MICHELE MARIE UZAN-MILOFSKY

LONDON SCHOOL OF ECONOMICS
AND POLITICAL SCIENCE

PhD - 2006

DAVID GAUTHIER'S

CONTRACTARIAN MORAL THEORY

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

A mes parents

ABSTRACT OF THESIS

Can we derive and internalise moral rules from pure rational agency? If such a project is possible then a contractarian moral theory is conceivable. Restricting the field of morality to social co-operation, David Gauthier is probably the pre-eminent 20th century champion of such a theory from his *Morals by Agreement,* published in 1986, to his most recent writings. The purpose of this thesis is present, analyse and argue against his project. The main argument developed is that a pre-conception of morality cripples his rational calculus and that moral rules cannot be internalised through rational calculus alone.

In the first part of the thesis, I give full voice to Gauthier. His work from 1963 to 1986 is presented and interpreted. In particular, in the first chapter I describe how, from his Hobbesian background and his research on practical rationality, Gauthier was led to develop a contractarian moral theory. Chapter two is then fully dedicated to developing the core features of *Morals by Agreement*, providing an interpretation that highlights its main strengths and weaknesses.

Gauthier's work has been extensively commented upon since 1986 and Gauthier actively participated in the debates generated by his writings. Although he stopped defending his *Morals by Agreement* from 1993, he remained faithful to the idea of a contractarian moral theory.

In the second part of the thesis, I first review in chapter three the criticisms made of *Morals by Agreement* as well as Gauthier's responses to them. By 1993, he had abandoned several key points of his theory. However, inspired by his Hobbesian background (chapter six), and within the framework of McClennen's concept of resoluteness (chapter five), Gauthier renewed his attempt to derive co-operation from pure rational agency. Chapter seven is a discussion of this latest attempt.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

When I started this PhD, I was told that it was a very solitary and challenging enterprise; I found this was not the case thanks to the expert supervision of Professor Richard Bradley. Richard let me lead my project redirecting me only when and where necessary. He was always available to address my doubts or answer my questions; he was quick to check my work and he gave me many valuable feedbacks. I enjoyed every discussion we had.

My parents have both been models of courage and abnegation. I owe them a good childhood, a solid education and countless hours of child minding! My strength comes from their unconditional love. I thank my father for giving me the love for books and the thirst for knowledge. I have often pushed myself to the limit to deserve his admiration. Thanks to my mother I experience and enjoy an exceptional friendship.

Mes parents ont tous deux été des modèles de courage et d'abnégation. Je leur dois une bonne enfance, une solide instruction et un nombre incalculable d'heures de garde d'enfants! Ma force me vient de leur inconditionnel amour. Je dois à mon père l'amour du livre et de la connaissance. J'ai souvent poussé mes limites pour mériter son admiration. Grace à ma mère je connais les joies d'une amitié exceptionnelle.

Last but not least I want to thank my husband Howard without whom this project would have never started. He encouraged me to apply and supported me all along. He was often the guinea pig of my ideas and his common sense has proved valuable in many animated philosophical discussions. Howard and our three beautiful treasures, Hanna, Gaby and Ben are all my wealth.

INTRODUCTION

Contractarianism is about an agreement between individuals who want to enter into social relationships with each others. It has been a branch of political philosophy since Glaucon challenged Socrates in Plato's *Republic*.[1] There are two issues at stake in an agreement: the first one is the agreement itself, its context and conditions, the second one is compliance with the agreement. If it is rational to agree on rules, practices or principles, it is less rational to abide by what was agreed. Indeed, it is a lot more advantageous to break the rules when (most of) the others comply with it.

In a famous paragraph, Hobbes described an imaginary Fool's reasoning.

'The Fool has said in his heart, there is no such thing as justice ... keep, or not keep covenants, was not against reason, when it conduced to one's benefit. He does not therein deny, that there be covenants; and that they are sometimes broken sometimes kept; and that such breach of them may be called injustice .. ; but he questions, whether injustice .. may not sometimes stand with that reason, which dictates to every man his own good; and particularly then, when it conduces to such a benefit, as shall put a man in a condition, to neglect not only the dispraise, and revilings, but also the power of other men'.[2]

To the Fool, Hobbes gave a political answer arguing for the need for a powerful sovereign. Hobbes did not trust men to keep their covenant and was convinced that there would be no viable society without a reliable enforcement system. Gauthier believes that men understand the language of reason and that, in their interest, they are able to internalise the constraints of the agreement. He believes that men can become moral by mere rational calculus, as long as they choose the appropriate mode of rational deliberation. In short he believes that it is possible to give a moral rather than a political answer to the Fool; he believes in a contractarian moral theory and proves it. The purpose of this thesis is to assess the success of his enterprise.

---

[1] See Plato's *Republic*, Book II
[2] *Leviathan*, pp 538-539

In this introduction we will define with more accuracy contractarianism and review its variety, main features and appeals. We will then see how Gauthier's contractarian theory fits in this broad picture prior to introducing the main arguments of this thesis.

About contractarianism

Origins and main features of contractarianism

It is usual to distinguish between two schools of contractarianism: the interest-based versus the right-based one. The first school is lead by Hobbes. Interest based contracts are like rational bargains between agents seeking mutual advantage; they start from a state of nature and they know their personal capacities and circumstances. The second school is usually identified with Locke, Kant or Rousseau. Agents in right based contracts have a shared interest in a common good and they seek reciprocity where everyone benefits from a fair and equal baseline.[3]

It is also not unusual to distinguish between political and moral contractarianism although both are often part of the same contractarian theory. This second distinction is less clear than the first one. I suggest distinguishing them as follows. Political contractarianism involves an agreement on the political rules and structures of society whereas moral contractarianism involves a further agreement on what is morally right. As we shall see below, Gauthier has given a whole new sense to the idea of contractarian moral theory.

P. Vallentyne describes contractarian moral theories as follows: they 'hold that an action, practice, law, or social structure is morally permissible just in case it, or the principles to which it conforms, would be (or has been) agreed to by the members of society under certain specified conditions.'[4]

This description lists the main features and shows the diversity of contractarianism in general and moral contractarianism in particular.

---

[3] For this short introduction, I borrowed the definitions and distinction from Samuel Freeman's 'Contractarianism'.

[4] 'Gauthier's Three Projects', p 3

• Contractarianism applies to actions, practices, laws or social structures. I assume that the distribution of wealth would also fit in this definition. For some theorists agreement is about the society the agents want to live in whereas for others, agreement is just about the distribution of the cooperative surplus.

• Contractarianism can be direct, i.e. agents agree directly on rules and social structure, or indirect, i.e. they agree first on moral principles and then on rules or social structure compatible with these principles.

• Contractarianism can be actual or hypothetical. It either describes what has happened or what agents *would* derive if they were to reflect on their social environment. Most modern contractarians describe a hypothetical contract. [5]

• The outcome of the agreement also depends on the original conditions to be specified by the theorist. For example, the agents in the pre-social conditions can either be bargainers in full knowledge of their capacities and circumstances, or deliberating agents who don't know yet their identity in the future society.

Appeals of contractarianism

Contractarianism has several appealing features. R. Sugden lists at least three main reasons to want contractarianism to succeed.[6]

Firstly, the contractarian approach does not presuppose any particular conception of the good life or of a good society.

'By thinking of society as a scheme of cooperation, we avoid having to ask what really is good in an absolute sense. Even if we cannot agree on common ends, we can cooperate for mutual benefit. Society, on this view, is merely a mechanism that we use in common to achieve our separate ends.'[7]

[5] The only noticeable contractarian who advocates actual agreement is G. Harman in his 'Justice and Moral Bargaining'.
[6] See 'The Contractarian Enterprise', pp 3-6
[7] 'The Contractarian Enterprise', p 4

Secondly, contractarian theories can, by design, be endorsed by every member of society. Contractarianism is based on public justifications: it does not only identify the characteristics of a just society, it also justifies that society to its members in terms that each can accept.

Thirdly, the contractarian approach 'limits the demands that society can make on the individual, or that individuals can make on each others'.[8] The contractarian idea is based on reciprocity. Social constraints are agreed by the agents themselves which means not only that they limit these constraints to what is necessary and sufficient but also that these constraints must be agreeable to all.

S. Freeman adds an extra appeal to this list: there is an intuitive force to the agreement. If we agree to something that is agreeable not only to us but also to all the other parties than there is an underlying promissory obligation to comply with what has been agreed. There is mutual grounds and reliance for holding all to the commitments.[9]

Gauthier's contractarianism

Where does Gauthier's contractarianism fit in this big picture? I believe we must first insist on the fact that his contractarianism has evolved over the years. But before we do emphasise the evolution, it is useful to concentrate first on the constant features of his theory.

Constant features of his contractarianism.

First of all Gauthier is both a Hobbes scholar and a Hobbesian. Gauthier wrote in 1969 a book which title was *The Logic of Leviathan.* In this book, he provides an original interpretation of Hobbes' masterpiece. Ever since, he has regularly published articles on Hobbes. His interpretation and his focus have considerably evolved over the years but he has remained faithful to the 17th century philosopher. *His work on Hobbes is an*

[8] 'The Contractarian Enterprise', p 5
[9] See 'Contractarianism', section 1: the role of agreement.

integral part of the development of his own theory. In this thesis, I emphasise the interactions and links between his evolving interpretation of Hobbes and his own theory. As I shall explain in my first chapter, I believe that Gauthier wants to give to Hobbes' Fool a moral rather than a political answer. This motivation is at the heart of his contractarianism.

I insist on the fact that I am not a Hobbes scholar myself. All the sections on Hobbes in this thesis are a presentation of Gauthier's research on the classical philosopher. I am not concerned with the accuracy or conformity of his work with the standard literature on Hobbes. I am only concerned with the relevance of his interpretation to his own contractarian moral theory.

Secondly, Gauthier is a moral rather than a political contractarian. Although his theory covers both fields, his primary ambition is to derive morality from rational agency. Gauthier believes that we live in a disenchanted world in which morality has lost its roots hence the distance he takes with traditional morality.[10] He wants to create morality as an artificial virtue rationally derived and acquired. Gauthier usually distinguishes two parts to his contractarian theory: a theory of justice in which agents rationally derive common moral standards and a moral theory that deals with the internalisation by the agents of these moral standards.[11] The moral theory is at the core of his research. Bargainers choose a moral principle on the basis of a calculation of self-interest. Then, as members of society, they abide by this principle in all future cases 'including those in which it is not in their self-interest so to abide. In agreeing to and abiding by the principle individuals are morally transformed.'[12] The purpose of his contractarianism is not only for agents to agree but it is also for agents to internalise and abide by the rules agreed. Morality is not a principle or a standard; it is primarily a human virtue.

The third constant feature of his theory is that the contract he refers to is *hypothetical.* Gauthier never describes what did happen but what would happen if agents were to reflect rationally on their social interactions. His contractarianism is justificatory and

[10] This is a theme he often refers to. See for example 'Why Contractarianism ?' for his development on this topic.
[11] About the distinction between the two theories within his contractarianism, see for example 'Public Reason', p 36
[12] Between Hobbes and Rawls', p 25. This quote was borrowed from R.P. Wolff who was writing about Rawls, but Gauthier made its own as a good characterisation of his project in moral theory.

normative rather than descriptive. Agents' starting point is a state of nature usually described as a vacuum of social norms and practices. In the state of nature, strategic interactions prevail and they yield a sub-optimal outcome. Agents come to realise that they could benefit more and interact better if they were to co-operate. Therefore, Gauthier *assumes* that they are *all* willing to exit the state of nature and that the type of society they want is a 'co-operative venture for mutual advantage'[13].

Last but not least, Gauthier is convinced that it is an intrinsic feature of human rationality to change and adapt to what is best or better for the agent. 'At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection.'[14] Through reason, agents are able to change their mode of rational deliberation. This characteristic of Gauthier's contractarianism is quite unique to him[15] and most certainly central to his moral theory. The rationality as assumed in economic and game theories is the rationality that prevails in the state of nature, in the realm of strategic interactions whereas a constrained mode of deliberation structure social interactions. *I shall argue that this change of rationality occurs while the agents are still in the state of nature and that it is morally loaded since it carries in itself all the demands of impartiality.*

Evolution in his contractarianism.

We can say that the first version of his contractarianism was prepared between 1963 and 1986, date at which he published his famous *Morals by Agreement.* This book is probably the first complete version of a contractarian moral theory. However, many of its parts are flawed and the theory has been heavily criticised. Gauthier initially tried to defend every single conception but stopped from 1993. Since then he has remained an ardent believer in moral contractarianism; he has continued to refer to *Morals by Agreement* as the backbone of his theory but he has taken his distance with some of the technicalities of the 1986 version.

---

[13] Gauthier borrowed this definition from J. Rawls's *A Theory of Justice*, p 4. Although his theory has evolved Gauthier has remained faithful to this definition and vision of society.
[14] *Morals by Agreement*, p 183
[15] Although, as we shall see Ned McClennen develops a similar argument.

I believe that Gauthier made three major changes in the 1990's. I also believe that these three changes combined together force Gauthier to take a serious Rawlsian turn in his approach.

Initially Gauthier was convinced that the *Prisoner's dilemma* was the modern way to express the Fool's challenge but he then realised that it was the wrong starting point. A contractarian moral theory or any theory of rational co-operation must solve an *assurance game* instead. The distinction is obviously crucial since an assurance game involves dynamic choice, intentions, plans and trust rather than mere strategic and static rational calculus.

Gauthier also seems to leave aside the theory of rational choice and the concept of rationality as utility maximisation. Gauthier distinguishes between the market strategic interactions and the constrained realm of politics. Gauthier abandons the language of utility, preference or maximisation and replaces it with deliberative rationality and resolute mode of deliberation. The bargainers are replaced by deliberators. Gauthier is more worried about the social structure than by the distribution of the cooperative surplus.

The last change is more a change of emphasis in his research on Hobbes. Gauthier deepens his understanding of Hobbes' public reason and he develops his research on the philosopher's theory of law. Both concepts are at the core of his new theory. Again, Gauthier is convinced that it is possible to give a moral rather than a political answer to the Fool: agents are able to internalise the constraints of public reason.

Main arguments of this thesis

Gauthier's agents are in a hypothetical state of nature where there are no rules and no co-operation. Agents come to realise that they need to exit this state of nature if they want to benefit from the advantages of cooperation and from a social structure. Gauthier assumes that the only way out the state of nature is via a change of mode of rational deliberation: agents must pursue mutual advantage rather than individual benefit. Only

then can they be accepted as partners in the social project and can they benefit from the advantages of social cooperation. They are 'fully rational'.

I believe that the concept of full rationality is artificial and does not come naturally to most agents. In cases of conflict between self and mutual interest, the agents will naturally follow the dictates of self interest. The 'overall' rational calculus that Gauthier suggests is not a sufficient motivation to change human nature.

To assume that the bargainers pursue mutual advantage amounts to allocating equal bargaining power to each of them. Gauthier claims that bargainers are all willing to exit the state of nature and to agree. *Therefore*, he claims, they are prepared to make concessions in order to be part of the bargaining process.

I claim that in doing so, Gauthier not only empties the bargaining process from its strategic interest but he also introduces a moral bias. Indeed, if as Gauthier defines it, morality is taken as impartial constraints on the pursuit of individual interest, than his conception of 'full rationality' carries all the demands of impartiality and constrains the agents' choices and behaviour. His conception of morality is engraved in his definition of full rationality.

Not only do I believe that Gauthier's moral standards are an input rather than an outcome of his theory, I also believe that he fails to demonstrate how agents actually internalise these moral standards by mere rational calculus; I believe that this rational calculus is flawed.

In order to assess Gauthier's success in establishing a full rounded contractarian moral theory, I will articulate my thesis around two parts.

In the first part, I will trace back the foundations of Gauthier's theory. In particular, the first chapter will show the emergence of a dual rationality from his interpretation of Hobbes' *Leviathan*. Between 1963 and 1986, Gauthier discovers game theory and the theory of rational choice, he is introduced to the prisoners' dilemma and step by step he

puts in place all the elements of his *Morals by Agreement.* This first chapter will set the framework of his work.

In the second chapter, I will present my own interpretation of *Morals by Agreement,* I will define concepts like morality, justice or rationality, I will expose the issues at stake and I will develop the core conceptions one by one. Each conception is of essential importance in establishing this first fully achieved contractarian moral theory. We will show how and why.

In the second part, I will review the developments that took place after 1986. *Morals by Agreement* has been heavily criticised and the third chapter will be fully dedicated to presenting the main criticisms. We will see how Gauthier defended his theory and the concessions he had to make. He had to abandon some of the key assumptions and to admit the weakness of some of his core conceptions. In chapter IV, I will develop my own criticism and present my reason for thinking that this first attempt of a contractarian moral theory failed.

Gauthier is not easily dismissed. He remains convinced that a moral contractarianism is possible. If he acknowledges that some features of *Morals by Agreement* are defeated, he believes that the core of his theory is viable. In his most recent publications, he reemphasises his belief in a dual rationality and provides elements to improve on the original version of his theory. Based on these articles and on my understanding of his approach, I will highlight his main changes in chapters V and VI and attempt to reconstruct his new version of a contractarian moral theory in chapter VII. I will also develop my main criticism against it.

In conclusion, I will review Gauthier's achievements and assess his contribution to the development of moral contractarianism. I believe that Gauthier has extensively contributed to the familiarisation of modern philosophers with the fabulous potential of this form of contractarianism. However, I believe that he has failed to bring his project to a viable state. Morality as an artificial virtue is not and will never be a stand-alone rational product. Moral standards or principles can possibly be rationally derived but they will never be internalised by mere rational calculus whichever rational calculus we apply.

PART 1

THE FOUNDATIONS OF GAUTHIER'S CONTRACTARIAN MORAL THEORY

Introduction

In this first part, I trace back Gauthier's intellectual process and the concepts that brought about *Morals by Agreement.* In chapter I, I highlight the influence of his research on Hobbes on the development of his own contractarian moral theory. I explain how his interpretation of the British philosopher guided his step into discovering the concept of a dual rationality. I also highlight the strong connection between Gauthier's research and the developments in moral and political philosophy of the second half of last century.

In chapter II, I detail and interpret *Morals by Agreement.* I focus on the logic and the articulation of his core conceptions, highlighting the links between them and the role of his basic assumptions. In the light of this interpretation, I will develop my own criticism of his theory in chapter IV of Part II.

CHAPTER I: THE BIRTH OF A CONTRACTARIAN MORAL THEORY

In this introductory chapter I would like to tell the story of the making of *Morals by Agreement.* The roots of Gauthier's contractarian moral theory are in a three-period story: the birth of a project, the finding of a framework and the use of new tools.

The first period was in the late fifties when Gauthier wrote his doctoral thesis on *Practical Reasoning.* At the time, Gauthier's main references were K. Baier's *The Moral Point of View* and R.M. Hare's *The Language of Morals.* The young Gauthier believed in the potential of instrumental reason. He distinguished between prudential and moral practical judgments; he then attempted to prove that all practical reasoning takes the same form whether moral or prudential and depends primarily on the wants of the agent doing the reasoning. In chapter VII of his first book he already suggested basing "morality on a set of principles derivable, as general practical judgments, from prudential grounds"[16]. This project would remain the core of his research.[17]

The second period started when Gauthier had completed his doctoral thesis and stayed in Oxford to further his research on practical reasoning. His interest in Hobbes probably took root then. Indeed, for moral practical reasoning to take the same form as prudential reasoning, the agent has to include the wants of others in his wants. This condition poses a problem in the case of obligations and Hobbes was one of the rare philosopher who had given this problem some thought and the beginning of a solution. During the sixties, Gauthier developed his understanding of Hobbes until he published *The Logic of Leviathan* in 1969. As we shall see, he amended this initial interpretation later.

Gauthier's original interest in Hobbes was entirely motivated by his own research on practical reasoning. However, he would find in Hobbes the framework of his own theory. If the project to derive morality from prudential reasoning sprang from his student years, the concept of a proper contractarian moral theory was born with his

[16] *Practical Reasoning* p 81.
[17] One will find further details on this initial project in the annex of this chapter.

research on Hobbes. In 1969, Gauthier identified failures in Hobbes's philosophy. He wanted to succeed where Hobbes failed. A new theory had to emerge within the Hobbesian framework. His work on Hobbes is an integral part of his own theory and we cannot understand one without understanding the other.

Then came what I would call the third period: the seventies. In 1971, Rawls published his theory of justice, in 1975 Buchanan wrote *The limits of Liberty* and in 1976 Harsanyi developed his general theory of rational behaviour. Game theory made its entry in the secluded world of moral and political philosophy.[18] This explosion of new theories and tools was obviously going to influence Gauthier who became very much aware of their tremendous potential. The 70's were the brewing years. Gauthier distinguished between a hypothetical state of nature and civil society. In the state of nature, only egoists can inhabit and they can only coordinate their actions. Rational cooperation and bargaining are steps toward civil society. Criticising Harsanyi, Gauthier created his arbitrator and on amending Rawls's lexical difference principle, he created his maximin principle. He also borrowed the economic concept of rationality as utility maximization to develop his own concept of constrained maximization. Of course game theory was the engine of his research. The famous prisoner dilemma was the first challenge Gauthier wanted to tackle. It is while trying to solve it that he came to think of the concept of dual rationality: a straightforward or natural rationality in the state of nature versus a constrained or conventional rationality once in society. Gauthier worked in every direction without any pre-conceived idea of where he wanted to go. Between 1970 and 1986, he developed most of the core conceptions that made the future *Morals by Agreement*[19]. His contractarian moral theory was born.

About Hobbes

In 1963, Gauthier defended the view that an obligation arises from a commitment and not inversely. It is because I committed myself to do X that I am obliged to do X. Gauthier claimed that according to Hobbes, we have a right of nature i.e. the right to all

[18] For those unfamiliar with all these theories and tools, please read the appendix.
[19] Gauthier had brushed a summary of his *Morals by Agreement* in 1979 in 'Bargaining Our Way Into Morality : A Do-It Yourself Primer'. In this article, one can find articulated the concept of the perfect free market as a morally free zone, the minimax and maximin principle, the concept of constrained maximization and the Lockean Proviso, every single concept contributing to demonstrate that morality so defined can be identified with rationality.

actions compatible with self-preservation. I can kill if it can protect my life. However this very right leads to war and thus to a situation that endangers my life. Self-preservation is men's good. Therefore it is in their best interest to surrender part of their right to secure this end. The covenant is a mean to this end. They must *really* renounce some of their rights to secure this end and this renunciation carries obligation. The making of a covenant is an obliging factor and this obliging factor becomes a sufficient reason for acting upon it if one is to reach the end one seeks.

A majority of contracting parties must keep the covenant otherwise it is in vain. This means that the covenant *must* oblige not that it *does* oblige. Hobbes is aware that there are situations where self-interest is in contradiction with the covenant. The main problem is there: the covenant rests on prudential grounds but the very same prudential grounds can lead men to break the covenant once agreed. That is where Gauthier departs from Hobbes since, at this stage, Hobbes needs to introduce the coercive power of the sovereign to secure the covenant. Unfortunately, "the effect of the sovereign is to by-pass the act of covenanting as an obliging factor"[20]. I don't do X not because I voluntarily committed myself not to do X but because I fear the sovereign's punishment if I do X.

From 1963 until 1969, Gauthier worked on the publication of a thorough research on Hobbes: *The Logic of Leviathan.* This second book was very much in the vein of *Practical Reasoning.* However it bore all the roots of his future project. Gauthier considered Hobbes as the "true parent of rational morality"[21] but he also believed that Hobbes chose the wrong premises on the state of nature and that he failed to bring to a satisfactory completion the concept of rational morality. His own theory is primarily built by rectifying the double failure he found in Hobbes.

In the 70's, Gauthier left aside his work on Hobbes to build up on his own theory. As said before, the core of *Morals by Agreement* was written during these years of research, taking on board the main developments of modern moral and political philosophy. In the late 70 and early 80's Gauthier came back to Hobbes in several articles and his reading was then vastly influenced by his own freshly derived views on rational

[20] *Practical Reasoning* p189
[21] 'Hobbes's Social Contract', p 72

morality. Gauthier could almost detect the concept of constrained maximization in the *Leviathan.*

In 1963, Gauthier wrote:

"Obligation cannot be effective under the conditions imposed by Hobbes... The objections urged against his theory serve to defeat any attempt to show that a prudential reason for undertaking an obligation suffices to ground performance of the action to which one obliges oneself. Prudential considerations may lead us to undertake obligations, as Hobbes clearly sees, but these obligations must provide independent, moral reasons for acting in order to oblige. In obliging oneself, one may introduce moral considerations not previously relevant to one's actions, and one may introduce them on prudential grounds. But the practical force of these considerations cannot depend on the ground for introducing them"[22].

Gauthier remained convinced that moral considerations had to back the prudential basis for any covenant to be kept. However, his opinion on Hobbes's achievement in this regard had changed by 1979:

"The *tour de force* in his [Hobbes's] theory is the reconciliation of maximizing rationality with constraining morality. How can one be rational in accepting the constraints of the laws of nature and so not exercising one's full right of nature? The answer requires Hobbes's account of right reason. For his true moral theory is a dual conventionalism, in which a conventional reason, superseding natural reason, justifies a conventional morality, constraining natural behaviour. And this dual conventionalism is Hobbes's enduring contribution to moral theory".[23]

The purpose of this chapter will be partly to understand what happened between these two quotes, i.e. to understand how Gauthier's research on rational morality influenced his interpretation of Hobbes. Our starting point will be Gauthier's *The Logic of Leviathan.*

Hobbes is a rationalist. The laws of nature are rules of reason deduced from assumptions about human nature and human conditions. God created Man and Man created the Leviathan i.e. "an artificial man of greater stature and strength than the natural, for whose protection and defence it was intended"[24]. How exactly was this Leviathan created? Hobbes articulates his argument in three steps. Firstly he describes what he calls the state of nature that informs us of his vision of human nature. Secondly, he explains how through a social compact, men can come out of the state of nature and

[22] *Practical Reasoning* p190-1
[23] 'Thomas Hobbes : Moral Theorist', p 547-8
[24] *Leviathan*, p 492

live together in peace. Thirdly he explains how to keep the social contract once agreed upon. We shall follow below these three steps in the following three sections[25].

I apologise in advance to the Hobbes scholars who might be frustrated by some of my simplifications. I have abstracted from Gauthier's book on Hobbes only what was relevant to his own theory, going quickly on some points and being more thorough on others. My purpose is not to develop Hobbes' philosophy but only to highlight its role in the development of Gauthier's own theory.

[25] These three steps are unfolded primarily in Part I chapters 13 to 15 of the *Leviathan*.

Section 1: About the state of nature and the original position

In this section, I will review Gauthier's description of Hobbes' state of nature and his criticisms of it prior to detailing his own developments on the sate of nature and the original position. Gauthier feels uncomfortable with Hobbes' psychological premises and prefers Locke's vision of human nature as appropriative. He also starts to argue for what will become important in his theory i.e. that, for the sake of the future cooperation, agents are prepared to accept constraints on their original position. Last but not least, I will highlight the influence of bargaining theory in his embryonic approach to the original position or base point.

Hobbes' state of nature in The Logic of Leviathan.

Hobbes starts by making a number of assumptions about human nature and the state of nature.

1. *Men are equal in mental and physical power.* Hobbes claims no man can be strong enough to feel secure; even if one man is stronger than the others, the others can always fight him by "machination or confederacy".[26]

2. This equality of insecurity combined with *scarcity of goods* leads to *conflicting desires.* If a man, through his abilities or labour, is in an enviable position, he can expect the fruit of his labour as well as his liberty or his life to be taken by another man. No man can secure himself or his position.

3. *Men are prudent and concerned with self preservation.* Hobbes assumes that men are pre-occupied not only with immediate but also with future self-preservation, self-preservation being an end in itself. It is the concept of prudence that will lead them to contract.

4. *The best strategy is to attack.* Men are naturally inclined to quarrel for competition, lack of trust or glory. Even if some would be happy to settle with modest means, there

[26] *Leviathan*, chapter 13, p 531

will always be some quarrelsome individuals in search of glory to attack them. In this context, defence and 'wait and see' strategies are unlikely to be successful.

From these assumptions, Hobbes concludes that the state of nature is a state of war of every man against every man. In the state of nature, there is neither property nor business nor prosperity. Nothing and nobody is safe; it is a state where the life of man is "solitary, poor, nasty, brutish and short"[27]. There is no social peace, no law and no justice.

Since all of Hobbes's moral and political philosophy is based on his assumptions concerning the nature of man, Gauthier tried to understand what lead Hobbes to make such assumptions and more importantly how from these assumptions about men, Hobbes can conclude that they do contract. Gauthier's focus is obviously primarily on assumption 3 since he believes that it is Hobbes's concept of rationality that is the root of his moral and political philosophy.

On the next page, one finds Gauthier's description of Hobbes's instrumental and theoretical reasoning, their interactions and their limits. A quick look at it should provide the necessary background to what follows.

He describes the Hobbesian human beings as self-maintaining engines in permanent motion since external bodies recurrently affect the organs of sense. We are permanently moved by our passions. Only power can secure success ... never ending power. From this description of man combined with an assumption of equality and scarcity of good, men are necessarily led to compete. Assumption 4 is almost a conclusion to the three previous assumptions. Competition is a by-product of the very nature of men and their innate concern for self-preservation. In search of self-preservation, we end up living in a state of war. When fully rational we act against our own best interest.

Hobbes is convinced that were the appropriate moral, political and social arrangements in place, men could maintain themselves free of competition and the cause of war would be removed. Even those who seek war for the sake of it could be tamed and trained.

[27] *Leviathan*, chapter 13, p 532

How can Hobbes derive normative (moral and political) conclusions from factual (psychological) premises? According to Hobbes, the basic nature of human motivation is self-preservation. Whatever is a condition for self-preservation is a means to man's end.

From premise	'X is necessary to self-preservation',
Hobbes concludes	'men must do X to secure what they want'
and then	'men, if rational, will do X'.

Gauthier considers that Hobbes is justified to claim that whoever accepts the psychological doctrine that self-preservation is man's basic end, must also accept its imperatives as rationally required. However, Gauthier's objections are more with the premises themselves. Firstly, he does not believe that self-preservation is a *necessary* and basic motive of human action. He agrees that self-preservation is very likely to be a motivation to act but it *does not have to* be the only one. Hobbes adopts a reductive account of human motivation. Other factors have to be taken into account in some situations. Secondly, he disagrees with Hobbes account of rationality: "those actions are most reasonable, that conduce to their [the agents'] ends". Again, he finds this account too restrictive since he considers that actions can be reasonable without conducing to the agents' ends.

Therefore Gauthier suggests modifying the inference as follows:

from	'X is a necessary means to self-preservation',
we conclude that	'men must do X to secure what they very probably want'
and then	'men very probably have some reason to do X'.

This provides a reason but not a *sufficient* reason for the imperative 'Do X'.

Sketch 1: Hobbes' practical and theoretical reasoning

DELIBERATION (Practical reasoning) Succession of alternate appetites	**MENTAL DISCOURSE** (Theoretical reasoning or Reason) Succession of alternate opinions
External objects act on the organs of sense and then on the brain to become motions of sense which then travel to the heart.	Absolute knowledge of facts i.e. knowledge of consequences from sense & memory recorded in definitions of words which proceeds into general affirmations and then into syllogisms
Motions of imagination in the **heart** (seat of vital motions)	Motions of imagination in the **brain**
If the vital motions weaken the movement than it is called pain; if it helps it, it is called pleasure. These motions becomes '**endeavours**' i.e. aversions or appetites. **Passions** are species of aversions or desires.	The last sum is the conclusion. These conclusions are **opinions**. They form a whole called science and science is power since it provides the means to ends.
The last appetite is the **WILL** ←	The last opinion is **JUDGMENT**

Reason aids deliberation at several levels:

1) At first level, it shows us the means to obtain or avoid what we will.
2) "As we determine further consequences of some acts, our desire to perform it or not alters. Our judgment of the several consequences of an action determines our last appetite, will" *The Logic of Leviathan* p12

COMMENTS ON THE ABOVE

- Deliberation is fallible since 'the will is always directed toward the promotion of our well being **as we conceive it** ... but passion may lead us to misconceive well being itself' *The Logic of Leviathan* p8. We can also mistake the means to the chosen end.
- 'In so far as we judge correctly and fully, we will what conduces most to our ends. And so correct reasoning enables will to be rational appetite' *The Logic of Leviathan* p12
- Unfortunately mental discourse is fallible too: the first ground of discourse might not be definitions or the definitions might not be joined together into syllogisms.
- The fallibility of mental discourse poses the problem of **right reason**. Gauthier notes that Hobbes is not clear on the determination of right reason.

