the existence of free-riding detection and punishment mechanisms, collective action would have been impossible. The size of the group is therefore essential in analysing the rationality of free-riding.

Competition for the use of a resource is not explicitly what Hobbes (1976) called a war of all against all, but this is a similar type of collective action failure. Hobbes's proposal for a strong government to protect property rights and provide security has been one of the most popular solutions to collective action failure and free-riding. In large groups like contemporary societies, a government or a similar institution is needed to punish free-riders and to ensure no one abuses the available resources. Realistic contemporary societies can usually be seen as a “large, heterogeneous, mobile community” (Mueller, 2003: 13), which makes free-riding more likely and more unlikely to be detected.

The core of the problem of free-riding and collective action failure lies in the fact that social welfare maximisation requires constraints on individual maximisation. Strictly speaking, it is in the rational individual's best interest to take advantage of the cooperative surplus to make a profit out of it immediately. In a complementary understanding, collective action failure arises when the future discount factor is so small that individuals prefer to maximise at the present rather than wait for a potentially higher utility in the future. However, free-riding can only occur when a natural resource exists already or a public good has been produced in the first place.

The rationality of free-riding depends on the fact that not everybody free-rides. It only pays to free-ride as long as free-riding is not the norm; when it becomes the social norm, the first free-rider will have gained more and therefore free-riding becomes a race to defect in a PD game. In the classic PD game the only rational strategy is defect; similarly in a ten iteration PD game, rational players will not reach the ninth iteration, but will defect immediately to avoid being taken advantage of. In any case, the absence of a central enforcer or of the possibility to detect and punish defectors in interactions taking place in small groups means free-riding is generally rational.

A cooperative surplus has to exist before it can be exploited and thus, in a Hobbesian state of nature there can be no meaningful free-riding. The collapse of all cooperation comes logically after the creation of a cooperative society. According to Hobbes, rationality will lead to agreement on the formation of a strong government that will lift humans from the state of nature and once the commonwealth has been established, will punish free-riders. So, in the state of nature long-term rationality prevails and individuals are able to see the advantages of cooperation and form cooperative structures and eventually a government. An effective government is needed then to enforce cooperation; to protect society from free-riders and to stop over-exploitation of natural resources.

Free-riding follows cooperation as it can occur only after rational agents have cooperated in the past. Hence, from a rational choice theory perspective in a convention where defection has become the established behaviour individual rationality cannot offer a solution; rational agents who free-ride are aware of the relative benefits of cooperation and defection and have chosen to defect. Once free-riding causes a failure of collective action, it is irrational for anyone to be the first to shift his strategy to cooperation. Free-riding conventions are locked in a permanent state of war of all against all unless there is a dramatic environmental shift or new information becomes

available. And in this sense there can be no normative prescription for an individual in such a convention. Obviously it does not pay to be a sucker in a world of nasties; in contemporary moral philosophy, any theory that allows for reasonable or rational agents, there cannot be a normative suggestion for one to cooperate in a world of defectors.

A cooperator within a group of free-riders cannot cause any real change to the structure of the convention or the behaviour of individuals. Therefore, there cannot be an argument, based on any concept of rationality against free-riding in a free-riding convention. Cooperative interactions may be imitated when they prove that they yield higher individual utilities over time; however, in a world of Fooles a rational agent will never initiate cooperative behaviour with someone who is a non-cooperator even in the case of first performer contracts. In this specific case one will choose to interact with a known cooperator or abstain from any interaction rather than interact with known defectors. Moral incentives against free-riding would have to assume agents of an unrealistically low degree of self-regard in order to claim that one should try to cooperate even when he is surrounded by defectors. For instance, Hume's theory allows for empathy and therefore implies that we are not always self-regarding and mutually unconcerned but even it does not recommend that we ought to be suckers in a world of fools (Hume, 2008). Hence, even normative arguments for cooperation in this context can only be very weak.

6.1.4 Conclusion

Conforming to a norm is assumed to maximise one's utility and thus be rational. In Gauthier's analysis, rational agents internalise constrained maximisation, without loss of their rationality. In both the Gauthier project and the conventional account there is no collective action failure. Agents' behaviour is assumed to be consistent with their original preference to abide by the normative rules or, put differently, to internalise constrained maximisation. However, rational agents must remain rational throughout their lives and be able to question their own behaviour. It cannot be assumed that their intention to maximise diminishes once they begin participating in a norm or constrain their maximisation. In social groups, and especially in large social groups, prudential maximisation is not always rational. An agent forms the intention to constrain her maximisation at the beginning of a series of interactions, as this behaviour is maximising at that time. It is possible but not certain that this will remain a maximising

behaviour for the duration of the interactions. A rational individual will look back in order to evaluate the interaction outcomes and will keep being informed about new developments in her environment.

In conclusion, according to the theory of collective action failure, it is rational for one to free-ride provided that the cooperative surplus will still be produced. Free-riding is rational on the condition that the cooperative surplus has been produced and will continue being produced, or put differently that not everybody is a free-rider. However, having accepted that individuals are rational and even equally rational, this thinking has to lead to the collapse of any cooperative structure in time. Rational agents within a convention who have access to similar information will eventually reach the same conclusions about maximising behaviour; hence, the assumption of equal rationality implies that conventions reach equilibria of some type, cooperative, non-cooperative or mixed. However, the latter are less likely once there is information about the success of cooperative equilibria since moving to a cooperative equilibrium is easier from a mixed than from a non-cooperative one. Within a convention, the rationality of free-riding is questionable; if free-riders maximise they will be imitated which would defeat their purpose. If their past behaviour is now known, other agents will avoid interacting with them. Therefore, the rationality of free-riding contradicts the main assumptions of conventional rationality; knowledge of past interactions and equal or roughly equal rationality.

The typical solution to the free-rider problem and the subsequent collective action failure is the introduction of an enforcing mechanism that will deter and punish free-riders, such as Hobbes's commonwealth (Hobbes, 1976) or a government as described in economic theory (Mueller, 2003). A strong government will force self-interested individuals to keep their agreements and when they do not, it will punish them. An alternative solution, coming from a conventional account of rational behaviour would suggest that cooperation is self-enforcing from within the convention. In order to support this argument, free-riding will have to be looked in the context of game-theoretical rationality. The following section consists of an analysis of free-riding from an informal game theoretical perspective, attempting to show how cooperation is maximising in repeated games between agents of equal rationality.

6.2 The prisoner's dilemma and the tragedy of the commons

Collective action and the provision of a cooperative surplus can be seen as an n-person

prisoner's dilemma. In the finite iterations two person PDs, the maximising strategy is to defect and the Nash equilibrium is always defect. Similarly, in a large group the Nash equilibrium is to defect. Any other strategy would imply that players do not behave rationally, or that their behaviour is being enforced. In the n-player case, even when interactions are perceived as infinitely repeated, the rational action would be to maximise individual utility by defecting before the other player does. By defecting, the rational individual can still receive his share of the cooperative surplus without contributing to its creation provided he is not caught.

6.2.1 Collective action games

The possibility of adopting a joint strategy during a game turns any interaction in a possible collective action game. The PD game can become a collective prisoners' dilemma implying that strategy decisions are decided by both agents. In a finite iteration PD game it is in each player's best interest to defect. In the context of collective action, rational agents maximise their utility when everybody else cooperates; or a sufficient number cooperate so as that there will be a cooperative surplus. Free-riding as an n-person PD game was discussed by Garrett Hardin in his *Tragedy of the Commons* (1968). In this story a group of farmers use an open access pasture to graze their animals. Given that everybody wants to maximise their individual utility, they will keep exploiting the land until it becomes useless for grazing. Individual maximisation contradicts collective maximisation; if everyone maximises, the group will be led to a Pareto-inferior equilibrium. The same principle applies when there is just one farmer with a high future discount rate. He will choose to maximise in the present and over-exploit the pasture, even if this means that he will not be able to use the pasture in the future (Taylor, 1987).

In this context, there are two variations of collective action failure. The typical one, where most individuals free-ride on the collective good, and the sole exploiter one where there is a single fully informed individual over-exploiting a resource, while the others continue to use the original allocation of the field. When the cooperators realise there is a free-rider they will also convert to free-riding but this requires full information for all individuals. Therefore, there can be an equilibrium where free-riding is limited to a small proportion of the population provided there is not complete information.

There are two weaknesses with the tragedy of the commons allegory. The pasture is assumed to be finite which is not always the case. Human history is

dominated by expansion and over-exploitation of new territories and inventions that make new methods of exploitation possible. A group of rational agents may very well decide collectively to exploit the pasture as quickly as possible and then move on to another one. Alternatively, there can be expectations for new inventions to replace the need for pastures or for more efficient methods of using the land.

A second criticism of the tragedy of the commons is the exclusive use of the PD game to describe collective action. Despite the fact that the tragedy of the commons is best described by the PD game, this does not mean that all collective action failure cases have the structure of a PD. Different games have different outcome matrices and thus different dominant strategies. The game of chicken can also be used to describe a class of collective action problems as effectively (Taylor, 1987). There are examples when the lack of a public good can be so devastating for a rational agent or a group of agents in the population that they will prefer to produce it by themselves and create positive externalities for the rest. Collective action failure can be examined through game structures such as chicken and assurance and not exclusively PD. The structure of the game pay-offs depends on the salience of the cooperative outcome and thus rational agents will not always avoid contribution.

In the previous chapter and especially in §5.1, Skyrms's (Skyrms, 2004) approach showed how the stag hunt can complement the PD game and be used more effectively to describe repeated interactions within a population. As he showed, both games can be used to describe similar situations and the same applies for collective action failures. A group of stag hunters are at risk of being contaminated by individual hare hunters, who in time will cause the collapse of the cooperative equilibrium. In the present context, the stag hunt may be more appropriate to discuss collective action failure as it takes into account the dynamics of a population and contains implications about the possible strategy changes over time. For instance, the social move from a stag hunt to a hare hunt happens over time and possibly over generations, which gives the time to rational individuals to reflect and evaluate their strategies and maybe adapt to the changing environment. Moreover, collective action can be seen as a hybrid game, where a player's strategies are those of a chicken game, and his opponent's strategies are those of an assurance game (Taylor, 1987).

It has been argued that in two person one-shot and finitely repeated games, cooperation or conditional cooperation is the maximising strategy, even without external enforcers (Taylor, 1987; Axelrod, 2006). Provided that others are disposed to cooperate

and have a small future discount factor, it is rational for one to cooperate. However, even when conditional cooperation leads to a cooperative equilibrium, there is no convincing explanation as to why, once at equilibrium, a rational agent will choose to cooperate instead of free-riding. Irrespective of how we choose to model interactions, in traditional game theory where individuals are utility maximisers, collective action fails unless there is an enforcement and punishment mechanism.

6.2.2 Solutions for collective action failure

There are two types of solutions to collective action failure. Solutions coming from within the group and solutions that are being enforced by a commonly accepted institution (Taylor, 1987). There is not much to be said about the latter. A government, of any form, is the common solution to most collective action problems. Free-riding becomes too costly only when there is detection and punishment and in this case free-riding becomes irrational. Internal solutions to collective action failure are not as straightforward to implement or as obviously effective. A group of rational agents that can foresee the advantages of cooperation and of the absence of free-riders may agree on the principle of mutual cooperation. Especially in small groups, it pays not to free-ride as detection and punishment by the other members of the group are likely and cheap; especially when interactions are repeated and take place within a small population, behaviour is known and can be punished. However the problem of enforcement persists in larger groups such as most modern societies. One possible solution is to view large groups as the summation of smaller ones. Most interactions take place among individuals who have a history of interactions, have interacted before and expect to interact again in the future. In this case free-riding is not a maximising strategy.

The above describes how existing social norms are sustained; most people cooperate with their immediate social circle but at the same time are disposed to free-ride when this will not affect them and their circle, and the likelihood of detection is small. As discussed earlier, in smaller groups behaviour is known and free-riders can be punished. Similarly, within given social circles behaviour can be known and free-riders excluded. Furthermore, Bergstrom (2002) has shown how in non-random interaction models such as the haystack, cooperative strategies become dominant; moreover and more famously, Olson (1965) has shown how smaller groups sustain cooperative equilibria as opposed to larger ones where free-riding is rational.

The tragedy of the commons is a good example in exhibiting how individual action and collective benefit are in conflict. At the same time, however, it is a simplifying description of social life and individual rationality. Agents, who play a one iteration PD game, reason differently than those who participate in a PD game with many iterations and maximise their utilities following different strategies. Real life interactions can be modelled theoretically as a PD game but this does not mean that they can always be accurately described by it. Similarly, social life and individual utility maximisation are modelled to an extent by the tragedy of the commons, but this does not mean that in the real world reasonable individuals would necessarily ignore the collective good when this is obviously linked to their individual utility.

These models show that individuals will prefer to further personal aims than collective ones if they are not constrained by exogenous enforcers. But the tragedy of the commons example ignores that individual rationality is defined by a plethora of parameters. For instance, some of the farmers in the pasture are likely to be neighbours that need to work together to drain a meadow (Hume, 2008). Maximising their utility in the pasture would have an effect on their utility at home. Even if the two utilities were somehow separated, each farmer would want to maximise his aggregate utility. That said, it seems more plausible to say though, that a person's utility function is one that consists of all the utilities derived from all his interactions.

Sugden (2004) argued that in repeated, n-person PD type games, behaviour is known and therefore free-riding is too costly a strategy to be adopted. Even in this case however, the maximising strategy depends on the strategies the majority of the population adopts; “in a world of nasties, just as in a world of suckers, cooperation never pays” (Sugden 2004: 125). Multilateral reciprocity, similarly to the Rawlsian notion of indirect reciprocity, calls for agents to cooperate with those who have a history of cooperation and defect when they interact with known defectors. The main premise is that all agents will prefer a cooperative to a non-cooperative equilibrium in a PD type game. However, the PD game distinction need not be exclusive. The value of Sugden's argument lies in the idea that rational agents look at others' past behaviour instead of guessing their disposition of relying on idealistic or metaphysical human traits; this applies to all human interactions and not just those that are possible to be described by a PD game. Even in the two person, one iteration PD game, collective welfare is maximised with cooperation. Cooperative behaviour in n-person, n-iteration games, is self-reinforcing: the greater the number of individuals who cooperate, the better the

outcome. Cooperation, despite being costly, produces a greater amount of cooperative surplus, and thus there is more to be distributed. Therefore, it is maximising for most to be at a cooperative equilibrium. Provided that there are enough cooperators, cooperation will be a stable equilibrium.

6.2.3 Conclusion

Collective action failure is a solid theoretical paradigm and free-riding is rational once we accept the premises of economic rational choice theory; given economic assumptions, collective action failure is a certainty. But theoretical premises in general and economic assumptions in particular are limiting when examining human behaviour and social structures. The absence of a government or a legal system would not lead to a collapse of social structures if these structures were based on pre-existing moral norms and conventions. Similarly, given adequate information about cooperative conventions, a defection equilibrium can move to cooperative one.

The purpose of the above paragraphs was to give a general description of the problem of free-riding and the subsequent collective action failure. Rational agents will free-ride if they know they cannot be detected and punished and if they interact in an environment where interactions are finite and their history is not known.

6.3 Free-riding within rational conventions

The above sections offered a description of the free-rider problem and the closely linked problem of collective action failure, while attempting to show that despite its significant strength as a theoretical paradigm it also has weaknesses. Rational agents will try to make the most out of every interaction, but at the same time their behaviour is influenced by established social norms and the behaviour of those around them. Given the unrealistic economic definition of rationality, free-riding is the maximising strategy. However, past social interactions and non-random repeated interactions within a social group, without loss of the assumption of rational agency, lead to different results.

The issue with free-riding is not whether cooperation in finitely repeated interactions is rational; it is not rational as long as the other party's history and disposition are not known. The question is whether it is possible to know at a high probability what others will do. Since maximising strategies and subsequent rational behaviour depend on the agent's environment at a given time, it is rational for one to cooperate when in a cooperative group and defect when surrounded by defectors. If

cooperation is somehow assured or interactions last for a sufficient number of iterations, it pays for one to cooperate. Cooperative interactions last longer and eventually yield a higher pay-off than free-riding. Rational agents who have cooperated in the past and expect to interact again in the future will choose to cooperate. Free-riders have to search for new cooperators after every interaction. If interactions with strangers in stable cooperative equilibria yield small pay-offs, free-riding is not as profitable by comparison with even finite cooperative interactions and the stability of the convention is reinforced.

Rational agents will follow optimal norms of behaviour when it minimises the costs of rational deliberation and bargaining. However, since we accept that agents within conventions retain their capacity for rational deliberation, we have also to allow for the possibility that rational individuals will consider free-riding. There is no reason why a rational agent will only deliberate at decision nodes or at specific time intervals. Committing to follow cooperative norms, even for short periods of time, is not an adequate measure against free-riding, as commitments can be broken if there is not an insurance mechanism within the rationale of the commitment in the first place. Driving on the right when everyone else drives on the left is obviously irrational. But running a red light late at night might be seen as rational.

There are two reasons why free-riding might not be rational when detection is unlikely. First the cost of being wrong is enormous. A defector risks not just his share of the cooperative surplus, but an exclusion from all future interactions within the established convention. This would force him to move to a new convention where defection is the norm or in cases of limited information, he can expect to interact with cooperators. Secondly, free-riding when in a cooperative convention requires costly deliberation that is avoided when following conventional rules. Rational agents will avoid deliberation about defecting as long as their current behaviour yields efficient outcomes and they have no reason to believe others will free-ride. Rationality dictates that individuals will maximise their utility by keeping their rational commitments, as new information retrieval is costly and the benefits from free-riding only temporary.

The structure of cooperative conventions can ensure their stability and enforce cooperation. Stable conventions of social behaviour are the outcome of repeated interactions that spread over a long period of time. Agents who have been accepted into the convention have shown their disposition to cooperate in a series of low pay-off interactions. Hence, they can be trusted to participate in interactions with higher

pay-offs. New participants in cooperative conventions are only allowed to interact in low pay-off interactions, free-riders will either have to spend time in order to reach higher pay-off interactions or settle for small benefits. The incremental increase of pay-offs ensures that possible free-riders are forced to behave conventionally for a given period of time so as to make free-riding eventually more costly. A history of successful cooperative iterations creates a type of trust and also acts as advertisement for agents' dispositions.

A convention is a stable equilibrium when there has been a large number of interactions within the convention and the actors that participate so as to ensure that everybody has devoted so much time participating in low pay-off interactions to make free-riding irrational. Skyrms has used the concept of the “secret handshake” (Skyrms, 2004:66) in a similar context. When communication between agents before interactions is possible and can be trusted, cooperation in PD type games becomes rational. Provided that cooperators can recognise others who are also disposed to cooperate, then cooperation becomes the equilibrium strategy. In order for this to be achieved cooperators use a secret handshake so as to exclude defectors. If an equilibrium is evolutionarily stable then even if there are defectors who imitate the secret handshake, they will not be able to destabilise it. By definition an evolutionarily stable equilibrium cannot be altered by invaders; in a population where a maximising strategy has been long established, interactions with agents who behave differently cannot influence the established behaviour. An incremental increase of pay-offs achieves the same goals but is more rigid. Defectors do not benefit from pretending to be cooperators because the number of iterations until a significant pay-off is reached is so high that it pays more to change into cooperating.

Furthermore, the secret handshake mechanism requires that individuals trust the signals they receive from potential interacting agents. If these signals are erroneous or are the result of an attempt to trick cooperators, then free-riders can maximise. However, this danger is minimised when agents choose with whom they will interact on the basis of their past behaviour. Therefore, it becomes highly unlikely that an individual who is disposed to free-ride by pretending to be a cooperator will be able to maximise in a convention where many low cost interactions are required before individuals are allowed to participate in high pay-off interactions. Therefore, the conventional account of social behaviour makes a disposition to free-ride very costly.

If an agent who has been behaving conventionally for the duration of his lifetime

comes across an interaction with huge pay-off he might still have the incentive to free-ride. In that case of course he will be risking much more than the ordinary free-rider; he will have spent much more time and he will have missed opportunities to free-ride in the past by committing to cooperative behaviour. As difficult as it might be for one to change long-term habits which are the outcome of costly rational deliberation and interactions, we should allow the possibility that rational agents have the capacity to do so. The argument against a behaviour shift of this sort is two-fold: the threat of detection and punishment has to be plausible enough and secondly the cost of changing one's behaviour will have to include information gathering which again will make the shift more costly. These parameters reduce the benefits from free-riding and are indirectly proportional to the length of previous interactions.

Individuals who maximise by behaving parasitically will not be able to benefit from a population following a cooperative convention. In this sense, straightforward maximisers will be excluded by cooperative conventions, even if it takes a number of interactions for their behaviour to become known. However, an honest cooperator may change his mind about the rationality of his cooperating behaviour. Rational agents who maximise by following cooperative conventions can realise that it is more beneficial for them to free-ride. This change of behaviour could be due to an “enlightenment”, a change of the agent's rationality. An endogenous change might make the individual realise that his “new rationality”, the newly adopted strategy, will lead to maximisation as opposed to the old one. However, this is only realistic when an agent has perceived a change in the environment that dictates a change in behaviour; individual rationality only changes when the environment or information available changes. Changes in preferences (for example, as a result of illness) do not need to mean a change in behaviour as the possibility of their realisation can be part of the agent's original life plan; although illness cannot be predicted, one can deliberate about how she would behave in extraordinary circumstances. Just as in *MbA*, one decides to become a CM because one realises that it is maximising; it is a life-changing decision and one that a rational agent is expected to follow throughout her life (assuming the existence of sufficient numbers of other CMs). In this case, given equal rationality and access to similar information within the convention, others should be expected to follow. Despite the fact that this agent may be viewed temporarily as a free-rider, his exclusion will only last for the duration of the shift to the new convention.

Detection of free-riders is paramount in the conventional account as well. Any

rational individual would rather free-ride when he is certain or almost certain that he will not be caught. The only assumption needed here in order to ensure detection is that free-riding becomes known, even when the free-rider himself does not. Within the context of conventions with equal information and repeated interactions amongst the same members, it is then plausible then to say that free-riding and free-riders can be detected rather easily. Therefore, the important thing is not that the free-rider is detected within the convention, but that his act is. The scenario of a hyper-rational free-rider taking advantage of different cooperators repeatedly is highly unlikely; within the convention there is information about the history of interactions and a free-rider will eventually be found out by taking into account past interactions.

A stable equilibrium that serves as the status-quo has to have developed a method of excluding or converting most free-riders. At the same time small numbers of free-riders cannot threaten the stability of a convention. Therefore, from an evolutionary standpoint, they can still exist within a stable convention consisting of rational agents, even when they are predominantly cooperators. In a rational choice theory context, though, stable conventions have to have established methods that make free-riding irrational. A high rate of detection and a credible threat of punishment would make the risk of free-riding too high to take. Information transmission through small groups make detection more plausible. Conventions do not have to be limited in small groups; information however becomes available from within small groups, where it is more readily available. For instance, information about a tax evader can become available through other conventions in which the tax evader participates and not exclusively through the tax agency. A tax evader will not be trusted by cooperators even if they are not directly harmed by his behaviour. It is not necessary for cooperators to avoid interactions with the tax evader as a form of punishment; they will do because it benefits them to avoid interacting with agents who have shown to be disposed to defect.

Being a rational maximiser, it will be too risky for one to interact with someone who is a known defector. Therefore, a free-rider will have to take these possible forms of punishment into account and calculate his utility from free-riding. Should these be deemed too costly, and given that free-riders are also utility maximisers, he should choose cooperation. It is important to note that the fear of detection in this case is as important as the possible punishment itself. As long as there is history of punishment of free-riders and detection rates are high, the threat of a punishment itself should be a strong enough incentive for cooperation. Rational agents who are disposed to cooperate

will not interact again with free-riders as this would not be maximising. By not interacting with them again they punish them, even though this might not be their primary intention. Therefore, an independent punishment mechanism may not be necessary. As long as individuals look to maximise through interactions, free-riders will be excluded from cooperative conventions and thus be punished. Punishment in this sense is a side effect of rational behaviour and works both as reward for cooperators and threat for potential free-riders.

To conclude, the availability of information and knowledge of the history of interactions can make free-riding too risky and possibly too costly to be a maximising strategy. In this context therefore, a rational individual has the responsibility to stay informed and to exclude defectors from future interactions.

6.4 A response to the Foole

The counter-argument to the above understanding of conventional behaviour would be that in cases where detection is very unlikely, difficult or costly it still pays to free-ride. In these cases a theory where moral behaviour or a version of constrained rational behaviour has intrinsic value is superior. A rational agent with parasitic behaviour can take advantage of one cooperative convention for short periods of time before moving on to another one. Hence, following a set of predefined moral principles that make free-riding unacceptable would make a stronger case for cooperative behaviour. A social arrangement according to which free-riding is always morally wrong is preferable to one where individuals ought merely to be rational. According to this understanding, arguing for moral behaviour has to include some kind of categorical imperative; a rule that is always true and should always be followed.

The absence of moral rules to regulate behaviour is related to the argument made by the Foole in the *Leviathan* (Hobbes, 1976). The Foole argues that it cannot be rational to comply with an agreement once the other party has fulfilled his duty. In essence, the Foole argues that there is a need for moral principles to regulate social interactions and that rationality alone is not enough. Thus, the Foole is purely individualistic and selfish and for him being rational cannot be reconciled with the possibility of adopting a joint strategy (Gauthier, 1986). The Foole would accept that in a world where rational agents are disposed to constrain their maximisation, it is rational to appear as a constrained maximiser in order to take advantage of them, “[f]or then he would not be excluded from the cooperative arrangements of his fellows” (Gauthier,

1986: 173).

The Foole's argument, as presented by Gauthier, supports the rationality of free-riding and moreover proposes that only free-riding behaviour is rational. In *Morals by Agreement* Gauthier introduces disposition translucency to counter the Foole's argument; an ideally rational agent "is directly aware of the disposition of his fellows" and hence "[d]eception is impossible" (Gauthier, 1986: 174).

The problems with disposition translucency have been analysed in previous chapters. Here, it is used so as to show that the free-rider's argument is null if dispositions are known. In the conventional account, dispositions can be known through an agent's history and past interactions within a convention. As a result, an argument very similar to Gauthier's can be made, with the bonus of being more realistic and plausible against the Foole's rationale. To reiterate, the Foole argues that although it might be rational to make agreements it is not always rational to keep them and hence compliance requires moral and not exclusively rational constraints on individual maximisation (Gauthier, 1990). As discussed previously, even the Foole needs a society where compliance is the norm; otherwise he would not be able to maximise. Free-riders need a society of cooperators in order to be able to free-ride.

The problem with this approach remains that it has to be based on pre-existing ideals or circular arguments. Suggesting that an agent who is disposed to free-ride should act cooperatively because it is the right thing to do has no force, having assumed that individuals are rational. Objective moral rules can only be meaningful when they are the outcome of individual rational deliberation. Anything else would have to limit the premises of individuals' rational capabilities and in general of the assumption of rational agency.

The conventional account can be seen as a weaker argument against parasitic behaviour than the ones offered by Hobbes and Gauthier since it requires explicitly rational motivation. But it is an account that allows for rational deliberation and is based exclusively on the capabilities of human nature. An agent acting conventionally would probably see the benefits of benefiting at the expense of others, once they have kept their side of the agreement. However, when real life restrictions are taken into consideration, costs of deliberation and of collecting new information for not acting conventionally, would be an incentive for conventional behaviour.

Even when there is consensus within a social convention or social contract as to how one ought to behave, this consensus is based on pre-existing ideas of right and

wrong. In this context then, our understanding of morality is biased. It is derived from specific cultural, social and psychological parameters; or put differently, pre-existing social conventions. Thus, it is difficult to convince someone from a different background that a certain set of moral norms is preferable to his and so there is little value in arguing over competing accounts of morality. The understanding of morality presented here is the outcome of rational interactions. At some point in history moral rules were a solution to problems of social interaction and by adopting them, a given society avoided conflict. These moral rules, although not necessarily the most efficient, solved more problems than they created. Over time the need for some moral rules may have become extinct. However, information costs and/or limited availability of information in a given environment, may have preserved rules without any practical use.

Moral imperatives can only be objectively justified if they are viewed from a rational perspective. The golden rule of helping those in need is no more powerful than a law calling for the amputation of a thief's hand unless we can base either one on rational arguments about individual maximisation and social welfare. Both these conventions of social behaviour have been stable and in this sense are rational given the specific environmental parameters. However, in order for different moral conventions to be compared they have to be based on notions of rationality. As has been discussed previously, rationality itself also depends on environmental parameters. Hence, moral conventions that have developed in similar circumstances are more easily reconciled than conventions that are fundamentally different.

In conclusion, the Foole's assertion that there can be no justice based on rationality (Hobbes, 1976) is to an extent correct; there can only be a justice as the outcome of repeated interactions between rational maximisers. Different conventions prescribe different accounts of justice depending on topical parameters. Thus, the characterisation of principles of justice is affected by local established social conventions. Since rational behaviour is bounded by these same social conventions, principles of justice that are accepted by rational agents are neither static nor universal.

6.5 Conclusion

The aim of this chapter was to examine the rationality of free-riding and discuss the contradictions arising from assuming rational, mutually unconcerned individuals who interact in a society that is “a cooperative venture for mutual advantage” (Rawls, 2005:

4). The conventional account in rationality and social interactions addresses some of these issues more efficiently than contractarian theories like the one introduced in *Morals by Agreement*. In addition, free-riding is less problematic in the context of the evolutionary account of social behaviour than it is for theories that are based on individual rationality without taking into account social dynamics. In this respect, a theory of social conventions that are seen as evolutionarily dynamic is better placed to deal with problems of collective action.

In the conventional account, dispositions are known since agents' histories are known. In addition, repeated interactions ensure that free-riders are punished through social exclusion and also that those who do not have a long history of cooperative interactions are only allowed to participate in interactions yielding small benefits. The gradual acceptance of agents in cooperative conventions makes it costly for agents with parasitic behaviour to take advantage of cooperators. Moreover, long-term cooperators who realise that their cooperative behaviour is no longer maximising are bounded by the same environmental parameters as their neighbours and thus one should expect that the strategy shift will eventually take over within the convention. Therefore, they only free-ride for a transitional period until the new strategy becomes the norm. In this understanding, social conventions serve as mechanisms of information sharing, which makes detection of free-riders or strategy shifts easier.

Hobbes's Foole is the archetypical free-rider. By arguing that it is impossible to argue that it is rational to comply always with one's agreements, he argues that rationality and morality cannot be reconciled and that a rational agent by definition should always make the most, disregarding others' benefit. Unless we accept moral constraints on rational maximisation or idealised assumptions such as disposition translucency, the Foole is right. However, Hobbes did not take into account that social interactions are repeated and free-riders can be found and punished, not necessarily by a government but by being excluded from future cooperative ventures. Therefore, rational interactions that are repeated, in an environment where information about individuals' history can be known, can lead to a mutually beneficial state. In other words, the Foole has rational incentives to cooperate as long as he knows that this action will be known and he expects to interact with the same people again in the future. Thus, the conventional account of social interactions can describe how rational individuals can interact for mutual advantage and reach a just outcome.

Given that rational strategies depend on environmental parameters and within a

convention agents have access to roughly similar information, we should expect that rational conventions are also conventions of justice. The next chapter will attempt to show how rational conventions being conventions of mutual advantage, also ensure the presence of justice. The rationality of the individual in combination with the evolutionary account of the establishment of conventions allows for a more flexible and realistic account not only of rationality but also of the subsequent conventions. Within evolutionarily stable, Pareto-efficient conventions a behaviour that is commonly accepted as moral, can be grounded on purely rational premises more efficiently than in the account presented in *Morals by Agreement*. In addition, the next chapter will focus on justice as mutual advantage, which requires rational interactions and is based on the assumption of rational agents. However it is problematic in respect to interactions between individuals that are extremely unequal, especially in the context presented in this and previous chapters of evolutionary conventions and mutually unconcerned agents. Therefore, it is essential to show in the next chapter that there can be a variation of the theory of justice as mutual advantage that includes the weak, without abolishing the assumption of individual rationality.

7. Social conventions and Justice

The priority of the individual over the collective that has been assumed and implied in the previous chapters has direct implications for the importance of individual rationality in this thesis. Assuming individual rationality and accepting it as the cornerstone of the argument presented here, implies that the only account of a theory of justice that can be justified as authoritative in this context is justice as mutual advantage.

The discussion in Chapter Six mostly dealt with the rationality of free-riding behaviour and examined whether free-riders, being rational agents, will accept conventional limitations on their behaviour. Individual moral behaviour should lead to a just society. In the following sections, rational agents will be shown to have incentives to adopt strategies that lead to a just outcome without loss of their rationality.

Rational strategies depend on the given environment – the information available and other agents' dispositions. Therefore, justice as mutual advantage, which is based on agents' rationality, will depend on the specific environment. Given the environmental limitations on agents' rationality there can be variations on the mutually accepted point of agreement. In other words, the central role rationality has in justice as mutual agreement makes it possible to have various just outcomes depending on environmental variety. Rationality will in a sense serve as the connecting point between conventional and moral behaviour. First, we will have to look at theories of justice as mutual advantage in contemporary political philosophy. This section will be followed by a discussion of justice in a conventional framework. The final section will examine what seems to be most important and challenging aspect of justice as mutual advantage: its requirements for the weak and the possibility of their inclusion in a rational social convention and by extension the social contract.

7.1 Rationality and Justice

The concept of justice within political and moral philosophy has various meanings and definitions. Depending on the accepted understanding of justice there are implications for its circumstances and rules. Broadly speaking, justice refers to constraints on rational maximising behaviour (Barry, 1991). These constraints may be externally enforced by others or be part of an individual's rational deliberation, as in Gauthier's theory. In the conventional account, social structures impose these constraints by reinforcing the behaviour that sustains them. Within established social conventions this

is in essence a cyclical relationship; rational conventions reinforce just behaviour by promoting conventional behaviour, since agents who behave conventionally are rewarded through more beneficial interactions whereas free-riders are punished by exclusion. Agents within maximising conventions will continue behaving conventionally, thus reinforcing the conventional structures that bound rational behaviour and thus implicitly constraining individual maximisation.

7.1.1 Justice

If we accept the moral priority of the individual over the collective or the outcome and the central role individual rationality plays in interactions, then it is imperative that justice at the very least includes some form of constrained maximisation. The assumption that individual rationality is the main motivation of behaviour means that there is little room for third party enforcement that violates the individual freedom to maximise.

According to Hume (2008) the need for justice arises when there is conflict of interests, whereas for Hobbes justice is keeping an agreement and also the impartial behaviour of a commonly accepted arbitrator or institution (Barry, 1991). Thus, justice can be described as a mechanism to resolve conflicts in a commonly accepted manner. In a situation where any number of individuals cooperate for the production of a surplus, a common understanding of a just distribution of this surplus is needed in order to avoid or resolve conflict. However, the method of deciding on the distribution of the surplus has to be accepted by all participants in order to be considered just. In order to avoid conflict after its production, all parties have to have agreed on the rules of justice. In a different understanding, rules of justice are required when the existing state of affairs can be seen as disadvantageous to some. In a distribution that is not proportional to the perceived contribution or that was based on coercion for instance, the disadvantaged can call for justice, the need for a distribution based on just premises. Just rules are to be used to correct any injustices in the status-quo.

From the above, we can see there are two, closely linked, uses of justice. Distributive and corrective justice. Distributive justice refers to the rules that are to be used to agree on a distribution of the cooperative surplus and the conditions under which it is produced, and corrective justice to the rules to be used to correct a status-quo that was reached by coercion or any other process that can be deemed unfair by one party (Alexander, 2007). Rules of justice can be derived using similar premises and

therefore both distributive and corrective justice can be based on the same principles. Hence, it is more important to establish the procedure with which justice is derived than the content of just rules.

Justice can refer to a just outcome of interactions or to a set of rules of justice according to which these interactions take place. An allocation of resources may be seen as just if it occurred in accordance with just rules or if it is considered just itself (Gauthier & Sugden 1993). Given the importance that has been attributed to individual rationality in the previous chapters, the nature and structure of interactions is central. Skyrms (2004) showed how social structure equilibria can be contagious; a maximising behaviour observed in a social group can infect individuals and hence shift the equilibrium point in a given society. In this manner the cooperative or non-cooperative equilibrium in a social group can influence the equilibria in the neighbouring social groups, depending on the dynamics of the existing social structures.

As discussed in Chapters Four (§ 4.4.2) and Five (§ 5.1), social structures on a local level are interdependent on the structures that define the social contract as a whole. Thus, a majority of cooperative social structures should sustain a cooperative social contract and vice versa an established cooperative social contract supports the circumstances for cooperative equilibria on a topical level. The same principle applies for just behaviour and equilibria of justice in social groups. A set of interactions among rational individuals that is commonly accepted as just, and therefore mutually beneficial in outcome, should lead to just outcomes on the level of the social contract. Rationality ensures that the justice of the rules of interactions will lead to a just outcome.

In this context, there is no need to distinguish between justice and fairness. What is accepted by rational agents as fair practice will lead to a just outcome. Interactions that are deemed fair, or mutually beneficial will be repeated and create rules of behaviour that are very likely to establish social conventions of behaviour. Therefore, in this sense, a stable social convention has to be a just social convention, or put differently, a social convention that is built on just rules. Interactions that leave all participants better off and satisfied with their share of the surplus, have to be repeated and will form stable structures of interactions, since we have assumed rational agents whose aim is to maximise their utility. In the same vein, interactions of this type are also just; everybody is better off as a result of the interaction and is willing to participate in similarly structured interactions in the future. Therefore, justice could be defined as a Nash equilibrium that is Pareto efficient (in the absence of coercion). In addition, a

Nash equilibrium that is not Pareto efficient can be a just situation provided that the equilibrium is a Pareto improvement over the status-quo. Hence, justice can be seen as a comparative measure and not exclusively as an ideal state of social affairs. A just situation can be improved by a more just one (Sen 2009) and in this respect, justice can be seen a process rather than a single equilibrium point.

Whether we view justice as an ideal point in history, a hypothetical structure to help us understand social behaviour, or a process of reaching a better society, it is essential that there is no form of coercion in a situation of justice. The absence of coercion is paramount, not because coercing someone to act against her will seems wrong, but because it cannot be reconciled with the assumption of equal individual rationality. If there is a need for one to be coerced to act in a certain way, then obviously her rational deliberation leads to a disposition to act differently. Rational interactions require both parties to be rational and behave in way that they think maximises their utility. Moreover, it is assumed that a rational agent would not start an interaction voluntarily if she knew she would be coerced to act in a certain way. Therefore, any just situation cannot be derived from or include coercion. In addition, when punishment takes the form of social exclusion, as is the case in the account presented here, there is no need to use coercive force for agents who behave unjustly or do not comply with an agreement. However, there has to be a type of coercion at least by way of protecting conventional equilibria of rational individual action. In this respect, we could distinguish coercion to force an action from coercive force that protects the possibility of an action taking place.

7.1.2 Justice as mutual advantage

Asserting that individuals are rational utility maximisers implies a specific understanding of justice. Justice as mutual advantage requires and assumes rational agents since it “begins with fully informed individuals...who are driven to pursue their own self-interest” (Matravers 2000: 156). An agent participates in an interaction only if she expects to benefit from it. If all interacting agents do the same, there will be a set of interactions that promote each individual's benefit and therefore the common good. However, the assumption of rationality is not as important for justice as mutual advantage as the one of equal rationality. If one agent cannot take advantage of others because of a type of hyper-rationality, then agents who do not behave strictly rationally do not affect the structure of a theory of justice as mutual advantage. Justice as mutual

advantage describes a bargaining procedure between rational individuals that leads to a commonly accepted outcome. Both agents have more to gain by participating in the cooperative procedure than they lose by constraining their self-interest.

Justice as mutual advantage assumes that individuals with conflicting interests participate in the production of a cooperative surplus which is mutually beneficial (Barry, 1991; Gauthier, 1986; Vanderschraaf, 2011). Given rational agents and absence of coercion, all voluntary interactions should lead to just outcomes. Hence, the connection between justice as mutual advantage and individual rationality is rather straightforward. Justice as mutual advantage by definition requires rationality as well as the absence of coercion. Moreover, if we accept that individuals are self-interested utility maximisers then the only account of justice that is meaningful is justice as mutual advantage.

It is essential that commonly accepted rules of justice contain a method of enforcing them. The only rational incentive for an agent's participating in an interaction is that her share of the cooperative surplus will increase her utility such that it will be greater than before, or in the absence of the interaction. Free-riding threatens cooperation when agents' future discount factor is very small, or when the distribution of the cooperative surplus is not seen as beneficial for all participants. Rational agents should be able to agree on such a rule based solely on their rationality. And rationality makes it essential that the commonly accepted rule must also be commonly beneficial. Conventional rationality as discussed earlier means that agents within a convention have access to similar information and are bound by similar environmental constraints. Therefore, their rational deliberation has to lead to similar outcomes with respect to maximising strategies. And in this respect they will agree on what is a just outcome and on the rules to enforce it.