Gauthier's state of nature and original position: Lockean rather than Hobbesian

Gauthier wants to change Hobbes state of nature. Like Locke, he believes that men seek appropriation more than self-preservation. Indeed, he demonstrates that the goods that men seek must fulfil three conditions: a- Cooperation increases supply; b- Possession creates distributive problems; c- There is unlimited demand or at least the demand is superior to supply. The only goods that fulfil these criteria are the goods of appropriation (the universal measure of property being money). Men have a strong desire to appropriate. The appetite for food, drink or sex is satisfiable in a way that the desire to appropriate is not. Men being by nature appropriators, they can never find a satisfactory level of satiation for appropriative activity. Therefore men find themselves in conflict with respect to their desire for property. "The competitive search for power is easily derived from the insatiable desire for appropriation."[28]Increasing competition means decreasing security of property. Thus "the state of nature leads to an outcome far worse for everyone than that would result were there a coercive force sufficient to curb each man's appetite for power and to channel his desire to appropriate into an arena which is competitive yet peaceful – the marketplace rather than the battlefield."[29] Men seek goods of appropriation being appropriators by nature and the market is the most natural and peaceful place for them to interact. However the perfect conditions for the free market to take place are not fulfilled[30]. Society has then to be rationally established and rationally maintained.

The state of nature is characterised by relations holding between any two persons where each acts on an independently selected principle of action. It is opposed to society which is characterised by relations holding between any two persons where both act on a mutually selected principle of action.[31] Gauthier also described society as "an instrument which individuals mutually accept in order to achieve, for each, benefits unattainable without such a collective instrument".[32]

---

[28] 'The Social Contract as Ideology', p 148
[29] 'The Social Contract as Ideology', p 148
[30] For full details on the market as a morally free zone see 'No Need for Morality : The Case of the Competitive Market'. This article is to be connected to the future chapter IV of *Morals By Agreement.*
[31] 'The Social Contract as Ideology', p 141
[32] 'Rational Cooperation', p 60

Coordination occurs in the state of nature and it is a natural extension of maximisation. In an article on coordination Gauthier developed on the unusual richness of coordination in the absence of bargaining or society. He wrote: coordination "might be interpreted as exhibiting one of egoism's successes. Even for straightforward maximisers, rudimentary practices of truthfulness and promise keeping can be defended, as resolving their coordination problems."[33]

The state of nature and the social contract are not real but merely hypothetical. The purpose of the philosopher is two-fold: he has to explain the transition from a state of nature to society and to find the principles of cooperative agreement that *rational* individuals *would* choose in a suitably characterized position of free choice.

The transition from the state of nature to society is done by rational cooperation and bargaining. Gauthier answers three questions concerning this transition: firstly, from which threshold does rational cooperation and therefore society become worthwhile for mere coordinators; secondly, what to do about the natural assets and original inequalities of individuals in the state of nature prior to cooperating; Thirdly, how to suitably characterise a position of free choice from which individuals can choose their future cooperative principles.

The first two aspects are obviously strongly connected since the initial assets are unlikely to be the same for all parties to the cooperation. Gauthier acknowledges and addresses this connection as follows:

"The minimal cooperative utility represents that point at which an individual prefers to act on the basis of mutual agreement rather than to act in a directly maximizing manner, or, we may say, prefers civil society to the state of nature. There is no reason to suppose that this point is the same for all men. To the extent to which one person can expect to do better for himself in the state of nature than another, he will demand a higher minimal return from cooperation. Thus the minimal cooperative utility reflects what we may call the *natural inequalities* among men. Rational cooperation takes these natural inequalities as given."[34]

Let us now visit Gauthier's various articles to find the answers to the three issues raised above.

---

[33] *Moral Dealing* p 5
[34] 'Rational Cooperation', p 59

About the threshold from which society becomes worthwhile.

Influenced by the development of bargaining theory, Gauthier sets three conditions for society to emerge[35]. He first introduces the concept of *natural outcome* i.e. the outcome of actions performed in the full state of nature. Society must improve on the natural outcome for everyone. He then defines an outcome as *optimal* if and only if any alternative outcome that would be better for some person would be worse for some other person. Society is possible if and only if the natural outcome is not optimal. Lastly, he defines an outcome as *stable*, or one in equilibrium if and only if no one can bring about an alternative outcome that is better for himself, by unilaterally changing his way of acting.[36] When there is no optimal outcome that is stable, not only is coordination no longer sufficient but rational cooperation and bargaining procedures become necessary.

About the initial endowment of the bargainers.

The next most urgent issue concerns the natural assets of the parties in cooperation. As we saw in the quote above, Gauthier takes the view that the initial assets of individuals are given. If A is strong, clever and hard worker, he would get more in the state of nature than B who is weak, slow and lazy. Cooperation has to improve on A's utility in the state of nature proportionally more than on B's utility in the state of nature.

When dealing with natural endowment, Gauthier claims:

"We must distinguish clearly between apportioning social benefits on an unequal basis, proportional to natural inequalities, and apportioning social benefits on an equal basis, after taking natural inequalities into account. The present theory requires the latter." [37]

If society is to be considered as a mutually accepted instrument whose function is to achieve benefits unattainable without it, then it has not only to assure each what they can attain for themselves in the state of nature but it also has to add a fair share of the benefits attainable only through cooperation.

---

[35] See 'The Social Contract as Ideology'
[36] 'The Social Contract as Ideology', p 142
[37] 'Rational Cooperation', p 60

Gauthier considers that Rawls, by using the veil of ignorance, relies on a pre-conception of morality:

"His use of the veil of ignorance is ultimately motivated by his insistence that the parties are to seek a conception of justice appropriate to 'free and equal moral persons'. The real differences and inequalities that characterize them are to be dismissed as morally irrelevant. Thus the principles of justice, in Rawls's view, must be related to a prior moral conception of the person."[38]

Rawls would not object to such a comment. He is the first to acknowledge his Kantian inheritance and the priority of the concept of right over the individuals' good (rational plans of life)[39]:

"In justice as fairness one does not take men's propensities and inclinations as given, whatever they are, and then seek the best way to fulfil them. Rather their desires and aspirations are restricted from the outset by the principles of justice which specify the boundaries that men's systems of ends must respect. We can express this by saying that in justice as fairness the concept of right is prior to that of the good."[40]

In order to demonstrate that Rawls's suggestion and his lexical difference principle[41] are incompatible with his contractual theory, Gauthier introduces two new definitions and one assumption. The *social surplus* is the proportion of primary goods produced through cooperation and that would not have been produced without it. The *social potential* of an individual is the maximum well being that he can expect under that optimal social arrangement that is most favourable to him. The social potential reflects individual natural endowments. Even if behind the veil of ignorance, she does not know her natural abilities, she knows that she has some natural talents that are different from those of the others so that, in the absence of an agreement, each would secure different levels of well being. "It is therefore possible for everyone to take account of the 'no agreement point' in her reasoning, even though no particular person knows how it will affect her."[42] The 'no agreement point' will become the 'base point' or 'status quo' and the social potential will become the target point.

---

[38] 'Justice as Social Choice', p176
[39] See appendix section 3, a
[40] *A Theory of Justice*, p 27-8
[41] Rawls' lexical difference principle states that "in a basic structure with n relevant representatives, first maximize the welfare of the worst off representative man; second, for equal welfare of the worst representative, maximize the welfare of the second worse off representative man, and so on until the last case which is, for equal welfare of all the preceding n-1 representatives, maximise the welfare of the welfare of the best off representative man" *A Theory of Justice*, p 83
[42] 'Justice and Natural Endowment', p 159

Gauthier claims that what is rational behind the veil of ignorance has to remain rational once the veil of ignorance is lifted for every single party to the contract, in whichever social position. That is where Rawls fails. When the veil of ignorance is lifted and each is aware of which primary goods she could have expected without agreement and which primary goods she can expect as a result of the agreement, she will want as many additional goods as possible for herself. She will not accept a smaller share of the social surplus simply to increase overall equality of benefit. Rawls lexical difference principle takes natural assets as common assets. Therefore, once the veil of ignorance is lifted, the naturally gifted person finds his talents and efforts primarily directed to the advantage of the naturally deprived. He will not find it rational to agree to principles that "require him to accept a lesser proportionate benefit, simply to increase overall equality of absolute benefit."[43]

Rational cooperation, as suggested by Gauthier, preserves *fixed social inequalities* among men but ensures *equality of opportunity* to each to realize their social potential.

"To accept any stronger form of equality would be to require some members of society to make proportionately greater sacrifices than others, and this would be rationally unacceptable to utility-maximizing individuals." [44]

About the original position

However, if Gauthier removes the veil of ignorance, he is aware that for the bargaining process to be fair, the pre-bargaining conditions must be fair. Through fairness of the bargaining background, Gauthier wants to secure the rationality of compliance once in society. He therefore imposes some pre-conditions on the original position. He requires that the principle of rational choice be defined for an ideal decision maker and for an ideal society.

"We achieve this ideal conception ... by taking each [bargainer] to be adequately informed not only about his own good but also about that of his fellows.

[43] 'Justice and Natural Endowment' p 164. Rawls had anticipated such an argument and his reply was as follows : 'The more advantaged, when they view the matter from a general perspective, recognise that the well-being of each depends on a scheme of social cooperation without which no one could have a satisfactory life; they recognise also that they can expect the willing cooperation of all only if the terms of the scheme are reasonable. So they regard themselves already compensated, as it were, by the advantages to which no one (including themselves) had a prior claim. They forego the idea of maximizing a weighted mean and regard the difference principle as a fair basis for regulating the basic structure.' *A Theory of Justice*, p 88

[44] 'Rational Cooperation', p 62

Communication among the persons must be full and free; no one is able to deceive another about anyone's interest or bluff successfully about what anyone is willing to do. The process of bargaining must be thought of as effectively cost-free... No one is in a position to benefit by his superior ability to outwit the others. Threats are useless... we require that the process of bargaining exhibit procedural equality and maximum competence among the persons who are to agree on the principles of justice... Each bargainer thus serves as an ideal representative of the particular person he will be in the social world to be shaped by the agreed principles of justice; thus fairness is assured at the procedural level."[45]

Gauthier's concern for ensuring a fair bargaining process goes further. We saw above that the rationale for an individual to bargain depends on his 'no agreement point'. It might not be beneficial for some individuals to interact at all or cooperate with others in view of their initial position. Therefore, Gauthier creates the 'base point' or 'status quo': this is the pre bargaining payoff. This base point represents what each person "could expect to gain from her own efforts in the absence of any agreed or cooperative interaction."[46] Gauthier specifies: "voluntary compliance with the terms of cooperation is rational in general only if the base point is not itself considered disadvantageous in relation to no interaction."[47]The original position could be such that discrepancies between potential bargainers would discourage some from interacting or to interact, taking as given the established discrepancies.

Gauthier specifies that for some purposes, the base point may be the final outcome, namely when bargaining fails, while for other purposes it may be the initial situation of the bargainers, that historical state of affairs prevailing prior to the bargaining and from which bargaining proceeds. In any case the status quo is established outside the bargain itself.

"If it is the result should bargaining fail, then it is determined by the structure of the situation in which the would-be bargainers must act, and by the states of affairs possible should they not bargain. If on the other hand the status quo is the initial situation, then it is determined by the historical background of the bargaining situation. In either case, it is given from the standpoint of bargaining itself."[48]

This condition on the base point reflects the same concern as the conditions on the bargaining process mentioned above: no interaction should benefit some parties to the agreement by worsening the situation of others. This conception of the state of nature

[45] 'Justice as Social Choice', p 177-178
[46] 'Justice as Social Choice', p 179
[47] 'Justice as Social Choice', p 182
[48] 'Social Choice and Distributive Justice', p 246

not only ensures the fairness of the process and moralises the base point for cooperation but it also secures compliance in social interactions. The future Lockean proviso is born.

Gauthier considerably modified Hobbes' state of nature. His account of human motivation owes more to the idea of an appropriative human nature than to self-preservation. The original position and natural assets are taken into account when bargaining. Gauthier adopts a Lockean original position rather than a Hobbesian one.

"Principles of justice may be conceived as the outcome of a rational bargain or agreement among all of the members of society, who consider how they would adopt *ex ante*, from the perspective of the state of nature, their fundamental terms of association. I argue that if terms of association are to gain voluntary and enduring recognition, and so provide a stable basis for society, they must be accepted from the perspective of a Lockean rather than a Hobbesian state of nature – That is a state of nature already constrained by a form of the Lockean proviso."[49]

The Lockean proviso and the willingness to rationalise and moralise the base point are already present in his mind. The rationality of the process depends on its fairness from the pre-bargaining position onwards.

[49] *Moral Dealing* p 6

Section 2: The making of the covenant.

In this section, we will first rapidly review the key features of Hobbes description of the making of the covenant. According to Hobbes, prudence, reason and the hope of securing property lead men to contract with each other. This aspect of Hobbes philosophy is usually very familiar to political philosophers and will become relevant later in this chapter. More importantly, we will then review Gauthier's treatment of the topic. Gauthier gives flesh to the making of the covenant by introducing a bargaining procedure. Nowhere in Gauthier's theory do we feel as much the influence of the philosophical environment in which he works than in his maximin concept. Gauthier refers to Nash, Arrow, Harsanyi, Sen, Rawls, Kalai and Smorodinsky or Nozick. He talks about social choice, welfare, distributive justice, Pareto-extension rule and bargaining procedure. One can say that his major contribution to the debate was to introduce in a comprehensive way game theory and bargaining theory into modern contractarianism.[50] However, Gauthier borrows concepts from these existing bodies of knowledge and adapt them to the needs of his own theory. One can question the legitimacy of such a strategy. Indeed, some of these concepts rely on specific assumptions without which they can lose their validity or properties. To borrow them and modify the assumptions to accommodate a contractarian framework can be a dangerous game to play.

Hobbes and the laws of nature.

According to Hobbes, men have a right of nature when in the state of nature, i.e. "the liberty each man has, to use his own power, as he will himself, for the preservation of his own nature; that is to say, of his own life; and consequently, of doing any thing, which in his own judgement, and reason, he shall conceive to be the aptest means thereunto".[51] However, through reason, men find a way out of this state of war. Hobbes edicts what he calls a law of nature i.e. "a precept or general rule, found out by reason, by which a man is forbidden to do that, which is destructive of his life, or takes away the means of preserving the same; and to omit that, by which he thinks it may be best

[50] See Braithwaite for a first attempt.
[51] *Leviathan*, chapter 14, p 533

preserved".[52] According to the first law of nature, "every man, ought to endeavour peace, as far as he has hope of obtaining it; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of war".[53]

The right of nature is described as a liberty whereas a law of nature is described as a general rule. The former spells out what is *permitted* whereas the latter spells out what is *required.* Both the right and the law of nature are summed up in the first law. One is *required* to seek peace wherever possible but, if it is not possible to obtain, one is *allowed* to defend himself and use means of war.

The second law of nature gives the nature and conditions of the social contract: "that a man be willing, when others are so too, as far-forth, as for peace, and defence of himself he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other men, as he would allow other men against himself".[54] In order to enter into and benefit from the agreement, men have to surrender part of their liberty at the *sine qua non* condition that the others are doing the same. In 1982, Gauthier interprets this point as follows: "The liberty that each person should continue to enjoy, so that peace may prevail, must be determined, not by each person's calculation of the benefit to himself which he may obtain by his own actions, but rather from each person's calculation of the *overall* benefit to himself that he may expect from the interaction of all"[55]. As we shall see, this interpretation takes into account two different concepts of rationality.

Then comes the third law of nature without which any covenant would be vain and empty words: "men perform their covenants made".[56] This is a law of justice. Justice is keeping the covenant. Injustice is the non performance of what was agreed. Whatsoever is not unjust is just. There would be no social contract and a state of war if it were not for the third law. Again in 1982, Gauthier interprets the connection between the second and the third law introducing the concept of intention. "We may say that if one does not intend to adhere to one's agreements, and so does not intend to follow the third law, then one does not sincerely enter the agreements, as required by the second law. To put

[52] *Leviathan*, chapter 14, p 533
[53] *Leviathan*, chapter 14, p 534
[54] *Leviathan*, chapter 14, p 534
[55] 'Three Against Justice' in *Moral Dealing*, p 133
[56] *Leviathan*, chapter 15, p 538

the second law into effect, then, one must also put the third law into effect; in agreeing to a limitation on one's liberty, one must intend that one adhere to the limitation"[57].

McClennen classifies the rest of the laws of nature as follows[58]: the fourth through the tenth prohibit holding attitudes that would make the social contract difficult to consummate. The eleventh through the nineteenth spell out procedures to adjudicate the disputes that are bound to arise in regard to our contracts.

The laws of nature spell out the precise terms of the social contract. They are the guarantor of a mutually advantageous state of affairs. Any move away from these laws would lead out of social peace back to the state of war, hence the necessity of ensuring that these laws are compulsory and enforced.

As will be explained below, Gauthier's two central arguments are as initially sketched in *Practical Reasoning.* Firstly, Gauthier argues that the obliging factor arises from the making of the covenant not from the covenant itself. "We do not limit our right of nature by rationally establishing the second law of nature. But once we have established it, we do limit our right of nature by acting in accordance with it. The limitation derives from our action not from the law, but our action is required by the law"[59]. The laws of nature are not laws but mere prescriptions. People should follow them but they have no force by themselves.

Secondly, Gauthier highlights a contradiction in Hobbes moral philosophy. Indeed, Hobbes's moral concepts are the right of nature, the laws of nature, obligation and justice. These moral concepts have to be interpreted by taking into account his psychological doctrine (i.e. self-preservation is man's basic end) if we want them to bear on his political philosophy. However, if we interpret them according to his psychological doctrine then Hobbes's arguments lack something distinctively moral. I have no *moral* obligation to keep my covenant if, once entered into, it becomes contrary to reason for me to keep it. As we shall see, Gauthier will later introduce a new distinction between the security and motivation problem in his interpretation of Hobbes's moral concepts.

---

[57] 'Three Against Justice' in *Moral Dealing*, p135
[58] In a set of lectures delivered at the L.S.E. in the winter 2002-3
[59] *The Logic of Leviathan* p 54

Gauthier and the making of the agreement.

Bargaining theory is an extension of game theory but the Hobbesian political philosophers rapidly seized the tremendous potential of such a theory. Indeed, bargaining theory claims that agents in full knowledge of their situation and endowment can agree on the redistribution of the cooperative surplus following a bargaining procedure. The theory sets the conditions for the cooperation and provides a unique solution[60] taking into account each party's bargaining power. Gauthier was immediately attracted by this theory but in order to appropriate it and use it in his own framework, he had first to restrict the field of morality and then modify the solution. If the first move is generally accepted, the second one is more problematic. In this sub-section, I have deliberately chosen to remove all technical discussions on Gauthier's maximin and minimax, both principles being described and explained at length in chapter II. The focus here will exclusively be on Gauthier's approach to the bargaining problem.

Gauthier's restrictive conception of morality

In 1974, Gauthier claimed that

"rational cooperation satisfies both a productive condition and a distributive condition, with respect to human well-being. It may then seem plausible to identify rational cooperation with morality, not in supposing that rational cooperation requires us to do whatever we ordinarily take to be morally right, but rather that it requires us to do what on reflection seems reasonable and justifiable in our moral practices. It enables us to subsume morality under rationality; what cannot be so subsumed we regard as irrational, and cease to consider moral. In particular, rational cooperation brings into relief those aspects of morality which we associate with justice and fairness."[61]

This quote reflects an important step in the evolution of his philosophy. Gauthier departs from the traditional boundaries of morality and the change is two-fold: firstly, traditional morality is no longer the starting point of our moral enquiry but the by-product of rational choice; secondly, the field of morality is restricted to the boundaries of distributive justice. We shall develop these two points below.

[60] See Nash's bargaining solution in the appendix section 1, b.
[61] 'Rational Cooperation', p58

The starting point is no longer traditional morality as we know it. Principles are rationally derived and then reconciled with moral practice. In 1974, Gauthier wrote about *A Theory of Justice*:

"The hypothesis, which makes the theory a genuine theory, is that an account can be given of these factors which both matches our considered judgments about them, and relates them so that the conception of justice would be embodied in the principles chosen by rational men in the initial situation."[62]

What is moral is what is derived through individual instrumental rationality and *collectively* agreed upon in order to obtain a mutually advantageous state of affair. The wants of others are included in my wants through agreement. We *all* agree on a principle of cooperation that is individually and rationally derived because it is beneficial to each of us. If it is rational for us to agree to it ex-ante (in the state of nature), it becomes rational for us to keep it ex-post (in society). Because I know that it is beneficial to every single party to the agreement, I have grounds to believe that they each will keep it and therefore I will keep it. Morality is what has been rationally agreed and will be rationally kept. We shall see in step 3 that Gauthier introduces a special form of rationality to fit the requirement of compliance ex-post. If Gauthier rejects Rawls's original position and conception of justice and fairness, he retains his method. Once the hypothetical original position has been properly characterised, whatever principle the cooperators in search of an optimal outcome will rationally choose represents the new boundaries of morality.

Rawls claimed "The Theory of Justice is a part, perhaps the most significant part of the theory of rational choice"[63]. In 1979[64], Gauthier attempts to demonstrate that such a claim is justified. The boundaries of morality are restricted to justice as fairness and not all of justice but merely *distributive* justice. The issues raised by a*cquisitive* justice are addressed by the Lockean proviso. By then, Gauthier clearly spells out that our "framework of moral concepts is outmoded"[65]. What is just is what we have rationally agreed to and what is just is moral. Distributive justice (i.e. constraints on preferences justified by an appeal to preferences) having been securely established within the realm rational choice can be removed from the area of speculative enquiry and the "age-old philosophical problems about the rationality of morality are solved for the case of

[62] 'Justice and Natural Endowment' p 151
[63] *A Theory of Justice*, p 15
[64] See 'Bargaining Our Way Into Morality : A Do-It Yourself Primer'
[65] 'Bargaining Our Way Into Morality : A Do-It Yourself Primer' p15

distributive justice"[66]... However external constraints on preferences which cannot be justified by an appeal to preference itself remains within the area of speculative enquiry and the philosophical problems posed by this field of morality remain unsolved. The field covered by morality is reduced to the boundaries of distributive justice. This leads Gauthier to a double conclusion: the optimistic conclusion is that the part of traditional morality that concerns distributive justice is now grounded on a strictly rational footing; the pessimistic conclusion is that an important part of the remainder of traditional morality cannot be placed on a similarly rational footing. "Having abandoned all religious or metaphysical props for morality, we are left with no justification for principles some of which, at least, we are unwilling to abandon."[67]

Gauthier wants to correct Rawls's framework. He has to ensure that what the parties to the 'covenant' agree to in the state of nature is rational and will remain rational for each party once in society. Gauthier rejects Harsanyi's ethical theory[68] for similar reasons. Indeed, in Harsanyi's theory, an individual not only does not know *who he will be* but also *who he is*. If an individual does not know who he is he is unable to express any preferences, he does not have a single unified standpoint from which to establish a preference ordering.[69] Besides, the principle chosen is beneficial to society as a whole but can be detrimental to some. If the actual outcome proves unfortunate to someone, on which rational ground should he adhere to this principle if non-adherence would allow him greater expected utility? Fairness and justice are dependent on the rationality of the agreement ex-ante. Therefore, the agreement cannot be chosen from behind a veil of ignorance but chosen by individuals in full knowledge of their position and initial endowment.

Gauthier's bargaining theory

When Gauthier was introduced to game theory, he rapidly realised the potential of bargaining theory: Hobbes' agreement had no specified content; bargaining theory

[66] 'Bargaining Our Way Into Morality : A Do-It Yourself Primer', p16
[67] 'Bargaining Our Way Into Morality : A Do-It Yourself Primer', p25
[68] See Appendix section 2, c.
[69] 'Given that one has no knowledge of one's own identity, one can function as an arbitrator selecting an outcome that is a fair compromise among the preferences of the individuals involved. But without preferences of one's own, one cannot act as an individual chooser... Harsanyi's argument fails, because it ensures impartiality in choice only by violating individuality.' 'On the refutation of Utilitarianism' p159

provided the means to determine a unique content to the agreement.[70] The parties to the agreement became bargainers and the making of the covenant became a bargaining situation where each agent defended her own interest in her own situation. The technical discussions on Gauthier's bargaining theory will be developed in chapters II and III. In this introductory section to his theory, we just need to know that the outcome of his bargaining is determined by what he calls the principle of maximin relative benefit or the principle of minimax relative concession. These principles are complementary. Voluntary agreement and distributive justice are reconciled when the maximin relative advantage that we can gain from an agreement is equivalent to the minimax concession necessary for it.

The first formulation of the maximin principle is published in 1974 in *Rational Cooperation.* In this article, Gauthier, who is in search for a solution to the prisoner's dilemma, presents a theory of rational cooperation as *a* theory of morality and justice in society for rational egoists (economic men). *Rational Cooperation* was written in the same year as *Justice and Natural Endowment: Toward a Critique of Rawls's ideological Framework.* Gauthier's maximin principle is clearly inspired both by Harsanyi's general theory of rational behaviour[71] and by Rawls' lexical difference principle and concept of morality as justice and fairness. From 1978, scattered in several articles[72], we can find a good sketch of his maximin. In most of these articles, he points out what he believes to be inadequate or wrong in Harsanyi, Arrow or Rawls prior to unfolding his own version. He systematically opposes his *game theoretic* approach to Harsanyi and Rawls's *decision theoretic* attempt to interpreting the social contract. Instead of Bayesian theory, he suggests bargaining theory. Against the social contract as individual decision, he suggests the social contract as collective bargain.

A first approach to his maximin is via an amendment to Rawls's principle. He demonstrates that by making two modifications to Rawls's lexical difference principle one can find an adequate maximin rule[73]. The first modification is the introduction of a *base point* (the *status quo*). Instead of maximizing the minimum utility, we maximize

---

[70] Nash had developed a unique solution to the bargaining problem given specific conditions and assumptions. For further details see the appendix section 1, b.
[71] See the appendix section 2.
[72] 'Critical Notice on Harsanyi', 'Social Choice and Distributive Justice', 'The Social Contract: Individual Decision or Collective Bargain ?'
[73] See the 'Critical Notice on Harsanyi'

the minimum increase in utility. The second modification is the introduction of a *target point* (represented by that outcome that maximizes one's utility). Gauthier suggests expressing the improvement of an individual's utility in relation to the status quo as a proportion of the improvement he would receive from his target, in relation to the status quo. Instead of maximizing the minimum increase in utility, we maximize the minimum proportionate increase.

Another approach is by generalisation of what he suggested in *Rational cooperation.* The bargaining position is characterized as follows. "We assign an expected utility for each party to the social contract to the state of nature, as status quo, and to each possible society, as possible bargaining outcome."[74] The method is described as follows.

"We now select a rational individual at random, and attempt to present an argument that would convince him to agree to the adoption of any member of a particular set of principles for cooperative action. A similar argument must be equally convincing to any other rational person. Therefore the argument means that it is possible to correlate with each party to the social contract, a set of principles for cooperative action, any member of which the particular party would rationally agree to adopt. We then consider the intersection of these sets' to obtain the content of the social contract".[75]

An individual chosen at random would choose the maximin relative benefit. This introduction to Gauthier's bargaining framework calls for some comment.

Firstly, one can note that Gauthier removes the veil of ignorance. Like the Hobbesian agents, Gauthier's choosers are individuals in full knowledge of their present and future situation. This means that what they rationally agree to in the state of nature will remain rationally agreeable in social interactions. The content of the social contract hereby derived is rational from every single standpoint whether in the hypothetical original position or retrospectively once in society.

Secondly, Gauthier imports the tools of bargaining theory into a contractarian framework. This import causes two sorts of problem. From a game theoretic point of view, Gauthier has to make some changes to the original orthodox bargaining theory. For example, as illustrated in the quote above, he assumes equal rationality of the bargainers. We will see in chapter III, that this change and others are major enough to alter the use of bargaining theory and threaten the legitimacy of his borrowing. From a

[74] 'The Social Contract: Individual Decision or Collective Bargain?', p56
[75] 'The Social Contract: Individual Decision or Collective Bargain?', p56

philosophical point of view, Gauthier's adoption of a bargaining procedure means that the agents such as children or disabled who are not or cannot participate in the production of the cooperative surplus are excluded from the bargaining process and therefore excluded from the agreement. This added restriction to his field of morality has been considered by many as unacceptable.

More importantly, Gauthier creates the maximin before the minimax. The first version of the minimax seems to appear in 1979.[76] Together, they are the two faces of the same coin but, as shall be explained later, the maximin represents the moral side whereas the minimax represents the rational side of the coin. Once again Gauthier is initially influenced by Rawls and Harsanyi's ethical approach. In *Morals by Agreement*, he will insist on the rationality of his solution presenting the minimax first and then reconciling rationality with morality by showing that the other face of the minimax is no other than the maximin.

We have now completed step two of the social contract. Parties to the social contract have chosen in a well-characterised position the cooperative principles they agree upon. The reasons to enter the agreement are rational whether from the state of nature or within society. But will they be sufficient to keep the agreement? We have now to turn to step three to address this issue.

[76] It appears in 'Bargaining Our Way Into Morality: A Do-It Yourself Primer'. As we saw above, in this article, Gauthier considerably reduces the field of morality to distributive justice. However in doing so, he is able to put morality on a rational ground through the minimax concession principle. The starting point this time is not what is mutually advantageous but rather what is mutually disadvantageous such as negative externalities. Parties to the bargain in this case realise that they have to compromise and make some concessions.

Section 3: Keeping the covenant

It is obvious that to enter into a social contract is only the first step towards social peace, property or prosperity. If the contracting parties don't keep the agreement once in society the contract is void, hence Hobbes' third law of nature. In this section, we will first introduce Hobbes' Fool's challenge prior to reviewing Gauthier's solution to it. In the introduction, we highlighted Gauthier's evolution in his interpretation of Hobbes. We shall understand the reason for this evolution in this section.

Hobbes' Fool: the problem of keeping the covenant and enforcement system.

In a very famous section, Hobbes develops an imaginary Fool's challenge.

"The Fool has said in his heart, there is no such thing as justice... keep, or not keep covenants, was not against reason, when it conduced to one's benefit. He does not therein deny, that there be covenants; and that they are sometimes broken sometimes kept; and that such breach of them may be called injustice... ; but he questions, whether injustice .. may not sometimes stand with that reason, which dictates to every man his own good; and particularly then, when it conduces to such a benefit, as shall put a man in a condition, to neglect not only the dispraise, and revilings, but also the power of other men".[77]

The Fool in question is not so foolish. Why should I keep a covenant when the benefit from breaking it can be so great that I can afford to neglect the impact of the breach not only on my reputation but also on the consequences for others? What about cases where it is more rational to break the covenant than to keep it? These types of cases are easy to imagine. The most common occurs when contractarian parties don't have to perform their obligations simultaneously. If one performs first, what incentive has the other party to perform next if the benefits from not performing are greater than the benefits from performing?

One has to be careful about the exact boundaries of the challenge. The Fool does not claim that it is not rational to enter into an agreement. On the contrary, it is in his best interest to commit himself to a mutually advantageous agreement and then to break it

[77] *Leviathan*, chapter 15, p 538

when everybody else keep the agreement. The Fool claims that to do so is probably unjust but a lot more beneficial and therefore rational.

Gauthier identifies a two-level problem: Firstly, at a practical level, there is a security problem. Men need to adhere to the covenants made. The laying down of rights creates obligations that are mere words until actions are taken upon them. Secondly, at a material level, there is a motivation problem. Men need sufficient motivation to keep the covenants that they had sufficient motivation to undertake. But in Hobbes's system, motivation and reason are linked.

What does Hobbes reply to the Fool? Firstly, he argues that one cannot always foresee all the consequences of ones choice and there is a part of uncertainty even in the most foreseeable event. Secondly, he claims that no man can survive on his own in the state of war. If one decides to betray the others in an agreement and benefit from their help and protection without returning the favour, he can either be excluded from the society in question (most likely outcome) or deceive them and be accepted in error (unlikely outcome). Whatever the outcome, it is neither prudent nor rational to betray: on the one hand, to be excluded is not beneficial since one cannot survive on ones own. On the other hand, when making the decision to betray, one cannot reckon upon being able to deceive. "Hobbes's argument is, then, that violation of covenant cannot be *expected* to be advantageous, although it may actually *be* advantageous"[78].

According to Gauthier, the motivation problem is partly addressed by the reply to the Fool. *If* the agents reason properly, they will realise that adherence to the covenants is expected to conduce to preservation. *But* misleading passions or erroneous reasoning leads men astray. Therefore, the solution to the motivation problem does not solve the security problem, hence the need for a sovereign. A mere promise or an agreement based solely on words is not sufficient to guarantee that people will keep their agreement. The covenant would be void. 'For he that performs first, has no assurance the other will perform after; because the bonds of words are too weak to bridle men's ambition, avarice, anger, and other passions, without the fear of some coercive power'[79]. Therefore, 'before the names of just, and unjust can have place, there must be

[78] *The Logic of Leviathan* p 84-5.
[79] *Leviathan*, chapter 14, p 536

some coercive power, to compel men equally to the performance of their covenants, by the terror of some punishment, greater than the benefit they expect by the breach of their covenant'.[80]

Hobbes's moral theory stands only if one can accept that reason is at the root of obligation. One can agree with Gauthier in linking motivation with reason. So the central question to address is: do I act upon the obligation I undertook because I have reason/motivation to do so or because I fear the coercive power? We saw that the solution to the motivation problem rested on the reply to the Fool. So the next question is: is the reply to the Fool really convincing? Yes only if one accepts that the Fool is unlikely to deceive others. However, the bigger the society the less likely a betrayal will be noticed. In weighing cost against benefit, the supposed Fool can still be prepared to take a small risk for a greater benefit. Why should we behave justly, when we know (or there is a strong probability) of behaving unjustly without being noticed and punished? So the motivation problem is not really solved by this reply. Hobbes' moral system is at stake.