The above discussion is based on the assumption that rationality and mutual advantage are linked; interactions among rational agents must lead to mutual advantage. Given rational agents, mutual advantage is defined by a Nash equilibrium that is also Pareto efficient. This in turn is the definition of justice. Rational behaviour depends on an agent's environment, others' expected behaviour and available information. Therefore, the possibility of justice depends on the environment which is primarily defined by agents' dispositions as revealed by their history. An interaction that leads to a sub-optimal Pareto equilibrium can be as just as an interaction that leads to Pareto optimality. Always defect is as just as always cooperate when we take into account

environmental parameters.

Gauthier, based on Hume, reached a similar conclusion: “In a world of Fooles...it would not be rational to be moral” (Gauthier 1986: 182). This can be problematic when the goal is to derive morality from rationality, as well as when it is asserted that justice can be the outcome of rational behaviour. In *Morals by Agreement* the problem is addressed by postulating that it is rational to be a constrained maximiser.

In a game theoretical framework, the issue of deriving morality from rationality or even linking the two takes the form of a discussion relating to the equilibrium selection problem. Both in game theory and in evolutionary game theory rational behaviour and evolutionary dynamics respectively lead to stable equilibria; the same applies to the conventional account presented here. The related literature (Sugden, 2001; Binmore, 1998) proposes various solutions to the problem of equilibrium selection by introducing aspects of individual morality or rationality in the analysis as opposed to rely exclusively to social dynamics. Traditional game theoretical models of repeated interactions show that there are many stable equilibria but do not propose a convincing mechanism for selecting the optimal, or even for distinguishing amongst them.

7.1.3 Conclusion

The following section will deal with the equilibrium selection problem in the context of repeated interactions between rational agents and the conventional account. Given inter-conventional communication and availability of information, an equilibrium that complies with the premises of individual rationality, as described in Chapter Three, must also comply with the principles of justice. The equilibrium selection problem relates to traditional contractarian theories in that it sets a different basis for examining and discussing the status-quo and the contract point.

7.2 Equilibrium selection

The theory of justice suggested in *Morals by Agreement* and the typical contractarian paradigm require a status-quo where the bargaining procedure starts and a description of the bargaining process itself. However, this implies a static description of interactions. Information and preferences remain the same during bargaining and depend on the original position. The conventional account of behaviour proposes a more realistic account of human behaviour. Information availability changes as more interactions occur and therefore preferences change accordingly. It has been suggested that

bargaining is part of the game that models interactions and therefore should not be seen as a separate set of interactions (Binmore 1998). In the following, social interactions will be assumed to include the game of bargaining and therefore it will be assumed that a single game describes all interactions, before, during and after the bargaining process.

7.2.1 Justice in conventions

In the conventional account of contractarianism, bargaining is not an essential requirement for a stable equilibrium. Individuals interact and reach a de-facto convention of behaviour that does not have to include a formal agreement or be the outcome of bargaining. Given rationality, a convention of behaviour that is not seen as mutually beneficial cannot be accepted by all participating parties and therefore cannot be a stable equilibrium. Similarly to rational theories of justice such as Gauthier's, all participants have to be benefited by the formation of a convention. But unlike the bargaining process in *Morals by Agreement* there is no need to make claims or to bargain. Actual behaviour and past interactions replace the need for promises and promise-keeping and interaction outcomes show whether equilibria shifts are possible. There is no need for one to promise to abide by the existing social contract rules; one's past behaviour will serve as a measure of one's acceptance in the social contract and the incremental increase of the benefits associated with interactions makes it too costly for new members to take advantage of the cooperative surplus. Thus, in the conventional account individuals are accepted or rejected by interaction structures based on their past behaviour and the development of the convention is incremental in order to allow for the exposing of defectors. That is in direct contrast to the theory in *Morals by Agreement* where rational agents are expected to guess others' dispositions before internalising constraints on their behaviour.

For Gauthier, given certain conditions, constrained maximisation leads to justice. In a group of constrained maximisers justice requires adopting strategies that are characterised by constrained maximisation. Interactions among rational agents who have internalised constrained maximisation leads to just outcomes. Similarly, in the conventional account presented here, justice requires agents to behave according to the rules of the established convention. Having assumed rational agents and absence of coercion the established convention will be one that promotes individual utility maximisation which in turn leads to group welfare maximisation. In this context, justice as well as rationality require that one accepts the established conventional rules, which

are the result of repeated interactions. For instance, in the divide the cake example the two agents can experiment with repeated divisions of the hypothetical cake. The fair division will be reached once both agents are satisfied with the outcome, irrespective of their share (Vanderschraaf, 1999). Agents acting according to the conventional account would experiment with small portions of the cake and would only actually divide the cake once they had reached an understanding on how it should be divided. Therefore in a sense bargaining still takes place, but just not as described in bargaining theories. Instead of claiming and consenting, individuals' actions show the most sustainable equilibrium point.

The main difference between the conventional description of the bargaining process and typical evolutionary game theoretical explanations is that evolutionary game theory does not take into account repeated interactions among the same agents. In Skyrms's analysis (2004) there is high correlation between interacting agents, but this does not imply that most interactions occur among agents who remember each other and their past. Therefore, this analysis does not take into account the possibility of repeated interactions in an evolutionary dynamic framework which is what makes constrained maximisation rational. In the conventional account equilibria are selected in roughly the same manner that evolutionary game theory suggests; but interacting agents have a memory of their recent interactions and decide their future behaviour based on their past. This will create dynamics of developing equilibria as similarly behaving individuals will be drawn together and form social conventions.

If the evolutionarily stable behaviour in divide the cake situations is to claim half (Skyrms, 2004), accepting agents' rational capabilities makes it possible to explain other non-equal distribution fairness norms that have been established and are present in many cultures (Binmore, 1998). For instance, the surplus is to be divided equally, but remembering that the last time the rule was applied in a division between a healthy adult and a sick infant will create a non-equal division equilibrium. Given that the health of the adult is not threatened significantly by claiming less and the infant will have an increased probability of surviving by claiming more than half, it is reasonable to assume that in similar interactions in the future the same behaviour will be followed. The marginal cost of claiming less for the adult is much less than for the infant and can be seen as incremental. Furthermore, it can be asserted that the survival of infants serves as insurance for adults. Hence, both parties maximise by following a non-equal distribution which is rational and efficient and thus easier to be replicated. Therefore, in

this sense the established commonly accepted behaviour at the eventual equilibrium point for each social group will depend on the specific circumstances. There can be equilibria that demand equality as the only form of justice and others where justice is what is required by the most powerful. All of these conventions may be stable and long-standing. A rational comparison of the various equilibria will show that the most efficient one will be replicated. In order for this comparison to take place rationality is essential, which in turn implies that information about other conventions has to be available. Otherwise, each isolated convention that has reached a stable equilibrium can also be said to be a just convention.

7.2.2 The equilibrium selection problem

The question that arises from the rationale of justice as equilibria of repeated interactions is how one ought to choose among competing stable conventions. However, this not quite the right way to put it; to an extent individuals do not choose among competing equilibria but among competing social structures that lead to social equilibria. And these structures as well as the individuals' maximising strategies depend on evolutionary dynamics. Therefore, evolutionary dynamics in conjunction with past established conventions define the set of equilibria that are available for selection at the topical level of social conventions and individual action.

The differences among various conventions can be explained by “factors such as historical precedents and accidents that lie outside of the description of the coordination game” (Samuelson, 1998: 6). At the same time, coordination games such as the driving game have two equilibria that are both Pareto-optimal and therefore rational agents are indifferent between them. Therefore, in this type of game it is irrelevant which equilibrium will be chosen. Similarly, in non-cooperative games an optimal equilibrium selection mechanism is essential for a just outcome. However, it is not imperative that the chosen equilibrium or the mechanism of selection is included in or described by the game parameters. When there are several stable equilibria that can be seen as just, it is difficult for a rational agent to select one and then to risk change if she is already in a convention. Obvious Pareto superior alternatives can mean that informed rational agents will change their behaviour so as to reach them. Hence, the availability of information is central to the feasibility of individual behaviour shifting the evolutionary direction and indirectly influencing the selection of an equilibrium. Selecting an optimal equilibrium requires information about other social states and a possibility of inter-conventional

comparison.

As implied earlier, when there is no information available in a given society about a Pareto superior convention then their existence does not play a role in accessing the status of social structure equilibria or the social contract. Asserting that a social state is just or not, is only affected by other possible equilibria that are known. For instance, a hypothetical society that has reached an ideal social contract can only be imitated by other societies if there is information about its status. An isolated ideal social contract cannot have an effect on the evolution of other social contracts or their classification as just or unjust. The selection of an optimal equilibrium is influenced by equilibria in other social groups, but for this to happen, information spreading and communication between conventions are essential.

A stable Nash equilibrium convention that is weakly Pareto efficient but not strictly optimal is still considered just, given the present information restrictions. In many respects the chosen equilibrium, or in other words the established convention, is the result of a plethora of historical and environmental factors and thus can be seen as random. A social contract that is represented by the game of life (Binmore, 1998) is affected by too many parameters to be included in a meaningful way in a formal description of interactions. However, given individual rationality and a reasonable degree of information availability it is possible to compare conventions and social contracts and decide which is closer to Pareto efficiency. Information availability and spreading becomes feasible and a plausible assumption in modern societies. At the very least it is much easier than it used to be; the effect of technological developments on information availability in contemporary societies will be examined more analytically in Chapter Eight. The conclusion from the above is that comparisons among various conventions and their efficiency is realistic and can lead to change in sub-optimal social conventions.

Equilibrium points depend on previously established equilibria, which are to be seen as established conventions. In turn these established conventions give rise to related types of new conventions. Therefore, the ideal convention is social structure specific and depends on the convention's historical evolution. Different societies and social groups will be more effective in various social conventions. In this context, there can be no universal ideal convention and Pareto optimality can be very different if we take into account the topical histories and cultures. Established conventions that facilitate justice in social interactions are therefore the result of previously established

social conventions. In other words, our understanding of fairness depends on group or society-specific norms that in turn lead to the selection of new equilibria of justice. For instance, in the meeting game Adam and Eve are to meet either at a boxing match or the ballet. If in a similar situation in the past they went to the boxing match, which maximises Adam's utility, reciprocity dictates that next time they will go to the ballet. Reciprocal behaviour is viewed as fair in most human cultures (Binmore, 1998), however that does not mean that it is generally fair or that it can be used as a justification for just behaviour in more complex and more realistic situations. In the meeting game both alternatives are stable Nash and Pareto-optimal equilibria; but in order to decide which corresponds to justice we have to take into account environmental parameters that are part of the game that represents the topical social (or for Binmore the game of life), but not of the specific game.

Using evolutionary game theoretical language means that “groups using a Pareto-superior equilibrium will therefore grow in size or number at the expense of groups using a Pareto-inferior equilibrium. Eventually the inferior groups will disappear” (Binmore, 1998: 204). Societies and social conventions that do better expand at the expense of those that do not perform as well. Performance in the context of cultural evolution can include anything that implies better living conditions for its members. Doing better does not have to be an exclusive or specific definition; it refers in essence to utility maximisation, but in real terms it can include being more peaceful, more affluent, or happier.

The difference between biological and cultural evolution is that human societies do not have to grow bigger in size or number to prove their evolutionary superiority. Social groups that are on Pareto-superior equilibria, on conventions that help more individuals maximise, attract agents who want to maximise. If one sees that one's neighbours (that can include neighbouring house, village, country or any other situation where information about different social conventions is available), do better, then it is only rational to try to imitate their behaviour and their customs if it is not possible to join their group. This change of behaviour happens obviously at the expense of Pareto-inferior social conventions that are gradually led to extinction. This description borrows terms from biological evolution theories, but it relies on individual rationality. Groups do not interact and do not expand in the sense that individuals forming these groups follow certain social conventions and thus cause changes on a group level. In this sense evolutionary accounts of justice can be compatible with justice as mutual

advantage. Conventions of behaviour that develop in groups and societies are the result of interactions among rational individuals. The same rational individuals are responsible for selecting among stable equilibria and remaining informed about the available alternative equilibria.

At the basis of the evolutionary rationale of social interactions remains the possibility of trial and error efforts to maximise. A rational agent who is informed about a seemingly better equilibrium can try its established behaviour in interactions with individuals of her convention in order to determine whether it is actually an improvement on the status quo. Given that information is roughly equal in her convention, one can expect others to imitate her. This description resembles Binmore's use of the concept of memes (Binmore, 1998; Dawkins, 2006). However, Binmore's account refers to the evolutionary account of interactions and is not based on the assumption of individual rationality. However, broadly understood “a meme is whatever gets replicated when people imitate their more successful neighbours” (Binmore, 1998: 294) and individuals in this context can be taken to be characterised by imperfect memory and processing ability which is in accordance with the assumption of bounded rationality. Thus, efficient conventions are replicated similarly to memes, without loss of agents' rational capabilities.

7.2.3 Conclusion

Discussing rational interactions is only useful in the absence of coercion. In cases where power differences are significant or in any case too great to allow for rational interactions, a rational agent can find it beneficial to maximise by forcing others to comply with his will. The discussion so far having assumed agents of roughly equal rationality and an evolutionary understanding of social structures, excluded the possibility of coercion and the existence of individuals who are significantly weaker than most others in a social group.

The main criticism of justice as mutual advantage is that it does not provide a convincing incentive for rational agents to be moral in cases of extreme power inequalities. The following section will discuss the possibility of providing a rational justification for moral behaviour in these cases by focusing on the vulnerable within a society and the requirements a theory of justice as mutual advantage can make of rational, mutually unconcerned agents.

7.3 Justice and the vulnerable

Plato's parable of the ring of Gyges exhibits the problem of justice very nicely. A person with a ring that makes him invisible is not bound by the same rules as others. The ring makes him immensely more powerful and he does not have anything to gain by constraining his maximisation. It creates an extreme power discrepancy such that justice as mutual advantage does not provide an incentive to behave morally. A theory of justice however is especially, if not exclusively, useful in case of extreme inequalities; “among equals, morality would be a necessary evil” (Gauthier, 1986: 315).

The following will discuss the requirements a theory of justice has in terms of including in the social contract those who cannot contribute. It will begin by addressing the question of who is to be considered weak in modern societies. This will be followed by an examination of how a theory of justice as mutual advantage can be expanded to include non-contributors.

In social contract theory handicapped individuals of various degrees are considered weak. A blind person and a person in a comatose state are both weak but at the same time very different in terms of their potential contributions to the social contract. The next subsection will focus on individuals who are severely disabled or temporarily not contributing (such as the very old and the very young) and discuss how the definition of vulnerability depends on social structure and not just on individual contribution. The requirements of justice as mutual advantage for those who cannot interact with the world or the mentally insane, who cannot contribute at all will be addressed in the following subsection where justice as mutual advantage and the vulnerable are examined more analytically.

7.3.1 The vulnerable

Vulnerability depends largely on social structures and also on technological means. In other words, the assessment about who is vulnerable is linked to a given social, physical and technological environment. A physically disabled person would be more vulnerable as a hunter-gatherer than in a modern society. Similarly, a dyslexic child would not be considered vulnerable in an illiterate society. A weak person in one society is not necessarily weak in a differently organised society. Hobbes, just as most modern contractarians, would not fare very well in the state of nature and “vulnerability is really best thought of as a matter of degree, varying both according to one’s particular society and to one’s individual circumstances” (Vanderschraaf, 2011: 9). The same applies to

social conventions within societies; the severity of one's weakness can be reduced or increased by social surroundings. This is in essence a Rawlsian argument; social institutions can cater for those less skilful or to the disabled and therefore reduce natural inequalities or the number of those deemed as vulnerable

Institutions that deal with levelling the field for the unlucky effectively are just (Rawls 2005). Similarly, "[i]t is unjust if society fails to adjust its institutions and social systems to accommodate the fact that some of its members are blind" (Matravers 2011: 139). Modern societies can make it possible for blind people to participate as equal members and the same applies for disabled people; at the same there have been societies that have excluded even lightly disabled individuals. Therefore, characterising vulnerability depends on the specific society and its efforts to include as many people as possible. In the same spirit, it is rational for one to make society as inclusive as possible so as to maximise the number of those contributing to the cooperative surplus, since the more the contributors the greater the surplus and therefore the greater the individual gain from its distribution. This can be taken to mean that severely disabled people who have never contributed or are not expected to contribute are not owed any moral, or non-rational, consideration. Maybe in order to preserve the requirement of justice as mutual advantage this is a necessary compromise.

In contractarian thought, and especially when justice is seen as justice as mutual advantage, as in *MbA,* the vulnerable and non-contributors are not "owed moral consideration" (Morris, 1991: 81).Therefore, one way of addressing the issue of whether the weak should be included in a social contract based on self-interest is to follow Gauthier in accepting that "[a]nimals, the unborn, the congenitally handicapped and defective, fall beyond the pale of a morality tied to mutuality" (Gauthier, 1986: 268). A second approach would be to argue that the weak participate in conventions of justice and therefore by extension they are moral agents. The following section will examine whether it is possible to include vulnerable and non-contributing members of a society in the realm of justice as mutual advantage, without abolishing the premises of rationality or the principles of justice as mutual advantage. The following two paragraphs will look at how technological and social developments have been making the inclusion of disabled people more likely. Then, the focus will turn to examining whether it can be rational to interact with the vulnerable and with non-contributors.

The definition of vulnerability includes increasingly fewer people as society changes and technology advances. A person born with any severe paralysing disease

today in the Western world can hope that through advantages in technology, his life will be much easier than it would have been even a few years ago. Moreover, disabled people today can be more socially and professionally active due to technological advances that have been accompanied by, and to a degree caused, cultural changes around the possibility of the social inclusion of the disabled. It is safe to claim that today disabled individuals are more active members of society than they were a few decades ago and that this trend will continue. In addition, the disabled are more active professionally than anyone would have expected a few years ago as a result of the technological progress and perception change. The disabled are increasingly included in social structures and expectations of their contributing, even in limited ways, have become the norm and the severity of disability less of a problem in matters of inclusion.

Severely physically disabled people can and are expected to contribute to the cooperative surplus as successfully as the able-bodied. “As successfully” does not have to mean “as effectively”; disabled individuals are not expected to contribute as much as the able-bodied, but they are expected to contribute to the best of their abilities. In this view, individuals with mental and physical disabilities can be expected to contribute to society. Social and technological developments have significant consequences in that they diminish the importance of physical ability as a requirement for one's contribution to the social output by comparison to the situation a few decades ago. As long as existing social norms and rules allow for and assist the inclusion of the physically vulnerable in society, it can be expected that they will contribute. There are rationally acceptable variations of net contribution to the cooperative surplus that can include the physically disabled without relaxing self-interested premises.

In addition to the arguments in the previous paragraphs about the fact that with social and technological change fewer people are excluded from contributing because of their disability, we will have to take into account those who are too severely disabled to contribute as well as interactions between generations. It is difficult to see how extreme cases, including those who are severely mentally or physically disabled, can participate in a meaningful way in the creation of the cooperative surplus. Although it is possible that in the future there will be treatments that make it possible, in the present day they cannot be accepted in a social contract that is based on rationality and mutual benefit alone. For this case, we would have to take into account the need for rational agents to sustain a cooperative reputation. One who interacts with someone who cannot contribute, loses from the series of interactions with the vulnerable but gains in terms of

reputation. A reputation as a cooperator makes it more likely that she will draw more cooperators and therefore increase the gains from cooperative interactions. Rational agents do not expect any immediate and direct benefit from interacting with agents who cannot contribute; they are benefited however from advertising their disposition and therefore attracting similarly disposed rational agents with whom interactions lead to optimal outcomes. In a sense, interactions with non-contributors are used as the basis for the real interactions for a rational agent – the ones yielding a benefit. The rationality of interacting with non-contributing agents, which also includes an argument for rational interactions with the weak who cannot contribute as effectively as the able-bodied, will be discussed more analytically in the next section on justice.

The elderly and the very young are a special case of the vulnerable. Even if they do not contribute at present they have contributed or are expected to in the future. In that respect, intergenerational interactions are an easier case for justice as mutual advantage than interactions with disabled agents. Moreover, an examination of intergenerational justice can be part of the discussion about the requirements of justice towards the vulnerable. The very young are expected to contribute to the social output and thus they do not pose significant trouble in terms of justice as mutual advantage. Gauthier (1986) argued that the too old to work have already contributed and therefore are entitled to their share of their cooperative surplus. The argument regarding a theory of justice for the very old or very young is very similar if not identical to the one for the very weak. However, it does pose less of a challenge to justice as mutual advantage as the young are expected to work at some point and therefore are to be included in the social contract as recipients only as an advance for their future contributions and in a form of insurance for the elderly; the rationale of including non-presently contributors as a form of insurance applies equally well to caring for the disabled in case one is found in their position. A rational agent has incentives to care for those who cannot care for themselves, so that if he becomes too weak to support himself, he will be able to count on others. The old, similarly to argument about the young future contributors, are to be included even though they do not contribute because they have contributed in the past and again as an insurance. Those who contribute for the too old to work can expect the same when they are too old to work. Younger generations can act as potential punisher; not because their utility functions include the very old but because if the convention of intergenerational reciprocity collapses, they will not be benefited by intergenerational reciprocity once they are too old to work.