Gauthier's answer to the Fool and constrained maximization

Gauthier identifies rationality with utility maximization. However, individual utility-maximization does not always bring about the optimal outcome. In *Reason and Maximization* published in 1974, he tackles this central problem that will lead to his answer to the Fool. Let us follow the articulation of his argument.

He describes the problem as follows: what is rational for one person is rational for every person. Every person must judge each person rational to maximize his own utility. If this is the case, then the policy of individual utility maximization must afford every person maximum compossible utility (i.e. the greatest utility each can receive, given the utilities received by the others). But the prisoner dilemma shows that individual utility maximization does not always yield an optimal outcome. Therefore every person does not correctly judge each person rational to maximize his own utility and so each does not correctly judge himself rational to maximize his own utility.

[80] *Leviathan*, chapter 15, p 538

To this objection Gauthier replies that the rationality of the agent is determined by assessing his intended actions *in relation to his point of view*. If person A is to assess the rationality of person B, then it is the relation of B's action to B's utility and not the relation of B's action to A's utility which is relevant. The argument fails to take the position of the agent. "A's maximum compossible utility is not the relevant criteria for assessing the rationality of the actions of persons other than himself."[81] Gauthier concludes that

"the straightforward identification of rationality with the aim of individual utility maximization, although not inconsistent, is inadequate, because it denies the possibility of agreements which require one or more of the parties to refrain from the maximization of individual utility, yet secure to each of the parties greater utility than is possible without such agreement. This inadequacy does not however, show that rationality is not connected with maximizing activity".[82]

The key to the problem lies in the distinction between independent actions (in state of nature) and interdependent actions (in civil society): "Independent rational persons will each *separately* adopt the same manner of action. Interdependent persons will *collectively* adopt a common manner of action." ... "An agreed way of acting is rational if it leads to an outcome which is optimal so far as the parties to the agreement are concerned".[83] He then identifies two conditions:

1. *Condition for independent action or condition of straightforward maximization*: "a person acting independently acts rationally only if the expected outcome of his action affords him a utility at least as great as that of the expected outcome of any action possible for him in the situation".

2. *Condition for interdependent actions or condition of constrained maximization*: "a person acting interdependently acts rationally only if the expected outcome of his action affords each person with whom his action is interdependent a utility such that there is no combination of possible actions, one for each person acting independently, with an expected outcome which affords each person other than himself at least as great utility, and himself a greater utility."[84] The policy derived from this condition is clearly

[81] 'Reason and Maximization', p 423
[82] 'Reason and Maximization', p 427
[83] 'Reason and Maximization', p 424-5
[84] 'Reason and Maximization', p 427

intended to maximize the agent's *overall* expected utility. It enables him to participate in agreements intended to secure optimal outcomes, when maximizing actions performed in the absence of agreement would lead to non-optimal outcomes.

Once in society, constrained maximization takes over from plain straightforward maximization. Both the rational policy of interdependent action (individual utility maximization) and the rational policy of interdependent action (agreed optimisation) satisfy the condition of constrained maximization that best expresses the identification of rational activity with maximizing activity.

The concept of a change of rationality once in society is not completely new. Rawls had touched on it. Indeed, in his theory, each individual pursues her own good i.e. a long term rational plan of life. Rawls notes:

"The motivation of the parties in the original position does not determine directly the motivation of people in a just society. For in the latter case, we assume that its members grow up and live under a just basic structure, as the two principles require; and then we try to work out what kind of conception of the good and moral sentiments people would acquire." [85]

Because the concept of right is anterior to the good once in society, the good is 'constrained' by it.

Gauthier develops on the theme: Can we really choose our conception of rationality? Yes, he replies, conceptions of rationality are not fixed in human nature, but rather the products of human socialization. "Far from supposing that the choice of a conception of rationality is unintelligible, I want to argue that the capacity to make such a choice is itself a necessary part of full rationality"[86]. And of course, if a person was to choose rationally his conception of rationality he would choose the conception of constrained maximization... even the economic man would.

The last touch to his new concept of morality as described in step 2 is brought about by this definition of rationality as constrained maximization. Morality is rational although not identified with mere prudence.

---

[85] *A Theory of Justice*, p 128.
[86] 'Reason and Maximization', p 431

"The moral man is no less concerned with his own well-being than is the prudent man, but he recognises that an exclusive attention to that well-being would prevent him from participation in mutually beneficial agreements... the prudent man considers it rational to *become* moral, but not rational to *be* moral. On prudential grounds he can justify the adoption of moral, rather than prudential grounds of action, but only if he does adopt moral grounds, and so becomes a moral man, can he justify a moral, rather than a prudential, policy of action."[87]
We saw above that morality was rationally derived and we see now that it is rationally maintained.

At this stage, Gauthier is still in his *Logic of Leviathan* framework and is convinced that Hobbes missed the point:

"In philosophical literature, the classic example of the man who is bound by his conception of reason is the Hobbesian man, the self-maintaining engine... Men recognise the rationality of entering society, but force, not reason is required to keep them in there ... Hobbesian man is unable to internalize the social requirement that he subordinate his direct pursuit of survival and well-being to the agreed pursuit of optimal outcomes which best ensure the survival and well-being of each person. Thus in our terms Hobbesian man actually remains in the state of nature; the civil power, the sovereign, can effect only the appearance of civil society, of interdependent action. The real difference between the state of nature and civil society must be a difference in man, and not merely in the external relations of men."[88]

The missing piece is a moral background.

We know from step 2 that, in the early 70's, a change is being processed in Gauthier's moral philosophy. His reconciliation with Hobbes's moral theory owes as much to his restrictive concept of morality as to his concept of rationality as constrained maximization. We have to wait until 1977 for Gauthier to read in Hobbes this dual rationality. Once more, let us follow him in *The Social Contract as Ideology.*

A dual rationality in Hobbes's philosophy

Gauthier starts by claiming that for Hobbes "reason does not in itself determine a system of rights and laws which relate men one to another in any way other than the natural relation of hostility. The 'rational' order corresponding to the unlimited right of

[87] 'Reason and Maximization', p 432-3
[88] 'Reason and Maximization', p 431

nature is the condition of war of every man with every man."[89] In other words, rationality is individual, instrumental and straightforward. But equipped with this form of rationality, men have no weapon to resolve their conflicts once in society. Hobbes needs to resort to an Arbitrator, a Judge of what 'right reason' should be, distinct and above individual self-interests.

If rationality is to be identified with utility maximization, we now face a new problem. Indeed, society must perform the functions of coordination, bargaining and constraint. Coordination and bargaining are compatible with maximization in a way that constraint is not. Internal constraint (constraint of conscience) especially poses the problem of contradicting the requirements of reason. Gauthier claims that such a contradiction can be solved by no means other than a distinction between natural and conventional reason as suggested in his *Reason and Maximization.*

Gauthier then quotes Hobbes:

'And when men that think themselves wiser than all others clamor and demand right Reason for judge; yet seek no more, but that things should be determined, by no other men's reason then their own, it is as intolerable in the society of men as it is in play after trump is turned, to use for trump on every occasion, that suite whereof they have most in their hand'.

Gauthier interprets this passage as follows:

'Trump is established by the social contract, as that convention required to achieve an optimal state of affairs, better for each than the natural outcome. But each man, guided by his own reason, uses for trump his own interest. And this is intolerable, for its undercuts the contract and makes society impossible.'[90]

In other words, Hobbes recognises the need for a conventional standard of right reason but he fails to establish this standard adhering to his individualist view. In the state of nature, this view is true "but as Hobbes himself recognizes, this is exactly what is intolerable in society or, indeed, intolerable if there is to be any society."[91]

A year later, Gauthier writes his *Thomas Hobbes: Moral Theorist.* In this article, Gauthier is then convinced that Hobbes has explicitly distinguished between right and conventional reason and has therefore established a moral theory. Gauthier distinguishes

[89] 'The Social Contract as Ideology', p 151
[90] 'The Social Contract as Ideology', p 155
[91] 'The Social Contract as Ideology', p 155

between reason and motivation. A covenant is *rationally stable* if each person has reason to adhere to it, provided others do. A covenant is *motivationally stable* if each is moved to adhere to it, provided others do. He confirms that in order to be motivationally stable Hobbes has to have recourse to a coercive power. The motivational problem cannot be solved on merely prudential grounds. But the rational problem can be solved using the concept of right reason.

In the state of nature, each man takes his own reason for right reason and can consider acting in accordance with his own reason as right. This *natural reason* enables men to achieve their primary end i.e. self-preservation and leads them to lay down their right of nature. But as they lay down their right of nature they also lay down this very natural reason as their aim switches from self-preservation to peace i.e. from individual to mutual advantage. A new conventional reason then supersedes natural reason. Since right reason is individually devised, it is no longer a reliable tool to harmonise all individual behaviours towards mutual advantage. Therefore, Hobbes needs to introduce an arbitrator (the sovereign) to decide on the common right reason, the *conventional reason*. Each has ground to accept it as long as it is common knowledge that most persons accept it and expect others to adhere to it. At last the Fool has his reply... says Gauthier. This Fool was using his natural reason within conventional boundaries when he should have laid it down with his right of nature. In 1982, Gauthier adds to this demonstration that the connection between right and reason is central to Hobbes's thought and that the Fool, in appealing to natural reason, fails to appreciate this tight conceptual connection[92].

What have we achieved and on what grounds does Hobbes's moral theory rest? The grounds are thin but here they are. Gauthier claims that Hobbes's morality arises with the laws of nature which establish a "uniquely dominant set of conventions for men who seeking preservation must seek peace"[93]. These laws restrict men's behaviour. Therefore Hobbes's morality is a set of conventions constraining each man's maximizing activity and distinguishing right from wrong. This concept of morality could be described as moral conventionalism but it is severely challenged by the Fool. Indeed, if morality is to be a rational and conventional constraint on natural behaviour,

[92] 'Three Against Justice', in *Moral Dealing*, p 143
[93] 'Thomas Hobbes : Moral Theorist', p 552

then it must be rationally stable. Since reason means maximisation of advantage, rationally stable means "most advantageous for each to follow, provided others do". But then rational and conventional morality cannot be considered as a constraint on one's behaviour. And if morality becomes a constraint, it is then irrational.

The solution to this dilemma is in the double change suggested above a) of end from self-preservation to peace and b) of reason from individual to conventional. In this context the second law of nature can be rephrased:

"As long as each person appeals solely to his natural reason, there can be no security to any man of living out the time that nature ordinarily allows. Thus a man must be willing when others are so too, as far as he shall think it necessary for peace, to lay down natural reason, and be contented with a standard of reason which allows him so much liberty against other men, as he would allow other men against himself"[94].
The convention is now rationally stable.

After *Morals by Agreement*, Gauthier will change his interpretation on Hobbes on a further point. He will claim that it is possible for a Hobbesian agent to commit herself on non-prudential grounds. In the *Logic of Leviathan,* following his initial belief as described in *Practical Reasoning*, Gauthier considers it impossible for a Hobbesian agent *to commit* himself, even in his own interest, to perform actions that would not be in his interest, *at the time of performance.* He argues then that there is no room in Hobbes's theory for non-prudential ground for obligation. In 1988, he is no longer so sure that the problem of commitment is clear to Hobbes.

"I now believe that the Hobbesian text gives no real guidance on this matter. A person's interest must enter at some point into an explanation of each of his voluntary actions; a commitment against interest, to perform an action against interest, would be incompatible with Hobbes's psychology. But this leaves room for a commitment based on interest. Whether Hobbes would have accepted it we cannot say, but it opens the door to a much more interesting and defensible interpretation of a large part of his moral and political philosophy"[95].

We shall see how in chapter VII.

[94] 'Thomas Hobbes : Moral Theorist', p 557
[95] 'Hobbes's Social Contract', p 79.

Conclusion

The present chapter has sketched an overview of Gauthier's work before *Morals by Agreement*. In doing so, my purpose was three-fold.

Firstly I intended to show how the project of his contractarian moral theory took form. In particular I felt it was important to highlight the emergence of his concept of a dual rationality so central to his theory. Gauthier is convinced that men, through reason, can change their instrumental reasoning. This change of rationality sustains the moral answer he wants to give to the Fool.

Secondly, this chapter intended to demonstrate that the Hobbesian background of the author influenced his research just as his research influenced his interpretation of Hobbes. Gauthier abandoned Hobbes's state of nature in favour of a Lockean one, he changed the boundaries of morality, he redefined the principle of the social contract and he distinguished two concepts of rationality but he retained Hobbes's central framework and rational morality. Gauthier gathered together all the main changes in the moral and political philosophy of his time to amend and improve on *The Leviathan*.

The third goal of this chapter was to introduce the future core conceptions of *Morals by Agreement* within their context. This book is not only the first achieved version of a contractarian moral theory, it also provides a good picture of the state of moral and political philosophy in the second half of the 20th century.

The major concepts that constitute *Morals by Agreement* have now been presented, we are at the beginning of the 80's and Gauthier is going to put together all the pieces of the jigsaw so that they become *a* theory.

Annex

Reason and practical reasoning.

The purpose of this annex is not to summarise Gauthier's thesis but to provide an overview of his original project since it will guide his future research. Indeed, Gauthier wants to merge prudential and moral reasoning and he believes that practical reasoning rather than deliberation can bring his project to completion. Even if, in my opinion, this first attempt is not successful, it highlights his ambition and original achievements. We shall first give some of his definitions prior to developing the core of his argument.

Some definitions and concepts

A *practical problem* is a problem about what to do. It is context dependent. A *practical question* is the verbal formulation of a practical problem. A *solution* is an action or set of actions performed by the person in the situation specified in the problem. A *resolution* is a decision taken by the person concerning what to do in the situation. An *answer* to the practical question is a judgement about what to do in the situation or an injunction to do something[96].

A *practical judgement* sets out the answer to a practical question in full detail. It is a conclusion on what we ought to do, reasoning from premises specifying the situation in which we act and our reasons for acting. It is usually formulated using the words 'ought', 'should' or 'best'. For purpose of simplification, Gauthier agrees that practical judgments will take the standard 'ought' form[97].

Decision and judgement can go in different direction. I can *judge* an action best for me and yet *decide* to follow another course of actions. Similarly, 'I ought to do x' does not equate 'I shall do x'.

---

[96] *Practical Reasoning*, p 5

[97] Although the use of ought is neither a necessary nor a sufficient condition of a practical judgement.

Going further, Gauthier distinguishes between reasons and motives. *Reasons* are the grounds for practical judgement whereas *motives* determine decisions. Practical judgment rests on weighing reasons for acting in the situation specified in the problem. The action one 'ought' to perform is the one supported by the weightiest reasons. Motives do not necessarily correspond to reasons. A rational action is an action governed by reasons i.e. an action resting on motives that corresponds to reasons.

Gauthier then introduces a distinction between prudential and moral practical judgments. You ought to do x can be interpreted in two ways: it could mean 'it is your duty to do x' or 'it is in your interest to do x'. In the case of duty the action is wanted for itself, in the case of best interest it is necessary as a means to an end. Gauthier will call the former 'moral judgment' and the latter 'prudential judgment'. He acknowledges that the distinction is not always easy to establish and he gives examples of conflicting or ambiguous practical judgments. Nevertheless, he settles for the following rough guideline to separate the two:

"a practical judgment which has a *moral* force is based on considerations independent of the will (purposes, aims, desires) of the prospective agent, and dependent solely on the intrinsic nature of the act itself, whereas a practical judgment which has a *prudential* force is based on considerations dependent on the will"[98].

Practical reasoning

Gauthier then proceeds to develop another distinction between deliberation and practical reasoning. Deliberation is a psychological process by which a person comes to resolve a problem whereas practical reasoning is a formal logical argument. The outcome of deliberation is a reason*ed* action i.e. an action for which one can provide a reason. But a reason*ed* action is not necessarily a reason*able* action; the reason provided can be the wrong reason because the agent 'may refuse to consider important features of the situation in which he acts' or he 'may ignore some of the consequences which his action will have'[99]. Reciprocally a reason*able* action needs not be reason*ed*. One can act on an impulse without prior reasoning and yet do the right action.

---

[98] *Practical Reasoning* p 20.
[99] *Practical Reasoning* p27.

Practical reasoning on the other hand provides an *explanation* for the action as well as a *justification*. In other words the reasons for acting are also the reasons why the agent acts. "If *his* reasons are not *the* reasons, or good reasons, for acting, no justification – or an insufficient justification - is provided"[100]. Does it make actions derived through practical reasoning reason*able*?

If practical reasoning is a formal logical argument then, Gauthier argues, the major premise must characterise the 'wants' of the agent. The wants should be understood as the realisation of states of affairs. They encompass "the possession of some object by the agent, the enjoyment of some experience by him, the attainment of some position or standard, or even the realisation of some situation not connected to him at all"[101]. These wants constitute the 'practical basis'. At least one constituent of the agent's practical basis must ground a practical premise. Such a premise has a practical force. At least one premise with practical force is necessary to any practical reasoning. Practical reasoning is about what we want consciously and on reflection.

An agent's wants may have conflicting objects that no state of affairs will satisfy simultaneously. In such a situation, the agent would have to weigh his wants and choose the one with the greatest practical force in order to maximise his satisfaction. In order to achieve this weighing exercise, one has to introduce comparative premises relating the desirability of the objects of the various wants and draw a conclusion regarding the most desirable state of affairs. Gauthier also suggests introducing future (foreseeable) wants in the major premises when weighing various objects and setting the main objective. Although they cannot impel one to act now, they can determine reasons for one to act now. Failure to take into account future wants would impede on the prudential ground of the practical reasoning.

But a central question remains: how are we to take into account the wants of others? The wants of others are not *my* wants and therefore cannot be grounds for my actions. If the wants of others are to provide me with reasons for acting, they have to be included in *my* wants. This is made possible only through one very simple scheme: not all my wants are dependent on me only. Some of my wants require the assistance of others. If I

[100] *Practical Reasoning* p27
[101] *Practical reasoning* p 30

want them to assist me in my wants, I must act in such a way that they are willing to do so. If I assist them in their wants they will assist me in my wants. Therefore, the wants of others have to be included in my wants.

Gauthier is aware of the complexity of such a scheme. The wants of others do not necessarily enter our wants so straightforwardly. It makes sense to me to help my neighbour today knowing that she might help me tomorrow in return. But why should I help a blind person to cross the road? The chances of this person being in a position to 'repay' me tomorrow are rather slim not to say non-existent. Some of our wants rest on principles, obligation, duties, laws or customs. We need now to turn to moral practical reasoning to see his answer.

Let us recapitulate: prudential practical reasoning can be described as follows.

"A person is confronted with a practical problem. From his practical basis, which comprises his wants (present and future), he selects those whose objects may in some way bear on his problem. He then formulates a set of statements characterising these objects as desirable to him, and comparing them in terms of this desirability. Taking further statements about his capacities and experiences, in so far as they may bear on the problem, about the situation he must act, and the consequences of actions possible in that situation, he may infer an answer to his problem, a judgment of what he ought to do"[102].

Practical conclusions are formulated by an imperative 'I ought to do x'. Imperatives may take many forms. They are not necessarily intended as commands. Rather they can take the form of advice, instruction or request.

Core of his argument.

The next step is the core of his argument since Gauthier attempts to demonstrate that all practical reasoning takes the same form whether prudential or moral. Moral practical judgments are derived using the same type of inferences as the ones used in prudential judgments. However, the basis of practical reasoning has to be enlarged. How?

The basis of the practical reasoning of any agent must include all wants of all persons. Gauthier almost immediately specifies that such an extension is not as large as it seems

[102] *Practical Reasoning* p 80.

since it is reduced by the condition of practicability i.e. the condition according to which it must be humanly possible to infer the conclusion from the premises. Therefore only 1) the wants relevant to the situation and 2) the wants of the beings whose points of view are to be considered in determining what one ought to do are taken into account.

Gauthier believes that unless 'a consideration is capable of moving some person directly to deliberate action – unless it is an object of his wants – it cannot serve anyone as a genuine reason for acting. My reasons for acting need not move me to action directly – but they must move someone, at some time'[103]. This conclusion can be interpreted as follows: 1) principles, obligations and duties must be included in my wants, or said differently again, they must be rationally derived; 2) some reasons for acting are potentially common to all. They are not specific to me.

This new approach to moral practical reasoning seems to blur the distinction that Gauthier originally made between prudential and moral judgments by giving a central role to the agents' will. However, we can recognise here what will become one of his main argument: instrumental reason is about the realisation of a state of affair that is rational for any agent to want to realise.

Gauthier then details his treatment of principles, duties and obligations.

As far as *principles* are concerned, Gauthier defines them as general practical judgments i.e. they are derived through anterior practical reasoning. They are sufficient to be reasons for acting and they can serve as the basis for practical reasoning. However they have to be reviewed according to the requirements of the situation. In particular, if two principles conflict, rational deliberation is required to sort out which principle would be the best reason for acting in the situation.

As far as *obligations* and *duties* are concerned, Gauthier follows G.E.M. Anscombe in reminding us that they are nothing else than survivals from the law conception of ethics, dating from the secularisation of Christian ethics. This forgotten historical origin has misled moral philosophers over the centuries. As we shall see, Gauthier prefers to follow Hobbes in his theory of obligation. He considers obligation as binding for

[103] *Practical Reasoning* p 94.

prudential purposes. We have the choice but we decide to bind ourselves because it is in our interest to do so. But is the undertaking of an obligation on a prudential basis sufficient to ground performance of the action to which one obliges oneself ? This question leads us directly onto Gauthier's work on Hobbes.

CHAPTER II: *MORALS BY AGREEMENT*

In this chapter I attempt to present Gauthier's masterpiece. I follow his steps as faithfully as possible and try to understand the theory as a whole. This book is dense enough to be interpreted in many different ways. I have decided to focus more exclusively on the very core of the theory i.e. the link between rationality and morality. In this chapter, I try as much as possible to give full voice to its author. I will turn to criticisms in the next chapter. There are places where some comments could be made but because they do not impede on the logic, coherence or clarity of his argument, I have chosen not to mention them and break the flow of the presentation.

In *Morals by Agreement*, David Gauthier defines moral principles as types of rational principles for making choices or decisions among possible actions that constrain the actor to pursue his own interest in an impartial way. His main purpose is to prove that moral duties are rationally grounded and that although morality has the appearance of a constraint on the pursuit of individual interests, its acceptance is truly advantageous.

Morals by Agreement offers a contractarian approach based on the core idea that once an individual is part of a group (or structure) his personal interest is likely to interact with the others'. An individual best interest becomes then connected with social welfare. Moral rules are those set of rational rules regulating an individual's personal interest when it is affected or affecting other individuals' personal interest. "A society could not command the willing allegiance of a rational person if, without appealing to her feelings for others, it afforded her no expectation of net benefit"[104].

Gauthier's purpose is therefore two-fold: firstly, it is to show why it is rational for people in a society to voluntarily agree ex-ante on a set of impartial moral rules and, secondly, it is to find the conditions that would guarantee that people do comply ex-post with the set of rules agreed upon. To this end he introduces *five core conceptions* that ensure the coherence of his theory

[104] *Morals by Agreement* p 11

1- Morally free zone. A morally free zone is a context within which the constraints of morality would have no place. It is usually assimilated to the idealized perfectly competitive market. There is no place, rationally, for constraints since mutual advantage is assured by the activity of each individual in pursuit of his own interest. There is no place, morally, for constraints due to the impartiality of the market's operations. People only need to pursue their personal interest to participate effectively in a venture for mutual advantage. However, the market would be a morally free zone only if it existed in its ideal form. But this ideal is not realisable since it rests on the assumptions of free activity, private ownership, private good and no externalities.

2- Minimax relative concession or maximin relative benefit. Where mutual benefit requires individual constraint, reconciliation is obtained through a rational agreement. The outcome of this agreement has to be mutually advantageous. Gauthier appeals then to the theory of rational bargaining in order to establish sufficient conditions for a mutually advantageous outcome. He introduces a measure of each person's stake in the bargain. The principle of minimax relative concession is the requirement that the greatest concession, measured as a proportion of the conceder's stake be as small as possible. The principle of maximin relative benefit requires that the least relative benefit, measured as a proportion of one's stake, be as great as possible. Individuals bargain (each knowing their situation in society) until a rational / fair agreement (i.e. common strategy) is reached; this agreement should comply with the principle of minimax relative concession.

3- Constrained maximization. The rationality of compliance ex-post to rules agreed ex-ante is not shown yet. Hence the necessity of introducing the third core notion of constrained maximization. A constrained maximiser is a person who is disposed to comply with mutually advantageous moral constraints provided she expects similar compliance from others. What if the others do not comply? Gauthier then attempt to show that under certain conditions, the net advantage that constrained maximisers gain from co-operation exceeds the exploitative benefits that other may expect when straightforward maximizing. Under these conditions, it is then rational to maximize in a constrained fashion. The essential point here is that constrained maximisers (CM) have to learn to spot the straightforward maximisers (SM) in order to exclude them from agreements. In order to accommodate this need, Gauthier introduces the assumption of

translucency (CM and SM have reasonable skills to spot each others). Under this assumption the CM have a good chance of cooperating between themselves and avoiding exploitation from the SM.

4- Proviso regarding the initial bargaining position. In order to ensure that an individual is rationally willing to enter into an agreement we have to look at his initial bargaining position. Gauthier thus introduces a constraint on the initial bargaining position that prohibits bettering one's position through interaction that worsens the position of another. Gauthier also clarifies a point: "our theory denies any place to rational constraint, and so to morality, outside the context of mutual benefit. A contractarian account of morals has no place for duties that are strictly redistributive in their effects, transferring but not increasing benefits, or duties that do not assume reciprocity from other persons"[105]. In other words, the initial endowment has to be fair in order for bargaining to be fair. Individual and Property rights are the outcome of the proviso not of the agreement. People arrive at the negotiation table with their rights.

5- Archimedean point. Now that all the conditions are set to ensure the rationality of agreeing ex-ante on a set of rules and complying ex-post to the rules agreed, we need to know how to choose. The missing piece of Gauthier's jigsaw is then the Archimedean point i.e. the point from which an individual can choose impartially and therefore morally on which rules to agree to. Although the individual has to choose impartially, so without knowing his social position, he knows that he will still pursue his personal best interest. Any individual best interest then becomes the common interest in cooperation. Unsurprisingly, anybody at the Archimedean point would choose the three principles unveiled above: 1) the proviso; 2) the market; 3) the minimax relative concession.

How do all these conceptions work together to form a theory? The first (the market as a morally free zone) and the last (the Archimedean point) are stand alone conceptions which play two different roles. The market highlights the role of impartiality and morality in the theory whereas the Archimedean point provides the outlook of an impartial actor on the whole.

---

[105] *Morals by Agreement* p16

Within the three core conceptions left, Gauthier distinguishes between the one that secure the rationality of making the agreement, namely the minimax principle and the ones that secure the rationality of keeping the agreement i.e. constrained maximisation and the proviso. Firstly, rational individuals defending their best interest and using common bargaining rules agree on a mutually advantageous moral principle. It is then indispensable that the agreement so reached be binding. To be so, Gauthier argues that the agreement must be *rational* to each agent once in social interactions and (procedurally) *fair* from the initial position onward. Gauthier claims that it is rational for individuals, once in social interactions, to change their rational disposition and constrain their utility maximisation when such a change is to their overall advantage. He also claims that rational individuals would agree to a rectification of the initial position through the proviso to obtain a fair procedure and therefore secure the compliance of all (or most) the agents to the agreement.

In this chapter, we shall first introduce the issues at stake and the angle from which Gauthier approaches his task. We will then present in a second section his theory as a whole. Each core conception will be developed and connected to the others. We shall then conclude on Gauthier's achievement and his vision of a moral society.

Section I: Issues at stake

In this section we shall highlight the core topics of the book and pose the problems that *Morals by Agreement* endeavours to solve. We shall first introduce the challenges raised by Gauthier's conception of justice prior to presenting Gauthier's conception of rationality as utility maximization and the difficulties posed by such a conception in strategic situations.

Justice as fairness

Two conceptions of justice

Brian Barry in a *Treatise on Social Justice*[106] distinguishes between two conceptions of justice: 'justice as impartiality' and 'justice as mutual advantage'. Both conceptions of justice have in common that they

- spring from conflict of interests between different people
- are the fruit of a rational agreement and
- represent justice as fairness.

In 'justice as mutual advantage', justice is "simply rational prudence pursued in contexts where the cooperation of other people is a condition of our being able to get what we want. Justice is the name we give to the constraints on themselves that rational self-interested people would agree to as the minimum price that has to be paid in order to obtain the cooperation of others."[107] Justice as mutual advantage is usually developed within the framework of bargaining theories. Justice is understood as the outcome of rational bargaining and reflects the bargaining power of the parties involved. The best representatives of such an approach are obviously Hobbes and his most recent follower, Gauthier. Barry presents this conception of justice as a 'two-stage theory'. In the first stage, the philosopher decides on a starting point representing the non agreement outcome (or 'baseline'). In a second stage, he suggests a formula that will take the

[106] *Theories of Justice*
[107] *Theories of Justice* p 6-7

bargainers from this 'state of nature' to an optimal outcome. Nash and Braithwaite are known for attempting such theories before Gauthier.[108]

In 'justice as impartiality', it is assumed that there exist independent reasons to behave justly that have nothing to do with the pursuit of self-interest. Holders of such theories must then explain the appeal to justice. According to this approach, "justice should be the content of an agreement that would be reached by rational people under conditions that do not allow for bargaining power to be translated into advantage."[109] Parties to the agreement do not attempt to look at things from their own point of view but seek to reach an agreement acceptable from all point of views. The reader would have recognised the Kantian approach in this conception of justice. Barry identifies this approach with the 'original position theories'. In other words, the parties to the agreement are placed in an original position, not knowing who they are or will be and they have to decide on a fair agreement. The most well known modern philosophers who have worked their way down this approach are Harsanyi and Rawls[110].

In other to understand and appreciate Gauthier's achievement, it seems interesting to focus now on the reasons to reject 'two-stages theories'. Barry presents a strong case against these bargaining theories. In substance he brings to the fore the following arguments.

- *Fairness of the outcome*: the justice so obtained is loaded by the bargaining powers of the parties to the agreement. If justice is to be understood as fairness, then the outcome is unlikely to agree with our intuitive notion of fairness since it would be to the favour of the parties with the greatest bargaining power or with the strongest 'threat advantage'.

- *Danger of deadlocks*: Bargaining can be a long, costly process and could end up in a deadlock if the parties fail to reach an agreement. Failure to reach an agreement would obviously occur if some or all the parties were not willing to make sufficient concessions. In case of deadlocks, players can fail to improve over their respective status quo (or non-agreement point).

[108] See the appendix section 1b for Nash solution.
[109] *Theories of Justice* p 7
[110] See the appendix sections 2c and 3a.

- *Danger of threats*: the parties can make threats that would be costly and counter-productive to enforce. In such cases, not only would the parties fail to improve but they would also worsen their situation over the non-agreement point. Yet, it can be rational in bargaining to carry out a threat in order to be 'taken seriously' on a next bargaining round i.e. reinforce a 'threat advantage' (obviously, this is true only if there is a next round).

One way to bring in impartiality and avoid deadlocks in bargaining would be to have recourse to an impartial arbitrator. However, Barry argues quite convincingly that arbitration outcomes simulate bargaining ones. The arbitrator, although impartial, takes into account the rational claims and positions of the parties. He is therefore lead to similar solutions that the parties would have reached had they sat at a bargaining table.

The argument about the (un)fairness of the outcome is important enough to justify a little detour. Braithwaite was inspired by Nash's bargaining theory but felt uncomfortable with the suggested solution that he found contrary to his intuitive notion of fairness. Indeed, Nash takes the parties' initial endowments as given and specifies a solution that improves on the non-agreement point (what the outcome would be if no agreement is reached) and does so in proportion to each parties' bargaining power.[111]

Braithwaite gave a problematic example[112]: Luke is a pianist and Matthew is a trumpeter. A flimsy wall separates their two flats and they have the same one hour a day to practice their respective instrument. When they both practice, it is a real cacophony and they obviously each prefer to practice solo. The Pareto frontier is between their respective maximum claims, i.e. each practising solo seven days a week. What would be a fair division of the practising time between the two musicians?

In this example there is no built-in non-agreement point. There is no status-quo or obvious starting point. Braithwaite therefore suggests basing the non-agreement point on 'optimal threats', like Nash would have done. These optimal threats are given by the respective preferences of the players. In our story, both musicians prefer to practise solo

[111] See the appendix, section 1, b for further details.
[112] *Theory of Games as Tool for the Moral Philosopher.*

but Matthew's second best is cacophony whereas Luke's second best is silence. Luke prefers not to practice than to practice at the same time as Matthew. Matthew has therefore a clear advantage over Luke, he has a 'threat advantage' that defines the non-agreement point.

Nash's solution[113] allocates nine tenths of the solo playing time to Matthew. Matthew plays nine evenings solo for one evening solo for Luke[114]. How can we assess the 'fairness' of such a division? Or rather what do we assess: the fairness of the procedure or the fairness of the outcome? Due to his preferences, Matthew has an advantage over Luke. This advantage has to be reflected in the solution otherwise Matthew would have no interest in sitting at the bargaining table. Matthew has to gain over the non-agreement point in order to agree to an outcome. Said otherwise, the outcome has to take into account his advantage over Luke. It would not be rational for Matthew to disregard his advantage when bargaining. But it would not be 'fair' either. If we want the bargaining to be fair, each player must gain from it *over* what the non-agreement point would bring her. The fairness of the outcome is then *derived* from the fairness of the bargaining procedure. We return to our initial question: what do we assess: the fairness of the procedure or the fairness of the outcome, independently of the process it comes from? Do we judge the outcome in the abstract or as a result of a bargaining process? [115]. As we shall see, Gauthier dedicates a lot of attention to this issue. Indeed, he considers that compliance with the agreed outcome of bargaining is necessarily threatened if the bargaining process is not fair.

As an introduction to Gauthier's conception of justice we shall provide below his definition of justice. We shall see that it leaves us with more questions than answers with regard to the issues raised above.

---

[113] See appendix section 1b.According to Nash, the solution on the Pareto frontier (between the two maximum claims) is where the product of both players' utilities is maximal. For a full demonstration of the Nash solution to Braithwaite's example see Barry's *Theories of Justice* p 33-36.

[114] Braithwaite suggested a formula bringing about a less 'extreme' solution (27 evenings of solo practice to Matthew against 16 to Luke).

[115]This debate echoes the debate raised between Luce & Raiffa and Harsanyi between arbitration and bargaining models. See appendix sections 1b and 2b.

Gauthier's conception of justice – A definition.

"Justice is the disposition not to take advantage of one's fellows, not to seek free goods or to impose uncompensated costs, provided that one supposes others similarly disposed... We may identify justice with the rational disposition to cooperative behaviour."[116]

- Justice is defined as a *disposition.* This is a central novelty of Gauthier's theory[117] and it is also the link between all its core conceptions. To be disposed to do something is to acquire the willingness to carry out *a plan.* A disposition acts as a *frame* of the mind and therefore as a frame of behaviour. The advantage of a disposition is that it carries through, from making an agreement to keeping it. The most natural question that comes to mind is: how is this disposition acquired (naturally or otherwise)?

- Justice is initially a disposition *not to* do a certain number of things. It therefore acts as a *constraint* on one's behaviour. We will see that in the market being a morally free zone, individuals are free to maximize their utility. Justice appears as a restriction on this freedom. Justice arises because of the existence of free riding and externalities i.e. it owes its existence to market failure. Once again, if justice is a constraint, what is the root of this constraint?

- There is one qualification / condition to the disposition: one must suppose that the others are likely disposed. One does not need to *know* whether the others are similarly disposed or not, one just needs to *suppose* they are. This distinction reveals the strategic aspect of justice; justice arises in interaction between individuals where there is uncertainty about others. Justice is therefore a *social disposition.*

- Justice is then identified with a disposition *to cooperate.* Cooperation seems to be the product of practical rationality but it also gives a positive side to justice. We saw above that justice is a constraint on behaviour, we learn now that the constraint has a purpose and the purpose is to frame behaviour in view to cooperate. Cooperation seems to be the link between practical rationality and justice.

---

[116] *Morals By Agreement*, p 113

[117] It is also a feature central to Mc McClennen's approach. See 'Pragmatic Rationality and Rules' for example.

Gauthier's definition of justice leaves us with enough questions to answer. But more difficulties arise when we turn to his conception of practical rationality. After detailing Gauthier's conception of rationality we shall consider the problem posed in strategic situations.

Rationality and utility maximization – Preference, utility and value

Chocolate cake or fruit salad?

Nicole and Andrew are on a date in a restaurant. The dessert menu comes. Nicole immediately spots 'Suicide by chocolate', a chocolate cake to die for. Nicole adores chocolate cakes. However, she believes that Andrew prefers slim girls with healthy eating habits. She likes Andrew very much and she really would like him to call her back for another date. So (reluctantly!) she chooses the fruit salad.

Nicole's choice definitely does not reflect her preferences since she would have strongly preferred to enjoy a delicious chocolate cake rather than a boring fruit salad. Nicole's choice only reflects her preference for *a state of affairs* in which she chooses a fruit salad and her chance of Andrew calling her back is high to a state of affair in which she chooses the chocolate cake and the probability of Andrew calling her back is low. In other words, she *values* the fruit salad (and a possible relationship with Andrew) more than the chocolate cake (and probably no Andrew).

Michelle, in an identical situation a few weeks later, chooses the chocolate cake. We don't know Michelle's *explanation* for her choice but we can suggest the following:
Explanation 1: Michelle holds different *beliefs* than Nicole's about Andrew's girl preferences. She is convinced that he prefers girls who know how to live and enjoy their food;
Explanation 2: She has no *information* at all about Andrew's preferences. She therefore feels no constraint on her choice of dessert and chooses the one she actually most prefers;
Explanation 3: After *consideration*, Michelle decides that, although she really likes Andrew and she believes he likes slim healthy girls, the *immediate* enjoyment of the

chocolate cake is *worth* more than a fruit salad and a higher probability of a relationship with Andrew.
Explanation 4: Michelle *considers* that she has to reveal her true preferences to Andrew straight away and let him decide to choose her as she is or to look elsewhere for a girl that would better suit his tastes.
Explanation 5: Michelle has already made her mind up that Andrew was not for her and that she may as well enjoy the dessert she most prefers since she no longer has anything to lose.

Parametric situations: preferences and values

Gauthier endorses the concept of rationality as utility maximization so popular in modern social and political philosophy. What is rational for an individual is what maximizes her utility, her utility being a *measure* (not an explanation) of her strength of preference. As such utility is also equated to value, to subjective and relative value. However, Gauthier is careful to specify that the equation between value and preference holds only for *coherent and considered preferences.*

Gauthier follows Harsanyi's treatment of utility maximization in parametric situations[118]. He distinguishes between conditions of certainty, risk and uncertainty. In conditions of certainty, preferences are ordinal, coherence in preference means completeness and transitivity. Under risk and uncertainty, preferences are cardinal, individuals maximize their *expected utilities*, using objective probabilities under risk and subjective probabilities under uncertainty. On top of completeness and transitivity, coherence of preferences means also independence and monotonicity.

It is irrational to express a preference (attitude) for blue and to choose (behaviour) red. To have considered preferences is to identify or coordinate attitude and behaviour. However, consideration goes further than this initial coherence and rationality. When choosing, one has to reveal a preference for a state of affair rather than for an object of choice, taking into account one's level of information about the situation and one's beliefs *at the time of decision.* If when choosing, an individual has a false belief or the wrong information, it is unfortunate but it is not a case of irrationality. The various

[118] See the appendix, sections 1a and 2a.

cases in the above example illustrate the concept of considered preferences adopted by Gauthier.

A difference in belief can alter the value of the object of choice and therefore the choice made. Nicole chose the fruit salad because she thought that Andrew prefers girls with healthy eating habits whereas Michelle in explanation 1 chose the chocolate cake because she held the opposite belief. Explanation 2 of Michelle's choice shows that the same holds for a relevant difference in the level or nature of information held by the decision maker.

In explanation 3 of Michelle's choice, we can imagine that Michelle later regrets her decision when she is waiting in vain for Andrew's call. However, it does not make her choice irrational during the date since she made it after consideration of all the relevant data. She preferred *then* the chocolate cake to the fruit salad even taking into account her beliefs about Andrew.

In both Nicole and Michelle's cases, the choice doesn't depend on the *intrinsic value* of the dessert. Neither seems to take into account the value of ingredients and labour expanded or the price of the actual dessert. The *absolute value* of the dessert is not taken into account either; Nicole does not choose the fruit salad because fruits are better for her health than chocolate, or at least not directly. Gauthier's conception of value is *subjective and relative*: subjective since subject to the rational assessment of the valuer and relative since dependent on his individual circumstances.

The value of the dessert *depends on the valuer*. In Michelle's explanation 3, she is in the same situation as Nicole a few weeks earlier, with the same level of information and the same belief. Yet she makes a different choice based on her own rational assessment of the situation. Unlike Nicole, she *prefers (or values more)* the immediate enjoyment of the cake than a higher probability of a relationship with Andrew.

The value of the dessert *depends also on the state of affair (or outcome) the valuer wants to bring about through her choice*. In explanation 5 of Michelle's choice, she is seeking a different outcome than Nicole since she is not interested in a relationship with Andrew. Her choice of dessert is different *because* the goal or outcome sought is

different. The choice is *explained* by the chooser's individual preference for a state of affair attached to the object of choice. Value is then a *measure of considered preference* for a state of affair.

Value is a measure of coherent and considered preference and, as such, can be equated with utility. To be rational is to maximize one's utility. In order to generalise such a statement, we have to extend it to strategic situations, in interaction, when the others' behaviour are not given. Two issues are at stake. Firstly, can we generalise the concept of value described above? Can we still say in interaction that value and utility are a measure of considered and coherent preferences? And if yes which conditions are required? Secondly, can we still equate rationality with utility maximisation in strategic situation? And if yes what does it mean and what does it imply?

Strategic situations

The equation of value with utility and therefore the identification of rationality with utility maximization are rather straightforward in parametric situations, but a lot less clear in strategic situations. The prisoner dilemma[119] is probably the most famous illustration of the contradictory consequences of straightforward maximisation. Gauthier describes the traditional ideal strategic case as follows:

"everyone knows that everyone knows them. But then each person's reasoning from these data to his own expectations and choices must be accessible to every other person. In effect individual choice must emerge from a common reasoning. Each person must view strategic choice both as a response to the choices of his fellows and as being responded to by those choices... Our sole departure from the fully ideal case will be to make it subjective."[120]

Gauthier then identifies three conditions on strategically rational choice:

"A: Each person's choice must be a *rational response* to the choices she expects the others to make.

---

[119] The police catch two thieves Jane and Jerry They are suspected of two offences but there is no proof for the most serious offence of the two. The police forbids communication between them and decides to give each of them *separately* the following choice: 1) he confesses the murder: if she does not confess then he gets away with one year for his cooperation with the police and she will be locked for ten years for the serious offence; if she does confess they both go to jail for five years; 2) he does not confess: if she does, she gets away with one year and he is locked in for ten; if she does not confess, they each spend two years in jail for the minor offence. If they are pure utility maximisers, they will both choose to confess, which will obviously lead them to their worst outcome.

[120] *Morals by Agreement*, p 60-61.

B: Each person must expect every other person's choice to satisfy condition A.
C: Each person must believe her choice and expectations to be reflected in the expectations of every other person."[121]

Most of Gauthier's theory will be to explain the concept of *rational response* referred to in condition A and to find an adequate formulation of this condition. It is now time to present the theory as a whole.

[121] *Morals by Agreement*, p 61

Section 2: The theory as a whole

Core conception 1: market as a morally free zone

A conception of morality

Gauthier's assessment of the market enables us to better understand his conception of morality and the role morality should have in society. Indeed the first core conception of his theory is to describe what a morally free zone could be. He conceives morality as "an impartial constraint on the direct pursuit of individual utility"[122]. He claims that morality is not needed wherever equilibrium and optimum coincide, i.e., wherever pursuit of individual gain coincides with mutual advantage without the agents *intending* to bring about this mutual advantage or without the agent *constraining* his utility maximization. Gauthier's argument is then two fold: firstly, the perfectly competitive market would be the ideal candidate for such a morally free zone; secondly, there is no such thing as a perfectly competitive market in real life. In the light of this two-fold demonstration, we should be able to draw a fuller picture of Gauthier's conception of morality.

To understand his demonstration we have to review what he identifies as being the underlying conditions of the market:

- The agents choose in a parametric environment. In other words the others' behaviours (or utility functions) are considered as given.
- The agents choose under conditions of certainty: the utility and production functions are given and therefore consumption and production fix the price to everybody's full knowledge.
- There is private consumption of products and factors of production. This in turn implies that:

  Individual factor endowment are given and

  There is free individual market activity
- There is private ownership. With this condition, Gauthier also assumes that agents are mutually unconcerned about each other.
- There are no externalities whether positive or negative.

[122] *Morals by Agreement* p 95

First leg of the demonstration

Gauthier first endeavours to demonstrate that, should these conditions be realised, the market would be a morally free zone. He first notes that free activity under certainty guarantees an equilibrium. Indeed, if one knows what the price of the good is or what service someone sells, how much her competitors can produce of it and how much consumers wants to consume, it is a mere calculation problem to know how much she should produce if she is rational (i.e. if she is a straightforward utility maximiser). Similarly, he notes that perfect competition[123] guarantees the coincidence of the optimum with the equilibrium so obtained. No one can be made better off unless someone else was made worse off.

That being noted he moves on to ask the following question: if morality is conceived as "an impartial constraint on the direct pursuit of individual utility", then "do the market interactions exhibit any form of partiality"? Does the market affect the agents in a differential way? His demonstration rests on three points: 1) free activity ensures that no one produces too much or is limited in his production. Any corrective device would bring partiality where it is not needed; 2) "The absence of externalities ensures that no one is affected whether beneficially or harmfully by any market activity to which she has not chosen to be party".[124] Impartiality is guaranteed by the equation of income with marginal contribution; 3) the optimality of the outcome implies that any move from it would make some better off at the expense of others.

Having demonstrated that at least three conditions of the market prevent partiality from polluting its functioning, he then concludes that the market is a morally free zone. Let us stop a moment and try to understand the role of impartiality in Gauthier's conception of morality.

He seems to equate morality with impartiality: a morally free zone is a zone where there is no partiality. So we first need to clarify the relation between morality and impartiality. The following quote should help us through this task: "We ask whether the

[123] One condition of perfect competition is that no one is able to affect market prices on her own.
[124] *Morals by Agreement* p 96

operation of the market exhibits any form of partiality that would justify from a moral standpoint the constraint needed to overcome it. Since we have already defended the claim that the market interaction is rational, were we to find it to be partial, we would have established at least a prima- facie conflict between moral and rational requirements. We shall deny that there is such conflict."[125]

In this quote he really seems to use morality and impartiality interchangeably: there is no morality without impartiality and wherever there is partiality there is a moral requirement to impose constraints in order to re-establish impartiality. So what is impartiality?

Let us come back to his demonstration to draw a picture of his conception of impartiality: 1) Free activity means that each individual is free to produce or consume as if she was a Robinson Crusoe on her island. There is no attention paid to any special capacities or circumstances. 2) The absence of externalities means that incomes are equal to respective marginal contributions, i.e. distribution depends solely on contribution and on no other particular factors. 3) Optimality means that no individuals are favoured to the detriment of others.

The concept of impartiality is still unclear. However, in these three conditions, *each individual is assumed to be rational* i.e. is *responsible for* looking after his own best interest. Impartiality only means that the best interest of some is not valued more than the best interest of others. Rationality and impartiality must cohabit. The three conditions above guarantee that, in market interactions, individual rationality (i.e. straightforward utility maximization) is compatible with impartiality.

The concept of morality is now clearer. Morality comes in the picture wherever straightforward utility maximisation impedes on impartiality i.e. wherever there is simultaneously individual rationality and partiality in interactions. There is no need for morality where impartiality prevails. Where there is no impartiality, morality comes in order to re-establish it. Morality is the artificial establishment of impartiality. This form of impartiality is artificial because it is not naturally compatible with individual straightforward utility maximization and because it derives from constraints on

[125] *Morals by Agreement* p 95

individuals' behaviour. In order to have become aware of the need for constraints and to discover what these constraints should be, individuals must have undertaken a change, a rational change. Could this change be from the natural rationality known as straightforward utility maximization to an 'artificial' form of rationality? We need to turn now to the rest of his demonstration to answer this question.

Second leg of the demonstration

In the second fold of his demonstration, Gauthier assesses the realism of each market condition listed above. He first notes that the market is unlikely to ever exemplify parametric situations under certainty. However, his most urgent concerns are about the conditions of private ownership and private consumption.

Firstly, it seems rather 'unfair' to consider the individual factor endowments as given. Indeed, the share of the outcome that each individual receives depends on his contribution but, in turn, his contribution depends on the factors he possesses and therefore on his endowment. To consider this endowment as given is therefore in itself a distortion.

Secondly, to assume that there is free individual activity is arbitrary for at least three reasons: 1) the overall productive capacity of a society is limited by the capacities of its members; 2) substitutability of capacities between agents can be a restriction; 3) some capacities do not necessarily meet a demand.

Last but not least, the existence of externalities upsets the matching between supply and demand. To assume that the consumption of one good by an individual precludes its consumption by another amounts to disregarding the existence of free goods (like air and water) and of public goods (like roads or street lights). A free rider is an individual who benefits from a good without contributing to its cost. A parasite is an individual who displaces the cost of his use of a good on to others. Gauthier's theory is also built up around the need to agree on how to share the costs and benefits of public goods. The problem posed by the existence of externalities is one of his central concerns.[126]

[126] This concern echoes Buchanan's. See appendix 1 Section 3b.

What about the condition of mutual unconcern? Are we really the monstrous and selfish individuals depicted by the advocates of *laissez-faire*? Gauthier's argument is that our concern for others is necessarily restricted to the people close to us. Social interactions cannot depend on such a restricted field of bonds. The assumption of mutual unconcern is therefore not only realistic but also necessary. This assumption also suits Gauthier's theory perfectly because mutual unconcern is a condition for impartiality and because it does not preclude particular relationships based on trust and mutual advantage to flourish on a voluntary basis.

Prior to moving on to his next core conception, we need to complete our picture of Gauthier's conception of morality. We saw above that morality was a form of artificial impartiality and we wondered if it was based on a change of rationality. In finding the root of morality, we should address this last issue.

Let us start with this first quote:

"morality has no application to market interaction under the conditions of perfect competition. Choice is neither morally right nor wrong, because the coincidence of utility maximization in free interaction removes both need and rationale for the constraints that morality provides ...Moral constraints arise only in the gap created by the conflict between the two rationality properties, when mutual benefit is not assured by the pursuit of individual gain. We assess outcomes as right or wrong when, but only when, maximizing one's utility given the actions of the others would fail to maximize it given the utilities of others."[127]

Here is our answer: the rationality of morality is based on 'mutual advantage' rather than on pursuit of individual gain. Partiality occurs when there is conflict between the rationality required to pursue one's individual gain and the one required to bring about mutual advantage. The reconciliation between these two types of rationalities then requires impartial constraints on individuals' utility maximization. Morality arises from the emergence of partiality and from the conflict between these two forms of rationalities.

What about justice? When introducing his conception of the market as a morally free zone, Gauthier refers to Smith's 'invisible hand'[128] and he quotes Smith: "a man is free as long as he does not violate the laws of justice". Gauthier then continues:

---

[127] *Morals by Agreement* p 93
[128] A. Smith *The wealth of Nations*

"But justice as the reader may recall, is conspicuously absent from Thomas Hobbes's account of the natural condition of mankind.... The absence of force and fraud is essential to the workings of the market. Before Smith's invisible hand can do its work, Hobbes's war of every man against every man must first be exorcised. And this, as we shall see means that the ideal of free interactions which Smith celebrates is not natural but artificial, arising, for rational persons, only within a framework of agreed constraints. In understanding the perfect market as a morally free zone we shall be led back to its underlying, antecedent morality."[129]

Morality, as constraints on behaviour, *applies to choices, actions and persons*[130]. Morality arises with a change in individuals' disposition. Individuals must acquire an *artificial virtue* that enables them to make the right choice and perform the right action to bring about the right outcome. *Justice* emerges when agents have acquired this artificial virtue. Justice and morality arise hand in hand as individuals reconcile the incompatible rationalities intrinsic to the state of nature. Justice and morality appear when individuals *decide* to agree. In other words, justice emerges when individuals *realise* that two rationales are in a *disadvantageous* competition and there is a need to reconcile them. As soon as individuals understand that, in their own best interest, they need to impose upon themselves some constraints in order to rectify the disadvantage caused by the pursuit of individual gain, they simultaneously acquire a disposition to abide by such constraints; they consciously decide on a change of rational approach or rather of rational *disposition*. We saw in the previous chapter that such a change of rationality was perfectly reasonable and acceptable in Gauthier's world. We see know how it is a necessary condition to cooperation.

The circumstances of justice

What Gauthier calls the circumstances of justice are "those features of the human situation that give rise to co-operation".[131] Co-operation arises from our '*awareness*' of those features. So what are they?

[129] *Morals by Agreement* p 85. As we shall see in Chapter III, the last sentence of this quote has been rather confusing to most commentators.
[130] This interpretation is confirmed by the following quote amongst others : 'Morality is concerned with actors, persons considered as doing and choosing, and as implicated in the consequences of their deeds and choices.' *Morals by Agreement*, p 235
[131] *Morals By Agreement*, p 116

• The first circumstance of justice is the awareness of self-bias in our character. Mutual unconcern is an extreme version of this condition. We cannot rely on any *natural* fellow feelings. Voluntary co-operation brings about another form of concern for each other

• The second condition is the presence of externalities. Because of externalities, the market equilibrium fails to be an optimum. The existence of free and public goods combined with utility maximization brings about a situation as described in the prisoners' dilemma and also the existence of free riders and parasites.

• The third condition is the awareness of scarcity of goods. On this point Gauthier differs from Hobbes. In Hobbes's world, goods are scarce but available in *fixed* quantities, hence the state of war intrinsic to the state of nature. For Hobbes, we are aware of each other as *competitors,* and we agree to cooperate in order to avoid destructive conflict. Gauthier prefers Hume's vision of the state of nature where goods are available in *variable* quantities depending on the production. Therefore, in Hume's world, we are aware of each other as *co-operators* in productive activity i.e. we see each other as 'potential sources of mutual benefit'.[132] Because we become aware of each other as co-operators, we are led to agree, not to avoid conflict, but to seek mutual advantage.

But Gauthier insists: the most important features are the awareness of externalities in our environment and awareness of self-bias in our character. Why? The answer is now obvious: the presence of externalities highlights the disadvantageous competition of the two forms of rationalities mentioned above between the pursuit of self-interest and the pursuit of mutual advantage. Because the natural tendency to be self-biased dominates our nature, it is predictable that we would be free riders and parasites in natural interactions. Such generalised behaviour would preclude the emergence of indispensable public goods. In order to be able to create such goods and benefit from them, individuals have to cooperate. Co-operation is then introduced as the rational response to market failure. It is interesting to note that in the rest of the book Gauthier systematically opposes 'natural and market' to 'co-operative' interactions. Natural and market are always used together. Co-operation is then an 'artificial rectification' of these 'natural' conditions.

---

[132] *Morals By Agreement*, p 115

Core conception 2: the principle of minimax relative concession

Internal rationality and choice of a joint strategy

We saw above that the change from natural and market conditions to co-operative interactions, the change from natural to artificial impartiality comes with a change of rationality: the original one is based on pursuit of individual gain whereas the social one is based on the pursuit of mutual advantage.

Gauthier now introduces a new distinction between external and internal rationality. *Internal rationality* concerns the rationality of agreement, the rationality of choosing a joint strategy. The problem of making a rational cooperative choice is solved by finding a principle that rationalizes the agreement "the way the principle of expected utility-maximisation rationalizes individual choice"[133]. On the other hand, *external rationality* no longer concerns the choice of a joint strategy but the rationality of keeping an agreement. External rationality is therefore concerned with the initial bargaining position that is fixed in such a way that it is rational to enter the agreement, to agree on a joint strategy and to act on the agreement ex-post. It is also concerned with the rationality of keeping the agreement once in co-operative interactions.

How are we to understand this distinction? Are we talking about two different forms of rationality? Are we talking about the rationality of a process or of individuals? We need first to establish that the concept of rationality applies to individuals. Individuals are rational or irrational. To say that a process or a strategy is rational is to say that a rational individual assessing it would find it acceptable from her perspective. Rationality is always a human feature. Now the distinction that Gauthier establishes seems to concern what is to be assessed. In the case of internal rationality what is assessed is the rationality of the choice made by parties to bargaining whereas in the case of external rationality what is assessed is the rationality of entering and keeping such a choice.

[133] *Morals By Agreement*, p 118

The most logical question that comes to mind is then: why do we need to distinguish the two? If the assessments are performed by the same individuals why do we need to distinguish the rationality according to what is to be assessed? Are the internal and external rationalities different and if yes what is different? Do the individuals doing the assessing undertake the change of rationality mentioned in core conception 1? If yes then we have a chronological problem. Indeed, the *external* rationality applies to the initial bargaining position, then the *internal* rationality applies to making the agreement and then the *external* rationality again applies to keeping the agreement. These rationality swings would be rather difficult to fit with reality. We need to follow Gauthier's steps before we can address these issues and answer these questions.

We saw in core conception 1 that in natural and market conditions, the existence of externalities leads the agent to want co-operation. "In order to take effective account of externalities, each person must choose her strategy to bring about a particular outcome determined by prior agreement. This agreement, if rational, will ensure optimality". Gauthier insists, this agreement might be implicit but it is not a mere fiction since "it gives rise to a new mode of interaction, which we identify with co-operation."[134] Through co-operation the agents can generate a co-operative surplus that would not be available otherwise. What individuals bargain over is the distribution of costs and benefits of this future co-operative surplus. The co-operative surplus represents the difference between the outcome of co-operation and the outcome of the initial bargaining position. The procedure of bargaining is then divided into two stages. In stage one, each party makes a claim and in stage two, some parties offer a concession by making a lower claim. The process continues until the parties agree on a mutually compatible outcome.

In this sub-section, we will develop core conception 2 in several stages. First we will dwell on Gauthier's conception of (internal) rationality in bargaining. Second, we will develop the minimax principle. Third we will try to understand his reconciliation of rationality and morality (or impartiality) prior to concluding on his approach about justice and bargaining. In this last part we will see how Gauthier addressed the issues raised by Barry against justice by mutual advantage.

[134] *Morals By Agreement*, p 117

The internal rationality of co-operation

Gauthier is extremely careful to set the boundaries of his bargainers' rationality: They are straightforward maximisers i.e. they try to maximise their utility given their expectations of the others' behaviour. However, his bargainers are straightforward maximisers in the context of co-operation, therefore their rationality is more sophisticated than the standard one. It is important that Gauthier derives the principle of co-operation from straightforward maximisation. Only then can he demonstrate that the principles so arrived at are also moral. If he had to rely on any other form of rationality to derive the co-operative principles, it would weaken his argument: morality would no longer be a by-product of practical rationality alone.

Gauthier combines three assumptions to develop his sophisticated conception of rationality.

- First Assumption (equal rationality): as we saw before the bargainers are rational and they assume the other parties to be likewise. This condition strongly echoes Harsanyi's principle of mutual expected rationality[135]. But where Harsanyi used it to develop on the similarity of subjective probabilities, Gauthier develops it to assume similarity and reciprocity of reasoning. No party would expect the others to accept what she would not accept herself.

- Second assumption (will to cooperate): bargainers are aware that co-operation allows them access to a co-operative surplus that would not be available otherwise. The parties are interested in co-operation in order to benefit from a surplus that would not exist without it. They are therefore interested in co-operation as an integral part of their best interest. Gauthier subtly substitutes the pursuit of individual gain with the pursuit of mutual advantage. The two goals become one and same in Gauthier's approach.

- Third assumption (full information): as mentioned in section 1, the parties are fully informed and they know the true preferences and the utilities of the others. There is

[135] See the Appendix section 2 b

therefore no room for bluffing and other distortions of preferences to gain bargaining power.

So now how does Gauthier combine these three assumptions to improve on the traditional conception of straightforward maximisation? Easily: co-operation being in our best interest, it is rational to pursue it. Since we assume the others to be similarly rational, we also each assume that the others pursue co-operation the way we do. When defending the conditions on rational bargaining, Gauthier argues:

"The rationale for these conditions turns on the benefit each person seeks to realise from the cooperative surplus. Each can increase his utility by cooperating; hence as a utility-maximiser each must find it rational to cooperate. And each recognises that everyone else must find it rational to co-operate."[136]

Gauthier seems to take this rationale as obvious. He therefore considers as *irrational* to do any of the following.

1. *To put co-operation in jeopardy* Deadlocks are counterproductive since they stop co-operation. Each party is discouraged from abusive claims. Anyway, from the first condition of 'symmetrical rationality', it is irrational to expect from the other parties concessions bigger than the ones we would be prepared to accept in their positions.

2. *To be excluded from co-operation.* Once again no co-operation means no share of the surplus. As rational individuals, we would not accept having to give a share of the surplus to someone who has not contributed to it. Similarly, we can not expect others to give us a share of the surplus when we have not contributed to it.

3. *To drive away others from the bargaining table.* Co-operation, by definition being a mutual project, is dependent on the contribution of all. It is not in our interest to discourage other parties from participating in it.

4. *To make threats.* A threat is counterproductive and therefore it would not be rational to carry it out. Since the others are rational and fully informed, they know that to carry out a threat is irrational and therefore, they would not believe us were we to make one. Since it is irrational to carry out a threat, it is irrational to make one.

[136] *Morals by Agreement*, p 143.

A 'rational' claim is therefore such that:

- Each party must claim the co-operative surplus that affords him maximum utility;
- nobody can claim a share of the surplus that he has not contributed to create,
- nobody can claim a share bigger than that it is possible for him to receive

The 'sophisticated' rationality of the bargainers is somewhat refrained not to say 'constrained'. They are unusually reasonable for straightforward maximisers. I intend to come back to this conception of rationality in chapter IV but we shall make a few comments here.

Firstly the second assumption is rather controversial. Even in the bargaining position, there are situations where the pursuit of best interest does clash with the pursuit of mutual advantage. This simplification overlooks some complex issues. This second condition also seems to rely on another subtle assumption: that all parties *want and need* co-operation *equally* as much. However, some could have more to lose from co-operation than others, despite the existence of a co-operative surplus. The minimax principle does not necessarily take into account such discrepancies between the parties.

Secondly, Gauthier assumes that the parties are fully informed and therefore that no party can bluff or distort her preferences in order to improve her position in bargaining. This assumption is not only very strong but it has also the disadvantage of undermining the strategic aspect of bargaining. The same applies to the possibility of making threats or of benefiting from a threat advantage. Besides, Gauthier has not proven that if it is irrational to carry out a threat, it follows that it is irrational to make a threat.

But let us continue. Which conditions of co-operative choice would these rational bargainers consider as acceptable?

- The first condition concerns the initial bargaining position. We know that the discussion about the initial bargaining position is postponed to much later. For the moment, we must only keep in mind that it cannot be identified with the non-co-operative outcome since such an outcome could have been obtained through force, fraud or previous free-riding and parasitism. The initial bargaining position is such that it is rational to enter in co-operation and to agree on a joint strategy. The set of utilities that constitutes the

initial bargaining position are taken as the initial endowment of the prospective co-operators.

- As we saw before, the second condition is that the object of rational co-operative choice must be an optimal outcome. If the bargainers choose independently as in natural and market conditions, they will choose a strategy according to what they expect the others to do. The outcome is then in equilibrium. If the bargainers have agreed on a common strategy, then they will choose within the strategy agreed the outcome that will bring them the best utility. The outcome is an optimum. Gauthier concludes that in co-operation, the core rationality property is optimality.

- The third condition of rational bargaining concerns the unicity of the rational optimal and the active involvement of the actors in selecting this unique outcome. Gauthier rejects Arrow's concept of social choice on the basis that nothing about Arrow's framework secures the active involvement of individuals.

- In the same vein, Gauthier rejects the utilitarian approach as developed by Harsanyi. Not only does welfare maximisation rely on interpersonal comparisons but it also disregards the structure of interactions and gives absolutely no say to co-operators who are mere passive recipients of goods. The active involvement of the bargainers is an indispensable requirement of a bargaining theory if we want the agreement on a joint strategy to be voluntary. Individuals cannot be bound by an agreement not voluntarily contracted.