One can be hopeful about the future; even severely disabled people can contribute to the cooperative surplus in modern, developed societies. It is reasonable to assume that with the help of technology, this trend will continue to expand to include more people irrespective of their disability. Therefore, over time the theoretical tradition of justice as mutual advantage has been expanding the circle of individuals who can be included in its realm.

In some respects vulnerability becomes a case of inequality in skills and contribution which can be addressed by current theories of justice as mutual advantage. However, even though this optimistic account is very reasonable, it does not provide an answer for contemporary severely disabled people who cannot make any contribution to the cooperative surplus.

7.3.2 Justice

A theory of justice as mutual advantage must give an account of why rational individuals should care for others, especially the weak, even when there is no immediate benefit in so doing. The core of any theory of justice has to include a justification for helping those who cannot reciprocate. Embedded in theories of justice as mutual advantage and practical rationality is the notion of mutual unconcern. Rational agents maximise their own utility functions without having an interest in the maximisation of others or put differently "utility functions are to be defined independently of one another" (Morris in Vallentyne, 1991: 81). The assumption of mutual unconcern is an essential requirement for rational agency and in turn for a discussion of justice as mutual advantage. If not, one could just assume that humans are empathetic, benevolent and mutually concerned and therefore justice would be the outcome of any social interaction. In the account of justice as mutual advantage, which is the one discussed here, rational agents should only participate in interactions as long as it maximises their own utility irrespective of what happens to others. The essence of the assumption of rationality is that agents have to be self-regarding and only interact with others when it helps them promote their own interests. Therefore, it is obvious that combining justice as mutual advantage and participation in interactions without direct mutual benefit can be problematic.

It is very difficult if not completely impossible to give an account of why conventions of caring for the weak were established in the first place. A very likely and plausible explanation can be that they serve as insurance. As mentioned in the previous

section, it is rational to care for the disabled, as well as the very old, as a form of insurance which will be realised if one becomes very old or disabled. Despite the difficulty of giving an empirical account of how conventions of caring for the weak have emerged, it is a fact that most modern societies have a justification and a method for doing so. Although it would have been very useful to be able to give a justification based on rationality of how modern social conventions include severely disabled infants for instance, it is not as important as recognising that this is an established convention. From an evolutionary account, the fact that conventions of helping the weak have survived and become established social norms indicates that they do serve a purpose that is socially beneficial. However, this is just an indication, albeit an important one; societies have moved from excluding the weak, to discriminating against them, to trying to include them as equals in the social contract. Therefore, we can claim that social cohesion and efficiency are not threatened by including those who are significantly weaker than the average.

Power discrepancies between contracting parties refer to differences in the ability to contribute to the cooperative surplus in similar terms. The reasons behind unequal contributions may be the result of differences such as wealth, natural skills or may also have to do with an agent's age, such as the too young or the too old to contribute. Also, significant differences include situations where a potential contracting party is so severely disabled that she cannot be expected to make any contribution at any time. The following discussion will focus on the latter case because it is the most challenging for theories of justice as mutual advantage. If there can be a convincing answer for cases of severely disabled individuals participating in the social contract, then cases of great power differences will be included.

Theories of justice as mutual advantage have been criticised for excluding the vulnerable (Barry, 1991); if the criticism is valid, then it is claimed they cannot be convincing as theories of justice – although contractarians like Gauthier disagree (as noted in section 7.3.1). However, a theory of justice as mutual advantage does not have to exclude the vulnerable. There are two distinct rational premises for justifying caring for the weak. First, a cooperative surplus is achieved even when there is an interaction with a disabled person, as long as he can contribute. Secondly, interacting with the weak, irrespectively of whether they can contribute or not, can be seen as insurance and can be justified through a rationale of indirect reciprocity.

A rational agent has reasons to participate in interactions characterised by

extreme power inequalities interactions because they require smaller contributions from the powerful party than interactions between equals; an interaction with a weak person requires a contribution that does not have to be the maximum possible but merely proportionate to the contribution of the weak. In this respect, an interaction with someone who cannot contribute as much can still be beneficial although not necessarily maximising. Hence, a rational agent should always choose to interact with someone of roughly equal capabilities. When this is not possible, interacting with someone weaker is also rational in the sense that there can be a cooperative surplus greater than would be through individual production. A handicapped person can still contribute to an interaction though she may not be as productive as an able-bodied individual.

The second rational justification for interacting with the weak is more directly linked to the understanding of social behaviour as conventional behaviour. Caring for the severely disabled can be justified as a form of insurance and a type of indirect reciprocity. Able-bodied individuals care for those who cannot support themselves, expecting that if they find themselves in the same position, others will do the same. These conventions are rational within small groups where free-riding can be identified and punished. And they can have only been established within small groups at first. Viewing modern societies as a collection of smaller social groups where reciprocity and cooperation are rational would mean that there can be a rational justification for this type of behaviour. Moreover, interactions with the weaker are rationally justifiable when we expect a form of indirect reciprocity; we “help those who cannot help themselves, so as to encourage those who can to help us” (Vanderschraaf, 1999: 349). The concept of indirect reciprocity is closely related with multilateral reciprocity, which was discussed in the previous chapter; agents who wish to cooperate choose agents who have cooperated in the past and avoid agents who have defected in the past (Sugden, 2004). In real life there is no ring of Gyges and at some point everyone will have to interact with someone who is significantly more powerful than the average. Interactions occur between parties of different power and most agents will find themselves interacting with much more powerful agents, more or less frequently.

The discussion in the paragraphs above mostly deals with cases of great but not extreme inequalities; the weak are considered to be contributors, even if they do not necessarily contribute as much. Although most of the arguments presented may apply in situations of extreme inequalities, in general they do not account for agents who cannot contribute at all and it cannot justify such interactions from the perspective of a

mutually unconcerned agent. A rather stronger justification for rational agents to interact with those who cannot contribute or cannot contribute as much, is that in the conventional account building a good reputation is vital as well as using these interactions as insurance. As mentioned earlier, one can use interactions where one's benefit is very small or non-existent to boost one's reputation and attract similarly disposed agents. A rational agent who has built a reputation for cooperation or even irrational behaviour will find that it is much easier to convince cooperators to interact with her in interactions with higher than normal benefits. At the same time, she will avoid interactions with agents who have not shown similar behaviour and therefore avoid being taken advantage of. The assumption that interactions are repeated and their history is known ensures that those who have a cooperative history will only cooperate among each other and avoid agents with a history of defection. The repetitiveness assumption also makes it possible to assert that the first stages of a series of interactions are a trial and error process during which interacting agents test the waters by engaging in small benefit and low risk interactions.

In addition, stable conventions have been assumed to build and develop incrementally; low significance interactions are used to build trust among rational agents and as a safety mechanism to make free-riding costly. In this context, someone who has a reputation of not taking advantage of the weak is much less likely to attempt to cheat his equals and be a free-rider and thus, will be more likely to be accepted in a cooperative convention. In the same spirit, a cooperative reputation is not very likely to attract defectors as the low importance and low payoff of initial interactions, make cooperative interactions too costly for agents who do not aim at establishing long term interactions. This seems to be the strongest case for including severely disabled individuals in the social contract. It is in fact a case of indirect reciprocity. A rational agent who interacts with a disabled can expect to be rewarded by the cooperative surplus of her increased future cooperative interactions. Similarly one can expect to be punished for avoiding interactions or taking advantage of the weak (Vanderschraaf, 2011). Given a reasonable capacity to retain information, rational agents in an established convention will punish those who have exhibited selfish behaviour and reward the altruists. An altruist may be irrational, but for the purposes of conventional rationality, and even for Gauthier's account of translucent dispositions, she meets the requirements set by the rationale of constrained maximisation both as introduced in *Morals by Agreement* and as presented in the conventional account of social behaviour.

In conclusion, in all the above cases it is as rational to interact with agents who cannot reciprocate. The account of conventional justice presented here can be problematic on two accounts: it does not explain how conventions of interacting with the non-contributor emerged in the first place and secondly it can be argued that having a reputation for cooperative interactions with non-contributors might harm one by attracting agents who are disposed to defect. The first point is addressed by complementing the account of cooperative reputation with the rationale of helping those in need as a form of insurance policy. Attracting defectors as a result of cooperating with the vulnerable is a danger that is dealt with to a great extent when we take into account the fact that, in the context of repeated interactions, it takes time for a series of cooperative interactions to yield high enough gains so as to make a defector's time and effort worthwhile.

Rational conventions that are based on such inclusive premises are more efficient than exclusive social conventions. Including the severely disabled in the cooperative surplus can only increase the total social output; especially if in the long run the costs for caring for the disabled are minimal by comparison to their potential contributions. Justice as mutual advantage does not have to mean that mutuality is restricted to interactions between two individuals. A rational agent can help someone weaker without expecting an immediate benefit from the specific individual. But having helped him has enhanced her reputation in order to make her future interactions more beneficial. Thus, when we conceive interactions as repeated within a society, indirect reciprocity is a valid argument for the rationality of behaving justly and within the rationale of the theory of justice as mutual advantage. In sum, creating a cooperative reputation leads to more cooperative interactions and thus greater benefits in the long run, but it needs to be viewed in conjunction with cooperation as insurance and whilst taking into account that cooperators will not engage in high benefit interactions with agents who have not proven their intentions.

A final point that needs to be included in the present examination of justice is the concept of desert. Justice is usually discussed in conjunction with desert (Rawls, 2005; Gauthier, 1986). In a just situation everybody gets what she deserves, or in other words, justice provides the regulation according to which a society should decide who deserves what. From the above understanding of justice as mutual advantage it may be concluded that the vulnerable do not receive anything as a result of thinking about what they deserve. In essence they do not deserve to be receivers of the cooperative surplus.

However, this does not have to imply that they do not receive anything from the division of the cooperative surplus. Their not being contributors is irrelevant since their participation in the social contract promotes the interests of those who are contributors. Therefore, participation in the social contract does not have to be tied to contribution to the cooperative surplus (Vanderschraaf, 2011).

7.3.3 Inter-conventional Justice

The discussion of justice presented in the previous paragraphs has not included an examination of possible conventions of the weak. Given extreme inequalities either in skills and wealth or mental and physical capabilities, it is theoretically possible, and a plausible alternative scenario, that rational agents will choose to interact with those of similar strengths. Conventions have been argued to be formed by individuals with access to similar information and thus similar outcomes of rational deliberation. The same can be argued for individuals whose capabilities are characterised by extreme inequalities. If coercion cannot be avoided in cases of extreme inequalities, the weak will choose to interact with agents of similar strength and therefore roughly similar maximising strategies. As a result, there will be social conventions including rational agents of similar strength who achieve optimal outcomes within the given environmental parameters. In this case we can have different types of social conventions within a social contract; social conventions consisting of the weak who simultaneously participate in conventions with stronger agents, should that help them maximise. Conventions of the weak is a realistic proposition, especially if we take into account how in contemporary societies patterns of behaviour change even within the same city. For instance, it is a common phenomenon in contemporary cities to have areas of low crime rates and other areas with very high criminality. These can be seen as different conventions within the boundaries of a single social contract. In this view, interactions between agents of different conventions are possible, though rarely maximising.

Interactions among individuals of different conventions might seem irrational at first view but, as discussed above, inclusive conventions can be more efficient; the same applies when we look at conventions in the context of the social contract and not just individuals in the context of social conventions. Individuals who cannot contribute in one convention can be useful in another; those who are too weak to contribute in a certain environment may be invaluable in a different one when they interact with different individuals.

A more interesting albeit not always as realistic case would be one where rational, non-coercive interactions among individuals of different conventions are not possible. Within an isolated convention rational deliberation can lead to practical outcomes that are incompatible with the rules of other established conventions. Assuming there was no communication between those two conventions in the past but that they can interact in the present, their fundamental differences can make rational interactions impossible. For instance, anthropological and historical evidence shows that European explorers in the 17th century encountered a couple of isolated Polynesian island societies that were not aware of each other. One had a completely peaceful culture of conflict resolution and the other was a typical hierarchical society lead by a warrior elite. The Europeans made the latter aware of the former resulting in the destruction of the peaceful society (Diamond, 2006). In this example, both isolated social contracts can be said to be based on rational conventions, given information and other environmental parameters. However, they had reached extremely different equilibria such that any interaction between them could not be rational. The same probably applies for interactions between the Polynesian and the European culture of the time. Their extreme inequalities in combination with the radically different culture would make rational interactions impossible.

In the modern world there are social states that are considered rational by their members but are viewed as fundamentally irrational by outsiders. Western European states consider many states that do not accept European enlightenment principles as irrational and vice versa. Similarly, within societies there are localised conventions of behaviour can be too different to make rational interactions a realistic possibility. It is obvious that in their circumstances one cannot rely on the rationality of individuals or on conventions for addressing conflicts of interests. This is even more true when between conventions there are significant power and cultural differences. Avoiding interactions is a solid idea theoretically, but not always feasible. Therefore, in cases like the above a third party enforcer may be necessary.

The notion that a third party is needed to solve problems between rational actors that cannot reach agreement is of course nothing new. A form of government intervention is needed, and in some cases is essential, when differences in conventional rationality are too great for an agreement to be reached. This does not have to mean that all differences are reconcilable nor that an arbitrator is always needed. It is possible that differences are so extreme that there can be no meaningful method of interaction even

with an impartial arbitrator. Furthermore, evolutionarily stable conventions cannot be altered significantly within short periods of time and therefore any government is powerless when its policies collide with the established conventions. The fact remains that impartial arbitrators cannot be ideal agents disconnected from reality and from the social conventions and structures that appointed them in the first place. Therefore, their behaviour should also be described by the conventional account and is limited and bounded by existing social structures. In cases of extreme inequalities like the two island societies above, an impartial arbitrator is not a more plausible solution than the one offered by the conventional account. Extreme inequalities meant that there is a possibility that there can be no agreement on a third party enforcer making coercion or abstention from interactions the only viable alternatives.

7.3.4 Conclusion

The above discussion has attempted to provide a rational justification for interacting with those who cannot contribute, or who cannot contribute as efficiently, to the creation of the social surplus. Indirect reciprocity and the need for a cooperative reputation make it rational for one to help the weak. Within the conventional account of rationality there can be a similar argument for the inclusion of the weak. Once a convention of inclusion in the social contract has been established then rational agents have reasons to follow it. The difficulty therefore lies in accepting that it is possible for such conventions to come about as a result of rational interactions. Once this argument is accepted, established conventions include stability dynamics that make their destabilisation very unlikely unless there are rational reasons for it. As long as inclusive equilibria are more efficient, then they cannot be destabilised. A rational agent within such a convention does not analyse each interaction individually; she behaves conventionally unless there is an environmental change that forces her to reconsider. Hence, an agent in a convention of justice behaves morally, not because it is in her best interest to do so but because it is in her best interest to behave conventionally.

7.4 Conclusion

The purpose of this chapter was to show that conventional rationality can support a realistic version of justice as mutual advantage. If we take into account the fact that all interactions are to be seen as repeated and that rational agents have incentives to act within stable and efficient conventions, justice can be reached through interactions

among individuals who want to maximise in the long run. Moreover, the outcome of interactions can be known at least within the boundaries of a social convention which means that agents can be punished or rewarded for their behaviour. Thus, rational agents will want to advertise their behaviour in order to maximise in their future interactions.

Interactions create obligations that can be described as moral; that is, they are not immediately maximising. An idealised individual who has never interacted with any other human being cannot have moral obligations. The beginning of a series of interactions means the beginning of one's moral obligations, towards the topical social convention and by extension towards society and not exclusively towards another person. Thus, obligations stem from a need for a social life; in an idealised account of humans, where a Robinson Crusoe is a possibility, morality and justice have no role. A human above the social realm, who has always been completely disassociated from society can have no meaningful debt to society and therefore no moral obligation to adhere to any principles of justice. In a realistic and plausible setting, humans need social interaction to maximise and in this respect need rules of justice.

This chapter attempted to show how principles of justice can be conceived in the context of social conventions and rationality. In this context, and provided we accept that social interactions are seen as infinitely repeated and that evolutionary dynamics determine social structures, there is a need for an equilibrium selection mechanism. In the conventional account this mechanism is provided by a combination of rational agency and the evolutionary dynamics of social structures. Provided there is information availability about other conventions, social equilibria points will converge towards Pareto optimality. Pareto optimality in conjunction with a Nash equilibrium define justice as mutual advantage in the conventional account. Adapting the definition of vulnerability to social and technological advancements as well as using indirect reciprocity to justify rational interactions with the vulnerable make justice as mutual advantage a valid theory of justice that complements the evolutionary account of rational conventions.

The realism and plausibility of the assertions and assumptions of this and previous chapters will be addressed in the next chapter. The conventional account of rationality will be shown to be realistic in the contemporary world where there can be cheap and efficient information spreading. Moreover, the theoretical paradigm that has been presented so far will be viewed in relation to actual behaviour as it is exhibited in contemporary societies.

8. Rational morality and social conventions in the real world

The previous chapters have been based on assumptions that can be seen as idealistic. Rationality, even bounded, and availability of information are not always present in the real world. The purpose of this chapter is to discuss how these assumptions and the account of conventions in general can be seen in the context of real contemporary social life. The focus will be on showing that although idealistic and simplifying to a degree, assumptions of rationality and information availability are not very far from the behaviour and the capabilities of real people.

At the same time, these theoretical concepts need to be adapted in order to be realistic and to correspond to the behaviour of actual persons. In addition, the analysis of the chapter will focus on showing to what extent the assumptions of rational behaviour and information availability are realistic as they are. The adaptations needed and the realism of the assumptions are obviously linked; the more realistic the assumptions the smaller the amount of adaptation needed. Bounded rationality requires individuals to be reasonable and not necessarily hyper-rational as in the economic account of rationality. Similarly, information availability and knowledge of the interaction history within a convention do not require unrealistic processing capacities. In contemporary societies the assumption of rational and informed individuals is plausible.

Information technology facilitates the plausibility of equal information and bounded rationality. In the contemporary world the assumption of equally informed agents is more realistic than ever before. In addition, it is plausible to assert that the rate of technological advancements will only increase, facilitating even more the spread of information as well as its availability to more people. From that perspective, social conventions as shaped by information availability obtain new meaning. They do not have to be limited by geographical boundaries, but rather they are shaped by individuals whose rational deliberation leads them to search for a certain type of information. Therefore, information technology is a great tool for quick information spreading while ensuring higher information accuracy. In this context, cultural memes can also be viewed from a new angle; they can spread quicker and go through a more rigorous trial and error process involving more agents.

Information technology, by making available greater amounts of accurate

information to more people, reinforces stable social conventions and thus social contracts and at the same time weakens those conventions that are based on coercion or on similar unsound grounds and hence that are not evolutionarily stable. Thus, in the modern world technological progress makes it increasingly plausible to assume equal information and rationality. In addition, the understanding of social conventions in the contemporary world in which they are bounded and affected by information technology limits, does not replace the traditional understanding of social conventions, in which their structure relies on geographical and cultural factors; on the contrary the two understandings are complementary and mutually reinforcing. Obviously information spreading is easier among people who speak the same language or deal with the same problems and therefore it is more likely that individuals in the same neighbourhood, city or country will be drawn to similar sources of information causing the collapse or the empowerment of the existing conventions. However, social conventions of behaviour are not formed exclusively by these factors because of information technology.

Moreover, the account of evolution and the establishment of social conventions is only one of the paradigms that can be used for explaining social behaviour. In that respect, there has to be a discussion about how they relate to real life through historical and anthropological examples. Evolutionary accounts of social structures and of individual behaviour within them are closer to reality as they can include historical and anthropological factors that have influenced cultural evolutions and the formation of social conventions.

Finally, the fact that the assumptions of rationality and conventional behaviour are realistic does not mean that agents always behave rationally within conventions. There are incidents of unexplained, irrational behaviour that can be attributed to a different, unique psychology of the actor and varying local environmental parameters that affect the outcome of rational deliberation, or more accurately influence the bounds of rationality. These inconsistencies are not as problematic for the account of human behaviour within social conventions as they are in a theory assuming idealistically rational agents, but nevertheless they will have to be discussed and explained.

The next section will deal with rationality and the extent to which it can be plausibly attributed to human beings. More specifically, it will examine rational behaviour with reference to social behaviour in the real world.

8.1 The realism of conventional rationality

In terms of rational agency, traditional game theory is based on idealistic hyper-rationality assumptions whereas evolutionary game theory uses assumptions that vary from bounded to the complete absence of rationality. Bounded rationality, within an evolutionary framework, is more realistic and more effective in examining how reasonable agents, who are mutually unconcerned and aim at maximising their utility, behave in the context of society and social structures. It is plausible to assert that individuals are rational utility maximisers, but that they are not capable of predicting accurately far into the future or of having complete memory of all past interactions. Moreover, it is also realistic to argue that societies and social structures evolve following patterns that are not directly linked to individual action. The following paragraphs will attempt to show that humans are boundedly rational and that the typical assumptions of bounded rationality are to be found in the real world.

8.1.1 Bounded rationality

There are a multitude of parameters that have to be taken into consideration when looking at how social structures come about and how they change. Most of these parameters can be influenced significantly by individual action and even more by the collective action of rational individuals. Therefore, it is essential that we use a framework that includes both explanation of collective behaviour, focusing on the behaviour of social groups, and also explanation of the individual behaviour of rational agents. All interactions are repeated in a society as individuals follow patterns of behaviour based on the established conventions. Within a society or a social group, it is unrealistic to assert that interactions are random; on the contrary, especially in human populations, interactions and their frequency depend on factors such as the outcome of past interactions and the agents' relative location. Real world societies are at equilibria which might be of various types, but individuals within these societies know what kind of behaviour the equilibrium requires of them. Therefore, a person can have certain expectations of others and also has a rough memory of how she fared against specific other individuals.

In other words, “when in Rome do like the Romans do”; those who do not in the conventional account of social interactions can be traced and excluded from social interactions (Andrés Guzmán & al, 2007). Rational agents conform to established social conventions and norms since their rational strategies have coevolved with the dynamics

of the existing social structures. A single visitor in Rome may be involved in one-off interactions and may be indifferent to the prospect of being excluded. In this sense, there are two groups: the Romans and the visitors. If many visitors perceive their interactions as one-off interactions, where cooperation is irrational, the Romans will behave accordingly. Then, Rome will develop a reputation as being hostile towards strangers, and over time there will be fewer visitors, harming both groups.

In an idealised model of the world, one's behaviour would become known and punished either in Rome or in one's native convention. In the real world this is not as easily achieved. However, it is plausible to assert that lone exploiters, visitors who break local conventions, can be punished even within the limited scope of their visit. Provided that they will interact with the local population, punishment is possible. For instance, a visitor to a foreign country who litters might not be fined or the fine might be insignificant, but he can expect lack of cooperation from the natives next time he asks for information. Therefore, by littering, agents do not maximise their utility even for short-term interactions given there is information about their actions.

Interactions are to be seen as repeated since individuals act within a given convention where there are established norms of behaviour and communication of information and there is information about their interactions outside the convention. The repetitiveness of interactions means that all interactions can be described as games where joint strategies yield higher utility to all players. For example, in the PD game, when both prisoners know that they will interact again in the same environment, they will coordinate their behaviour so as to maximise their utility. Whether this is still a PD game or not, in the formal sense, is not important. What matters in this context is that it more accurately describes rational interactions. Repeated interactions and the capability of agents to learn and remember, within reason, shows that an account of bounded rationality is more plausible and effective in describing rational behaviour. In addition, the above understanding of game theory as a descriptive tool of social interactions gives an account of existing social structures that are cooperative and a result of having adopted joint and not individual strategies.

The account of bounded rationality that has been used to support the concept of rational conventions, as opposed to economic rationality, is not unrealistic. People learn from their experiences and they want to maximise their benefit as often as possible. Moreover, it is common and expected that the same people interact more than once; it is more realistic to claim that individuals encounter each other in repeated interactions

than not. People live in societies, towns and neighbours and therefore, it is much more likely that most of their interactions will be with their neighbours and people living in the same town, than with people living in a different country. At the same time low individual rationality does not necessarily lead to sub-optimal equilibria (Young 2001). Assumptions of rational agents with near complete information are usually seen as idealised assumptions; however, through recent technological developments full information has become a more plausible assumption. Furthermore, it is reasonable to assert that most people within a convention have access to the same information and that information spreads quickly within the bounds of the given convention.

A social convention is bound and formed by information availability, a common conception of maximising behaviour, and a general common understanding of how social problems can be resolved. It is in a sense tautological to say that within a convention agents have access to similar information and similar rational capabilities. If they did not, the convention would not have formed in the first place. Similar accounts of rationality lead to similar deliberative outcomes about how to maximise individual utility. For instance, some conventions may call for maximisation through debate while others through force; in the former case intellectual skills are needed, whereas in the latter physical power is required. In both conventions, despite their fundamental differences, there is a common understanding of how to maximise. And within the limits of each convention, it is rational to behave conventionally. Therefore, if we take the existence of social conventions as a given, the assumption of agents with similar capabilities to reason and access to similar information, is anything but an idealised assumption.

Bounded or conventional rationality and economic rationality are identical in that agents are assumed to be mutually unconcerned utility maximisers. However, conventional rationality does not require complete memory, infinite processing power, or a capacity to predict the future with high degrees of accuracy. Hence, the conventional account of rationality is closer to the way real humans reason. In fact, including mutual unconcern in bounded rationality may be seen as an unnecessary addition since individuals are not always mutually unconcerned; they develop bonds and behave in ways that are frequently not explained by rational choice theory. In reality humans care for family and friends without a need for constraints on their rationality. Most humans are not always mutually unconcerned with respect to any person with whom they interact, which at least sometimes makes them behave irrationally according

to the economic account of rationality, which requires that a rational agent should maximise her utility irrespective of whom she interacts with or other environmental parameters. Assuming that they are mutually unconcerned however, means that moral responsibilities deriving from an assumption of mutual unconcern will also include interactions between agents who are mutually concerned. Offering an account of rationality based incentives for moral behaviour can only strengthen the argument presented here. Therefore, in a realistic setting mutual unconcern, although not always present in interactions, does not threaten the assumption of rational agency.

This subsection aimed to show that the account of rationality used in previous chapters is realistic and can be used to describe to a great extent the actual behaviour of humans in societies. The following subsection will expand on that and take into consideration how the assumption of rationality is influenced by the existence of conventions. Social conventions as presented before will be shown to offer a realistic account of social structures.

8.1.2 Rationality in conventions

Conventional rational morality aspires to be a descriptive account of social behaviour, at least to an extent. But it is not limited to that; its normative dimension has been implied but is also rather clear. An individual has to be rational, which means that she has to stay open to receiving new information and actively seek new knowledge about the possibilities of conventional maximisation. A rational conventional agent has to be ready to be critical of her convention should there be a rational justification for doing so. At the same time a rational agent who has accepted the rationality of her convention has to behave conventionally other things being equal. Put differently, a rational agent ought to behave rationally. This sounds tautological and superfluous but there is a need for a type of active rationality. Conventional rationality demands that agents use their assumed rationality to confirm that they are part of a rational, utility maximising convention. This thinking creates explicit normative obligations for conventional agents. Realistically, one cannot continue behaving in the same way when those around have shifted to a new behavioural pattern. In the real world people behave in a certain way only if it pays; should they realise that their behaviour is no longer maximising they change it. For instance, a rational committed party voter will only continue voting if the party's political positions remain similar to hers. If either her or the party's positions change, she will also change her voting behaviour. This is not always the case,

but it seems reasonable to assert that it is the case more often than not and whenever is not, it merely shows how conventions influence rational behaviour.

Conventional behaviour is plausible and can describe rational behaviour with high degrees of accuracy. However, there are several problems when we try to apply conventional accounts of rationality to the real world. First of all, although it is plausible to say that everyone living in a region or within a social group has access to the same information, this is not a sufficient condition for maximisation. Individuals choose to make use of some pieces of information and to ignore others. Using different information, both in terms of quality and quantity, results in adopting different strategies. Although the information processing power in all adult humans can be said to be roughly the same, some people are better at it. Thus there are inequalities arising from the ability to use the available information efficiently. Moreover, in the real world information is often expensive and/or not widely available. Again, those who can afford it or know where to find it have an advantage in interactions. In this context, therefore, the discrepancy between the real world and the theory lies on information availability. It is plausible to assert that information is rather readily available, especially since the creation and development of the internet, but real life experience shows that to be less than accurate.

Another, closely related, assumption that can be criticised as idealistic, is the ability of individuals and groups to learn. Long-established conventions resist change and people within those conventions are more likely than not to prefer a status-quo that is not Pareto optimal than the uncertainty of change. Accounts of conventional rationality can deal with these problems. Rationality is viewed at the level of social convention, where information is roughly similar, learning is easier and trying new strategies is not irreversible. Therefore the realism of the assumptions of bounded rationality is related to the realism of the assumptions of social conventions which will be discussed next.

This section attempted to show how the assumptions of rationality and rational conventions that were presented in the previous chapters are plausible and realistic. Thus, the theoretical analysis of conventions as they are formulated through rational interactions offers realistic description of human social behaviour. The following section will attempt to do the same for the account of social conventions presented in the previous chapters.

8.2 Social conventions in the real world

Established rational conventions serve as boundaries of action by constraining irrational and non-conventional behaviour. In cases where a type of behaviour is rational but not conventional, the role of the established convention would be to preserve conventional behaviour until it becomes clear to a large number of convention followers that their behaviour has become irrational. In that respect conventions are conservative; they resist change in order to protect maximisation. If for instance a new higher welfare convention becomes known to an individual or to a minority, the established convention – through its structures, formalised rules and majority of support – will protect conventional behaviour and punish defectors. As it is not always apparent to all agents within a convention that alternative strategies are preferable, there will be discrepancies as to which is the maximising strategy. As there are information asymmetries within a convention, there will be a latency in the convention moving to a new equilibrium.

8.2.1 Rational conventions

Usually there is conflict between the actions required by individual and collective rationality. A social group maximises social welfare by limiting individual maximisation. Rational conventions are established to reconcile the two and to an extent achieve both. Social welfare maximisation depends on the conventional behaviour of its members. Conventions that are based on the rationality of their members constrain the maximisation of the individual in order to maximise social welfare; through the maximisation of social welfare, individual utility is optimised. When some members of the convention adopt non-conventional strategies, they will have to be punished in order for the convention to be sustained. Rational conventions ensure that individual utility is maximised within the constraints of limited resources and social welfare. In other words, rational conventions ensure that one individual or a group of individuals cannot destabilise conventional equilibrium.

In the real world people follow established conventions of social behaviour as a means of maximising their utility within the given social and physical environment. Conventional behaviour does not necessarily imply uniform behaviour by all, but a common understanding of what is accepted. Individuals within a society behave in a certain manner and according to the established social norms, which is what defines a society in the first place. Following those norms is rational since it means they avoid punishment for defecting.

Arguably, an individual or a group of individuals who do not agree with the established conventional rules have little choice but to comply. They might be able to deviate slightly, but outright breaking of conventional rules will result in their punishment and social exclusion. For instance a rational agent may believe it is irrational to pay taxes either because it does not allow his utility maximisation or even because he thinks that social welfare maximisation can be achieved differently. However, unless he is able to convince others about the fact that their convention can achieve higher social welfare through not paying taxes, he does not have a rational choice not to pay. Similarly, in the tragedy of the commons example (Hardin 1968), an established convention according to which everyone maximises their individual utility will also be maximising for society, at least for the short term. Depending on the size of the pasture and the number of farmers this can last for several generations. This then will be viable and rational both on an individual level and collectively. A rational individual who will be able to predict that the pasture will be depleted, does not have a rational alternative to continuing to follow conventional behaviour. If he stops using the pasture, he will not change the end result while his utility will be reduced. When individual rationality conflicts the conventional rationality, the individual who cannot extend a new behaviour to his neighbours will have to behave 'irrationally' until others accept his reasoning.

8.2.2 Real social conventions

An appropriate example of how conventional analysis can be used in a real world case is the Rwanda genocide in 1994. Farming land in Rwanda had been expanding for decades. People continued doing what maximised their and previous generations' utility in the past; therefore it was rational to try to farm on as much land as possible. This in turn led to a population increase up to the point where the land could not support so many people. Since more land was becoming available through methods like deforestation, there was no rational incentive for farmers to modernise or try new crops (Diamond, 2005).

Farmers in Rwanda during the 1980's were behaving rationally, both as individuals and a society. Within the limitations of conventional rationality this outcome could not be foreseen. Even if scientists in Rwanda had all the data available and could predict that there would be a Malthusian food crisis, it would have been extremely difficult for them to predict the genocide. Despite the fact that crises like the 1994 one

had occurred before in Rwanda, they were limited by comparison and usually attributed to tribal competition and local culture rather than an over-exploitation of natural resources and lack of economic and social planning (Diamond 2005). People in Rwanda were being boundedly rational within the limitations of the available information and knowledge. Based on the history of interactions and the rules of the established convention, foreseeing the catastrophe was realistically impossible.

A boundedly rational agent could have analysed the situation and reached a useful conclusion – at least it seems so after the effect. Previous violent outbursts as a result of land and food shortages were not considered serious enough to convince people who had been farming for generations on the same lands using similar methods that their tried and tested lifestyle would lead to a genocide. As discussed previously, the longer a social convention has been established for and the more successful it has been, the more difficult it is to change. Successful farming and population growth was controlled by local, non-catastrophic violence that included redistributing land and reducing the population. The 1994 genocide was the peak of a series of smaller scale violent events stemming from the same roots.

These factors created a relative indifference to violence that was seen as a cultural or racist phenomenon unrelated to individuals' living standards. A rational agent in pre-1994 Rwanda would have to constrain his maximisation; live off smaller areas of land and follow family planning principles. This case exhibits how conventional rationality can define individual preferences and limit agents' freedom to act. Moreover, it shows that great inequalities within a social contract have to be dealt with through a third party. Furthermore, cynically speaking the genocide itself can be seen as an evolutionary step. The overpopulation was controlled, land was redistributed and the violence had such a great effect on the local culture as to ensure that future similar events will be avoided by a better informed and more knowledgeable population.

Usually in the real world, the case with established social conventions is that they have been active over many generations and therefore it is more difficult for a single rational agent to have accurate information and to reason adequately about individual and social maximisation. Incremental change over long periods of time is very difficult, if not impossible, to detect and poses significant problems to rational agents trying to understand their environment. Looking at historic societies can offer significant insights into the behaviour of social conventions and rational agents.

Easter Island had a complex society and culture before being first visited by

European explorers in the second half of the 18th century. However, by the time of the first contact with Europeans, Easter Island inhabitants had been reduced in number and their society was in decay. Archaeological evidence indicates that the famous Easter Island monuments were erected at a time of affluence for religious purposes. The island had rich forests that provided timber that made building such monuments possible while supporting a healthy economy and society. A combination of specific climatic conditions and human behaviour lead to the almost complete deforestation of Easter Island by the end of the 18th century. Rational or even reasonable people should have been able to adjust their behaviour to the changing physical environment of the island. Deforestation has to have taken place over several generations and its devastating outcomes must have been apparent at some point before all the trees of the island were extinguished. The same applies of course in the case of human intervention. If we accept that deforestation was a result of the islanders' over-consumption, we have to question their inability to be even marginally prudent. Easter Island's climate is more fragile by comparison to similar islands and therefore, more sensitive to human activity. The physical climate is not as interesting for this discussion as Easter Islanders' response to a changing environment; or rather the lack of response.

Most archaeological evidence suggests that the Easter Island statues were erected to exhibit political power or as religious symbols (Diamond 2005). At the time of their construction they promoted social cohesion and peace while showing off the local chief's power. In that respect they were essential. The islanders' religious and social culture required these sculptures to continue being built even when it became apparent that it was becoming unsustainable in terms of resources. The motives for building monuments became so socially entrenched that they overpowered rational calculation. A political and religious system that had become so successful and long-lived would pose problems to rational sceptics. The gradual deforestation and its incremental effects on social life were not powerful enough factors to cause social change, before the lack of trees enforced that change and lead to decay (Diamond, 2005).

The Easter Island geographical isolation created more problems as the local society could not have asked for outside help or discovered information about more successful social conventions that they could then imitate. Just as in the case of Rwanda, the Easter Island society failed to adapt to a changing environment. What was rational originally became catastrophic as social structures failed to evolve and the physical

environment could not sustain the same behaviour. Although it might seem that the physical environment was central to the Easter Island case, that is not true. The physical resources of the island as well as its isolation played a role, but the important factor was that the local society failed to perceive that social changes were needed; and that failure to evolve culturally had to do first and foremost with the structure of the given society and only then its physical environment.

The above cases of Rwanda and Easter Island show how rational agents ought to take into account historical evidence, learn about similar societies and adapt to environmental changes in order to maximise. The same principles apply to successful social conventions. Preserving conventional behaviour can be destructive or productive, depending on an accurate understanding of the environment. It becomes obvious then than accurate information and data, as well as communication, are paramount for the adoption of rational strategies. Given that human rational capabilities cannot change significantly, a rational agent and a rational convention must take into account as many parameters as possible in order to sustain maximisation.

Both in Rwanda and Easter Island, “human activities dramatically altered the environment, and this in turn changed the course of cultural evolution” (Ehrlich 2002). Both these examples focus on the environmental impact of human behaviour and on how societies fail to adapt to a changing physical environment. However, they also apply as realistic paradigms of how rational agents and sustainable social conventions need to continue evaluating their behaviour in order to remain rational. An agent must take into account her social and physical environment in order to be rational. If changes in the environment occur, it might very well be rational at a given time to follow strategies that several years or decades later would be catastrophically irrational. Moreover, these examples show how information availability is essential for rationality and also poses the problem of evolutionary time.

Even in cultural evolution environmental changes occur slowly by comparison to the duration of a human life and, just as in biological evolution, social conventions change over several generations. This raises questions about the realism and usefulness of a rational agent model in a cultural evolutionary context. The problems of evolutionary time in political and moral philosophy and the realism of information availability will be discussed in the following paragraphs, where there will be special emphasis on the changes in the establishment and stability of social conventions caused by technological advancements. The next section will discuss how the assumption of

equal information and information availability is more realistic in the contemporary world than it used to be and how that affects the possibility of equal rationality.

8.3 Information availability

Cultural evolution refers to the changes of social conventions that lead to changes in the social contract; in addition, it encompasses changes in the available information and its processing. Hence, cultural evolution is intrinsically linked with information availability and communication. Rational capabilities cannot be assumed to change over time, but rational deliberation yields different outcomes as the available information is being enriched with past experiences. Thus, cultural evolution depends on the available information, or put differently "[c]ulture is information stored in people's heads, which can be transmitted among individuals" (Henrich et al., 2001: 2).

Especially in a cultural evolutionary context, information availability has to be taken into consideration in accessing rational behaviour; available information determines rational strategies and as a result, individuals in similar cultures will adopt similar strategies that are maximising. Therefore, cultural evolution to a great extent refers to information availability, which in turn is essential for rational deliberation. Information increases as our history becomes longer and the means for its storage and transmission become more efficient and reliable. Modern people have been able to make better decisions because of the fact that they know more than people in the past, "if progress is real...[it is] because we are born to a richer heritage" (Durant, 2010: 102). Hence, we can claim that the accumulation of knowledge works to our advantage as history evolves. Technological developments and social changes have been making the spreading of this knowledge increasingly easier and therefore have been affecting the outcomes of rational deliberation.

8.3.1 Information availability and rationality

The ease of information spreading in the contemporary world shows knowledge of a plethora of social conventions and social contracts, a fact that would have been impossible without the technological advancements of the past couple of centuries. Being able to travel further in the 19th and 20th centuries as well the availability of more accurate information in the last a hundred years, gives us the capability to compare and contrast our conventions with those in every other part of the world. This ability creates responsibilities from a rational choice perspective; if we find out about

more efficient equilibria we ought to imitate the behaviour observed there. In the same spirit, if western developed societies have a moral obligation to help those in need anywhere in the world, it is because of the mere fact that today more than ever before they know about their situation and they have the ability to change it (Barry, 1991). Similarly, the internet and the possibilities it has created for communication and learning have profound implications for political philosophy and the assumption of rationality.

Information availability is a realistic assumption in the contemporary world and has significant effects on our understanding of personal responsibility. An agent can learn which is the maximising behaviour in a given environment and which social conventions maximise social welfare. The more the internet becomes part of social life the more diverse information becomes easily available and the closer societies get to being societies of fully, or at least equally, informed individuals. Moreover, the body-mind separation (Matthews in Hoven and Weckert, 2009), makes it more plausible to assert that there can be societies where physical skills are not linked to contribution to the cooperative surplus. It is plausible to argue then that the information age brings us closer to a society of individuals with roughly equal capabilities. The expansion of information technologies and the fact that they have become more easily accessible to more people creates an equality of information availability, which can lead to agents of roughly equal rational capabilities. Equal rationality ensures rational, mutually beneficial interactions and subsequently efficient social conventions.