The principle of co-operation

The object of rational co-operation must be an *optimal* outcome not an equilibrium. The co-operators agree on a joint strategy chosen through bargaining. Such a procedure secures the involvement of the co-operators and therefore their voluntary acceptance. It goes without saying that the initial endowments of the parties are not at stake and that no bargainer can be left at the end of bargaining with less than he had to start with. Bargaining is about the distribution of costs and benefits of the surplus generated through co-operation. It is a two stage process: first, the parties make a claim. Then in

the second stage they offer a concession by withdrawing a portion of their original claim in order to reach a concession point such that all claims are compatible and the outcome is mutually advantageous. Bearing in mind the above conditions of rationality, we need now to know which concessions a rational individual would make.

Let u* be the utility afforded by an agent in the initial bargaining position, u# be the utility of his claim and u the utility obtained with the outcome of a joint strategy. The *absolute magnitude* of his concession is the difference between the utility of his claim and the utility the party obtains at the end of bargaining (u# - u). Such a measure does not afford any ground for comparing the range of concessions between parties. Gauthier therefore suggests reliance on a relative measurement of concession. In order to obtain it, he first introduces the absolute magnitude of a *complete concession* i.e. the difference between the utility of his claim and the utility of his original position (u# - u*). The *relative magnitude of his concession* is then the proportion that his absolute magnitude bears to a complete concession [(u# - u) / (u# -u*)]. The relative concession of no concession is zero, the relative concession of a full concession is one.

The relative magnitude of a concession is a very useful tool since it does not rely on any interpersonal comparisons and it is independent of the choice of the utility scale. Each party, being fully informed, is able to compute the relative magnitude of the others' concessions and to compare them to hers.

The Zeuthen principle[137] then provides a rule to decide which concession is rational and therefore which outcome is acceptable. According to this principle, the person with a lesser relative concession must concede. Applied by Gauthier to bargaining, it becomes the *minimax relative concession*: "given a range of outcomes, each of which requires concessions by some or all persons if it is to be selected, then an outcome be selected only if the greatest or *maximum* relative concession it requires, is as small as possible, or a *minimum*, that is, is no greater than the maximum relative concession required by every other outcome."[138] Each bargainer is willing to entertain a feasible concession

[137] See the appendix section 2b
[138] *Morals by Agreement* p 137

point[139] (i.e. he is willing to make the concession it requires from her) as long as the others are willing to make the concessions it requires from them.

According to Gauthier, the minimax relative concession has the following merits.

• It avoids the controversial debate on interpersonal comparisons of utilities since it allows interpersonal comparison of the proportion of each person's potential gain that he must concede instead.

• Gauthier specifies that we could mistakenly assume from the description of the minimax that it requires equal relative concession from the bargainers. It is not always the case. However, what is true is the fact that if there is an optimum requiring an equal concession from each bargainer, then the minimax would select it.

Gauthier is now able to spell out the *Principle of Minimax Relative Concession (the Principle)* that provides the conditions on rational bargaining:

(1) Each party must make a rational claim
(2) Given rational claims, each party must suppose that there is a feasible concession point that every rational person is willing to entertain.
(3) Each party is willing to make a concession in relation to a concession point as long as its relative magnitude is no greater than what he supposes any rational person would be willing to make.
(4) No person is willing to make a concession in relation to a concession point if he is not required to do so by conditions (2) and (3).[140]

The first and third conditions of the Principle are respectively a consequence and a corollary of the assumption of equal rationality and the second assumption for internal rationality given above. According to Gauthier, the second condition follows from the willingness of the parties to bring about co-operation. The fourth condition is common sense: nobody would make an unduly big concession. It is against any form of rationality.

---

[139] A feasible concession point is a concession point in the outcome space. The outcome space is a closed convex figure in the utility space.
[140] *Morals By Agreement*, p 143

Gauthier can then advance the following claims:

• Conditions (2) and (3) suffice to bring about an agreement on a feasible point. His demonstration is based primarily on condition (2): as long as the bargainers *suppose* that there is a feasible concession point, they are willing to reach it since they want co-operation above all.

• The principle of minimax relative concession expresses the principle of expected utility maximisation in the context of bargaining. Expected utility maximisation rationalises individual choice, the Principle rationalises mutual choice.

• The principle of minimax relative concession determines the formal content of a rational bargain. It gives to each actor a strategy to choose the best utility from in co-operative interactions.

Gauthier seems to assume that there exists a solution to the bargaining problem in each bargaining situation. The rational behind this assumption is again that no co-operation means no surplus; people are prepared to entertain rational concessions in order to benefit from a share of the surplus. If I can follow the rational of this explanation, I am not convinced that Gauthier has proven either the existence of a solution or its unicity. Gauthier's demonstration of the existence of an agreement is largely dependent on the second condition on bargaining (the object of rational cooperative choice is an optimum) which itself follows from the first one on rational claims (each party must claim the cooperative surplus that afford her maximum utility) which itself relies on the second assumption of bargaining (will of the bargainers to cooperate). The existence of an agreement therefore depends on the sophisticated rationality of the bargainers carefully bounded by Gauthier. Again, I believe it worth coming back on this conception of rationality later in chapter IV.

When rationality meets impartiality

Prior to developing his argument for the maximin relative benefit, some comments are required:

In most of Gauthier's writings up to 1979, the making of an agreement was based on the maximin relative benefit principle, in reaction to Rawls's lexical principle. He first developed the moral principle that he then tried to reconcile with rationality. As we shall see later, this approach will support his idea of Archimedean point. In *Morals by Agreement,* Gauthier changes strategy. He develops his bargaining theory, basing it on the Zeuthen principle, and obtains the minimax relative concession principle. He illustrates the rationality of the bargaining through the minimax relative concession principle and only then reconciles this *rational* approach with the *impartial* approach that leads to the maximin relative benefit principle. I believe that this change of strategy is meant to substantiate his theory, grounding it firmly in a theory of rationality.

As a corollary to the above comment, we know that Gauthier had his own view on fairness when developing his theory. Before presenting the maximin relative benefit he wrote that co-operation, to avoid the immoral effects of externalities, "must ensure that the ratio between the benefit the co-operator receives and the contribution she makes is, so far as possible, constant, the same for all."[141] He knows what his principle is aiming at and it is very much in line with his original project as described in chapter I: "we may suppose that the basic ethical requirement is to ensure that utilities are equitably distributed among all individuals."[142] However, although he knows what he is aiming at, *he is able to base his principle on the concept of rationality alone, without any pre-reference to its fairness.*

Gauthier specifies about impartiality: "we shall address impartiality, as we have addressed rationality, from the standpoint of the individual actually involved in bargaining."[143]The joint strategy selected will be impartial if it is acceptable from every standpoint by every persons involved. One could say that this impartiality of the process is a mere consequence of the conditions of rationality imposed on the process. Indeed

[141] *Morals By Agreement,* p 152
[142] 'Critical Notice On Harsanyi', p 705
[143] *Morals by Agreement,* p 151

we remember that one of these conditions was the active involvement of the bargainers as an indispensable requirement if the agreement on a joint strategy was to be voluntary. Gauthier's bargaining theory is built around this central idea. Knowing that the bargainers are rational, it comes as no surprise that they find acceptable whatever they have voluntarily agreed to. We shall see later how he reconciles this form of impartiality with the impartiality of a hypothetical arbitrator exterior to the bargaining process, who assesses the outcome from his standpoint.

Gauthier distinguishes two types of goods.

1) In the case of single transferable good produced in fixed quantity and divisible among the co-operators, he demonstrates that the maximin relative concession principle leads to an equal distribution of the surplus. The bargainers contribute proportionally to their means but they benefit evenly from the surplus. However, such a distribution applies only if the contribution of each co-operator is *indispensable* to the production of the surplus. If a co-operator can be replaced by any (or at least one) other than this distribution does not apply.

We touch here on a very important point that was already mentioned before: Gauthier seems to assume that each bargainer is needed by and needs the others *equally*. Co-operation as conceived by Gauthier is at stake in this implicit assumption for at least two reasons: firstly, this assumption supports the second assumption. For co-operation to become a priority as integral part of each party's best interest, it has to be needed and wanted by all the parties roughly equally. Secondly, it also explains the uselessness of strategic games in bargaining. Nobody needs to strengthen her bargaining power by bluff or threats since everybody knows that everybody needs all the others as partners and wants co-operation as much as themselves. The problem is that this tacit assumption is rather difficult to accommodate with real case situations.

2) If co-operation does not result in a fixed quantity of a fully transferable good, then the share of the surplus each party can get is measured in terms of benefit. Gauthier specifies: "if a fair or impartial distribution of the co-operative surplus relates the

benefit each person receives to the contribution he makes, each person's fair share of the surplus is determined by making shares proportional to claim."[144]

Let once again u* be the utility afforded by an agent in the initial bargaining position, u# be the utility of his claim and u the utility obtained with the outcome of a joint strategy. Then his claim on the co-operative surplus is (u# - u*) and his share of the cooperative surplus is (u -u*). His relative benefit is then the proportion [(u-u*) / (u# - u*)]. This relative benefit of the initial bargaining position is 0. The relative benefit in situation where his claim is satisfied is 1. For the shares to be proportional to claim, each must receive the maximum of the minimum *(maximin) relative benefit.* We fall back on a rather egalitarian formula completely in line with his conception of fairness which is to ensure that utilities are equitably distributed among all individuals.

Whatever is conceded takes away from the benefit enjoyed hence the unsurprising result: relative benefit and relative concession sum to unit [(u# - u) / (u# -u*)] + [(u-u*) / (u# -u*)] = 1. In other words, bargaining is about finding the right balance between concessions and benefit such that co-operation is advantageous to all.

There have been extensive commentaries on Gauthier's minimax and maximin principles and we will come back to his critics' comments in chapter III. At this stage, we just want to review how he has addressed Barry's criticisms of justice as mutual advantage.

We wondered if what was assessed was the fairness of the process or the fairness of the outcome independently from the process. Gauthier's answer is very clear: "given the original position, co-operation is just if the joint strategy on which it is based is the outcome of a fair bargain among the co-operators. But the fairness of the bargaining process does not correct any partiality that may be present in the initial position; indeed it would simply transmit the partiality from the initial position to the joint strategy selected."[145] So Gauthier is clear, the fairness of the process is an integral part of the fairness of the outcome. But Gauthier goes back further than the other bargaining theorists since the fairness of the process also applies to the initial bargaining position.

[144] *Morals by Agreement*, p 154
[145] *Morals by Agreement*, p 151

If the initial position is 'unfair' i.e. obtained through fraud or force or if it is spoilt by parasitism or free riding, the outcome of the bargaining process will be unfair or impartial even if the process itself is fair.

The parties do not play any strategic games and do not try to develop their bargaining powers. Therefore the intuitive 'unfairness' of the Nash's solution is replaced by a rather egalitarian solution: Gauthier's minimax allows Matthew to play only 56% of the time against the 93% suggested by Nash formula[146].

The very sophisticated rationality that Gauthier attributes to his bargainers allows him to label deadlocks or threats as irrational. Any rational individual would see the individual gain he could derive from a share of the co-operative surplus and nobody would be foolish enough to play strategic games that would put such an advantage at risk. We have already commented on this conception of rationality and we shall come back to it later in chapter IV.

One of Barry's argument against justice as mutual advantage was that arbitration would simulate the outcome of bargaining which somehow defeats the purpose of having an arbitrator. But arbitration is needed when there is danger of deadlocks. Gauthier's bargainers are so 'rational' that they don't need an arbitrator. They don't bluff, they don't lie, they don't threaten, they are perfectly informed, they suppose there is a possible agreement and they aim at it.

Now that parties to bargaining have rationally agreed on a joint strategy based on straightforward utility maximisation, Gauthier turns to what he calls the *external rationality* of co-operation i.e. *the rationality of keeping the agreement.* He tackles this issue from two angles: first he explains on which rational basis, parties can ground their compliance (constrained maximisation); second, he comes back to the initial bargaining position (Lockean proviso) considering that if this starting point is unfair, the full process is unfair and therefore rationally unstable.

[146] See Barry's *Theories of Justice*, chapters 1 and 2

Core conception 3: Constrained maximisation
External rationality and keeping the agreement

Gauthier's answer to the Fool and the solution to the problem of extending the concept of value developed in parametric situations to strategic situations are contained in core conception 3, that of constrained maximisation.

Parties have agreed on a principle of mutually advantageous co-operation that should rectify the distortions caused by externalities in natural and market conditions. The problem is now to ensure compliance with the agreement rationally contracted. The conception of constrained maximisation rests on two complementary concepts: a disposition to comply and translucency. One concept is of no use without the other. Indeed, on the one hand, by acquiring a disposition to comply with the agreement, a party is accepted into the co-operative group and is allowed to benefit from the co-operative surplus. A party that is so disposed is called a constrained maximiser (CM). On the other hand, a party who is disposed to cheat will be excluded from co-operation and will not be allowed to benefit from it. Such a party is called a straightforward maximiser (SM). However a SM is excluded from co-operation *only if* the others can spot her as such. If a SM is wrongly taken for a CM and accepted in the group, then there is exploitation. That is where translucency comes in the picture: the parties develop a skill to identify correctly each others as CM or SM. The development of translucency amongst the parties also acts as a motivation to acquire the disposition to comply. If I know that people are likely to 'read' my intentions correctly than I am a lot more inclined to acquire the most beneficial intention i.e. the intention to comply with the co-operative agreement. It then becomes rational to acquire the disposition to comply. Constrained maximisation is therefore the type of rationality suited to co-operation and to strategic situations. Constrained maximisation is the external rationality needed to keep the agreement.

A disposition to comply.

In order to fully appreciate Gauthier's achievement, we need to come back to the Fool's challenge. We remember from chapter I that the Fool never contested that it was rational to enter an agreement. His argument was rather that, once the agreement

contracted, it could be more beneficial and therefore rational not to keep it when the others were keeping it. Or said differently again, the Fool compares the utility he would gain were he to comply to his utility were he to unilaterally defect; if the latter is bigger, he considers it worthwhile not to comply.

Hobbes' reply to the Fool was as follows: when disposing himself to cheat, the Fool takes the risk of being discovered and of being excluded from the society of men. A Fool has more to lose from being excluded than from complying:

"A person disposed to violate his covenants cannot be admitted as a party to co-operative arrangements by those who are both rational and aware of his disposition, and such a person cannot rationally expect to reap the benefits available to co-operators ... The disposition to decide whether or not to adhere to one's covenants by appealing to directly utility maximising considerations, is itself disadvantageous, if known, or sufficiently suspected, because it excludes one from participating, with those who suspect one's disposition, in those co-operative arrangements in which the benefits to be realised require each to forego utility maximisation ... This will prove to be the key to our demonstration that a fully rational utility-maximiser disposes himself to compliance with his rationally undertaken covenants."[147]

The most immediate reading of this quote is as follows: the disposition not to comply "is itself disadvantageous *if known or sufficiently suspected*". It is in someone's best interest to dispose herself to comply *if* her disposition not to comply can be spotted by the other parties to the contract. To be spotted as a straightforward utility maximiser in a society of co-operators entails exclusion and is therefore detrimental to the one so disposed. It becomes rational to dispose oneself to comply. Compliance to a covenant must be accompanied by a change of rationality such that reason is no longer connected to direct benefit in performance but rather to the benefit in the disposition to perform.

According to Gauthier, Hobbes distinguishes the two forms of rationalities (*natural reason* versus *right reason*) but considers men incapable of internalising this change. Hobbes' individuals remain straightforward maximisers and the change of rationality has to be forced upon them by the sovereign through punishment and prophylaxis. The answer to the Fool is political rather than moral.

Gauthier on the other hand claims that it is possible to give a moral response to the Fool, i.e. that it is possible for each individual to internalise the required change of rationality

[147] *Morals by Agreement*, p 162

and understand that it is in her best interest to constrain her maximisation by disposing herself to comply. But once again the disposition to comply becomes advantageous and therefore rational only if '*known and sufficiently suspected*'. The issue at stake is central to Gauthier's claim: would Gauthier's individuals still acquire the disposition to comply were their disposition to comply or not to comply be impossible or difficult to identify? Said otherwise, is their change of rationality coming from within themselves or is it the fruit of an external motivation? Let him develop his claim.

We saw in core conception 2, that parties to bargaining would base their agreement on the principle of minimax relative concession. They agree on a joint strategy that if followed by all should bring about a mutually advantageous state of affair. A constrained maximiser maximises his utility within the joint strategy agreed upon, i.e. he maximises his utility given the utility afforded to the others, whereas a straightforward maximiser maximises his utility given the others' strategies. A constrained maximiser bases her actions on a joint strategy, a straightforward maximiser employs an individual strategy.

Gauthier is fully aware of the fact that there will always be some straightforward maximisers hidden amongst the co-operators and that they will make the outcome less optimal than what it should have been had everybody been constrained maximisers. The outcome in Gauthier's scenario of co-operation is therefore *nearly* fair and optimal. Once we accept that not all parties will keep the agreement a new piece of information enters the decision process of the potential co-operators: their expectation of the others' compliance. A potential co-operator has to calculate her utility given her estimate of the degree of co-operation of the others. To each estimate corresponds a joint strategy. The actual outcome of everyone's actions does not have to be identical to the outcome of the ideal joint strategy were everyone to comply, as long as it does not fall too far short of the latter. Said otherwise, a constrained maximiser is not disposed to co-operate at any price, she must have grounds to believe that she is amongst like-disposed persons before she actually constrains her behaviour. To be *rational*, a conditional disposition to comply must fulfil some criteria.

Gauthier characterises the constrained maximiser as follows. She is someone who **is** conditionally disposed to base her actions on a joint strategy or practice **and** who acts

on this conditional disposition. The only condition for this disposition to develop is that the utility she expects were everyone to base their action on the joint strategy be no less than what she would expect were everyone to employ individual strategies. In other words, the utility she expects must approach what she would expect from the co-operative outcome determined by minimax relative concession.[148]

It is interesting to note that the comparison point is universal non-co-operation. The threshold seems to be rather low and therefore constrained maximisation is likely to be recommended in most cases. However, there are cases where such a threshold might not work. Let us illustrate our claim with an example.

As I write these pages, the Olympic Games are taking place in Athens. A lot of athletes are facing a challenging dilemma: should they take steroids to improve their performance or not? Let us imagine that all the athletes and their coach sit at a table and agree that nobody would take steroids for the race. If, back in the changing room, one athlete suspects competitor A will defect, what will he do? If he feels that he had a good chance of coming first, he can now expect to be only second. If he felt he could have come second, he can now only dream of a bronze medal. Worse, if he thought he had a fair chance of coming third, he now has no hope of getting on the podium at all. Besides, if he suspects competitor A will defect, he will assume that he is not the only one to have doubts about him. If each of the other parties to the agreement has the same doubts, he can easily imagine that the doubt will trigger their temptation to defect too. How many will resist? When you prepare for four years for one race, you want to maximise your chance of wining whatever you feel you deserve. In this case, the suspicion of one defection is rationally sufficient to justify straightforward rather than constrained maximisation. Gauthier has not yet addressed the Fool's challenge.

It can be, on occasion, disadvantageous to constrain one's maximising behaviour if one has wrongly estimated others' compliance level. Therefore, the rationality of compliance cannot rest on this leg alone as we could have been led to believe from Gauthier's characterisation of a constrained maximiser given above. We can only say that having a reasonable expectation of others' disposition to comply is a necessary

[148] *Morals by Agreement*, p 167.

condition of rationality but it is not sufficient. Gauthier now needs to introduce the moral dimension of compliance.

One has to maintain one's disposition to comply even when it might not be advantageous. Constrained maximisation is not about gaining the trust of others by regular compliance in order to benefit in the future, it is about developing a disposition to comply. Only those who are so disposed can be accepted as parties to a co-operative agreement and benefit from mutual advantage. A constrained maximiser disregards the individual strategies available to him because he cultivates within himself a sense of justice when amongst like-minded people. His disposition to comply is motivated by his conviction that co-operation is mutually beneficial. He assumes that the others being equally rational have acquired a similar disposition. *This disposition has to pre-exist interactions and must constrain his behaviour.* Those who have acquired the conviction that co-operation is more beneficial, have in fact *internalised the change of rationality* from the pursuit of individual gain to the pursuit of mutual advantage. They are prepared to constrain their behaviour on this basis alone.

However, even if the disposition to comply must pre-exist interactions, the rationality of keeping this disposition depends on the compliance of others. We fall back on our initial condition: a disposition to comply is rational *only if known or sufficiently suspected* by others. The rationality of disposing oneself to comply depends on this second condition: the 'disposition spotting skill'. That is what Gauthier calls 'translucency'.

Before we move on to presenting Gauthier's concept of translucency, we note that there is a full aspect of the theory which does not seem to have been addressed. Gauthier's argument is that, when spotted, straightforward maximisers are excluded. When reasoning about which disposition to choose, a person argues as follows: "suppose I adopt straightforward maximisation. Then I must expect others to employ maximizing individual strategies in interacting with me"[149]. The question is really how does one use at the same time, individual strategies when interacting with straightforward maximisers, and a joint strategy when interacting with constrained maximisers? In the above example, if an athlete suspects competitor A of defecting and wants to employ an individual strategy in interacting with him, as a result he also uses an individual strategy

[149] *Morals by Agreement*, p 172

in all his interactions including interactions with potential co-operators. It is either he takes steroids or he does not. (Or in another example, it could be: either I pay my tax or I don't). More generally, the issue of *how concretely to exclude straightforward maximisers* seems to have been left unattended. This is rather unfortunate since Gauthier insists that the rationality of disposing oneself to comply rests also on the effective exclusion of the straightforward maximisers. Indeed, if I can employ individual strategies without any fear of exclusion than why should I dispose myself to employ a joint strategy? Minimax relative concession and constrained maximisation are rational responses to the problem posed by externalities and public goods in the market and nature conditions. It would be relevant to know how (and at what cost) one can prevent a straightforward maximiser from using a road or a lighthouse. Gauthier has not achieved much if, in order to apply his theory, Hobbes' enforcement system is replaced by an exclusion system. The cost of such an exclusion system would surely make the outcome sub-optimal.

Translucency

Leaving aside the concrete reality of exclusion, we can now turn to translucency. Gauthier acknowledges that he developed concept of translucency in reaction to a comment that D. Parfit made about *Reason and Maximisation.*[150] We saw in chapter 1 how Gauthier initially presented constrained maximisation. Parfit noted, in this original version, the implicit supposition that constrained and straightforward maximisers could clearly identify each other for what they were, as if they were transparent. In *Morals by Agreement,* Gauthier modifies this unrealistic assumption and creates the concept of translucency. Translucency is less than transparency but more than opacity. The disposition of a translucent person can be identified with a reasonable degree of accuracy but not with certainty. *Gauthier assumes that there is such a feature as translucency.* It is, in itself, a very strong assumption. We are provided with full details on how the 'disposition spotting skill' contributes to the rationality of constrained maximisation, but we lack information on the existence and the development of such feature.

[150] See *Reasons and Persons* in his discussion of the 'self-interest theory', especially pp 18-19.

If we accept that individuals are only translucent, then it is possible to mistake a SM for a CM and therefore it is possible to find situations of defection and exploitation. Co-operation ineluctably yields only nearly fair and optimal outcomes. Given the possibility of error, we need to know from what threshold it is rational to dispose oneself to comply. To address this issue, Gauthier creates a measure of translucency. Let p be the probability that CMs will achieve mutual recognition and successfully co-operate and q be the probability that CMs fail to recognise SMs but will be recognised by them and therefore will be exploited. Then the ratio p/q can be a measure of translucency: the more CMs are translucent, the more likely will they achieve co-operation and the less likely will they suffer from exploitation. The ratio works as follows:

- When there are more CMs, they achieve an outcome closer to the optimal one suggested by the minimax relative concession rule. They can afford a certain level of exploitation without losing out too much. Therefore, p/q can be small. Said differently, it is not very important if they are not translucent enough to identify each other correctly since they will still benefit from the high level of cooperation.

- The fewer the CMs, the further will the outcome be from the ideal joint strategy and greater the ratio p/q must be for co-operation to be rational. Indeed, the CMs cannot afford exploitation as much and they must display a reasonable level of translucency in order to recognise each other with as little failure as possible.

Realistically, and Gauthier labels it clearly, his argument appeals "implicitly to the requirement that co-operation yield nearly fair and optimal outcome."[151] An individual is more likely to dispose herself to comply if she feels she is amongst like-minded people and therefore she is likely to achieve a nearly fair and optimal outcome. He introduces a distinction between narrowly and broadly compliant persons. A *narrowly compliant* person is disposed to co-operate in ways that if followed by all would *yield nearly fair and optimal outcomes,* whereas a *broadly compliant* person is disposed to co-operate in ways that if followed by all would merely *yield her some benefit in relation to universal non-co-operation.* The broadly compliant person is the constrained maximiser as she was characterised above. Gauthier specifies that such a person is an easy target for unscrupulous SMs. So although she can still reap some benefit from her

[151] *Morals by Agreement,* p 178

disposition, she probably does not perform as well as a narrowly compliant person who is seeking co-operation wherever it is mutually beneficial on terms equally fair and rational to all. With the introduction of translucency and the problem posed by the difficulty for SMs and CMs to identify each others correctly, Gauthier has to narrow down his characterisation of constrained maximisation in order to maintain the rationality of the concept. If we assume equal rationality amongst individuals, only narrow compliance is rational.[152]

Gauthier also insists on the rationality of developing the 'disposition spotting skill'. As SMs and CMs improve their skill at detecting each other, CMs benefit more and more from co-operation and suffer less and less from exploitation. The probability p increases while q remains constant. The overall rationality of constrained maximisation benefits from the development of translucency. As people will develop the skill of correctly identifying the others as CMs or SMs, CMs will be able to co-operate more often together and avoid exploitation. Not only will they benefit from mutual advantage but the disposition to comply will come more naturally to them. They will develop a sense of guilt when behaving as SMs. A population of CMs gains in stability.

In order to have a clear picture of this core conception, it seems appropriate to recapitulate the *conditions of rationality required in constrained maximisation.*

(1) A constrained maximiser must estimate at a reasonable level the degree of the others' compliance. If a constrained maximiser feels that he is amongst like minded-people, he expects to achieve a nearly fair and optimal outcome and to gain from mutual cooperation.

(2) A disposition to comply or not to comply must be identifi*able* and identifi*ed* with a certain degree of accuracy. The more CMs, the less necessary it is to be translucent, the less CMs, the more relevant it is for them to be able to recognise each others without fail.

[152] See Gauthier's demonstration in *Morals by Agreement* p 226-227

(3) Last but not least, SMs must be effectively excluded from co-operation when spotted. Gauthier only mentions this rather obvious condition[153]. But, as we saw above, this condition, although essential and rather problematic, seems to have been underdeveloped.

Has Gauthier addressed the problem raised in section 1 about the identification of rationality with utility maximisation in strategic situations?

In parametric situations, rationality is identified with utility maximisation at the level of particular choices. In strategic situations, Gauthier suggests identifying rationality with utility maximisation at the level of a disposition to choose. Through his disposition, a constrained maximiser finds himself in a situation yielding better outcomes than is yielded by the situation of a straightforward maximiser. "In parametric contexts, the disposition to make straightforwardly maximising choices is uncontroversially utility-maximising. We may therefore employ the device of a parametric choice among dispositions to make constrained choices, rather than straightforwardly maximising choices, is utility maximising."[154]

The pursuit of mutual advantage being in our best interest is represented in our *coherent and considered preferences*. In strategic situations utility remains a measure of these preferences. If we remember our chocolate cake example, we had characterised them as follows:

- They reveal a preference for a state of affair rather than for a direct object of choice;
- They are based on our level of information and our beliefs at the time of decision;
- They depend on the valuer (subjective value);
- They depend on the state of affair she wants to bring about at the time of decision (relative value).

If an individual has chosen to bring about a state of affair where she co-operates and benefits from the mutual advantages **and** she believes that (most of) the others are like-

[153] *Morals by Agreement*, p 182
[154] *Morals by Agreement*, p 183

minded, **then** by constraining her utility maximisation, she reveals her considered and coherent preferences through her choices. *Rationality can still be equated with utility maximisation but with constrained rather than straightforward utility maximisation.*

We also remember the three conditions on strategically rational choice and in particular condition A: Each person's choice must be a *rational response* to the choices she expects the others to make. We can now replace it with condition A':

"Each person's choice must be *a fair optimising response* to the choice he expects the others to make, *provided such a response is available to him; otherwise, his choice must be a utility –maximising response.* A just person is disposed to interact with others on the basis of condition A'."[155]

As initially announced optimality is a necessary condition of rationality but it depends on the existence of the fair optimum. The second condition of rationality is therefore the degree of compliance of the others.

Has Gauthier provided a moral answer to the Fool?

Let us reformulate the Fool's challenge: some agents have rationally agreed to a contract that, if kept, can be mutually advantageous to all. Why would an agent comply with this covenant rather than individually defect? Would it be fair to answer as follows?

A Hobbesian's agent would comply because he knows it is in his interest, but more generally because he dreads being spotted and punished by the sovereign. *The expectation of punishment makes it more costly to defect than to comply and therefore makes it rational to comply.*

A constrained maximiser complies with this agreement because she disposes herself to do so; and she disposes herself to do so because she dreads being spotted and excluded. *The expectation of exclusion makes it more costly to defect than to comply and therefore makes it rational to dispose herself to comply.*

[155] *Morals by Agreement*, p 157

If this answer is correct, and Gauthier's response to the Fool was in the blurred (and probably institutionalised) reality of exclusion, Gauthier would have failed to give a moral answer to the Fool. We must look elsewhere for his answer. The rationality of constrained maximisation depends on our estimate of the others' compliance, on the concept of translucency and on the effective exclusion of SM. But all this technical probability apparatus can only feed or discourage an existing pre-disposition. For people to find it rational to comply, they must have internalised the change of rationality that sustains morality. Some of us have realised the benefit we would all gain were we all to comply and have fully assimilated the rationality of compliance. We *assume* that the others, being equally rational, must have reached a similar conclusion. We are disposed to comply on this basis alone. They have acquired the disposition to comply or, to put it differently, they have become just persons; they possess this artificial virtue called morality. The change of rationality from the pursuit of individual gain to the pursuit of mutual advantage has taken place within them and this change in turn supports their compliance. Based on the assumption of equal rationality, agents decide to comply because they have internalised the change of rationality. That is Gauthier's moral answer to the Fool. The fact that the others' *actual* behaviours subsequently strengthens or weakens their disposition is only secondary to the argument. The technical apparatus meant to demonstrate the rationality of compliance makes sense only once some have chosen morality, based on their belief that the others, being equally rational, have made the same rational choice. They have accepted to constrain their behaviour in line with the agreement rationally contracted. The change of rationality from individual to mutual gain triggers the sense of justice of some agents. The development of this sense of justice amongst individuals, in turn, feeds the practical rationality of compliance.

This moral response is reemphasized later in *Morals by Agreement* when Gauthier addresses Glaucon's challenge.

"Glaucon's claims that were the just man to put on the ring of Gyges he would behave no differently from the unjust man. In so claiming Glaucon thinks of the just man merely as someone who recognises the need to accept certain constraints, but whose emotions are in no way engaged by them. The 'just man' thus lacks any sense of justice, any capacity to be moved by considerations of justice as such... This shows only that he is not truly the just man. Properly understood, the just man is the person who recognising a certain course of action to be just, finds her feelings engaged by that

recognition and so finds herself moved to adhere to that course of action because of its justice."[156]

Gauthier's response to the Fool therefore should read:

A constrained maximiser complies with the agreement because she disposes herself to do so; and she disposes herself to do so *because she has internalised the change of rationality required for cooperation to exist.* Depending on the others' compliance, her disposition will be strengthened or weakened.

The internalisation of the change of rationality from individual gain to mutual advantage creates the disposition to comply. The principle of co-operation (minimax) obtained through bargaining provides the object of compliance; the minimax substantiates co-operation.

Could this be the difference between internal and external rationality? Before we can answer this question we need to complement our picture of external rationality with the next core conception that of the initial bargaining position.

Core conception 4: The Lockean Proviso
External rationality and fairness of the initial bargaining position.

Before we develop this core conception, it seems appropriate to make a couple of comments on it. Firstly, it comes in the fourth position which is rather strange when logically it should have been the starting point of the theory. The initial bargaining position should have been logically and chronologically anterior to bargaining. This anomaly is not accidental. Secondly, we have noted that Gauthier's strategy so far has been to establish the rationality of a core conception and only then to reconcile rationality with impartiality. In presenting the Lockean proviso, he first focuses on demonstrating its impartiality prior to demonstrating its rationality. Once again this change is not accidental.

[156] *Morals by Agreement*, p328

This sub-section will hopefully make sense of these two introductory points. We shall first present the Lockean proviso prior to developing on its sustaining morality and rationality. We shall then be fully equipped to come back to our last issue: why does Gauthier need to distinguish between internal and external rationality?