The account of economic rationality is probably only observed in the real world in corporative behaviour. Corporations and businesses keep detailed records of their past performance and use all the means at their disposal to accurately predict the future. Agents of bounded rationality may not have the detailed and accurate account of their past or the predictive means that companies do. However, just like companies, all agents within a convention have access to roughly the same amounts of information. Therefore, using their rational capabilities, they can make decisions in the same framework. If there could be an argument for a for-profit business to constrain its maximisation, it would also apply to rational actor models. The following subsection will look some real world cases were businesses decided to constrain their maximisation as a method of maximising their profits.

8.3.2 Information spreading in contemporary societies

Freemium is a successful business model where paying for the product is, in a sense, voluntary since many, if not a majority of consumers, do not pay (Pujol 2010). Some consumers, usually businesses and professionals, can select to pay for more advanced services and thus, make it possible for freemium companies to give away their product to those who are not willing to pay. Free-riders in this model are welcome in that they help the company advertise its product and expand – they serve as cooperators with apparent dispositions. This model is primarily seen in companies that are active on the internet (such as Dropbox), but there is no reason for it to be limited there. As long as there is a service or a product it is easy to see how constraining maximisation is profitable in economic terms. The same applies for more traditional consumables, although it is not as straightforward. Free samples have been an established advertising practice for many years (Pujol, 2010). It is not too far-fetched to say that there can be companies that will offer a proportion of their product at prices that are very low or even below production, so as to increase their consumer base. A typical case is low- cost airlines that sell some tickets at very low prices, provided there is a form of cooperation by consumers, such as booking long in advance or waive some services. Despite restrictions such as the need for low marginal costs and a relatively long-lived and large consumer base, some freemium companies have shown that it is possible to be as competitive as traditionally organised companies.

However, the relative success of business models like freemium does not mean that traditional businesses cannot be equally or more profitable. Word of mouth effects that are vital for freemium companies also apply to traditional companies. However and more importantly, the freemium model shows that it is viable for a for-profit company to give away small quantities of its product and to make profit from large consumers only. Inequalities are embedded in the model and the weak – those using the product for free – are essential for the company's profits, since they are used as advertisers. Having a good reputation is essential since information spreading (i.e. free advertisement) as well as the number of paying customers depend on the quality of the product. Since it is safe to assume that only rational individuals will use or buy the product, there will be a strictly mutually advantageous relationship with the company. The vulnerable, the non-paying consumers, are vital for information spreading which makes the model viable.

Similar principles apply to all economic transactions as information availability

creates individual responsibility. Micro-financing was made possible to a large extent because of the possibility of informing poorer people about the prospects and the benefits of their cooperation (Vatta, 2003). The relative success of micro-finance exhibits how extreme power inequalities do not always make rational interactions impossible. The weak can form groups, or conventions, in order to interact with those much stronger. In this light we can claim that there are economic models that inform moral conventions of cooperation and constrained maximisation. Therefore, at least within some economic models, there is room for embedded principles of fairness.

A fair price in contemporary societies and economies is one that is determined by the free market mechanism. A market mechanism that works more efficiently will produce prices that meet the fairness criterion. In order for this to happen, full or at least equal information is essential for both buyers and sellers. Information for both products and sellers and buyers has to be freely available, in order to be used for assessment in future interactions. Just as reputations matter in repeated games within social conventions, in an ideal market past interactions count towards a seller's or buyer's reputation and determine success or failure. This is even more likely in today's information age where the availability of information is easier and cheaper; for instance, ebay can store seller's and buyer's ratings for all their transactions, which are then used as an accurate indication of their behaviour.

An ideally competitive market is indeed a morally free-zone as Gauthier described it; the development of the internet has not created a perfect market yet, but in many respects, topical markets on the internet fit the description of an ideal market. In that respect, it is plausible to claim that a near-ideal market is a realistic possibility and moreover that, within it, rational interactions will maximise both social welfare and individual utility without the need for moral constraints.

8.4 The evolutionary time-frame

Petit (Hoven & Weckert, 2009) argues that internet interactions do not encompass the trust that face to face interactions do, but this is obviously wrong. Internet interactions can be monitored and recorded much more easily and efficiently and therefore they create an accurate record of each person's history. Defectors find it more difficult to hide and therefore an assumption of disposition translucency becomes more realistic. Information technology improves the availability of accurate information and communication and therefore, cultural memes are transmitted and tested faster. Hence

memes, as pieces of information that are replicated, and are fundamental in adopting rational strategies, become known more easily and quickly (Binmore 1998).

As Dawkins (2006) described, memes evolve similarly to genes as successful ones are replicated more frequently. Memes' transmission and adoption is time-consuming, even if it is not as slow as gene evolution. Biological evolution is incremental and as a result it is irrelevant to individual maximisation. Cultural evolution is of course faster but it still may take several generations depending on environmental parameters. Therefore, it is not always straightforward to say how rational actors can benefit from evolutionary processes when the evolutionary equilibrium may not occur during their natural life, since their benefit maybe incremental.

In this respect, information technology is related to cultural evolution in that it accelerates it and thus makes it relevant to human behaviour. Cultural evolution is based on information availability and spreading which has been facilitated by the internet. New technologies and especially information technology make information sharing and spreading much faster. Therefore, social change is more likely to occur faster as information about more successful social conventions becomes more readily and cost-effectively available. Through information technology the cultural evolutionary time shrinks to fit human life span.

Moreover, cultural evolution takes place in steps which can be seen as social welfare curves of Pareto optimality. Evolutionary incremental changes require topical changes that are important on an individual level; a rational agent will prefer a social improvement, even a small-scale one to the status-quo. Incremental social changes that have an impact over generations also improve social conventions equilibria and individual utility functions in the short term. In addition, reciprocal behaviour is rational when there is adequate information about it. Rational reciprocity is maximising in inter-generational contexts as well. As was discussed in the previous chapter, there can be a rational interest for the unborn and for future generations. If a generation exhibits no interest in their offspring, then in time they cannot expect to be included in the established social contract (once they are too old to contribute).

Social conventions are influenced by the equilibria in neighbouring conventions. An equilibrium of defection in a PD type game can destabilise a cooperative convention when there are interactions among their members. An individual's future tragedy is reinforced by her recent and present interactions. Non-cooperative interactions will lead to non-cooperative behaviour in the future (Skyrms, 2004). Therefore, a convention may

move in a variety of directions depending on local interactions on the boundaries of two neighbouring conventions. In continental Europe driving on the right was a result of specific historic circumstances. Walking on the right was first established after the French revolution after which “one can a see a gradual but steady shift...in favour of the right-hand rule” (Young, 2001) until 1967 when Sweden became the last country to change its driving rules. The evolution of even minor conventions in real life can take centuries, depending on topical endogenous parameters, such as a revolution in the case of France or the established conventions of neighbouring countries in the case of Sweden. Arguably, in today's interdependent world where communication and information exchange is much more frequent, a similar convention shift would occur much faster. However, more deeply rooted conventions may not change any more quickly than they would have two hundred years ago, despite obvious benefits of homogeneity.

Thus, within a social contract there are social conventions that evolve with varying speeds. Individual best strategies depend on and influence the changing social conventions and in this way there is a form of parallel evolution of the social contract, its conventions and the best individual strategies. This coevolution of individual behaviour and social conventions and the extent to which individuals affect equilibria will be discussed next.

8.5 The individual and the social contract

The fundamental cornerstone of each social group and society is the social contract. The following paragraphs will attempt to examine to what degree the assumed interdependence of the social contract and social conventions, and the central role of individual action, make sense in the real world.

In the discussion in this and the previous chapters, the social contract is seen as consisting of social conventions that arise from repeated interactions. At the same time, social structures define and bound rational agents' strategy sets and social behaviour. Therefore, there is a bi-directional relationship between social structures, namely social conventions and the social contract, and individual action. A change in the social contract has to occur through a gradual change in its social conventions in order to be sustained. Thus, the social contract is seen as the equilibrium in a repeated game which is reached only when its sub-games have reached their respective topical equilibria.

A shift in the super-game equilibrium will not necessarily cause all the sub-game

equilibria to collapse, especially when they are evolutionarily stable. However, for a social contract to shift to a new evolutionarily stable equilibrium, most of its component social conventions will have to reach new evolutionarily stable equilibria; the coevolution of the social contract and its conventions does not require the collapse of either, but it can be achieved as an adjustment to the existing rules of behaviour. Cultural evolution can explain how social contracts came about and why they are stable, but it does not include an analysis of their optimisation (Young 2001) since, in the evolutionary analysis, an equilibrium can be evolutionarily stable but not optimal. Equilibrium optimality depends on the optimality of the constitutive social conventions. In turn, social conventions depend on individual behaviour. Rationality in interactions can cause equilibria to shift towards more efficient or optimal states.

Conventions such as driving on the left cannot change over an election cycle or even a generation, when they have been established for centuries. Of course, the driving convention is only an over-simplified convention that does not affect social welfare. More important and more deeply rooted conventions are more complicated structurally and so more difficult to shift. For instance, the imposition of democracy in societies without a democratic culture will not lead to a democratic state or a society with democratic values; democracy depends on a multitude of institutions and behaviours throughout society and the political realm, and a third party imposition that requires drastic convention change in short periods cannot be effective. Holding an election when there is not a social culture of democratic principles, or in other words established topical conventions that are not democratic, cannot make a society democratic.

This poses questions about the level of change individual behaviour can cause. However, given the accelerated time of cultural evolution and information spreading that is possible nowadays, it is realistic to assert that a local convention can bring about change at a social contract level within reasonable time. For instance, in many parts of Africa female genital mutilation was an established convention until recently, despite social and economic progress and a rise in literacy. This changed after a small number of villages decided collectively to abandon the practice, which triggered the change in several other village clusters, to the point that it became illegal in Senegal only a year after the first village meeting (Bowles, 2009). Rational individuals can cause large scale change by establishing higher utility social conventions in their local interactions.

Rational actors within conventions have been assumed to be allowed to behave according to their rational deliberation even if that leads them to non-conventional

strategies. This makes evolution of conventions and the process by which a society reaches higher welfare states possible. For this to be achieved, freedom of individual action and absence of coercion are essential so as that rational agents can use all the information available to them, learn from their own past and adapt to environmental changes, in order to make maximising decisions. Agents who behave non-conventionally are punished by exclusion but any other authoritative form of behaviour enforcement would mean that the evolutionary process would be skewed. It is imperative to keep in mind that cultural evolution includes errors and therefore, irrational behaviour and non-optimal equilibria are useful as lessons of what to avoid in the future and also as a method to try new strategies that might prove more efficient. Learning through trial and error applies both to rational agents and social conventions and in order to have an evolutionary process of social conventions it is vital that individuals experiment and make errors. Therefore, individual behaviour does affect the evolution of conventions and indirectly the evolution of the social contract. Individual rational strategies evolve in parallel with the evolution of the social contract.

The question of how one ought to act relates, in the context presented here, to the issue of whether individual action can influence social structures or society determines individual behaviour. The essence of the question is whether an individual has any real power over collective decisions and more importantly over social structures and their evolution. To an extent, Karl Marx was right to claim that man makes his own history in predefined circumstances (Hollis, 1994). If the social contract consists of social conventions that rational agents should follow, then it might seem there is little room for personal responsibility and individual rationality once these conventions are established. However, the social contract through social conventions is dynamic and individual behaviour can cause change on a social level. The individual has the responsibility to be rational through adaptation, imitation and learning (Young, 2001).

A further problem with the realism of this account of individual action within a society that is described by evolutionary conventions is the presence of extreme inequalities. Even though the stability of the social contract is not necessarily threatened, extreme inequalities as well as the use of coercion threaten its cohesion and optimality. Information symmetries level the field to a degree but in real life there are significant inequalities within a society, or in other words among members of the same social contract. When there are extreme inequalities – as is often the case in the real world – rational interactions become impossible and therefore there is a need for formal

institutions to regulate interactions. Any institution whose purpose would be to reduce those inequalities would therefore have to provide the means for equal access to information and ensure inter-conventional discrepancies do not increase inequalities. In that respect, institutions are required to facilitate information spreading and to ameliorate extreme inequalities, especially between conventions.

Interactions among rational utility maximisers should lead to commonly accepted outcomes that are Pareto efficient, but this is not always the case. Especially when there are information discrepancies that make rational deliberation reach different outcomes. There are, in this sense, discrepancies in rational deliberation outcomes which nullify the assumptions of equal bounded rationality and rational utility maximisers. In these cases, there is a need for a third party that addresses issues of incompatibility between social conventions. The more homogeneous a society, the less the need for a third party enforcer. For instance in a society where social conventions have converged so that rational interactions are possible does not need an external enforcer. However, this only occurs locally and within conventions. Social conventions that are homogeneous enough to ensure acceptance of the status-quo or agreement on any deviation are more plausible.

Social institutions are in themselves conventions and are bound by the same rules as lower level conventions like the driving game. Therefore, any type of institution that has evolved without external enforcement will ensure the best available results in terms of social welfare within the specific environmental parameters. Administrative and legal institutions have evolved in similar ways as the driving conventions. Therefore, they cannot be changed arbitrarily, without underlying conventions shifting as well. Social conventions that define cultural norms affect formal conventions such as the government and the legal system. Hence, they are supportive of the formal institutions and they have to shift before any institutional change becomes realistic.

The convergence of social conventions and individual rationality may very well lead to consensus over important to matters. This is a direct implication of asserting that rationality can be common to all and given the same information rational agents will reach the same conclusions. This not always empirically true. Rational individuals with access to the same information often make different decisions. Therefore, convergence of individual behaviour is not always a given. In a world of competing conventions and individual behaviour, it is easier to argue that there can be an evolutionary process of convention and behaviour convergence. Societies with different historical backgrounds

choose different conventions of social behaviour. One cannot change the past but one has to adapt to it in the best possible way. A maximising strategy is maximising given a specific environment and in diverse environments there will be different maximising strategies. Absence of coercion means that one culture cannot enforce specific equilibria on others using universal moral laws as justification. Each convention ought to reach an optimal equilibrium through its own specific historical evolution. Therefore, there is not much point or need to discuss an optimal social contract. There can be more than one optimal social contract, given that optimality depends on the underlying conventions. Different conventions support different social contracts and imitating more successful social contracts without the appropriate conventions cannot be sustainable. On this view the social contract cannot be seen as a mechanism for social peace that should be designed for optimality. Although it is apparent that "mechanism design is based on the obvious principle that decision-making should be decentralized to the people who have the necessary knowledge and experience" (Binmore, 2005: 136), deciding who those people are depends on the values of each society.

Top-down or bottom-up accounts of social and individual behaviour (Hollis, 1994), are therefore misplaced. One has to take into account a complementary view of coevolution of social structures and individual behaviour for an accurate description of how human societies and individuals within them, operate.

8.6 Conclusion

Paraphrasing Hobbes (1976), it is rational to be nasty in a world of nasties. That implies that an individual has rational and not moral obligations, or differently that a rational agent has moral obligations only when surrounded by agents who behave morally. In a social convention of defectors, it is moral to defect. Thus, the analysis presented here does not provide a strong normative argument for moral behaviour. But it does aspire to present a realistic account of moral conventions and the importance of rational agency and individual action in influencing their direction. Hume (2008) unrealistically suggested that people should move to a different country if they disagree with the local convention. In the modern world this is more plausible than it was in Hume's time, but still not very realistic. However, in modern societies individuals can participate in conventions without being limited by geography and they also have more power to influence their local convention. Change and the establishment of optimal equilibria may be incremental and slow, but it can be realised and thus individuals have an

obligation to promote and maintain it through their actions.

The rationale behind any political philosophy thesis is to provide an argument for a better social arrangement; and the incentive behind contemporary political philosophy is related to extreme social inequalities and a more or less widespread view that it is in societies' power to rectify this situation. Historically, top-down approaches where social problems are solved by strong centralised institutions have not proven efficient in alleviating extreme inequalities. Centralisation inhibits information availability and spreading and, therefore, the possibility of roughly equal rationality. Moreover, centralisation by controlling information makes rational behaviour a structural and not an agency problem, which has apparent consequences for individual behaviour within social conventions. The coevolution of individual agency and social structures and the subsequent understanding of individual moral obligations, makes the argument presented here primarily one in moral philosophy. The political philosophy implications are mostly implied and can be derived from the discussion of the role of the individual in relation to the role of society.

The discussion in this chapter is mostly normative in terms of individual behaviour and descriptive in terms of social structures. There is a bi-directional relationship between the two so that there can be argued to be coevolution between what is considered a rational strategy and how this affects social equilibria and how social conventions provide a framework for individual rationality. In this framework, social conventions of fairness are indispensable for social contracts of justice.

9. Conclusion

The aim of this thesis has been to reinforce Gauthier's argument for a rational morality. This has been done through the use of an account of social interactions that is based on social conventions. Despite the fact that *Morals by Agreement* offers a convincing theory of rational morality, it has some significant weaknesses (as discussed in §2.6). In particular, the internalisation of constrained maximisation is unrealistic and the smuggling in of non-rational considerations in the theory undermines its rationalist credentials.

The conventional account presented here bypasses these problems by suggesting that agents acting within a society adopt joint strategies since social interactions take place in an environment where interactions are repeated. The force of the account of conventional rationality lies in that it does not need a mechanism for internalising a type of behaviour or for guessing others' future behaviour. On the contrary, conventional interactions promote and support a given behaviour in the light of past interactions. Therefore, conventional rationality offers a more plausible theoretical argument and a realistic description of actual social life while being impervious to criticism about moral constraints on rational behaviour.

Social conventions, their dynamics and their structures, have been described through an evolutionary account of social structure. Social interactions and social structures have been viewed as the result of repeated interactions between rational agents. These interactions can by analysed via traditional game theory, as in *Morals by Agreement*. However, the complexity of the multitude of interactions occurring in a society demands a paradigm that describes population interactions and not just a model of interactions between two agents. In that respect, the evolutionary account adds a dimension of realism to social explanation and to the description of human interactions.

The tension between evolutionary theory and rational choice theory is obvious; the former refers to groups or populations and the latter to the behaviour of individuals. The methodological analysis of Chapter Four – and in particular §4.2 and §4.3 – deals with questions related to the possibility of a theoretical model that includes both rational choice theory and the conventional and evolutionary accounts of social behaviour. It was shown that methodological individualism and holism can be reconciled given certain conditions such as bounded rationality. The combination of these methodological paradigms shows that the conventional account is a realistic account of individual and

social behaviour, taking into consideration both individual rational agency and the benefits gained from adopting joint strategies in a social group context.

Chapter Four emphasised the plausibility of individual rational behaviour in an evolutionary theory context, while Chapter Five showed that, under realistic conditions, the concept of constrained maximisation as presented in *Morals by Agreement* can be replaced by conventional rationality. Hence, morality has been shown to be an outcome of repeated interactions between rational agents within social conventions. In this understanding, moral behaviour depends on the local equilibrium, since the established social convention determines moral behaviour and therefore morality is convention-specific. The conventional account of social behaviour means that various rational conventions may reach different accounts of moral behaviour; since these conventions are the outcome of rational interactions their respective understanding of morality is defensible from a rational choice perspective. Thus, the conventional account of behaviour makes it possible to link rational and moral behaviour and to show that morality can be the outcome of interactions between rational, mutually concerned actors.

Moreover, as demonstrated in Chapter Seven, conventions of rational behaviour combine nicely with theories of justice as mutual advantage. Accounts of justice as mutual advantage have been criticised because they do not make strong claims about interactions with the vulnerable. Rational conventions can answer this criticism (as was shown in §7.3). The conventional account of behaviour is also an account of justice as mutual advantage in the context of repeated interactions between rational agents. In Chapter Seven, it was shown that justice is linked to rational behaviour and the outcome of repeated rational interactions, a claim similar to the one made by traditional theories of justice as mutual advantage. In addition, the account of social conventions makes it possible to claim that interactions with the vulnerable are rational and therefore can be included in a theory of justice as mutual advantage.

Thus, as discussed in the Introduction, a methodological basis has been provided to show that it is possible to base morality and justice on the rational behaviour of mutually unconcerned actors. Relaxing the assumptions of individual rationality and accepting that humans are boundedly rational within a framework of social conventions allows us to show how morality is derived from rationality. In addition, the conventional understanding of social life has led to the conclusion that there can be more than one acceptable account of moral behaviour.