The Lockean proviso

So far impartiality has been assessed by the individuals involved in bargaining and not by an external observer. This standpoint intricately links impartiality with rationality. Indeed, rational individuals who voluntarily agree, in full knowledge of their circumstances and capacities, are responsible for their own best interest. As such they are the guarantors for the impartiality of the agreement ... unless they are coerced or victims of force or fraud. Therefore, it is essential to Gauthier's theory that rational individuals make their choice free from coercion of any sort. In this context, the initial bargaining position must be cleansed of the effects of previous force, fraud, free riding or parasitism. As we shall see, any discrepancies in bargaining power between future parties to an agreement are reduced as a consequence of this cleansing job.

Locke's theory of property states the following:

"He that subdued, tilled, and sowed any part of it [earth], thereby annexed to it something that was his property, which another had no title to, nor could without injury take from him. Nor was his appropriation of any parcel of land, by improving it, any prejudice to any other man, since there was *still enough, and as good left; and more than the yet unprovided could use.*"[157]

Gauthier's Lockean proviso, following Nozick's, is based on this sentence in italics and becomes: the proviso "prohibits worsening the situation of others except where this is necessary to avoid worsening one's own position."[158] A bit of explanation is required. To worsen or better someone's situation is always in relation to a base point. The most obvious base point is then someone's situation *in my absence*. What would be your situation had I not been here or interacting with you: if it would have been better, than I

[157] *Second Treatise of Government*, Chapter 5, paragraphs 32/33, p 633
[158] *Morals by Agreement*, p 203

worsened your situation; if it would have been worse, than I bettered your situation[159]. Similarly, my situation is worsened (or bettered) by you if it would have been better had you not be here. There are particular cases to this broad rule. For example, if you are the tax inspector in charge of my file and you are away from your desk each time that I try to contact you, your absence worsens my situation. It is not you in particular that I try to reach but the tax inspector in charge of my file who happens (to my sorrow!) to be you. Within this institutional framework, your mere absence does worsen my situation. In general, Gauthier states that the proviso allows us to better our situation as long as we don't worsen anybody else's.

In this pre-co-operation environment, each has to maximise his utility given the assumption of mutual unconcern. Each assesses her own situation separately, taking only immediate reciprocity into consideration. The proviso imposes a constraint on this utility maximisation according to the following rule. An individual should first try to better or at least not worsen his situation, without worsening that of anyone else. If that is not possible, then he should try to better or at least not worsen his situation, trying to minimise the worsening of the situation of others. If that is not possible, he must then try to minimise worsening his situation when interaction is not to his advantage.

The proviso forbids any unnecessary worsening of others' situation. This is obviously a departure from Hobbes' state of nature where every man had a right to everything 'even to one another's body'. The constraint is a huge requirement. Let us imagine that we are in a non co-operative environment, and you are a lot stronger and brighter than me, why should you stop yourself from abusing your natural advantage over me? Both the moral and the rational motivations of the proviso need now to be presented.

Proviso and impartiality.

The proviso is the constraint "by which we move from a Hobbesian state of nature, in which there are no exclusive rights whatsoever but only liberties, to the initial position

[159] Gauthier assumes a causal implication between my absence and the bettering or worsening of your situation.

for social interaction."[160] It is a four-step move. The first two steps are about the conversion of the pure state of nature. The last two steps concern the transition from natural to co-operative interactions.

Step 1: The proviso affords each person *exclusive right* to the use of his body and its power, his physical and mental capacities and imposes *a duty* not to use another's power since that would worsen their situation.

Step 2: I am entitled to the possession of whatever I produce using my power and capacities. However, my situation is not worsened if someone seizes the fruits of my labour but compensates me for it or if someone uses something for which I had no intended use. Therefore there is *no exclusive right to possession.* The compensation required is 'full' i.e. sufficient to cover my loss. 'Market' compensation, which is superior to 'full' compensation, can only be given in step 4, once out of the state of nature.

Step 3: Gauthier distinguishes between natural and social interactions. In natural interactions, the incidental imposition of costs on others does not violate the proviso. However, once we view others as potential co-operators (even still in the state of nature), whatever betters the situation of one by worsening the situation of others calls for compensation if the proviso is not to be violated. When moving from natural to market interactions, costs must be internalised amongst those interacting.

Step 4: This last step ensures the conversion from the state of nature into society, from common use to *exclusive right of possession* to land or other good. This exclusivity means that individuals feel secure to maximise the use of their possessions. Each specialises and everyone's needs are met through market exchange. Division of labour becomes mutually beneficial. However, not everyone benefits equally from this change. Gauthier needs next to confirm that this inequality does not generate partiality.

Before we move on to his argumentation, we must note that we have arrived, in the state of nature, at a definition of the actors in terms of *initial factor endowment as rights to their persons and property. These rights have to pre-exist bargaining and social*

[160] *Morals by Agreement*, p 209

interactions. Indeed, bargaining pre-supposes a clear definition of each actor's initial factor endowment. This definition is provided by the proviso. The proviso is a constraint on utility maximisation. It remains to be demonstrated that this constraint is impartial.

To argue for the impartiality of the proviso, Gauthier gives the example of 16 Crusoes living on 16 different islands. Each is either clever or stupid, either energetic or lazy, either strong or weak, either living on a well or an ill supplied island. All the possible combinations are represented from the clever, energetic, strong Crusoe living on a well supplied island to the stupid, lazy, weak Crusoe living on an ill supplied island. Gauthier claims that, although such distribution is completely arbitrary, it is no one's responsibility to redistribute in a more egalitarian way the natural assets so that the weak, stupid and lazy Crusoe benefits from the help of the clever, strong, energetic one. Any redistribution would be classified as free riding or parasitism. Gauthier's initial bargaining position is determined as if the parties were Crusoes on their islands using to their best ability their natural endowment.

"Each human being is an actor with certain preferences and certain physical and mental capacities which in the absence of her fellows, she naturally directs to the fulfilment of her preferences. This provides a basis in no way arbitrary, from which we may examine and assess interaction, introducing such conceptions as bettering and worsening... A principle that did not take this basis as normatively fundamental would not relate impartially to human beings as actors."[161]

It is interesting to note that impartiality is based on the concept of bettering and worsening, which in turn is based on individuals' natural power and capacities. We remember that in Core Conception One, impartiality was assessed rather on whether individuals were differentially affected or not.

Let us imagine that due to your natural endowment (for example force or intelligence), you could successfully use some predatory power on me. Does the proviso mean that you cannot use your power? And if you do, does the proviso mean that you have to compensate me before bargaining? On the one hand, you are able to worsen my position by bettering yours. On the other hand, our natural endowments are such that you are able to prey on me. Why should you refrain from using predatory powers you did not

[161] *Morals By Agreement*, p 221

choose to have? Why should you compensate me for having used these natural powers? The proviso in deciding who is better off and who is worse off does affect us differentially.

One can understand that the parties who have suffered from predation in the state of nature would not find it rational to comply with an agreement based on such coercion. But what about the predators: why would they sit at a bargaining table if, before they do, they are asked to surrender some of their predatory powers and gains? And if they do bargain on this basis, why should they comply afterwards? It is now time to turn to the rational defence of the proviso.

Proviso and rationality.

I believe that it is worth taking Gauthier's example and follow his reasoning faithfully.

We are two fisherfolks. "You, the upstreamer, discharge your wastes into the running water of the river, thus causing pollution, and so costs for me, the downstreamer. This benefits you in interaction with me, and so brings the proviso into play; you lack the right to pollute."[162]

There are then two alternatives: either your use of the river for waste disposal is the most efficient method *overall* or it is not. Two optimal responses are available.

Optimal response 1: If it is, you should continue to use it but compensate me for the costs it brings me. If it is not you should use and finance some other method of waste disposal.

Optimal response 2: If it is, you continue to use it. If it is not "I should pay you the difference in your costs necessary to induce you to adopt the most efficient method, *since this payment must be less than the cost to me of your pollution.*"[163]

[162] *Morals By Agreement*, p 223
[163] *Morals By Agreement*, p 223

Before we pursue this example, two small comments can be made. Firstly, it is surprising that in the state of nature, we should consider the *overall* efficiency of a process. Gauthier has emphasised before that prior to any agreement each uses individual strategies independently worked out. To decide whether a method is efficient overall requires some mutually devised reference point. Secondly we note that optimal response 2 is the perfect illustration of how a threat advantage can make a substantial difference in the outcome. The upstreamer is clearly in a stronger position than the downstreamer and can therefore dictate his preference. It is interesting to see how Gauthier will defend the rationality of not making use of such an obvious threat advantage. So let us continue his argument.

Why would the upstreamer keep the proviso? Here is Gauthier's answer:

"Co-operation has, as it sole and sufficient rationale, the maximisation of expected utility. Thus in bargaining, the claim advanced and the concession offered by each person depend on his endeavour to maximise his utility, together with his recognition of the similar endeavour of every other person. The principle of minimax relative concessions determines the outcome of co-operative interaction in such a way that shares in the cooperative surplus are related to contributions to its production in the same way for all. Of course not every particular interaction, considered apart from a practice of cooperation, will benefit each party to it proportionately to his contribution. *Everyone may expect to gain from certain arrangements for mutual assistance even though on any given occasion the recipient of assistance gains and the donors loses. It is the practice, and not the occasion, that must satisfy minimax relative concession.* Think now of the fisherfolk. I take a net loss if you dump your waste in the river. Disposing of waste by the method least costly to the disposer, ignoring all effects on others, *is not a practice offering expected benefit to each member of society. The particular interaction cannot then be defended by relating it to the practice that satisfies minimax relative concession.* Hence it violates the requirement, fundamental to rational co-operation, of mutual benefit proportionate to contribution."[164]

This very long quote calls for some comments.

Firstly, Gauthier has in effect demonstrated the rationality of the proviso through its compatibility with the minimax relative concession principle. The minimax principle has to pre-exist the proviso for the proviso to be rational. But we now face a surprising conclusion. Indeed, the initial bargaining position being the fruit of the proviso and the proviso being subsequent to the minimax, parties to the bargaining process are choosing the principle not knowing their initial bargaining position, *as if* they were behind a veil

[164] *Morals By Agreement*, pp 224-5

of ignorance. Gauthier appeals to their assumed equal rationality not to their best interest to justify the minimax. This approach is a lot more Kantian than Hobbesian and not really compatible with Gauthier's contractarian framework.[165]

The second issue also concerns the outlook on the minimax relative principle developed here. Indeed, we understand that Gauthier refers to an overall rationality: individuals do not assess the benefit of each situation within co-operation but the overall benefit that co-operation can bring to them. Sometimes they gain sometimes they lose. It does not matter as long as, overall, they gain: hence the need to be *disposed* to comply. The disposition to comply is the only true answer to the Fool's proposed unilateral defection on those occasions where compliance is not beneficial.

Lastly, Gauthier seems to appeal to another level of rationality namely *overall social rationality*. Co-operation, in being beneficial to all, is rational for all. Precisely because it is not the rationality of each particular case which is assessed but the overall rationality of co-operation, is it possible to have losers and winners in each situation as long as everybody gains from cooperative interaction. *Gauthier does not evaluate how rational it is to adopt the proviso for each of the fisherfolk, but rather how rational it is to adopt the proviso in this particular interaction overall.* In this interaction cooperation would be in jeopardy if the proviso was not applied. It is because cooperation is at threat and both parties supposedly want cooperation equally that the proviso is rationally justified.

Why would the upstreamer obey the proviso? Why should a predator refrain from using his natural endowment? The above answer leaves us short of an *individual* rational answer for each actor. So Gauthier continues. If a predator, free rider or parasite *wants* co-operation with others, or if future co-operation with others *will be beneficial* then she has to constrain her behaviour in the state of nature and either refrain from predation or compensate (before bargaining) prospective partners in co-operation if predation has occurred. The rationality of the proviso for the predators depends on this threshold between what they have to lose if they refrain from making full use of their power in the state of nature and what they have to gain from future co-operation. The last word is to expected utility maximisation.

[165] This point will be developed at length in chapter IV.

"For utility maximisers, the link between co-operation and mutual benefit must take precedence over the link between co-operation and impartiality or fairness... the proviso constrains the initial bargaining position to the extent, that such constraint is compatible with the co-operative outcome affording each person the expectation of a utility greater than that afforded by the non-cooperative outcome. It is rational to comply with a co-operative joint strategy if and only if its expected outcome is (nearly) optimal and as fair as its compatible with mutual benefit. We abandon neither the proviso nor narrow compliance, but we subordinate them to the requirement of mutual benefit."[166]

What about the distinction between internal and external rationality?

Let us recapitulate. The two central ideas of this core conception are as described by Gauthier himself: the Lockean proviso "moralises and rationalises the state of nature – but only insofar as we conceive the state of nature as giving way to society"[167].

The fairness of the process is considered from the initial bargaining position onwards. If the initial position is 'unfair', the outcome of the bargaining process will be unfair. If the outcome is unfair then the parties will not comply with it. *On the fairness of the initial bargaining position depends the rationality of keeping the agreement.* Fairness is therefore a pre-condition to the agreement not its by-product.

"It is both rational and just for each individual to accept a certain constraint on natural interaction and on the determination of his initial factor endowment, as a condition of being voluntarily acceptable to his fellows as a party to co-operative and markets arrangements – to social interaction. This constraint is part of morals by agreement, not in being an object of an agreement among rational individuals, but in being a pre-condition to such an agreement."[168]

We saw above that morality was an impartial constraint on the direct pursuit of individual utility. In the state of nature, there is no agreement on a joint strategy yet and we each maximise our individual utility. Gauthier demonstrates that the proviso is an impartial constraint on this individual utility maximisation. The proviso guarantees this artificial impartiality that replaces the natural impartiality prevailing in natural and market conditions. *Morality, in the form of this impartial constraint, has therefore to pre-exist the agreement.* This brings us to the second central idea of the proviso.

---

[166] *Morals by Agreement*, p 229-230
[167] *Morals by Agreement*, p 193
[168] *Morals by Agreement*, p 192

Parties accept a constraint on their utility maximisation in the initial position *only with co-operation and social interactions in sight.* It is *because* they are expecting to mutually benefit from co-operation that they voluntarily accept the proviso. In the state of nature, they constrain their behaviour with their future partners in co-operation in order to be accepted in the co-operative agreement. The proviso is therefore a rational requirement if we want the agents to comply with the agreement made. And it is a rational requirement because it guarantees fairness. We are back to the scenario of core conception 3. Once individuals have internalised the change of rationality from the pursuit of individual gain to the search for mutual advantage, they accept to constrain their behaviour. They constrain their individual utility maximisation with the prospect of benefiting from a co-operative surplus. It is this prospect that motivates their constraint.

Gauthier goes further: parties in the initial bargaining position don't just vaguely know that it is beneficial to co-operate, they know about the exact terms of their future co-operation since they know about the minimax relative concession principle and about constrained maximisation. "Without the prospect of agreement and society, there would be no morality, and the proviso would have no rationale. Fortunately, the prospect of society is realised for us; our concern is then to understand the rationale of the morality that sustains it."[169] They have to work from the principle of co-operation back to the proviso. As with core conception 3, the internalisation of the change of rationality motivates their disposition to obtain co-operation and the minimax relative principle substantiates the co-operation they have to comply with.

We are now able to make sense of the two introductory points made both on the structure of the book and the structure of the chapter on the Lockean proviso. Gauthier needs first to introduce the minimax and the concept of constrained maximisation to establish the moral need for the proviso. It is in turn the role played by the proviso (in securing impartiality) that supports its own rationality.

At last, we can also better understand the distinction between internal and external rationality. Once individuals have realised that it is in their best interest to pursue

[169] *Morals by Agreement*, p 193.

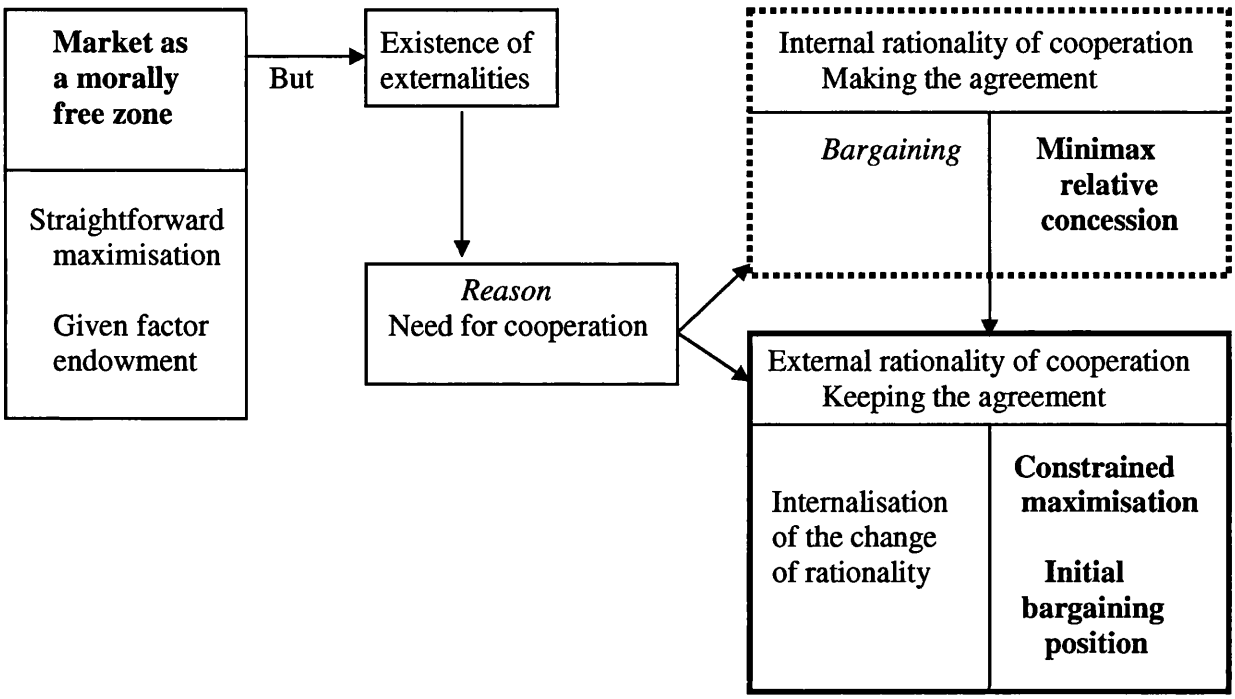
mutual benefit rather than individual gain, once they have realised the need for co-operation, two rationales are in place.

The internal rationality of co-operation deals with making the contract. Sophisticated straightforward maximisers negotiate their respective share of the co-operative surplus. Morality has no place in this process. *The principle of co-operation is a product of pure practical rationality. Internal rationality is the rationality that sustains the pursuit of individual gain.*

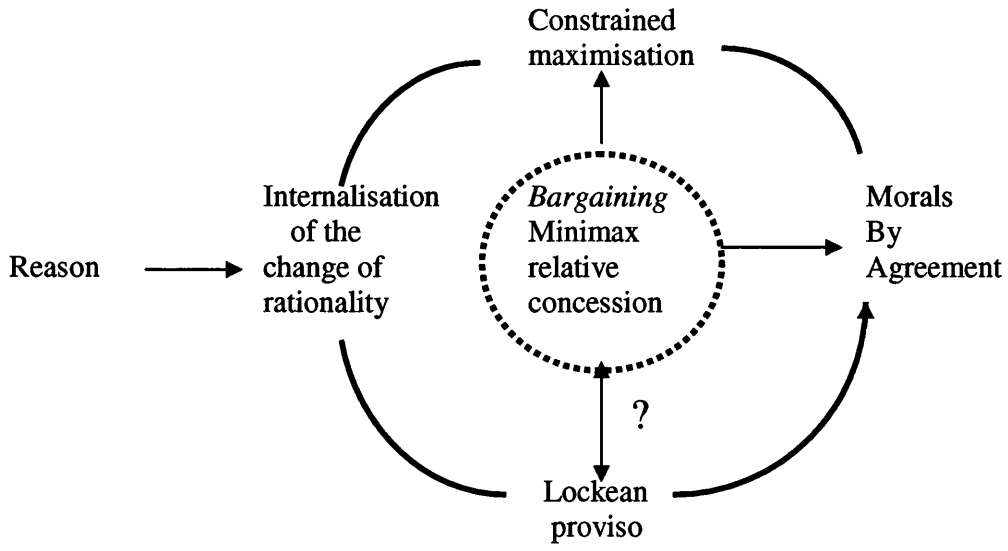
The external rationality deals with keeping the contract. Parties have internalised the change of rationality required for co-operation to blossom. They have acquired a disposition to obtain and maintain co-operation. Morality has emerged from this change of rational disposition. *External rationality is the rationality that sustains morality.* Parties constrain their utility maximisation both in the initial position and in co-operative interactions. The principle of co-operation, fruit of the internal rationality, merely provides them with the object of compliance.

We must remember that for Gauthier, the contract is hypothetical and that anybody at any time can go through the above mental exercise. The two schemas below should help the reader to visualise our explanation. However, where we can accept in principle the possibility of such a mental exercise, it is a lot more difficult to accept the suggested distinction. This dual rationality seems unrealistic and unnecessary. If, individuals have changed their rational disposition, then the change must shine through all their choices and decisions whether before, during or after bargaining. Besides, we still feel that the relation between the proviso and the minimax relative principle has not been fully explained yet. The proviso has to pre-exist the principle of co-operation. Therefore its rationality cannot be explained using the minimax. There should be a stand alone rationale for the proviso. In chapter IV, we come back to this central theme.

Schema 1: Core theory



Schema 2: External and internal rationalities



Core conception 5: The Archimedean point

Rationality and morality

This core conception stands out from the rest of the theory. The theory could have stopped at the proviso but Gauthier wanted to reconcile justice as mutual advantage with justice as impartiality. "From the Archimedean point one has the moral capacity to shape society."[170] His starting point is therefore impartiality rather than rationality. His ambition is to demonstrate that a rational ideal actor who has to choose the framework of co-operation arrives at the first four core conceptions. However, because she chooses from an impartial standpoint, her choice expresses the norms of justice.

We find in the Archimedean point the approach he had until 1979: he first presents his disagreement with Harsanyi's average utility principle and with Rawls' lexical difference principle, prior to presenting his own standpoint. As mentioned before, when taking this approach he arrives at the maximin relative benefit principle rather than at the minimax relative concession principle. However, the Archimedean point is now enriched with a proper ideal actor and this actor chooses not only the maximin principle but the proviso, the market as a morally free zone in conditions of perfect competition and the concept of compliance and translucency.

The ideal actor

Recall his original approach of the maximin in his *Social Contract*. Here was the way he described the bargaining position and the standpoint of choice:

"We assign an expected utility for each party to the social contract to the state of nature, as status quo, and to each possible society, as possible bargaining outcome.... We now select a rational individual at random, and attempt to present an argument that would convince him to agree to the adoption of any member of a particular set of principles for cooperative action. A similar argument must be equally convincing to any other rational person. Therefore the argument means that it is possible to correlate with each party to the social contract, a set of principles for cooperative action, any member of which the particular party would rationally agree to adopt. We then consider the intersection of these sets' to obtain the content of the social contract."[171]

---

[170] *Morals by Agreement*, p 233

[171] 'The Social Contract : Individual Decision or Collective Bargain ?' p56

This time the individual selected at random becomes the ideal actor who has to choose a just society. Let's read her description.

"The ideal actor is of course rational and generally informed, but she seeks her greatest utility without being aware of the particular circumstances in which she acts, the particular capacities with which she acts, and the particular preferences for whose fulfilment she acts.
But the ignorance of the ideal actor extends only to her inability to identify herself as a particular person within society. About the nature of the society and its members her knowledge is as complete as can be. She knows the limits and variations of human capacities and interests. She knows the range of feasible social structure, and the individual roles afforded by each... She chooses a society in ignorance of the life she will find in it, and she chooses ... only by considering the ways in which the preferences, capacities and circumstances can fit within the feasible societies."[172]

We remember that Gauthier reproached both Harsanyi and Rawls for choosing an *impersonal* individual in the Archimedean point. For both theorists, the Archimedean standpoint was occupied by individuals that could not be real persons and that could not have preferences or capacities or circumstances to relate their preferences to. Gauthier insists on that very point and makes the individual in the Archimedean point an *ideal actor*, ideal because of her knowledge and actor because she is not a mere observer, or an emphatic sympathiser but is implicated in the actual choice among principles of interactions. *As an actor, she is a chooser.*

"Each must be able to identify with the ideal actor, so that each may recognise, in the choice made from the Archimedean point, the choice he could have made, had he been subject to the constraints in knowledge that define that point. The ideal actor must therefore choose, not as if she had an equal chance of being each of the person affected by her choice, but as if she were each of those persons."[173]

We could object that if the ideal actor must choose *as if* she was each of us, can't we say that she merely simulates the bargaining process by computing it? Instead of each of us defending our position, we have this ideal actor having all the relevant data to do the choosing job on each of our behalf. By being simultaneously each of us, she is doing, on her own, the job that we are each doing from our own position.

Each of us could be an ideal actor in theory. We would just have to choose a society amongst the feasible ones and then imagine that we are any other person selected at a

[172] *Morals by Agreement*, p 235-6
[173] *Morals by Agreement*, p 255

random and ask ourselves whether, were we in her position, we would or not make the same choice and select the same principles. The impartiality of the ideal actor relies precisely on each of us being able to identify with her and agree with her choices.

The ideal actor now greatly resembles *the individual selected at random* of the *Social Contract*. Like him she is a utility maximiser, fully informed and she could be any and each of us. It is easy to see what we are getting at with our objection. If the ideal actor merely computes all our positions in order to choose principles of co-operation, what is the difference between what she does and what we have been doing up till now in the four previous core conceptions? Said otherwise, should it come to us as a surprise if she derives the same principles as we did? What extra features does she possess that could make the difference? Does she refer to a meta-level of rationality?

Gauthier answers this last question straight away: no, the ideal actor maximises her utility given her particular capacities, assets or circumstances. "There is no other level of rationality involved."[174]

In a way, his description of the Archimedean standpoint actually rejoins our objection. Read:

"We may think of the Archimedean point as a point of convergence; beginning from individuals choosing, each from his own perspective, principles for social interaction – principles which will of course reflect the chooser's concern to maximize his own utility – we alter the perspectives until we find that the same principle would be chosen by all. Thus each person is able to place himself in the Archimedean point by considering the circumstances under which his choice among principles for social interaction would converge with the choices of his fellows."[175]

When do we start to have converging choices? We start to have converging choices when we all aim at the same goal. Being the ideal actor in the Archimedean standpoint, she must achieve convergence and convergence is achieved through aiming at the *common goal* rather than at each individual's goal (when in conflict with the common one).

"They [the principles of interactions] address themselves to the *intent that is common to all*... The convergence of choice in the Archimedean point achieves this, ensuring that

[174] *Morals by Agreement*, p 256
[175] *Morals by Agreement*, p 255

the intent common to all will be considered *and no individual intent can be considered.*"[176]

That was the last element missing for the ideal actor to be fully one of us. *The intent that is common to all is to bring about co-operation since it is mutually beneficial.* We have just spent the last three conceptions explaining how this central idea was at the root of each one of them.

- In the bargaining position, parties do not bluff, threaten or lie because it would put at risk their chances of being part of the co-operative venture.
- The parties acquired a disposition to comply with the agreement rationally contracted because this agreement sets the conditions for the beneficial co-operation they all want and need.
- Parties who are willing to participate in the co-operative venture will also refrain from coercive, free riding or parasitic activities in the state of nature or compensate their prospective partners in co-operation if they do.

However - and we probably hold here the difference we were looking for between her job and ours - Gauthier insists: the ideal actor has to ensure that *no individual intent can be considered.* Said otherwise, the ideal actor is required to disregard the intent of the straightforward maximisers. Only the intentions of the constrained maximisers are to be taken into account. In this way the outcome obtained by the ideal actor will be different from the outcome obtained through our interactions, it will be 'cleansed' of individual players.

What principles would the ideal actor choose?

"Want for others what you want for yourself": that could be the ideal actor's maxim. If she wants benefit and freedom for herself, then she must want *mutual benefit and freedom for all.* She would reject the anarchic Hobbesian state of nature where we can each pursue our own interest without limits. *She would choose instead the proviso* according to which we can each pursue our own interest as long as we don't worsen the

[176] *Morals by Agreement*, p 256, emphasises added.

others' position. Once mutually accepted by a group, the proviso enables this group to become a society.

Given freedom constrained by the proviso a market emerges. In the absence of externalities and in conditions of perfect competition, the ideal actor would choose the *market* since it is the only option compatible with optimality. Any other distribution of costs and benefits would involve displacement of costs.

The presence of externalities leads to an outcome that is sub-optimal with respect to the market. Gauthier gives the example of two individuals of similar talents and energy who come from two different backgrounds. One would obviously do better than the other. Their return might correspond to their contribution to the production of goods. However, the structure in which they evolve is unfair "because it fails to relate benefits to the contributions each person *would have made* had each enjoyed similar opportunities."[177] The ideal actor, when choosing a social structure, must therefore ensure that no individuals benefit differentially from it. The principle chosen from the Archimedean point must provide that each person expects a share of the fruits of social interaction that is in relation to the contribution he would have made in the social structure most favourable to the development of his talents and capacities. To harmonise these contributions and express their common measure within the same social structure, Gauthier suggests equating it to the claim each would make on the surplus and calls it the maximum social benefit. In this context, the ideal actor wanting to benefit from co-operation would choose to maximise minimum relative benefit i.e. *the minimax relative benefit.*

The only conception that cannot be chosen from the Archimedean point is constrained maximisation since the rationality of compliance depends on the characteristics and circumstances of each individuals involved in interaction. The ideal actor can only choose the processes that make narrow compliance rational. The hope we had earlier that from the Archimedean point, the intent of straightforward maximisers could be disregarded now collapses. We are left with nothing to distinguish the ideal actor from the parties to interactions.

---

[177] *Morals by Agreement*, p 263

We had been lead from chapter I to have high expectations of Gauthier's Archimedean point and ideal actor. But I believe that in giving so much 'flesh' to his ideal actor, Gauthier has failed to distinguish her from the bargainers and parties to interaction described in the first four core conceptions. In doing so I believe that he has also failed to reconcile justice as mutual advantage with justice as impartiality.... unless his justice as mutual advantage was already embedded in impartiality.

The story is not over yet. Gauthier dedicates a full chapter to cover topics as varied as the problem of inequalities in natural assets, the problem of rents, inter-nation relationships or inter-generational problems such as inheritance or investment. But as he says it himself: "we embark on the present chapter in a spirit of exploration ...ready to admit that, even if our theory of morals by agreement be fully acceptable, much that we say here must be tentative and controversial."[178] And indeed a lot of the views expressed in this rather less theoretical chapter are very controversial. Each of the topics covered could be the object of a debate in itself. To give only a couple of examples, Gauthier claims that factor rent should be divided among co-operators, or that an outsider able to make much better use of a land than its present occupiers has a right to settle on the land as long as he keeps the proviso and compensates the inhabitants.

I choose not to develop on any of these debates which, anyway, add nothing to the theory developed above. However, it is relevant to know that, according to Gauthier:

- The proviso allows for inheritance and provides room for property rights passed by bequest

- Investment is not only allowed but recommended as it increases the absolute value of the surplus from generation to generations.

[178] *Morals by Agreement*, p 269-270

Conclusion

We have just travelled a long way. Gauthier's theory has taken us from pure practical rationality to morality. From mere utility maximisation, we have discovered the minimax relative concession principle, constrained maximisation and Lockean proviso. We have explored Gauthier's conception of morality, justice, impartiality and rationality.

We remember from chapter I that Gauthier considers Hobbes as the 'true parent of rational morality'. However, Gauthier also believes that Hobbes chose the wrong premises on the state of nature and that he failed to bring to a satisfactory completion the concept of rational morality. I claimed then that his own theory was primarily built by rectifying the double failure he found in Hobbes. I believe that *Morals by Agreement* is meant to be an improved and modernised version of the *Leviathan* where these two major weaknesses are rectified using the tools of modern moral and political philosophy.

I hope that this second chapter has now substantiated my claim. We have just seen how Gauthier replaced the Hobbesian state of nature by a Lockean one, how he replaced the search of self preservation by the pursuit of mutual benefit and how he has emphasised the difference between natural and conventional reason using the internalisation of conventional reason as the basis for rational morality. Has he really succeeded in providing a convincing theory? Has he really improved the Leviathan? The answer to these questions is postponed to later.

One thing is for sure: *Morals by Agreement* has become an unavoidable classic of modern moral and political philosophy. It has been extensively commented upon and prior to making any comments on its long lasting impacts, it is fair to turn first to the criticisms made of it and Gauthier's reply to them.

PART 2

IS GAUTHIER'S CONTRACTARIAN MORAL THEORY POSSIBLE?

Introduction

Morals by Agreement generated an abundant literature. At first, David Gauthier actively participated in the debate that followed the publication of his theory, but from 1993 onward, he started to explore new horizons. Faithful to the idea of a contractarian moral theory, he first replaced his constrained maximiser with a resolute planner using McClennen's concept of resoluteness. He then came back to his interpretation of Hobbes, modifying his initial understanding of it. Both these changes profoundly modified his original contractarian moral theory. However, the core idea remained the same: Gauthier wanted to derive morality, or at least distributive justice, from pure rational grounds. This second part assesses his achievement with regards to this goal, whether in *Morals by Agreement* or in the numerous articles he wrote after 1986.

The present part is organised as follows. In chapter III, I first review the debates that took place after the publication of *Morals by Agreement* and Gauthier's reply to his critics. I conclude on what is left of this original version of his theory. I then develop in chapter IV my own criticism in the continuity of the interpretation provided in chapter II. I first demonstrate that Gauthier's bargainers aren't and cannot be straightforward maximisers and that they are constrained maximisers. I then evaluate the consequences of such a claim on Gauthier's achievement and argue that the bargainers' disposition involve a moral attitude that jeopardises his core project.

I then turn to Gauthier's latest articles and new developments. In chapter V, I review his concept of a resolute planner and demonstrate against both his and McClennen's claims, that sophistication is a better strategy than resoluteness in assurance games. I also argue that resoluteness cannot be a *stand-alone* rational concept. In chapter VI, I review his latest interpretation of Hobbes. Unsurprisingly, Gauthier looks for renewed inspiration in his Hobbesian background. In Chapter I, we saw how much his interpretation of the *Leviathan* influenced his research. Again, this time his original (if not controversial)

interpretation of both Hobbes' contractarian theory of law and concept of public reason will refresh his research.

From both these changes has emerged a new skimmed theory which the backbone remains *Morals by Agreement*. In chapter VII, I attempt to reconstruct the latest version of his contractarian moral theory. I insist on the fact that Gauthier has continued to refer to *Morals by Agreement* for the broad lines of his contractarian theory and he has never published any new version of it. The reconstruction drafted in the last chapter is only a suggestion of what his latest theory *could be*, based on the acknowledged shortcomings of *Morals by Agreement* and his latest publications. I conclude that this latest version carries the same weakness as the one highlighted in chapter IV: in assuming a change of rationality, Gauthier introduces morality at the outset of the agreement.

CHAPTER III: GAUTHIER AND HIS CRITICS

Gauthier's theory has been extensively commented on and criticised.[179] David Gauthier initially defended *Morals by Agreement*, but the theory so perfectly articulated in *Morals by Agreement* had almost entirely collapsed by 1993. Indeed Gauthier had to give up on most of his assumptions (equal rationality and mutual unconcern) as well as on most of his core conceptions. The market as a morally free zone was proven to be sustained by morality, the minimum relative concession principle did not resist the game theorists' attacks and constrained maximisation was challenged by McClennen's concept of resoluteness. Gauthier expressed some doubts about the Lockean proviso further to comments about its fairness. The only core conception that is still standing is the Archimedean point.[180]

From 1993 onwards, Gauthier stopped defending his masterpiece and started to concentrate on two new fields of research: firstly, the concepts of resoluteness, rational deliberation and rational commitment; secondly, the more practical application in politics of his work on individual rationality and commitment.

It would have been almost impossible to mention all the objections made against *Morals by Agreement*. I have made a selection and have chosen to present arguments that either contribute to explaining where and why Gauthier failed or arguments that could be adapted to fit into current debates.

---

[179] These criticisms and comments are to be found primarily in the following publications:
+ In 1987, *Ethics* was the first journal to react to *Morals by Agreement* with a symposium on David Gauthier.
+ In 1988, the revue *Social Philosophy and Policy* dedicated a full volume to *Gauthier's New Social Contract.* The same articles were edited the same year by Paul, Miller and Ahrens in *The new Social Contract: Essays on David Gauthier.*
+ In the same year again, *The Canadian Journal of Philosophy* published a symposium on *Morals by Agreement.*
+ *Contractarianism and Rational Choice* from P. Vallentyne is an excellent selection of essays on *Morals by Agreement.* It was initially published in 1991.
+ In 1993, Gauthier co-edited with Robert Sugden *Rationality, Justice and the Social Contract.*
+ Morris & Ripstein *Practical Rationality & Preference: Essays for David Gauthier* was published in 2001 and is the most recent publication on the topic.

[180] Quotes and references on this paragraph's statements are obviously provided in the body of this chapter.

Twenty years after it was first published, one can wonder what is left of Gauthier's classic. In this chapter, we will review the debates that have taken place around *Morals by Agreement,* assumption by assumption, core conception by core conception. The only core conception that we will not review is the Archimedean point since it has hardly been commented on. In conclusion, I review what is left of *Morals by Agreement.* Surprisingly, despite the apparent collapsing of his theory, its core ideas are still very much alive. Even if his approach has changed, Gauthier continues to defend the concepts of hypothetical contract, rational morality and morality as constrained behaviour. Prior to developing on the criticisms made of *Morals by Agreement,* it seems appropriate to open this chapter with some comments on Gauthier's moral theory.

About Gauthier's moral theory

Morals by Agreement applies to cool rational adults who can contribute something to society ... that leaves out a good fringe of the population! What type of morality does he obtain and does it have anything to do with morality anymore? The following quotes draw a very good picture of what is reproached in Gauthier's theory.

A. Baier writes: "His morality ... simply drops a very large section of what has traditionally passed as morality"[181]. We saw in chapter I that Gauthier is not worried about reconciling his morality with traditional morality. However, the objection is a lot deeper. Baier continues:

"He had already made it clear that any commitment to care for the handicapped and dependent elderly is not one of the constraints of his modest morality[182] - those who never had or no longer have anything left to bring to the moral bazaars and markets are to get nothing from them, on pain of the charge of parasitism or free ridership."[183]

Similarly, J. Hampton argues that Gauthier's theory doesn't capture the nature of morality but she diagnoses it differently:

"*regardless* of whether or not one can engage in beneficial cooperative interactions with another, one owes that person respectful treatment simply in virtue of the fact that he is a person. Not all value is subjective; in particular, the value which human beings have is objective, and demands one's respect, whether that human being is an infant with whom

[181] 'Pilgrim's progress', p 326
[182] See *Morals by Agreement,* p 268
[183] 'Pilgrim's progress', p 327

one will never have reason to cooperate, an elderly man past his prime, or an adult whose talents one finds of no particular use. It is because Gauthier does not assume that human beings have an objective worth that he sometimes gets unintuitive and unacceptable results in his moral theorising."[184]

D. Braybrooke raises a similar issue but expands it beyond the level of human beings: "Gauthier's contract, like the morality entailed by it, leaves out of account people who are not in a position to contribute to producing any part of the cooperative surplus." If a rich nation has nothing to gain from cooperating with a poorer nation, do we have to understand that it is moral for the rich nation not to interact with the poorer one? "Many (I among them) will feel that this possibility of indifference – or of worse – toward other people is an enormous limitation on the morality that here emerges from the deduction project." We shall comment on the assumption of mutual unconcern later. We just want to emphasise here the point made by Braybrooke. "Why should we think that all that any of us want to find in morality or justice will be put there by reason alone?"[185] Where Hampton diagnosed a wrong account of value, Braybrooke explains Gauthier's counter-intuitive and limited morality by the cool and sole use of rationality. We are not mere rational machines. We feel as well as we think.

We saw before that Gauthier clearly puts a distance between conventional morality and his rationally derived morality. This poses a problem: his moral theory cannot be tested since, in his view, the truth of moral judgements is not known independently of the plausibility of the moral theory.[186] This means that Gauthier cannot afford to rely on non-demonstrated or non-ascertained assumptions or conceptions. Everything the theory rests on has to be proven or at least generally accepted. As we shall now see, Gauthier is no way near such an achievement.

[184] 'Can We Agree on Morals?', p 352
[185] 'Contract Theory's Fanciest Flight', p 756
[186] See C. Morris 'Relation Between Self-interest and Justice', p130-131

Section 1: The main assumptions and the concept of coherent and considered preferences

The assumption of equal rationality

"I should admit ... that were I to become convinced that an appeal to equal rationality was either a concealed moral appeal, or inadmissible on some other ground, then I should have to abandon much of the core argument of *Morals By Agreement*."[187]

David Gauthier, 1988

"The appeal to equal rationality is part of the flawed argument that I am now persuaded I must abandon."[188] David Gauthier, 1993

Why is the assumption of equal rationality so essential to *Morals by Agreement*? And why does Gauthier have to abandon it?

The importance of the assumption of equal rationality

As we saw in chapter II, the assumption of equal rationality supports most of the core conceptions.

1- Equal rationality is one of the conditions on rational bargaining. Since each utility maximiser seeks to minimise her concession, then no one can expect any other rational person to be willing to make concessions that she would not be willing to make herself.[189] The minimum relative concession principle is entirely dependent on this assumption.

2- Equal rationality prevents bluffing and threatening in bargaining between fully informed individuals. No one would make a threat or attempt to bluff. Being fully

[187] 'Morality, Rational Choice, and Semantic Representation', p 186
[188] 'Uniting Separate Persons', p 180
[189] *Morals by Agreement*, p 143-144

informed, no one would believe anyone who claims that she will act in a non-utility maximising way.[190]

3- Equal rationality is indispensable to support his defence of narrow compliance against broad compliance. It cannot be rational for everyone to be disposed to broad compliance. If it is not rational for everyone, it is not rational for anyone. On the contrary, a rational person finds it rational to comply given the minimum relative concession principle and the proviso. She expects others to adhere to the agreement and, given equal rationality, considers it rational to adhere herself.[191]

4- Equal rationality is the foundation of Gauthier's concept of justice: "We do claim that justice, the disposition not to take advantage of one's fellows, is the virtue appropriate to co-operation, voluntarily accepted by equally rational persons. Morals arise in and from the rational agreement of equals."[192]

5- Gauthier uses the assumption of equal rationality to show that rationality requires us to accept constraints which apply impartially to all of us. The assumption obviously backs up the concept of Archimedean point. The ideal chooser providing the impartial standpoint must assume the equal rationality of the agents.[193]

Objections to this controversial assumption

G. Harman was the first one to react to the assumption of equal rationality. In substance, he made the two following points. Firstly, he argued that equal rationality does not mean equal bargaining power. Unless equal bargaining power means that each adopts the same attitude toward any given level of relative concession, equal rationality cannot imply equal relative concession.[194] Secondly, he questioned the validity of the equation between equal rationality and equal compliance. Equal rationality cannot mean similar psychology or situation of those involved and therefore cannot be a relevant assumption for real people in the real world.[195]

---

[190] *Morals by Agreement*, p 155
[191] *Morals by Agreement*, p 226
[192] *Morals by Agreement*, p 232
[193] *Morals by Agreement*, p 235
[194] 'Rationality in Agreement', p 6-7
[195] 'Rationality in Agreement', p 9-10 and 14

The first point was taken over by R. Goodin[196]. His argument is that the assumption of equal rationality insidiously introduces impartiality in the premises. Like Harman, he claims that equal rationality does not mean equal bargaining power. Real people in the real world have different tastes and different resources.

A collector can be prepared to pay a disproportionate amount for an item missing in his collection. Given his taste, it is perfectly rational to make a relatively bigger concession than the seller on the price of the item. Similarly, it is perfectly rational for people with lesser resources to make greater concessions than richer people. The only difference between the collector and the poor is that the poor concedes relatively more out of necessity not out of desire. Bargainers who are better endowed can last longer in a bargain and hold a stronger position. Someone with lower resources can settle for a less advantageous bargain not because she is less rational but because she has more to lose if she doesn't. Differing attitudes towards risks lead perfectly rational people to make very unequal relative concessions. Attitudes towards risk are not a matter of personal tastes or personality: for those who have less, it is rational to be reluctant to take risks.

In all these cases, 'preferences' for greater concessions are rational with regards to the bargainers' particular taste or situation. These 'preferences' are forced upon them.

"The effect of such preferences, once people have internalised them, is to make it rational for agents to behave in the ways described. If it works to their relative disadvantage in bargaining games to have these preferences, then that is just their bad luck... But a model that commends this course of action can hardly be said to be 'morally impartial' as between rich and poor."[197]

A very similar argument could be built up against the assumption of equal rationality with regards to compliance. Equal rationality does not mean equal *interest* in compliance. It does not mean either that there is a unique degree of compliance which is rationally prescribed[198] Even if we can all see the rationality in complying with mutually advantageous rules, not all rules are equally important to each of us. The father of young children has probably a lot more interest in the compliance of all with road rules than a bachelor. He has a lot more too lose if these laws are not kept. As Harman

[196] 'Equal Rationality and Initial Endowment', p 119 -122
[197] 'Equal Rationality and Initial Endowment', p 122
[198] See R. Sugden's 'Is the Contractarian Enterprise Possible?' for a good defence of this argument.

labels it, real people have different psychologies and are in different situations. The bachelor might find these rules a bit boring and useless. The father looks at them as essential to his children's safety.

As we saw at the beginning of this section, Gauthier abandons the assumption of equal rationality in 1993.

The assumption of mutual unconcern and morality

Many have commented that the assumption of mutual unconcern is false and does not correspond to reality: emotions and affectivity do enter into our preferences and can play a role in co-operation[199]. We do care about our family and about our friends. Our concern for them does affect our preferences. Our emotions can sometimes move us in favour of co-operation naturally without any rational grounding. Our feelings can affect our motivations at the bargaining table.[200]The assumption leaves all these human aspects aside. In doing so, it fails to connect the theory to reality and precludes an accurate description of 'real' or 'true' preferences. Such a false assumption introduces a damaging bias into a contractarian theory based on bargaining and involving real people in their real life situation.

To understand the nature of the objection, let us make a little detour by *Morals by Agreement* to see how Gauthier introduces this assumption.

Gauthier does not deny that we do exhibit a concern for people close to us but our positive fellow feelings are restricted to a very limited circle of individuals. He seems to imply that therefore the *impact* of such fellow feelings is necessarily limited. Since in the vast majority of cases, co-operation involves people we don't know and therefore we don't care about, the assumption is an adequate simplification of reality.[201]

---

[199] See A. Baier in 'Pilgrim's progress' ('Our actual psychology diverges importantly from that required of the hypothetical progenitors of Gauthier's liberal individual.' p 326). See also C. Morris in 'Relation Between Self-interest and Justice', p 129, L. Thomas 'Rationality and Affectivity' or P. Vallentyne 'Contractarianism and the Assumption of Mutual Unconcern', p 72

[200] P. Vallentyne, 'Contractarianism and the Assumption of Mutual Unconcern', p 73

[201] *Morals by Agreement*, p 100-101

His argument actually goes further. This assumption is not a mere simplification. It is made necessary by the very fact that our fellow feelings are so restricted in scope. We cannot *rely on* such feelings to build up a moral theory and therefore we are lead to assume a lack of mutual concern if we want to free the theory from marginal and partial cases. The assumption makes possible a broader application of the theory[202].

Nevertheless, in his enthusiasm, Gauthier then turns the argument:

"We shall show that moral constraint is not only compatible with mutual unconcern, but indeed rationally required given this unconcern and the typical structures of interaction. Thus we propose to defend the liberating idea of a society that imposes no affective bonds on its member."[203]

The starting point is no longer reality and the existence of acknowledged human affective bonds, but rather what is expected from and for a moral theory. Unconcern is a given data. Morality must not depend on affection, so it speaks to reason rather than to contingent emotions and feelings.[204]

As if the above was not enough, Gauthier turns to Kant to justify the assumption. Kant, he claims, insists that morality cannot depend on such particular psychological phenomena however humane and however universal.[205]

Most of the objectors strongly contested Gauthier's reference to Kant[206]. Kant can assume mutual unconcern within his theoretical framework, but Gauthier's use of rationality is instrumental and such an assumption is out of place in his theory. What about Gauthier's other arguments in favour of the assumption of mutual unconcern? Unconvincing they say!

Can we use it as a simplifying assumption? No. The most compelling reason suggested against this argument was provided by C. Morris. Indeed, he claims that what is simplified is the complexity of interests in each others' interests which in itself is

---

[202] *Morals by Agreement*, p 102
[203] *Morals by Agreement*, p 102
[204] *Morals by Agreement*, p 238
[205] *Morals by Agreement*, p 103 or p 238
[206] See for example P. Vallentyne in 'Contractarianism and the Assumption of Mutual Unconcern', p 72-3 or C. Morris in 'Relation Between Self-interest and Justice' p 127-8

morally relevant.[207]. Gauthier wants to exclude morality from his premises in order to derive morality from exclusively rational grounds. However, this assumption of mutual unconcern introduces a major flaw in this strategy. In denying the existence of positive (love or care) as well as negative (envy, hatred or spite) feelings, it eliminates one of the major threats to morality in human nature. The assumption is not only false it also introduces a moral bias in the outcome of the theory. The bias runs from the outcome of bargaining through compliance with it. Out of love or hatred, I can bargain against my interest or fail to comply with an agreed norm.

More worrying is the fact that some will exploit others by using their feelings. Gauthier claims that in assuming that we are unconcerned, we can re-assess on purely rational grounds some of our existing practices as well as those relationships involving affections. The morality obtained can then be freed from inherited conventional norms.[208] Therefore, for the purpose of evaluation, we are entitled to refer to mutual unconcern. Such an argument obviously seems appealing and could address some of the feminist's concerns. Morris objects that such an option is not open to Gauthier who refers to a subjective account of value.[209] Indeed, such an account of value denies any relation of priority amongst preferences. In particular, self-interested preferences are not given priority over other-regarding preferences. However, for a successful evaluation, one must assume that these premises have an evaluative priority over the conclusion. Without appealing to objective value, no priority relation can be sustained.

In response to these objections, Gauthier argues that moral loading of the assumption would have to be positive rather than by default. After all, he is not suggesting that the preferences should have a moral content (such as an appeal to just behaviour in preferences). As such the assumption of mutual unconcern could be removed if its sole purpose was to constrain individual preferences. "It is important to show that the argument for morality is compatible with persons having strictly nontuistic preferences, but to do this, we need not assume that our preferences are in fact predominantly nontuistic."[210]

---

[207] See C. Morris, 'Relation Between Self-interest and Justice', p 129
[208] See *Morals by Agreement*, p 351
[209] See C. Morris, 'Relation Between Self-interest and Justice, pp142-144
[210] See 'Morality, Rational Choice and Semantic Representation', p 215-216

Gauthier is more concerned by the second level of Morris' objection. He agrees that his account of value does limit his use of the assumption of mutual unconcern. He therefore suggests revising the role of the assumption in his theory. The assumption is now provided with a forward-looking justification that strikingly resembles the justification of the proviso. Individuals choose their practices and institutions based on their nontuistic preferences but with the prospect of benefiting later from such preferences as long as others do the same. Once the agreement is made, individuals are free to act upon their tuistic preferences as long as they fit into the institutions chosen.[211] In order to address Morris's objection, Gauthier is led to prioritise preferences. Such a modification is not without consequences for two of his concepts: considered preferences and subjective value. Let us illustrate with an example the problem posed by this revised account.

In my younger days, I was a dedicated student. One evening I was preparing for an exam that I had to sit on the next day. My plan was to read all my notes once more and go to bed for an early night. I was reading when a heart broken friend knocked at my door seeking comfort. What was I suppose to do? Let her in and comfort her, running the risk of a late night, no revision and therefore failed exam or send her back home with her sorrow but with a chance for me to better perform at the exam on the next day? It was her interest against mine. My considered preference was to comfort her. I valued more a state of affair in which she would stay but with a risk for me of failing my exam on the next day to a state of affair in which I would send her off but with a higher chance of passing my exam on the next day. I had reflected enough on my preference. Based on past experience, I knew that, if I were to send her home straight away, I would feel a sense of guilt on the next day that would impede my concentration at the exam anyway. Gauthier's revised account does not allow for such considered preferences.

In 1991, he modifies his views once more. Originally he defined morality as a rational constraint on the pursuit of individual interest. In 1991, he acknowledges that such a formulation is misleading. "I want to defend morality as a rational constraint on the pursuit of one's aims and objectives, whether or not these objectives have any connection with one's interest or one's personal well-being."[212] The formal aim of a

[211] See 'Morality, Rational Choice and Semantic Representation', p 216
[212] 'Rational Constraint: Some last Words', p 323

rational individual becomes the maximum realisation of her substantive aims. We shall come back later to this new concept of substantive aim and its impact on the theory. For the moment, we just need to note that, in effect, this new formulation of morality amounts to a **removal of the assumption of mutual unconcern**.

Our emotions and feelings do affect our preferences. So does our environment. Now, we need to turn to another sort of issue raised by our affective bonds and emotions. I call it the circularity problem.

The circularity problem and rationality of considered preferences

Gauthier acknowledged that we have emotions and feelings that could enter in our preferences but already in 1988 suggested channelling them towards co-operation and framing them within rational agreement:

"In my argument, preferences are fixed... as objects of rational reflection and then dispositions are defined in relation to them. But what if these dispositions are themselves objects of preference? ...I must allow for the possibility of such preferences ... for since human beings have, in addition to beliefs and desires, emotions that affect their willingness and ability to behave in different ways, the cultivation of appropriate feelings for our fellows may be essential to inducing behaviour that is maximally conductive to mutual preference fulfilment. The cultivation of appropriate fellow feelings thus becomes a suitable object of rational agreement."[213]

With such a statement, Gauthier leaves us with a new issue. What comes first: the change in our preference or the bargaining from a pre-social position? Do we bargain with brand new preferences cautiously acquired by appropriate cultivation of our feelings for others or do we have to agree first on which preference to acquire and then cultivate? I shall call this problem the circularity problem. As Harman suggests: "Gauthier abstracts away from the actual dynamic character of human convention... Surely in real life it is not possible to separate cooperating from bargaining."[214]

Gauthier is aware of this circularity problem but he continues to argue in favour of the hypothetical contract. In his opinion, a theory based on a hypothetical contract can still be used in an on-going society. As a theoretical tool, it enables us to step back and

[213] 'Moral Artifice', p 401
[214] G. Harman, 'Rationality in Agreement', p 13-14

reflect on the society we live in. When assessing our existing institutions and practice we can decide whether it is rational to maintain them or preferable to change them. [215] Gauthier continues: "The same reflective capacity, I claim, leads from those principles that would be agreed to, in existing social circumstances, to those that would receive ex-ante agreement, prior to all society."[216]

This defence of the hypothetical contract only confirms the circularity problem met by his concept of coherent and considered preferences. If we assess our practices and institutions from within, our considered preferences are necessarily influenced by our environment and our existing system of evaluation. As C. Morris notes:

"A choice would not be a fundamental evaluation of the institution, for its standards are not independent of the domain of evaluation. For a rational choice evaluation of a social practice or system to be fundamental the preferences are to be more than coherent and considered. They must also be independent of the practice or system. In the absence of such independence, all that can be shown by rational choice is that the preferences are, broadly speaking, consistent with the practice or system." [217]

It seems that to maintain his claim, Gauthier is forced to take a Rawlsian turn:

"rational persons will revert from actual to hypothetical agreement, considering what principles they would have agreed to from an initial position not structured by arbitrary constraints... Once awareness of what persons would have agreed to apart from the arbitrary contingencies of actual society, becomes known, existing institutions and practices in contravention of those agreements will be destabilised."[218]

In such a defence, Gauthier is forced to move away from his initial description of real people bargaining from their situation in real life. Individuals are now required to decide not only on the way to distribute the co-operative surplus but also on which society they want to live in. They have to agree on social practices and institutions. In order to be able to do so, they are required to abstract from their social contingencies, abstract from any received models or values and use exclusively their rational skills.

Such a move seems incompatible with his theory of bargaining. **Either** the reflective choosers are real people influenced by their practices and institutions in which case their considered preferences are biased **or** they are choosing from behind a sort of veil of

[215] See 'Morality, rational Choice and Semantic Representation', p 179 - 182
[216] 'Morality, rational Choice and Semantic Representation', p 184
[217] C. Morris 'Relation Between Self-interest and Justice', p 140.
[218] 'Morality, rational Choice and Semantic Representation', p 185

ignorance in which case Gauthier has to work his way down a Kantian theory of justice. The instrumental rationality referred to in these cases is not the same.

A. Weale raised a similar issue in his 1993 article[219]. In this article, he argues that Gauthier presents us with a theory of social union (involving some common final ends between its members) to be reconciled with the separateness of persons. In attempting this reconciliation, Gauthier ends up blurring the type of instrumental rationality he refers to. Indeed, common ends or some social norms and institutions become the condition for individual rationality rather than a derivative from it: it is rational to conform to these norms and institutions if one wants to benefit from the cooperative surplus.

'It would seem that the only conception of rationality capable of moving individuals to conformity in this context is not one that rests upon the practical advantages of individual conformity, but is one which instead appeals to the rationality of being able to will only that which could be rationally willed by all. Whether Durkheim and Kant can be so invoked to succeed where Hobbes and the theory of games fail is of course an open question'[220]

If Gauthier wants his theory to apply to real people in real life situations, he has to take on board some aspects of human psychology in describing their preferences: people do care about each others (or at least some others) and they are influenced by their environment. This double influence from our emotions and from our environment obviously raises a difficult question about the rationality of our preferences.

Considered preferences are based on sufficient and adequate experience and reflection. K. Baier notes that all we can say is that considered preferences are the fruit of careful *reasoning* but if based on false beliefs, they can be *irrational.*[221] Preferences based on eccentric or brain washed beliefs are prime candidate for irrationality.[222]

In response to this objection, Gauthier re-emphasises two points. Firstly, he reminds us that considered preferences are attitudinal preferences. If one fails to behave according to the outcome of one's adequate reflection, it is a case of weakness of will. Coherent and considered preferences are those preferences for which attitude and behaviour are

[219] 'Justice, Social Union and the Separateness of Persons'.
[220] 'Justice, Social Union and the Separateness of Persons', pp 93-94
[221] See 'Rationality, Value and Preference', pp 37-40
[222] See also J. Fishkin, 'Bargaining, Justice and Justification', p 54-55 for a similar point. Fishkin calls it the indoctrination problem.

coherent. Secondly, he insists on the self-critical dimension of practical rationality. Rational persons subject their preferences to reflective assessment. Indoctrinated or brainwashed people have preferences which are not autonomous and therefore cannot be 'considered' in Gauthier's sense.[223]

Can we be satisfied with such a defence? I am not. I believe that we are all brainwashed whether by the media or by our politicians or religious leaders. What is 'rational' in France can be completely 'irrational' in Korea or in Jordan or even in the United States[224]. Real individuals are framed by their language, social background or political system. It is an illusion to believe that their preferences, even considered, are autonomous. At best we can be aware of our lack of autonomy but, even if we are, we meet two difficulties. The first one is to be aware of the full extent of the influence exerted over us. The second one is to exercise enough imagination to step back and reflect on what we really want.

In earlier days, Gauthier acknowledged that social contract theory was framed in Western ideology.[225] If he can accept that we are influenced by our cultural and historical background, how can he claim that we can have autonomous preferences and be able of self critical assessment? Now, if he believes that it is possible to educate preferences so our emotions and feelings can enhance co-operation, his theory needs to be amended.

[223] 'Morality, Rational Choice and Semantic Representation', p 191-195
[224] Have you ever tried to explain French labour laws to an American citizen? I have. She thought we were mad.
[225] See 'Social Contract as Ideology'.

Section 2: The market as a morally free zone

This core conception is marginal in Gauthier's theory and, as such, it has been hardly commented upon. The main comments were about the following ambiguous quote from *Morals by Agreement* that caught our attention in chapter II:

"The absence of force and fraud is essential to the workings of the market. Before Smith's invisible hand can do its work, Hobbes's war of every man against every man must first be exorcised. And this, as we shall see, means that the ideal of free interactions which Smith celebrates is not natural but artificial, arising, for rational persons, only within a framework of agreed constraints. In understanding the perfect market as a morally free zone we shall be led back to its underlying, antecedent morality."[226]

Most commentators who showed an interest in this core conception noted this ambiguity. If the market is a morally free zone, how are we to understand that it has an underlying antecedent morality? The answer was unanimous: the market is not a morally free zone.[227]

One could defend this quote as follows. Gauthier had just explained that there is no such thing in real life as the perfect market. What Smith describes as the perfect market is only what Gauthier calls cooperative interactions. The perfect market being unrealistic, the nearest approximation or realistic simulation is agreed co-operation. Gauthier's subsequent theory is dedicated to explaining how such an agreed co-operation possesses 'underlying antecedent morality'. The last sentence could therefore read, 'in understanding the failure of the perfect market to exist and function as a morally free zone, we shall be led back to understanding how morality needs to be introduced to allow the existence and the functioning of its nearest simulation (i.e. agreed co-operation).' The above was my initial understanding of it.

Nevertheless, on closer examination, this interpretation does not hold. Gauthier distinguishes between market activity and co-operative interactions implying that the market regulates private goods transactions and agreed co-operation regulates public

[226] *Morals by Agreement* p 85.

[227] See for example P. Danielson's 'The Visible Hand of Morality', p 367 who claims that at best, the market can be described as a co-operation free zone, D. Braybrooke's 'Contract Theory's Fanciest Flight', p 757-760, K. Baier's 'Rationality, Value and Preference', footnote 2, p 18 or J. Buchanan's 'The Gauthier Enterprise', p 89-91.

goods' ones.[228] Such a distinction is not sustainable. Any transactions whether on private or public goods is supported by rules, laws and institutions. "Individual rights must be guaranteed; contracts must be enforced; fraud in exchange must be prevented...market relationship offers the exemplar of rational morality, rather than a morally free zone".[229]

The social contract must apply to all types of transactions. In keeping private goods out of the social scope, two issues are left unattended. Firstly, Gauthier does not provide a principle for dividing the utilities obtained by agreed co-operation from utilities obtained by the free running of the market. Secondly, if co-operation is constrained, utilities received through the market might increase, thus raising the issue of distributing market's benefits.[230]

Gauthier rapidly seized the problem. In 1988, he wrote

"I now think that choice of the market, as of other determinate principles and institutions, cannot be shown to follow from a consideration simply of the reasoning of **the Archimedean chooser**, but depends on information about the determinate effects of differing possible ways of organizing social and economic life. She **views society as "a single co-operative enterprise", but in so viewing it, she does not view the market in isolation** as I supposed her to in chapter. VIII.4.3."[231]

Having conceded that much, Gauthier is still reluctant to completely let go his concept of the market as a morally free zone. In replying to Buchanan, he is very vague about the extent of his concession. He only concludes:

"For we may suppose that morality replaces the forcibly imposed external constraints ... by voluntarily accepted internal constraints that ... open up an area, albeit limited, of genuine freedom, this area being of course the market... He [Buchanan] and I can agree that it is unfortunate that so much recent thought and practice has been devoted to inventing new ways in which pervasive and overriding political direction can be alleged to be necessary after all."[232]

This passage is difficult to understand. I believe that the only sustainable interpretation of Gauthier's position about the market is now as follows. He accepts that agents must internalise some constraints in order to interact successfully on the market. From this

---

[228] See for example 'Moral Artifice', p 394
[229] J. Buchanan' s 'The Gauthier Enterprise', p 89
[230] These issues were raised by D. Braybrooke in 'Contract Theory's Fanciest Flight'. The first one is on p 757 and the second one on p 758
[231] 'Moral Artifice', p 414
[232] 'Morality, Rational Choice and Semantic Representation', p 204

point of view, market interactions are no different from co-operative ones. Internalised - and therefore moral - constraints must back the market. However, in the absence of these moral constraints, political solutions have been put in place. In any case, **the market can no longer be considered as a morally free zone**.

Section 3: Theory of bargaining and Minimum Relative Concession Principle

In this section, I will give an overview of the objections raised against the minimum relative concession principle (MRC hereafter) prior to developing on Gauthier's latest view. Gauthier's bargaining theory has left many commentators perplexed. Vallentyne argues that it is best understood as a consensus theory[233]. I personally believe that it is best understood as a bargaining theory but between constrained maximisers. I develop this argument in chapter IV. Most of the objections below hardly moved Gauthier. The only one that eventually made him surrender was the attacks from 'orthodox' game theorists as exemplified by K. Binmore.[234]

MRC and fairness

Many commentators have objected that the MRC relied on a pre-conception of fairness. We raised that issue in chapter II §2, CC2. Hampton argues as follows:

"The mushiness of our intuitions on these matters gives us good reason to turn to a contractarian methodology in the hopes that it will persuasively single out a unique distributive principle. Gauthier insists that this methodology singles out the MRC principle, but my argument against the MRC principle suggests that this contract language only disguises an implicit appeal to our intuitions about fairness.... Depending on what intuitions one has, one will prefer a certain method of reward, and it is easy to imagine (and define) contractors in a hypothetical contract situation who share these intuitions and preferences. But such imagining hardly counts as proof."[235]

Gauthier's reply on this issue is rather blunt: "MRC may or may not have intuitive appeal, but its defence is to be found not in moral intuitions, but in an analysis of rational bargaining."[236] Is MRC's defence to be found in an analysis of rational bargaining? I am not the only one to think that it is not.