Future developments

The methodological analysis deployed in this thesis can be expanded provided that our understanding of the limits of humanities and the possibilities of interdisciplinary analysis change. In order for this to happen, the integration of humanities and social science needs to continue and intensify (without excluding contributions from science, such as the formal modelling of game theory and evolutionary game theory). The use of game theoretical models has enriched moral and political philosophy and given rise to a wealth of literature led by *Morals by Agreement*.

It is reasonable to claim that advancements in formal game theory will offer new ways to look at philosophical issues. Binmore admits that there is much to be done in game theory, especially with respect to coalitions within groups (Binmore, 1998). Obviously, such developments can have a profound effect on an analysis that attempts to combine individual rationality and moral behaviour. Therefore, it is safe to say, moral philosophy can be enhanced significantly through breakthroughs in social science. This has been demonstrated in this thesis where I have presented a game theoretical model that is able to account for groups' behaviour as well as individual maximisation.

Game Theory and the Social Contract (Binmore 1998) is an innovative work that in many respects shows the future direction of moral philosophy. Contractarianism in conjunction with a realistic model of analysis of human behaviour – such as is offered in game theory – can broaden the horizons of moral and political philosophy and enrich our understanding of human behaviour and society. *Chaos in Game Dynamics* (Skyrms, 1992) and other similar works – for example, *Chaos Theory in the Social Sciences* (Kiel & Elliott, 1997) – give a clear indication of the direction of travel of these types of enquiries. Complexity theory in behavioural science and the humanities is in its infancy now, just as traditional game theory was before the publication of *The Theory of Games as a Tool for the Moral Philosopher* (Braithwaite, 2009) and Gauthier's work. However, in the future it can enrich social and behavioural science, and by extension moral and political philosophy by offering an account of human behaviour that considers all, or at least most of, the complexities of real social life.

As discussed in Chapter Four, one major point of tension with the analysis of this thesis is methodological. The challenge, remember, was to combine the methodological individualism of game theory with the holistic approach of evolutionary theory. Sections §4.2 and §4.3, in particular, addressed the issue and explained how it could be resolved. In order to expand and advance theories of moral contractarianism,

we have to consider their limitations and the possibility of inserting concepts from holistic explanations of social behaviour. The argument proposed in the previous chapters is based to an extent on an understanding of human behaviour and rationality that is the result of considering two distinct methodological paradigms. Having shown that it is feasible to talk about rationality and at the same time about social equilibria (Chapters Four and Five), the conventional account of behaviour can be offered as a convincing theoretical paradigm of rational morality.

In addition to these methodological tensions, the ideas of practical rationality and rational morality as expressed both in this thesis and in *Morals by Agreement* can have practical applications in the organisation of social and economic life. Concepts from economic theory have flooded the humanities and as a result scientific methods threaten analytical fields such as moral and political philosophy. In a sense, Gauthier follows this path by importing rational choice theory into moral philosophy (Gauthier, 1986). However, as was discussed in the Introduction and in Chapter Four, the premises of moral and political philosophy, especially in the work of both Rawls (2005) and Gauthier, remains unaffected. Rational choice theory and game theory serve as tools of examining interactions and not as a substitute for moral philosophy. The fact that many of the arguments made by contemporary political philosophers were first made centuries ago by Hobbes and Hume, without being put in the terms of the modern formal approach, confirms this point.

Economic imperialism does not have to be a one way street. The conclusions we draw from fields such as analytic philosophy can be used in economic life. For example, the social change that was initiated by advancements in information technology has made companies that had succeeded by following traditional economic models more sensitive to competition from companies following alternative methods of production. Complete information is a basic assumption in neoclassical economics and near complete information is now a realistic possibility. This is a development that may have great impact on economic behaviour in the real world and lead to a reconsideration of traditional theoretical models. Individual may choose a behaviour that is closer to constrained maximisation, than typical straightforward maximisation as described in economic models. In that respect, it is plausible to say that moral norms can be used to inform economic behaviour and market activity.

Profitable companies such as eBay have not followed the traditional economic model. eBay started off as an oddity where trust between users was essential and

evolved into a market similar to Gauthier's ideal market (Gauthier, 1986; Cohen, 2003). Complete information about others' past interactions is known and reputation in all interactions matters. Moreover, Freemium companies such as Dropbox, welcome free-riders in economic terms, as long as they do not free-ride in spreading the word; or put differently, helping the company establish a reputation as a cooperator. The same applies to open source software companies such as Canonical that offer their product for free to private individuals while charging business users. They aim to improve their reputation while attracting individuals to make small (or great) modifications to the original product without expecting payment.

Arguably, these cases primarily concern software companies where distribution of the product is easy and almost costless. However, the point of these examples is to show that it is plausible to claim that norms of rational morality can inform economic behaviour, not just theoretically but in the real world as well. In doing so, maximisation is not compromised, thus proving that a version of Gauthier's constrained maximisation is plausible and realisable. The incentive for the creation of companies like these was not moral, and it did not have to be, but still the result from the operation has been maximisation. Moreover, the fact that these are examples of alternative economic models does not reduce their importance for moral philosophy and for showing that assumptions about the link between morality and rationality can be realised and be effective in the real world. This applies especially in today's societies where information is cheaper and more accurate than ever before for an increasing number of people irrespective of their economic situation.

Advancements in the methodology of behavioural sciences and the humanities can be the result of the realisation that the interdependence of various fields of study of human behaviour is not just useful but essential if we are to have an accurate account of human interactions. The implications of the possible methodological breakthroughs for moral and political philosophy will be enormous given that they will affect greatly our understanding of rationality and moral behaviour. Philosophy, and specifically moral and political philosophy, is “a voyage that is not, and cannot be, completed, but that finds a temporary harbour” (Gauthier, 1986: preface). It seems that especially in the humanities and philosophy we cannot and should not expect an end, but rather a process of incremental steps, taking us closer to a more accurate understanding of individual behaviour and social life. Therefore, we can be satisfied that moral philosophy has advanced tremendously in the 20th century with works such as *Morals by Agreement*,

but at the same time we should expect that our present understanding is temporary and simply a step in the process of intellectual progress.

Bibliography

Alexander JM (2000) Evolutionary Explanations of Distributive Justice. *Philosophy of Science*, 67(3), 490–516.

Alexander JM (2009) Evolutionary Game Theory. Available from: http://plato.stanford.edu/entries/game-evolutionary/ (accessed 17 March 2010).

Alexander JM (2007) *The structural evolution of morality*. Cambridge University Press.

Andrés Guzmán R, Rodríguez-Sickert C and Rowthorn R (2007) When in Rome, do as the Romans do: the coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior*, 28(2), 112–117.

Arthur WB, Durlauf SN and Lane DA (1997) *The economy as an evolving complex system II*. Addison-Wesley.

Axelrod R (1986) An Evolutionary Approach to Norms. *The American Political Science Review*, 80(4), 1095–1111.

Axelrod R (1981) The Emergence of Cooperation among Egoists. *The American Political Science Review*, 75(2), 306–318.

Axelrod R and Cohen MD (2000) *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. Simon and Schuster.

Axelrod R and Keohane RO (1985) Achieving Cooperation under Anarchy: Strategies and Institutions. *World Politics*, 38(1), 226–254.

Axelrod RM (2006) *The evolution of cooperation*. Basic Books.

Barry B (1991) *Theories of justice*. University of California Press.

Bergstrom TC (2002) Evolution of Social Behavior: Individual and Group Selection. *The Journal of Economic Perspectives*, 16(2), 67–88.

Bicchieri C (2006) *The grammar of society*. Cambridge University Press.

Bicchieri C, Jeffrey RC and Skyrms B (1997) *The dynamics of norms*. Cambridge University Press.

Binmore K (2001) Evolutionary Social Theory: Reply to Robert Sugden. *The Economic Journal*, 111(469), F244–F248.

Binmore K (2007) Rational Decisions in Large Worlds. *Annales d'Économie et de Statistique*, (86), 25–41.

Binmore K (1989) Social Contract I: Harsanyi and Rawls. *The Economic Journal*, 99(395), 84–102.

Binmore K (1999) Why Experiment in Economics? *The Economic Journal*, 109(453), F16–F24.

Binmore K (1998) *Game Theory and the Social Contract: Just playing*. MIT Press.

Binmore K (1994) *Game Theory and the Social Contract: Playing Fair v. 1*. MIT Press.

Binmore K (2005) *Natural justice*. Oxford University Press.

Binmore K (2007d) *Playing for real: a text on game theory*. Oxford University Press.

Binmore K and Samuelson L (1999) Evolutionary Drift and Equilibrium Selection. *The Review of Economic Studies*, 66(2), 363–393.

Bowles S (2011) Cultivation of cereals by the first farmers was not more productive than foraging. *Proceedings of the National Academy of Sciences of the United States of America*, 108(12), 4760–4765.

Bowles S (2009) The Coevolution of Institutions and Preferences: History and Theory. *Working paper*.

Bowles S and Gintis Herbert (2011) *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press.

Boyd Robert and Richerson PJ (1988) *Culture and the Evolutionary Process*. University of Chicago Press.

Braithwaite RB (2009) *Theory of Games as a Tool for the Moral Philosopher*. 1st ed. Cambridge University Press.

Broome J (1999) *Ethics out of economics*. Cambridge University Press.

Campbell R and Sowden L (1985) *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. UBC Press.

Cohen A (2003) *The Perfect Store: Inside EBay*. New Ed. Piatkus.

Coleman JL and Morris CW (1998) *Rational commitment and social justice: essays for Gregory Kavka*. Cambridge University Press.

Cubitt RP and Sugden R (1998) The Selection of Preferences Through Imitation. *The Review of Economic Studies*, 65(4), 761–771.

Cudd A (2007) Contractarianism. Available from: http://plato.stanford.edu/entries/contractarianism/#3 (accessed 24 November 2010).

D'Agostino F and Gaus G (2011) Contemporary Approaches to the Social Contract. Available from: http://plato.stanford.edu/entries/contractarianism-contemporary/ (accessed 24 November 2010).

Daniels N (2011) Reflective Equilibrium. Available from: http://plato.stanford.edu/entries/reflective-equilibrium/ (accessed 12 April 2010).

Danielson P (1998) *Modeling rationality, morality, and evolution*. Oxford University Press US.

Danielson, P (2002). Artificial Morality: Virtuous Robots for Virtual Games. Routledge.

D'Arms J, Batterman R and Krzyzstof Gorny (1998) Game Theoretic Explanations and the Evolution of Justice. *Philosophy of Science*, 65(1), 76–102.

Dawkins R (2006) *The selfish gene*. Oxford University Press.

Diamond J (1994) Ecological Collapses of Past Civilizations. *Proceedings of the American Philosophical Society*, 138(3), 363–370.

Diamond J (2000) How to Organize a Rich and Successful Group: Lessons from Natural Experiments in History. *Bulletin of the American Academy of Arts and Sciences*, 53(4), 20–33.

Diamond J (1997) Location, Location, Location: The First Farmers. *Science*, New Series, 278(5341), 1243–1244.

Diamond J (2005) *Collapse*. Viking.

Dimock S (2010) Defending Non-Tuism. *Canadian Journal of Philosophy*, 29(2), 251–273.

Durant W (2010) *The Lessons of History*. Simon & Schuster.

Ehrlich PR (2002) *Human Natures: Genes, Cultures, and the Human Prospect*. Reissue. Penguin Putnam.

Elster J (1985) Rationality, Morality, and Collective Action. *Ethics*, 96(1), 136–155.

Frank RH (1988) *Passions Within Reason: The Strategic Role of Emotions*. 1st ed. W W Norton & Co Inc

Frey RG and Morris CW (1993) *Value, welfare, and morality*. Cambridge University Press.

Gaus GF (2007) On Justifying the Moral Rights of the Moderns: A Case of Old Wine in New Bottles. *Social Philosophy and Policy*, 24(01), 84–119.

Gaus GF (2011) *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge, Cambridge University Press.

Gauthier D (1994) Assure and Threaten. *Ethics*, 104(4), 690–721.

Gauthier D (1972) Brandt on Egoism. *The Journal of Philosophy*, 69(20), 697–698.

Gauthier D (1979a) David Hume, Contractarian. *The Philosophical Review*, 88(1), 3–38.

Gauthier D (1984) Deterrence, Maximization, and Rationality. *Ethics*, 94(3), 474–495.

Gauthier D (1988a) In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality). *Proceedings of the Aristotelian Society*, New Series, 89, 179–194.

Gauthier D (1988b) Moral Artifice. *Canadian Journal of Philosophy*, 18(2), 385–418.

Gauthier D (1986) *Morals by agreement*. Clarendon Press.

Gauthier D (1987) Reason to Be Moral? *Synthese*, 72(1), 5–27.

Gauthier D (1997) Resolute Choice and Rational Deliberation: A Critique and a Defense. *Noûs*, 31(1), 1–25.

Gauthier D (1974b) The Impossibility of Rational Egoism. *The Journal of Philosophy*, 71(14), 439–456.

Gauthier D (1977) The Social Contract as Ideology. *Philosophy and Public Affairs*, 6(2), 130–164.

Gauthier D (1979b) Thomas Hobbes: Moral Theorist. *The Journal of Philosophy*, 76(10), 547–559.

Gauthier D (1990) *Moral dealing*. Cornell University Press.

Gauthier D and Sugden R (eds) (1993) *Rationality, Justice and the Social Contract: Themes from 'Morals by Agreement'*. The University of Michigan Press.

Geiger G (1985) Review: Is Life a Non-Zero-Sum Game? *Politics and the Life Sciences*, 4(1), 80–81.

Gigerenzer Gerd (2002) *Bounded Rationality: The Adaptive Toolbox*. New Ed. MIT Press.

Gintis H (2006) *Moral Sentiments and Material Interests: The Foundation of Cooperation in Economic Life*. New ed. MIT Press.

Gintis H (2009a) *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction (Second Edition)*. Princeton University Press.

Gintis H (2009b) *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press.

Gosseries A and Meyer LH (2009) *Intergenerational justice*. Oxford University Press.

Gould SJ (2008) *The Panda's Thumb: More Reflections in Natural History*. Paw Prints.

Hamlin A (1987) Review of The Economics of Rights, Co-operation and Welfare, by R. Sugden. Economic Journal (97), 751–3.

Hardin G (1968) The Tragedy of the Commons. *Science*, New Series, 162(3859),

1243–1248.

Hardin R (2003) The Free Rider Problem. Available from: http://plato.stanford.edu/entries/free-rider/ (accessed 24 February 2011).

Harsanyi JC, Leinfellner W and Köhler E (1998) *Game theory, experience, rationality*. Springer.

Hartogh GD (1993) The Rationality of Conditional Cooperation. *Erkenntnis (1975-)*, 38(3), 405–427.

Hausman D (1989) Are Markets Morally Free Zones? *Philosophy and Public Affairs*, (18), 317–33.

Heap SH (1989) *Rationality in economics*. B. Blackwell.

Heath J (2011) Methodological Individualism. Spring 2011. In: Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy*, Available from: http://plato.stanford.edu/archives/spr2011/entries/methodological-individualism/ (accessed 27 August 2012).

Henrich J, Albers W, Boyd R., et al. (2001) What is the role of culture in bounded rationality? In: *Bounded rationality The adaptive toolbox Dahlem Workshop Report*, pp. 343–359.

Hilgevoord J and Uffink J (2012) The Uncertainty Principle. Summer 2012. In: Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy*, Available from: http://plato.stanford.edu/archives/sum2012/entries/qt-uncertainty/ (accessed 22 August 2012).Hobbes T (1976) *Leviathan*. Forgotten Books.

Hobbes, T., 1976. Leviathan. Forgotten Books.

Hofbauer J and Sigmund K (1998) *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Hollis (2007) *Rational Economic Man*. Cambridge University Press.

Hollis M (1996) *Reason in action: essays in the philosophy of social science*. Cambridge University Press.

Hollis M (1994) *The philosophy of social science: an introduction*. Cambridge University Press.

Hoven J van den and Weckert J (eds) (2009) *Information Technology and Moral Philosophy*. 1st ed. Cambridge University Press.

Hume D (2008) *A Treatise of Human Nature*. NuVision Publications, LLC.

Hume D (1987) *Essays, moral, political, and literary*. LibertyClassics.

Jon Elster (1982) The Case for Methodological Individualism. *Theory and Society*, 11(4), 453–482.

Katz LD (2000) *Evolutionary Origins of Morality*. Imprint Academic.

Kavka GS (1983) The Toxin Puzzle. *Analysis*, 43(1), 33–36.

Kiel LD and Elliott EW (1997) *Chaos theory in the social sciences*. University of Michigan Press.

Kitcher P (2011) *The Ethical Project*. Harvard University Press.

Knight C and Stemplowska Z (eds) (2011) *Responsibility and Distributive Justice*. OUP Oxford.

Kymlicka W (1990) *Contemporary political philosophy: an introduction*. Clarendon Press.

Levin J (2010) Functionalism. Summer 2010. In: Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy*, Available from: http://plato.stanford.edu/archives/sum2010/entries/functionalism/ (accessed 27 August 2012).

Locke J (1988) *Two Treatises of Government*. 3rd ed. Cambridge University Press.

Martin and McIntyre (1994) Readings in the philosophy of social science. Cambridge, Mass. ; London, MIT Press.

Matravers M (2000) *Justice and punishment: the rationale of coercion*. Oxford University Press.

Matravers M (2003) *Scanlon and contractualism*. Psychology Press.

Matsui A (1996) On Cultural Evolution: Social Norms, Rational Behavior, and Evolutionary Game Theory. *Journal of the Japanese and International Economies*, 10(3), 262–294.

McClennen EF (1990) *Rationality and dynamic choice: foundational explorations*. Cambridge University Press.

Morris CW and Ripstein A (2001) *Practical rationality and preference: essays for David Gauthier*. Cambridge University Press.

Mueller DC (2003) *Public Choice III*. Cambridge University Press.

Murray M (1999) How to Blackmail a Contractarian. *Public Affairs Quarterly*, 13(4), 347–361.

Murray M (2007) *The Moral Wager: Evolution and Contract*. 1st ed. Springer.

Murray RM, Murray M and Narveson J (2007) *Liberty, games and contracts: Jan Narveson and the defence of libertarianism*. Ashgate Publishing, Ltd.

Narveson J (1999) *Moral matters*. Broadview Press.

Narveson J and Dimock S (2000) *Liberalism: new essays on liberal themes*. Springer.

Nida-Rümelin J and Spohn W (2000) *Rationality, rules, and structure*. Springer.

Okasha S (2006) *Evolution and the levels of selection*. Oxford University Press.

Okasha S (2002) *Philosophy of science: a very short introduction*. Oxford University Press.

Olson M (1965) *The logic of collective action: public goods and the theory of groups*. Harvard University Press.

Ostrom E (2000) Collective Action and the Evolution of Social Norms. *The Journal of Economic Perspectives*, 14(3), 137–158.

Parkin J (2011) Straw Men and Political Philosophy: The Case of Hobbes. *Political Studies*, 59(3), 564–579.

Plato (2006) *The Republic*. Yale University Press.

Popper K (2002) *The Logic of Scientific Discovery*. Routledge.

Pujol N (2010) Freemium: Attributes of an Emerging Business Model. *SSRN eLibrary, Available* from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1718663 (accessed 27 July 2012).

Rawls J (2005) *A theory of justice*. Harvard University Press. Rawls J (2003) *Justice as fairness: a restatement*. Harvard University Press.

Raz J (1999) *Practical Reason and Norms*. Oxford University Press.

Rescorla M (2011) Convention. Spring 2011. In: Zalta EN (ed.), *The Stanford*

Encyclopedia of Philosophy, Available from: http://plato.stanford.edu/archives/spr2011/entries/convention/ (accessed 28 August 2012).

Richards D (2000) *Political complexity*. University of Michigan Press.

Ridley M (1994) *The Red Queen: Sex and the Evolution of Human Nature*. Penguin UK.

Riley J (2006) Genes, Memes and Justice. *Analyse & Kritik*, 28/2006, 32–56.

Ross, H. L. (1994). *Confronting Drunk Driving: Social Policy for Saving Lives*. Yale University Press.

Rosas A (2010) Evolutionary game theory meets social science: Is there a unifying rule for human cooperation? *Journal of Theoretical Biology, Available* from: http://www.ncbi.nlm.nih.gov/pubmed/20167223 (accessed 27 February 2010).

Rousseau J-J and Cole GDH (2008) *The Social Contract*. Cosimo, Inc.

Rubinstein A (1998) *Modeling bounded rationality*. MIT Press.

Samuelson L (1998) *Evolutionary Games and Equilibrium Selection*. MIT Press.

Schelling TC (2006) *Micromotives and macrobehavior*. Norton.

Sen A (1988) *On ethics and economics*. Wiley-Blackwell.

Sen A (2009) *The Idea of Justice*. Harvard University Press.

Sidgwick H (1907) *The Methods Of Ethics*. Hackett.

Simpson M (2004) Brian Skyrms, The Stag Hunt and the Evolution of Social Structure. *Ethics*, 115(1), 166–169.

Skyrms B (1992) Chaos in Game Dynamics. *Journal of Logic, Language, and Information*, 1(2), 111–130.

Skyrms B (2010) *Signals: Evolution, Learning, and Information*. OUP Oxford.

Skyrms B (2004) *The stag hunt and the evolution of social structure*. Cambridge University Press.

Smith A and Haakonssen K (2002) *The theory of moral sentiments*. Cambridge University Press.

Smith JM (1982) *Evolution and the theory of games*. Cambridge University Press.

Sugden R (1995a) A Theory of Focal Points. *The Economic Journal*, 105(430), 533–550.

Sugden R (1990) Contractarianism and Norms. *Ethics*, 100(4), 768–786.

Sugden R (1991) Review: Impartiality and Mutual Advantage. *Ethics*, 101(3), 634–643.

Sugden R (2001a) Review: Ken Binmore's Evolutionary Social Theory. *The Economic Journal*, 111(469), F213–F243.

Sugden R (1989) Spontaneous Order. *The Journal of Economic Perspectives*, 3(4), 85–97.

Sugden R (2004) *The economics of rights, co-operation, and welfare*. Palgrave Macmillan.

Sugden R (2001b) The evolutionary turn in game theory. *Journal of Economic Methodology*, Journal of Economic Methodology, 8(1), 113–130.

Sugden R (1998) The Role of Inductive Reasoning in the Evolution of Conventions. *Law and Philosophy*, 17(4), 377–410.

Sugden R (2004b) What Public Choice and Philosophy Should Not Learn from One Another. *American Journal of Economics and Sociology*, 63(1), 207–211.

Sugden R and Weale A (1979) A Contractual Reformulation of Certain Aspects of Welfare Economics. *Economica*, New Series, 46(182), 111–123.

Taylor M (1987) *The possibility of cooperation*. Cambridge University Press.

Taylor PD and Jonker LB (1978) Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2), 145–156.

Thornton, M. (1991) Alcohol Prohibition Was A Failure. Retrieved March 14, 2013, from http://www.cato.org/pubs/pas/pa-157.html

Ullmann-Margalit E (1977) *The emergence of norms*. Clarendon Press.

Vallentyne P (1989) Contractarianism and the Assumption of Mutual Unconcern. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 56(2), 187–192.

Vallentyne P (ed.) (1991) *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*. New York, Cambridge University Press.

Vanderschraaf P (1999) Game Theory, Evolution, and Justice. *Philosophy & Public Affairs*, 28(4), 325–358.

Vanderschraaf P (2011) Justice as mutual advantage and the vulnerable. *Politics, Philosophy & Economics*, 10(2), 119 –147.

Varoufakis Y (2008) Game Theory: Can it Unify the Social Sciences? *Organization Studies*, 29(8-9), 1255–1277.

Vatta K (2003) Microfinance and Poverty Alleviation. *Economic and Political Weekly*, 38(5), 432–433.

Verbeek B (2002) *Instrumental rationality and moral philosophy: an essay on the virtues of cooperation*. Springer.

Ward H (1979) A Behavioural Model of Bargaining. *British Journal of Political Science*, 9(2), 201–218.

Weibull JW (1995) *Evolutionary Game Theory*. Cambridge, Mass, MIT Press.

Weirich P (2011) Exclusion from the social contract. *Politics, Philosophy & Economics*, 10(2), 148 –169.

Young HP (1993a) An Evolutionary Model of Bargaining. *Journal of Economic Theory*, 59(1), 145–168.

Young HP (2001) *Individual strategy and social structure*. Princeton University Press.

Young HP (1993b) The Evolution of Conventions. *Econometrica*, 61(1), 57–84.

Alphabetical Index

A
agreement 18, 19, 23, 24, 26, 27, 28, 29, 30, 32, 35, 41, 42, 43, 55, 64, 65, 71, 73, 89, 102, 104, 105, 106, 107, 111, 112, 113, 114, 115, 116, 123, 124, 134, 146, 151, 152, 157, 173, 174, 196
Archimedean Point.....18, 32, 33, 34
Association.....60
ational morality.....2, 17, 31, 32, 35, 38, 43, 61, 75, 176, 181
B
bargain.…18, 22, 23, 24, 25, 26, 27, 28, 30, 34, 36, 37, 39, 40, 42, 43, 44, 49, 55, 56, 63, 64, 65, 66, 77, 105, 106, 107, 112, 113, 114, 115, 116, 121, 122, 123, 124, 125, 129, 142, 155, 156, 157, 158
bargaining.....38
Binmore 40, 43, 45, 47, 48, 49, 50, 62, 63, 64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 77, 113, 117, 122, 123, 124, 157, 158, 160, 161, 162, 210
biological.....
    evolution.....47, 48, 62, 63, 65, 66, 67, 73, 91, 108, 121, 122, 125, 133, 161, 187
C
Cheap talk.....57
Coevolution.....59, 60, 206
concession..18, 19, 23, 24, 25, 27, 28, 29, 31, 32, 34, 35, 36, 37, 38, 39, 40, 42, 43, 101, 106, 115, 124
constrained.....
    maximisation…2, 17, 18, 19, 29, 30, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41, 42, 44, 45, 52, 55, 56, 58, 61, 74, 75, 76, 77, 80, 81, 82, 91, 97, 98, 99, 100, 101, 102, 104, 124, 126, 127, 129, 135, 152, 157, 158, 170, 191
    social contract.....43, 49, 52, 53, 56, 59, 60, 61, 62, 64, 67, 105, 107, 108, 109
constrained maximisation.....2
constrained maximisers.….2, 17, 29, 30, 31, 41, 42, 52, 56, 58, 75, 77, 81, 99, 100, 102, 157
contractarian.…18, 35, 38, 46, 52, 53, 60, 61, 62, 63, 66, 67, 72, 73, 74, 76, 77, 80, 103, 104, 105, 106, 107, 108, 112, 113, 114, 115, 116, 118, 121, 122, 123, 127, 129, 156, 157, 163, 206
convention. 1, 2, 17, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 62, 66, 67, 68, 70, 74, 75, 77, 82, 86, 87, 88, 89, 90, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 115, 116, 117, 118, 119, 120, 121, 122, 124, 125, 126, 127, 128, 129, 130, 131, 134, 135, 136, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 164, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 187, 188, 189, 191, 192, 193, 194, 195, 196, 197, 198, 210
cooperation. .22, 23, 25, 27, 28, 29, 30, 41, 42, 43, 45, 48, 49, 52, 54, 55, 56, 57, 65, 68, 77, 81, 84, 87, 89, 90, 93, 94, 95, 96, 97, 102, 105, 112, 119, 120, 121, 129, 134, 136, 138, 139, 140, 141, 142, 143, 145, 155, 169, 170, 179, 191, 205, 210, 211
cooperative equilibria.....50, 52, 57, 61, 73, 79, 95, 125, 131, 139, 142, 153
coordination problems.....65, 66, 70
cultural.....
    evolution.…2, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 86, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 106, 107, 108, 109, 110, 112, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 129, 130, 133, 143, 145,

149, 150, 156, 158, 159, 160, 161, 162, 168, 174, 175, 177, 178, 185, 187, 188, 191, 192, 193, 194, 195, 196, 197, 198, 205, 206, 210
D
Dawkins.....65, 120, 162, 192
disposition2, 30, 31, 41, 42, 58, 76, 90, 97, 100, 101, 102, 114, 143, 147, 149, 151, 155, 157, 170, 190
division.....23, 28, 34, 55, 56, 124, 158, 172
dynamic. 45, 49, 50, 52, 53, 56, 60, 74, 75, 76, 77, 78, 85, 86, 87, 88, 91, 92, 93, 94, 96, 97, 98, 99, 101, 103, 104, 106, 107, 108, 109, 111, 112, 115, 116, 120, 125, 127, 132, 138, 149, 153, 156, 158, 159, 174, 175, 195, 205, 209, 211
E
EGT.....78, 79, 94, 99, 102
environment.....19, 28, 29, 47, 55, 76, 82, 83, 84, 85, 86, 87, 88, 89, 93, 95, 96, 99, 107, 108, 111, 114, 118, 121, 125, 128, 133, 134, 136, 138, 141, 144, 148, 149, 151, 155, 156, 160, 161, 163, 172, 173, 174, 177, 179, 183, 185, 186, 187, 189, 195, 196, 197
equilibria selection.....55, 103, 121
equilibrium...19, 21, 22, 25, 26, 28, 31, 34, 44, 49, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 63, 64, 65, 66, 69, 73, 78, 84, 88, 89, 93, 94, 95, 96, 99, 100, 101, 102, 108, 109, 111, 112, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 127, 128, 131, 132, 137, 138, 139, 140, 141, 143, 145, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 175, 178, 183, 192, 193, 194, 197, 206
Equilibrium.....22, 156, 160, 205, 206, 210
equilibrium 54, 58, 93, 95, 101, 102, 132, 140, 153
established.....
Aristotle.....53
nature.....38, 53, 63, 66, 107, 108, 112, 115, 134, 147, 153, 163
evolution.....2, 45, 47, 48, 49, 50, 52, 56, 63, 66, 67, 68, 71, 101, 193
evolutionary.....
game theory...40, 43, 45, 46, 47, 49, 50, 51, 61, 62, 63, 64, 67, 68, 72, 73, 74, 77, 78, 79, 80, 81, 86, 87, 91, 92, 93, 94, 95, 96, 98, 99, 101, 102, 104, 107, 112, 139, 156, 158, 178, 179, 205, 210
Evolutionary game theory.....45, 46, 72, 78, 91, 92, 93, 101, 210
externalities.....20, 21, 22, 43
F
factor endowments.....20, 22, 23, 26
fairness.....32, 33, 49, 62, 63, 66, 69, 70, 71, 73, 101, 123, 124, 153, 158, 161, 191, 198, 209
free.....
factor endowments.....23
market.....18, 19, 20, 21, 22, 23, 25, 28, 33, 34, 43, 85, 191
Functionalism.....86, 209
future discount factor.....70, 76, 84, 132, 134, 139, 155
G
Gauthier…2, 17, 18, 19, 21, 22, 24, 26, 27, 29, 30, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 49, 50, 51, 52, 55, 56, 58, 59, 61, 69, 70, 72, 73, 74, 75, 76, 77, 79, 80, 81, 90, 97, 99, 100, 101, 102, 104, 105, 106, 112, 113, 124, 125, 126, 127, 129, 130, 135, 146, 147, 151, 153, 155, 156, 157, 163, 166, 170, 171, 191, 209, 211
Gauthier project.....38
Gintis.....46, 70, 74, 206
government.....18, 19, 67, 105, 133, 134, 141, 149, 173, 174, 196

H
Harsanyi.....63, 71, 205
Hayek.....50
Hobbes. 17, 18, 19, 29, 30, 38, 41, 46, 50, 53, 61, 64, 105, 112, 126, 127, 129, 130, 133, 134, 136, 146, 147, 148, 149, 152, 163, 197, 207, 209
holism.....46, 47, 86, 104
Hollis.....46, 80, 86, 87
Holly Smith.....41
homo economicus.....46, 64, 80, 82
Hume.....46, 50, 61, 94, 95, 105, 126, 135, 140, 152, 156, 197, 207
I
individual.....
rationality 18, 25, 35, 46, 48, 53, 61, 69, 74, 76, 80, 85, 86, 90, 91, 98, 101, 102, 105, 120, 123, 125, 127, 128, 132, 134, 140, 144, 149, 151, 152, 153, 160, 180, 184, 195, 196
individualism.....46, 47, 86, 104, 121, 208
K
Kavka.....41, 206
L
Leviathan.....18, 29, 38, 46, 63, 112, 146, 208
local interaction.....55, 56, 99, 193, 194
location.....52, 55, 56, 58, 61, 93, 94, 99, 137, 153, 178
Lockean Proviso.....18, 26, 27, 28, 29, 36, 39, 43, 123
M
market.....19, 50
maximisation. 2, 17, 18, 19, 20, 24, 25, 28, 29, 30, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41, 42, 44, 45, 47, 48, 52, 53, 55, 56, 58, 61, 70, 74, 75, 76, 77, 78, 80, 81, 82, 89, 91, 96, 97, 98, 99, 100, 101, 102, 104, 105, 106, 111, 113, 120, 124, 126, 127, 128, 129, 131, 133, 134, 135, 137, 140, 144, 146, 147, 148, 149, 152, 157, 158, 161, 163, 167, 170, 180, 181, 182, 183, 184, 185, 187, 189, 190, 191
maximising.....
behaviour...2, 17, 18, 19, 21, 22, 23, 28, 29, 30, 35, 36, 37, 38, 39, 45, 46, 49, 50, 51, 54, 55, 56, 57, 59, 60, 61, 62, 65, 68, 70, 71, 72, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 95, 98, 99, 111, 112, 114, 116, 117, 120, 121, 124, 126, 128, 130, 131, 142, 143, 144, 145, 146, 147, 151, 152, 153, 154, 155, 156, 157, 161, 162, 169, 172, 173, 175, 176, 179, 180, 181, 182, 183, 186, 187, 188, 189, 192, 193, 194, 195, 196, 197, 198
established convention 51, 88, 89, 90, 95, 97, 108, 111, 128, 142, 157, 159, 160, 168, 170, 173, 174, 178, 182, 183, 184, 185, 194
justice25, 26, 34, 36, 43, 45, 49, 50, 51, 52, 62, 95, 102, 111, 112, 123, 124, 125, 148, 150, 151, 152, 153, 154, 155, 156, 157, 159, 160, 161, 162, 163, 164, 166, 167, 168, 171, 172, 174, 175, 198, 205, 206, 207, 209
repeated games.....49, 51, 65, 70, 77, 92, 98, 120, 121, 136, 138, 191
maximising strategy.....50, 97, 100, 141, 146, 183
MbA.....17, 18, 19, 23, 24, 26, 27, 29, 31, 34, 35, 36, 38, 52
minimax relative concession ……18, 19, 24, 25, 28, 29, 31, 32, 35, 36, 37, 38, 39, 40, 42
moral behaviour…1, 2, 17, 19, 31, 35, 36, 37, 38, 39, 43, 44, 45, 50, 51, 65, 75, 77, 104, 105, 112, 117, 146, 151, 162, 181, 197
moral norms.....43, 44, 49, 66, 141, 148
morally free zone.....18, 19, 21, 22, 34, 43

Morals.....1
Morals by Agreement 2, 17, 18, 23, 24, 32, 34, 35, 39, 40, 42, 51, 52, 58, 70, 73, 75, 76, 80, 82, 100, 102, 105, 106, 111, 112, 124, 126, 129, 147, 150, 156, 157, 170, 207, 211
N
neighbours.....52, 55, 56, 84, 90, 93, 94, 98, 119, 120, 140, 149, 161, 162, 184
O
optimality.....17, 19, 21, 22, 25, 44, 113, 124, 155, 160, 175, 192, 195, 197
original factor endowment.....21
P
Pareto 19, 64, 66, 73, 89, 93, 102, 105, 111, 113, 117, 119, 121, 122, 123, 124, 131, 132, 137, 150, 153, 154, 155, 159, 160, 161, 175, 182, 192, 196
perfectly competitive.....
market.....19, 21
population.....2, 53, 78, 91, 92, 93, 94, 96, 99, 102, 178
prisoner's dilemma.....30, 54, 68, 77, 84, 89, 120, 137
prisonerÃs dilemma 48, 49, 53, 97, 136
R
rational. 1, 2, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 55, 56, 58, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 109, 110, 111, 112, 113, 114, 115, 116, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 206, 207, 208, 209, 210, 211
behaviour1, 2, 17, 22, 25, 26, 29, 30, 31, 35, 36, 37, 38, 39, 42, 43, 44, 45, 46, 47, 48, 49, 51, 52, 56, 59, 60, 61, 62, 65, 68, 69, 71, 75, 77, 78, 79, 80, 81, 82, 86, 87, 88, 89, 90, 91, 92, 94, 96, 97, 98, 99, 100, 102, 104, 105, 107, 108, 109, 110, 116, 117, 119, 120, 121, 125, 126, 127, 129, 130, 131, 132, 133, 134, 135, 136, 137, 139, 140, 141, 142, 146, 147, 148, 149, 150, 151, 152, 156, 157, 158, 159, 161, 162, 169, 170, 174, 175, 176, 177, 178, 179, 181, 183, 184, 185, 186, 187, 189, 191, 193, 194, 195, 196, 197
Rational agents.....2
Rational agents 23, 25, 28, 55, 88, 98, 114, 123, 136, 141, 142, 144, 155, 166, 167
rationality.....38
Rawls.....18, 32, 33, 62, 63, 67, 71, 105, 140, 164, 171, 205
Repeated interactions.....2, 76, 83, 97, 98, 101, 179
Robinson Crusoe.....22, 59, 87, 175
Rousseau.....46, 53
S
Sayre-McCord.....42
self-interest 18, 34, 53
shadow of the future.....54
Skyrms…43, 45, 47, 48, 49, 50, 52, 53, 55, 56, 58, 59, 60, 61, 62, 66, 67, 68, 69, 72, 73, 74, 76, 90, 92, 94, 95, 96, 97, 119, 125, 133, 138, 143, 153, 158, 205, 210
social contract. .2, 43, 49, 52, 53, 54, 61, 63, 64, 65, 66, 67, 69, 71, 72, 73, 77, 103, 104, 107, 108, 109, 110, 111, 112, 113, 115, 116, 117, 119, 120, 121, 122, 123, 124, 125, 128, 147, 151, 153, 157, 160, 163, 165, 166, 168, 170, 172, 173, 174, 177, 185, 188, 192,

193, 194, 195, 197, 198, 211
social conventions2, 17, 44, 46, 47, 48, 49, 50, 62, 66, 70, 77, 88, 90, 92, 103, 104, 105, 106, 107, 109, 110, 111, 116, 118, 119, 121, 125, 126, 127, 128, 148, 149, 151, 153, 158, 159, 160, 161, 164, 168, 171, 172, 174, 175, 176, 177, 180, 182, 184, 185, 187, 188, 189, 191, 192, 193, 194, 195, 196, 198
social structure2, 32, 33, 34, 44, 45, 47, 49, 52, 53, 55, 56, 60, 61, 62, 68, 69, 74, 76, 77, 78, 86, 90, 94, 96, 97, 99, 108, 115, 117, 119, 125, 127, 130, 141, 151, 153, 159, 160, 162, 163, 165, 174, 175, 177, 178, 179, 181, 193, 195, 197, 198, 210, 211
society....21, 22, 30, 32, 33, 34, 42, 43, 44, 46, 47, 60, 61, 63, 64, 68, 86, 87, 88, 89, 95, 99, 101, 102, 104, 105, 109, 111, 117, 119, 122, 126, 127, 128, 131, 132, 134, 148, 151, 153, 154, 160, 161, 162, 163, 164, 165, 171, 173, 175, 178, 183, 184, 185, 186, 187, 189, 193, 194, 195, 196, 198, 205
Spontaneous order.....50
stability.2, 25, 28, 53, 54, 56, 60, 64, 66, 86, 99, 109, 111, 120, 121, 142, 145, 174, 187, 195
Stable equilibria.....50
stag hunt.49, 52, 53, 54, 55, 56, 58, 59, 60, 61, 68, 72, 97, 100, 101, 110, 120, 125, 128, 138, 210
strategy...17, 24, 29, 30, 31, 48, 49, 50, 55, 56, 59, 60, 61, 76, 78, 82, 83, 87, 88, 90, 92, 93, 94, 96, 97, 98, 100, 101, 110, 113, 119, 120, 127, 128, 129, 130, 133, 134, 137, 138, 139, 140, 141, 143, 144, 146, 149, 183, 192, 193, 197, 198, 211
sub-game.....65, 73, 81, 108, 109, 117, 118, 193
Sugden.…37, 40, 43, 44, 45, 47, 48, 49, 50, 51, 52, 62, 66, 67, 68, 69, 70, 72, 73, 74, 77, 79, 83, 89, 95, 97, 104, 124, 125, 140, 153, 205, 206, 207
Sugden, 1986.....51
Sugden, 2001.....79
super-game.....65, 108, 109, 117, 118, 193
T
toxin puzzle.....41
translucency.....30, 31, 41, 42, 56, 58, 75, 76, 99, 100, 147, 149
Y
Young.....74, 108, 166, 195
Z
Zeuthen.....23