As argued above, his conception of rational bargaining is already morally loaded. Gauthier explains: "[a]ny distribution of the surplus is advantageous, and *an equal distribution reflects their equal bargaining power in reaching agreement.*"[237] Gauthier assumes (or even ensures) that the parties to bargaining have equal bargaining power

---

[233] See 'Gauthier's three projects', footnote 9 p 5
[234] A. Rubinstein's 'Perfect Equilibrium in a Bargaining Model' and K. Binmore's 'Bargaining and Morality'
[235] 'Can We Agree on Morals?' p 338. J. Mendola makes a very similar comment in 'Two kinds of Rationality', p 772-773.
[236] 'Moral Artifice', p 394
[237] 'Moral Artifice', p 392

and MRC naturally follows from it. However, to assume equal bargaining power is to skip a full aspect of bargaining, it is to miss out on what makes bargaining so unpredictable. More than anything it is contrary to reality. To be mutually advantageous does not necessarily imply that needs are identically reciprocal. As we saw above, the division of the surplus depends on many factors such as the bargainers' particular circumstances, bargaining skills and/or respective concession thresholds.

Similarly, P. Danielson claims: Gauthier "defends a principle of rational bargaining, against the widely held conviction that bargaining situations admit of no strictly rational solution."[238] There is very little rationale involved in how bargainers split the surplus. G. Harman argues in the same direction. He claims that in identifying MRC as the unique bargaining solution, Gauthier probably means that MRC is the most salient solution. If it is so, he makes two points. Firstly, saliency appeals to psychological rather than rational factors. Secondly, what is salient to one person might not be salient to another. In any case, MRC is not determined by the rationality of the participants, contrary to Gauthier's arguments.[239]

About MRC and fairness, Kraus and Coleman also raised a very different issue. They argued that unfair bargains are not always unstable and that some unfair bargains are rational.[240] We shall come back to that point later when discussing the proviso.

MRC and the number of players.

MRC is perfectly meaningful in two-person interactions, but problems arise once the number of persons interacting increases. We can identify three categories of problems: alternative coalitions, non cooperative behaviour and information required.

About alternative coalition

J. Hampton first raised the issue of alternative coalitions: "the rational solution to a bargain depends not only upon how much each could get by herself but also upon how

[238] 'The visible Hand of Morality', p 359
[239] G. Harman, 'Rationality in Agreement', p 8
[240] See 'Morality and the Theory of Rational Choice', p 726 -730. Harman makes a similar point in 'Rationality in Agreement', p 11 - 12

much each could get if she joined alternative coalitions."[241] Why should individuals be prevented from forming alternative cooperative coalitions? Why should there be only one societal investment opportunity for each individual? Such a restriction cannot conceal the fact that even within one society people have multiple cooperative opportunities.

Gauthier admits that in general, in multi-person interactions, there is a problem in determining the appropriate cooperative infrastructure.[242] However, he claims elsewhere that usually community bonds flourish naturally (historically or otherwise).[243] He also admits that the contribution of each person to the overall cooperative surplus is not always easily discerned. Although he acknowledges the difficulty of applying MRC to multi-person interactions, he sees "no reason to doubt that an appropriate multi-stage application of MRC offers a satisfactory resolution."[244]

About non cooperative behaviour

Gauthier dismisses as easily Buchanan's double objection regarding the set of players and non co-operative behaviour. Firstly, Buchanan notes that co-operative behaviour is not all inclusive. Two firms could co-operate and create a duopoly to the disadvantage of consumers.[245] Similarly, we know that rich countries co-operate and do fix the price of commodities to their advantage. Their co-operation is mutually advantageous but detrimental to poorer countries. Gauthier answers: "Given that several persons can, by interacting, achieve a cooperative surplus, cooperative behaviour has value for them. That it may have disvalue in relation to other persons need not be worrying."[246] He continues by reminding us that, after all, it is not advantageous to society if the prisoners co-operate in the famous prisoners' dilemma. Once again, Gauthier is not bothered with the normative significance of his theory neither does he worry about its compatibility with conventional morality.

---

[241] 'Can We Agree on Morals?', p 342
[242] 'Moral Artifice', p 397
[243] 'Morality, Rational Choice and Semantic Representation', p 204-205
[244] 'Moral Artifice', p 398
[245] 'The Gauthier Enterprise', p 76-79
[246] 'Morality, Rational Choice and Semantic Representation', p 204-205

Secondly, Buchanan argues (but does not demonstrate) that the bigger the number of players, the less likely the prospect of co-operative behaviour. "In order for the Gauthier rule to dictate continued adherence to a cooperative strategy as numbers in the interaction increase, the probability of any one player adopting the co-operative strategy must increase, which seems to counter common-sense notions about the way persons behave."[247] To this argument Gauthier responds that community bonds flourish naturally (historically or otherwise). Within these boundaries, the theory does not require universal but only widespread compliance. If the moral practice is mutually advantageous, widespread compliance is realistic to expect.

Personally, I am absolutely not convinced by Gauthier's answer. He continues his argument as follows:

"If tax revenues are wisely spent in the protective and productive ways suggested in Buchanan's *The Limits of Liberty*, then each may expect to do better voluntarily paying his taxes in circumstances in which most of his fellows do the same, than in a society in which everyone pays only what is coercively extracted from him."[248]

Given the scale of a country (i.e. involving a large number of individuals), the likelihood of all (or even a majority) agreeing with the ways the tax revenues are spent is already rather slim. The likelihood of a majority paying voluntarily is even slimmer. The more people involved, the more anonymous one is and the more marginal one's contribution is with regards to the big total. If one knows that one's defection will go unnoticed, one will defect in the absence of an enforcement system. I can only agree with Buchanan that numbers play against co-operation. This objection is obviously relevant to discussions about resolute choice.

About the realism of MRC

In addition to the above, a couple of philosophers have objected to the realistic application of MRC. For example, R. Hardin demonstrates that MRC is not applicable to a real economy. Co-operation has a cost which is not taken into account in MRC. The larger the society, the more difficult it is to distinguish contributions, cooperative surplus and a fortiori shares of it in proportion to contributions. Similarly, very few individuals can make a difference in their social contributions and therefore nobody can

[247] 'The Gauthier Enterprise', p 80
[248] 'Morality, Rational Choice and Semantic Representation', p 206

really threaten to withdraw from social interaction at a cost to the others.[249] Gauthier's bargaining theory is not applicable to real society.

D. Braybrooke's comment is in the same vein. He notes that Gauthier makes a fantastic demand on information which no contracting parties or even government could ever meet. Gauthier's agents are required to identify the social structure under which they would reach their fullest personal development and he wonders how they would ever be able to collect and compute the information needed to make such a choice. In any case, he notes that MRC certainly does not enable us to make such a choice.[250]

MRC and 'orthodox' game theory

We know from above that, in raising the issue of multi-person interactions, J. Hampton had opened a first breach in MRC. In 1991, two game theorists[251] made a second indent on MRC. They demonstrated that for the general n-person case, MRC should be replaced by a lexicographic principle which can be formulated as a maximin principle in relative utility gains or a minimax relative principle in relative concession. Gauthier admits that their principle meets the formal inadequacy of MRC for multi-person bargaining, but he is aware that this formulation, like his own, "ignores the structure of the interactions by which what is distributed in bargaining has been produced." He acknowledges: "**I am now convinced that MRC needs at least some modification.**"[252] We have to wait until 1993 to read that MRC could not stand in its present form. Let us go back a few years to understand why.[253]

In the 1950's, Nash developed two approaches to solve the bargaining problem. The two approaches, given specified conditions, converge towards the same bargaining solution. Together they form what is known as the Nash Program. The bargaining problem can be described as follows. Two individuals have to choose between several possible contractual agreements. Both have an interest in reaching an agreement and

---

[249] R. Hardin's 'Bargaining for Justice', p 66 - 70
[250] D. Braybrooke, 'Contract Theory's Fanciest Flight', p 760
[251] W. Gaertner and M. Klemish-Alert, 'Gauthier's Approach to Distributive Justice and Other Bargaining Solutions'.
[252] 'Rational Constraint: Some Last Words', p 325
[253] See the Appendix, section 1, b

they are both assumed to behave rationally. The question to answer is what contract will they agree on?

According to the **axiomatic approach**, the bargaining process does not need to be provided. However, the game is described by a set of possible outcomes. "One states as axioms several properties that it would seem natural for the solution to have and then one discovers the axioms actually determine the solution uniquely."[254] In the **strategic (or non-cooperative) approach**, the bargaining process of the players is specified. Do players make proposals in turn or simultaneously? What is the cost of time? ... Once the negotiation situation is fully described, it can be modelled as a 'non-cooperative' game. This game is then analysed using standard game theory tools.

Gauthier's bargaining theory is axiomatic. It does not specify the bargaining process and therefore avoid the problems inherent to the strategic approach. However his solution is highly contested by game theorists. Gauthier's solution stands alone and is not backed by the strategic approach. K. Binmore argues that Nash axiomatic solution can be defended *only because* it is securely grounded in non-cooperative game theory. The solutions generated by the Nash axiomatic approach can also be generated by the strategic approach, given plausible assumptions about the negotiations process.[255]

Such defence cannot apply to Gauthier. Not only, Gauthier does not describe the negotiations process, but his theory departs so unrealistically from the standard game theory that no non-cooperative model could ever describe it.[256] Binmore lists the departures from the standard theory that would prevent any strategic modelling.

- Gauthier relies on the out-dated Zeuthen principle. Most game theorists consider it an *ad hoc* criterion no longer used in strategic or axiomatic approaches.
- Gauthier assumes *equal rationality* of the bargainers. Binmore notes that such an assumption amounts to assuming that bargainers choose in equivalent if not identical circumstances. This assumption is unrealistic and difficult to describe in a non-cooperative model.

---

[254] Nash, 'Two-Person Cooperative Games', p 129
[255] See 'Bargaining and Morality', pp 134 and 141-142.
[256] See 'Bargaining and Morality', especially pages 134 to 140.

- The concept of *constrained maximisation* departs from the "most fundamental principle of non-cooperative game theory – namely, that players will not use a strongly dominated strategy."[257] Defection in the Prisoner dilemma is almost a tautology in game theory. In demonstrating why, Binmore raises the two following corollary points.
- Firstly, he notes that constrained maximisation is made possible by the assumption of *translucency*. Binmore does not see the point of examining a model containing such an idealised assumption.
- Secondly, he notes that Gauthier attributes players with the *power to make commitments* in a game. In doing so, Gauthier disregards another game-theoretic tautology; commitments in game theory have to be modelled formally as moves within the games, *providing that a convincing enforcement mechanism is in place and described.*

Most of the above became clear to Gauthier when he heard a game theorist called Ariel Rubinstein at a conference. If he is 'unmoved' by Binmore's attacks against his concept of constrained maximisation[258], he understands the importance of the strategic approach.

Ariel Rubinstein wrote in 1982 an article entitled *Perfect Equilibrium in a Bargaining Model.* In this article, he shows that Nash solution is the solution to non-cooperative games in some fairly general cases. Based on simple bargaining rules, Rubinstein's theory offers a minimalist defence of the Nash bargaining solution.

Despite his reservations about Nash's independence of irrelevant alternatives axiom and about the rejection of the Zeuthen principle, **Gauthier no longer wishes to defend MRC** and accepts that what he said in *Morals by Agreement* about bargaining "is simply undercut by the non-cooperative approach."[259] He admits: "The real work of defending MRC as a bargaining outcome ... requires arguing that in the circumstances of the social contract, MRC coincides with the Nash bargaining solution."[260]

---

[257] 'Bargaining and Morality', p 136.
[258] 'Uniting Separate Persons', p 186
[259] 'Uniting Separate Persons', p 177
[260] 'Uniting Separate Persons', p 178

Section 4: Constrained maximisation and Resolute Choice

In this section, we will first review the main objection to this core conception prior to turning to McClennen's concept of resolute choice. Gauthier has not officially abandoned constrained maximisation. However, since 1988, he has worked almost exclusively within the framework of McClennen's concept of resoluteness using as well Bratman's research on plans and intentions, and Parfit's work on irrationality.

Introductory objection.

The main objection to constrained maximisation is about the gap left open by Gauthier between disposing oneself to comply and acting upon such a disposition. Harman is amongst the first to spot this weakness:

"Although I agree that under idealised conditions, it could be rational to dispose oneself to be a conditional co-operator and then rational actually to act from that disposition, I do not agree with Gauthier's assumption that, necessarily, if it is rational to acquire a disposition to do D in circumstances C, then it is rational actually to do D in circumstances C. This assumption is not universally true."[261]

The problem is that in the context of Gauthier's contractual morality, dispositions must be irrevocable.

Gauthier does not demonstrate that if it is in our interest to *acquire* a disposition to do D in circumstances C, it follows that it is in our interest to do D in circumstances C. It can be rational for me to acquire the disposition to comply at date *t*. At *t*, I am genuinely committed to comply with our agreement. However, at the time of acting *t+1*, I might not be prepared to act upon my disposition. Indeed, it might be irrational for me to do so. Let us review a few illustration cases (the list below is not exhaustive).

Case 1: Change of situation

The most usual case is a change of situation that renders compliance irrational. I agreed to meet you at a date t, but I had to accept an interview for the job of my dream on the

[261] 'Rationality in Agreement', p 5. H. Smith calls this assumption the *causal efficacy thesis*: 'the thesis that forming an intention to do A will cause the performance of A.' (See 'Deriving Morality from Rationality', p 235). Like Harman, she argues that Gauthier has not proven this assumption true.

same day. I could not foresee this change of circumstances and I have to cancel our appointment.

Case 2: Change of perception

The *perception* of the situation can change between *t* and *t+1*. At the time of acting *t+1*, I don't see the agreement in the same light as I did at *t* when I disposed myself to comply with it. To act upon my original disposition appears irrational to me *now*. Let us illustrate this case with a classical example.

My car breaks in the middle of the desert. You pass by with a car in good condition. You offer to give me a lift out of the desert on the condition that I commit myself to giving you a generous lump sum when I can access my bank account. In the middle of nowhere, without water or food, I see you as my saver and genuinely commit myself to giving you the money. I believe my life is worth the amount you want and I am disposed to pay you. However, once out of danger, after I have drunk and eaten, my outlook is different. I think that after all, someone else could have passed by and helped me. I don't see you anymore as my saver. I wonder how I could have committed myself to give you such a ludicrous amount of money for such an insignificant favour. Gauthier claims that I must give you the money as promised but he has not demonstrated that it is rational to do so.[262]

Case 3: Exception to the best strategy

The strategy 'comply with agreement' has been beneficial until now and I have genuinely acquired a disposition to comply. However, I am now in a situation where adhering to the strategy chosen and therefore complying is irrational. The loss that I would incur were I to comply would more than offset all the past (and future) benefits made (to make) with the strategy.[263]

[262] H. Smith labels this attitude the "rationality of perseverance" principle and notes that Gauthier offers no positive argument in favour of this principle. See 'Deriving Morality from Rationality', p 244.
[263] D. Parfit illustrates such a situation with the case of a *threat ignorer* who can be killed if he ignores the next threat. See 'Bombs, Coconuts, or Rational Irrationality' p 82-83

This case is obviously not considered by Gauthier. He has explicitly excluded repeated Prisoner's Dilemma situations.[264] However, as Harman notes, we seldom gain confidence in one another in one shot Prisoner's Dilemmas. The disposition to cooperate normally comes into play in repeated prisoner's dilemma situations where co-operation can be defended by an appeal to long term self interest.[265] Again Gauthier excludes that type of justification.[266]

This attitude towards iterated games is particularly troublesome in threat cases. Is it rational to carry out a threat if it was rational to make it? It can be rational to make a threat and it is usually rational to carry it out when one want to be taken seriously in future threats. The cost encountered in carrying out the threat then is balanced out by future gain prospects. This justification does not apply to non iterated games.

Case 4: Deception and other alternative rational strategies

CM and narrow compliance are not uniquely rational for equally rational agents.[267] However, knowing that Gauthier has given up on the assumption of equal rationality, these objections should not worry us. More preoccupying are the objections in favour of deception. The Fool, who was shown the door, comes back by the window.

At *t*, it is rational for me to commit myself to comply in order to gain others' trust and compliance but at *t+1* it is not rational for me to comply. I can foresee at *t* that I will defect at *t+1* but I decide to try and deceive others, to appear as a CM when I remain really a SM..

Gauthier argues that I am not able to acquire the disposition to comply if I know that I will not comply at the time of acting. However, there are two objections to his argument. Firstly, agents are translucent and not transparent. Translucency allows for a degree of error which Gauthier has not catered for. He argues that rationality requires developing a skill to spot SM but one could argue that rationality equally requires

[264] See *Morals by Agreement*, p 169-170
[265] See *Rationality in Agreement*, p 5
[266] See *Morals by Agreement*, p 169-170
[267] See Danielson's 'The Visible Hand of Morality', pp 373 – 383 in which he suggests 'Reciprocal Cooperation' as an alternative strategy that dominates CM. See also Kraus and Coleman's 'Morality and Rational Choice', pp 736 – 745. They argue against the unique rationality of narrow compliance.

developing the skill of deception, pretending to be a CM when one is really a SM.[268] As a corollary to this argument, one notes that in order to become a CM, individuals must carry out a SM's reasoning. The decision to acquire a CM's disposition is the outcome of a SM's calculus.[269] Where is the threshold? It is now time to turn to McClennen.

McClennen's friendly alternative to constrained maximisation

McClennen's objections to constrained maximisation are quite lethal. However, one must not be mistaken: McClennen fully supports Gauthier's project of building a rationally based contractarian theory of morality.[270] His concept of resolute choice fits such a project. As we shall see in chapter VII, McClennen's objection actually opens a new door for Gauthier's project.

- *McClennen's objections to constrained maximisation.*

Firstly he notes that, in demonstrating that it is rational to acquire a disposition to cooperate[271], Gauthier assumes that agents are transparent. In his calculus, agents know with certainty who is a SM and who is a CM and therefore calculate their expected utility accordingly. Indeed, the calculus provides transparent agents with a compelling reason to become CM.

However, McClennen notes that the same reasoning applied to mere translucent agents could fail to demonstrate the rationality of constrained maximisation. The person who is willing to adopt the disposition to cooperate will face a trade-off between expected gains when cooperating with like-minded people and expected losses when interacting with SMs in disguise. Whether one does better in acquiring a disposition to co-operate depends then on

"one's estimate of the proportion of deceptive SMs to CMs in society, the relative frequency of one's encounters with members of each group, the probability of being mistaken about whether a given person is a deceptive SM or CM, the magnitude of the

[268] See G. Sayre-McCord, 'Deception and Reasons to be Moral' in which he develops the 'trans-opaque' strategy: SM send all sorts of misleading signals to be trusted as CM.
[269] See for example McClennen's 'Constrained maximisation and Resolute Choice', footnote 8, p 98.
[270] See 'Constrained Maximisation and Resolute Choice', p 108-109
[271] See *Morals by Agreement*, p 172

gains and losses in each case, *and* one's estimate of one's ability to effectively play the role of a deceptive SM."[272]

This quote reveals the full extent of the weakness of constrained maximisation. There are too many factors to take into account in such a calculation for it to be straightforwardly rational. McClennen includes as well the advantage of deception in such a calculus reminding us that a future CM first reasons like a SM.

In view of the numbers of uncertain factors involved, McClennen prefers to assimilate the decision problem to become or not a CM to a coordination problem. Like Parfit[273], McClennen concludes on this first point that when agents are translucent and not transparent, the best strategy remains to appear to develop a disposition to co-operate and to continue to choose in a strategic manner. Obviously if everybody does the same, co-operation never takes root. Gauthier fails to demonstrate the rationality of co-operation for translucent agents and instead feeds the Fool's argument in favour of deception.

McClennen second argument against constrained maximisation is a lot more subtle. Like other critics mentioned above, McClennen notes the divorce between the disposition and acting upon the disposition. The so-called divorce, he claims, can be traced back to the difference in reasoning in two choice situations - namely *ex-ante*, when choosing a disposition, and *ex-post*, at the time of action. Ex-ante, transparent agents expect higher returns from disposing themselves to co-operate whereas, ex-post, their rational choice of action is to behave like SM. Foreseeing this discrepancy, transparent agents are unable to acquire the disposition at the first place. Being transparent, their lack of disposition will be revealed to others and co-operation will be forgotten.

The reason for such a divorce, argues McClennen, is that in parametric reasoning, a decision made at t to adopt a disposition can have no carrying power over time since the outcomes still available at $t+1$ are the same as they were at t. I cannot dispose myself at t not to choose a certain outcome A at $t+1$, when I know already at t that A not only will still be available at $t+1$ but will also be the most rational outcome to choose then.

---

[272] 'Constrained Maximisation and Resolute Choice', p 103
[273] See 'Bombs and Coconuts, or Rational Irrationality'.

McClennen has now set the context and can safely introduce his concept of resolute choice.

- *McClennen's concept of resolute choice*

In the Prisoners' Dilemma scenario, we all know that to co-operate is the best mutual outcome but the traditional account of rationality leads each agent to defect. In doing so, they realise their worst mutual outcome. In knowledge of their inherent rational weakness, agents have come to put in place costly surveillance and enforcement devices in order to secure mutually advantageous co-operation.

McClennen's key argument is then as follows: if rational individuals are willing to tie themselves using costly devices in order to secure mutual benefits, why can't they motivate themselves to secure these very benefits free of charge. They would achieve the same benefits at lower costs, i.e. they would make even greater benefits. As expected utility maximisers, such a prospect can only appeal to them.

This pragmatic approach bears two consequences on preferences. Firstly, it means that these 'context dependent preferences'[274] for actions are rooted in the disposition to seek greater benefits. Preferences for co-operation are therefore endogenously derived. Secondly, it means that agents have a holistic rather than an incremental outlook. They choose an outcome that maximises their over-arching preferences rather than an outcome that maximises their preferences at each stage of the decision process.

"In place of Gauthier's model of an agent who, it seems, must reason parametrically at each point in time, one can consider a different model altogether, predicated on the notion that the agent is a being who continues over time, with concerns that have some continuity to them. Such an agent can be understood to view himself as deliberating over alternative plans..., as choosing some particular plan, and then proceeding, at least in the normal course of events, to make specific choices (at different points in time) that serve to execute or implement the plan chosen. What is characteristic of such an agent is that his *ex-post* preferences among available actions are disciplined or shaped by what he judges, from the perspective of plans taken as wholes to be the best plan to pursue... Such an agent can be described as a resolute chooser."[275]

---

[274] 'Constrained Maximisation and Resolute Choice', p 110-111
[275] 'Constrained Maximisation and Resolute Choice', p 111-112

To illustrate this point, let's take an (extensively used) example, borrowed from Elster[276]. Ulysses knows that he has to pass by the sirens' islands on his way to Ithaca. His preference is not to stop and go straight home. However, he knows that if he hears the sirens he will prefer to stop on their island. Knowing that, Ulysses can decide at the beginning of his trip that he will bind himself to the mast in order to make sure that he will pass by the island without stopping. *Myopic* Ulysses resolves to sail by the island but succumbs to his preference to listen to the sirens when he passes by their island. *Sophisticated* Ulysses binds himself to the mast and sails straight home. *Resolute* Ulysses resolves not to stop on the island and does not stop.

- *Gauthier's response.*

Gauthier very rapidly accepts the objection:

"McClennen's account of resolute choice is more systematically developed than my conception of constrained maximisation and it has wider applicability. In embracing straightforward maximisation for individual decision making, I have, as I now realise, ignored preference change and I should want to extend the scope of constrained maximisation."[277]

From this 1988 quote, we can see that Gauthier has taken the full measure of McClennen's theory's potential. 1993 is probably the last year he mentions constrained maximisation and it is already within the resolute choice framework.

"I remain convinced that this [constrained maximisation] remains the most fruitful idea in *Morals by Agreement*". Indeed, he will remain faithful to the core idea of morality as self-imposed constraints but he continues:

"A constrained maximiser, as I shall use the term in a way that generalises somewhat from my use in *Morals by Agreement*, is someone who takes her reasons for acting, not only directly from the utilities of possible outcomes she may bring about, but also from her plans and commitments."[278]

The seed is planted and the tree will not stop growing from then onwards. As we shall see in chapter VII, Gauthier will develop this theme extensively, trying to save what can be considered as the very heart of *Morals by Agreement*. Gauthier remains convinced that rational morality is viable.

---

[276] *Ulysses and the Sirens: Studies in Rationality and Irrationality.*
[277] 'Morality, Rational Choice, and Semantic Representation', p 209.
[278] 'Uniting Separate Persons', p 185-186

Section 5: The Lockean Proviso

As we saw in chapter II, the rationale of the proviso is forward looking. The prospect of benefiting later from co-operation and society makes it rational for individuals to constrain their behaviour in the pre-bargaining position. The argument is two fold. Firstly, Gauthier argues that individuals would not be accepted as parties in bargaining if they were not willing to comply with the proviso requirements. Secondly, he argues that people would not comply with a bargain that started from an unfair initial position. In ensuring the fairness of the initial position (and therefore contributing to the fairness of the bargaining process) the proviso secures individuals' future compliance to the agreement so obtained.

Two aspects of this argument have been contested. It has been argued that Gauthier's Lockean proviso is not necessarily fair or that it can generate an unfair starting point. It has also been demonstrated that it can be rational to comply with a bargain starting from an unfair initial point.

Prior to developing both these arguments, I would like to bring to the fore an objection. A. Baier wrote a rather violent diatribe against *Morals by Agreement* which I find mostly unfair. However, she noted something about the proviso that did strike me too when I read chapter VIII. Gauthier is rather silent about people's *social* endowment: "It is as if the adolescent Emile is to suffer amnesia, forget his parents, wet nurses, and tutor, perhaps to see himself as sprung fully factored from the head of Zeuthen-Nash."[279] What about education, social or historical background and other social inputs? Is education considered as an externality? We mentioned similar issues above in section 1 (about mutual unconcern and the circularity problem of considered preferences). However, this omission from Gauthier is unexpected in a chapter fully dedicated to the initial bargaining position.

Proviso and forward looking rational defence

Gauthier suggests a no-coercion baseline: whatever has been obtained through coercing others or being coerced by them should not be on the bargaining table or available for

[279] 'Pilgrim's Progress', p 320

redistribution. He argues that if coercion is built into bargaining, parties will not *voluntarily* comply and costly enforcement will have to be put in place. In removing coercion and unfairness from the starting point, Gauthier's revised proviso ensures voluntary compliance and reduces enforcement costs. The proviso's *rationale* is therefore forward looking.

Two arguments are often suggested against this claim. Firstly, it is argued that it is not always irrational to agree to and comply with an arrangement that does not satisfy the proviso. Secondly, this claim seems to rest on moral rather than rational grounds. Let us develop both these arguments.

Which irrationality?

Is it really irrational to voluntarily comply with a joint strategy agreed to from a coercive position? After all bargaining is about distributing a surplus. If the surplus is big enough and allows each to make a big enough gain from the status quo, than it does not matter whether the status quo was coerced or not. If the parties do gain from the bargain, they will voluntarily comply. Departures from the no-coercion baseline do not necessarily entail enforcement costs, at least as far as purely rational individuals are concerned. So on what basis does Gauthier justify the irrationality of coercion?

Gauthier equates the irrationality of coercion with the irrationality of unproductive transfers.

"An unproductive transfer brings no new goods; it simply redistributes some existing goods from one person to another. Thus it involves a utility cost for which no benefit is received, a utility gain for which no service should be provided."[280]

Gauthier claims that, by nature, unproductive transfers cannot be voluntary and can only be obtained through coercion. As such they cannot be rational.

'False!' reply some objectors. Unproductive transfers can have a rational function and therefore be rational. Indeed they can secure the stability of the bargain. For example, when the bargaining situation of the parties is not equal, a transfer that is apparently

[280] *Morals by Agreement*, p 197

unproductive can stop the less favoured party from attacking the most favoured one.[281] More generally, what Gauthier calls an unproductive transfer can merely balance out a structural threat advantage. A transfer enables one to offset the threat advantage of another and bargain on a more equal footing with her.

Rationality, psychology or morality?

"Why should coercion being built into the baseline necessarily diminish the value of bargains based on it, in the eyes of purely rational egoists?" asks R. Goodin. [282] Goodin's answer is rather straightforward: "what is really at work here is an argument not from irrationality of coercion but rather from its immorality... It is moral offence, not rational miscalculation, that drives the model here."[283] Goodin sees two moral inputs in the proviso.

Firstly, Goodin argues that people must see themselves being coerced when Gauthier sees them being coerced. Resentment and resistance to coercion are psychological rather than rational factors. Human beings resent and resist coercion not because it is irrational but because it is immoral. The argument for the cost of enforcement rests on moral rather than rational factors.[284]

Secondly, Gauthier distinguishes between worsening someone's situation and failing to better it. The proviso prohibits the former not the latter. Goodin notes that such a distinction is itself morally loaded. It has no appeal to a rational egoist. Why should a rational person make a difference between worsening someone's situation and failing to better it? It seems that Gauthier merely imposes his own conception of fairness as a starting point.

---

[281] See G. Harman's 'Rationality in Agreement', p 11

[282] 'Equal Rationality and Initial Endowments', p 125

[283] 'Equal Rationality and Initial Endowments', p 126. G. Harman makes a similar claim but does not develop it (see *Rationality in Agreement*, p 11).

[284] Although I agree with most of Goodin's argument, I believe that Gauthier's point is not that resentment is rationally derived. His claim is that it is based on a feeling of unfairness that Goodin calls moral offence. This resentment then *generates* defection and resistance to the agreement. In knowledge of such a process, Gauthier argues that it *becomes rational* to remove unfairness from the baseline in order to secure future compliance.

Gauthier writes: "**I agree that the proviso is a moral premise**. The question then is whether it is rational to accept it as a constraint on bargaining to the social contract."[285] Gauthier is aware that he inputs a conception of fairness in the proviso. The forward looking rational of the proviso depends on it. Unfortunately, as we shall now see, his conception of justice is far from having gained general acceptance.

Proviso and fairness.

Is Gauthier's revised Lockean proviso fair or does it secure a fair starting point to bargaining? According to most commentators on the topic, the answer is no. Through his revised Lockean proviso, Gauthier artificially attempts to rectify the unfairness of original or natural threat advantages. His attempt is not only *procedurally* unfair; it is also *structurally* unfair either because it does not preclude coercion in the initial bargaining position or because it can partially affect the bargaining outcome. Knowing that Gauthier's initial position was built up in reaction to J. Buchanan's one, it seems appropriate to open this sub-section with Buchanan's response.

Procedural fairness

Buchanan deals with Gauthier's example of masters and slaves very bluntly. If the slaves are in this situation, it is "only because of some inability to enforce more favourable terms of existence."[286] More generally, if one coerced another in the pre-bargaining situation it is because one is stronger or better (in a way or another) than the other. Why should it be rational or even fair to rectify the existing pre-bargaining situation? Buchanan's core argument is that any rectificatory redistribution involves violation of the contractarian or agreement based criteria for fairness.[287]

Gauthier rectifies the existing pre-bargaining situation by using a secession criterion. He tries to establish what would have been the situation of each of the bargaining parties if they had been living separately on isolated islands. The use of this criterion to correct past injustices does not only seem unrealistic, it is also counter-intuitive. Buchanan

[285] 'Uniting Separate Persons', p 183
[286] 'The Gauthier Enterprise', p 84
[287] 'The Gauthier Enterprise', p 87

argues that even in a coerced situation, individuals are likely to perform better as part of a group rather than in isolation.

Buchanan does not claim that it is wrong to investigate the several historical stages of a pre-bargaining situation. He only claims that it is against the contractarian approach to label as unfair some past stages and to rectify past injustices. Any process of rectification of the existing pre-bargaining situation is procedurally anti-contractarian. "If fairness criteria have been violated at earlier stages of the process that generated that which exists, do these historical violations in themselves offer justification for violations in some process of rectification?"[288]

Gauthier is unmoved: "failure to rectify past injustice destabilises present arrangements."[289] For such an argument to remain valid, Gauthier's revised proviso must rectify all past injustices and establish a fair starting point. We shall now see that he fails to do so.

Structural fairness

Indeed, many critics have argued that Gauthier's Lockean proviso is also *structurally* unfair. If it fits perfectly Gauthier's own example of masters and slaves, it is completely inappropriate in many other instances. The list of counter examples is endless.

Gauthier's proviso does not preclude coercion in the initial bargaining position.[290] Some situations are coercively structured. One might be led into an unfair bargain without violating the proviso. Coercion springs from a natural 'threat advantage'. When I was in the desert without any hope of being rescued, I was in a very weak bargaining position upon your arrival. You were not responsible for my distress. Therefore you did not violate the proviso when you took advantage of your strong bargaining position and asked for a ludicrous amount of money. I needed you, you did not need me. The proviso can do nothing to rectify that type of naturally coercive situation. Similarly, a poor person who needs to work to survive must be prepared to accept exploitative wages. An

[288] 'The Gauthier Enterprise', p 87
[289] 'Morality, Rational Choice, and Semantic Representation'.
[290] See for example J. Fishkin 'Bargaining, Justice and Justification', p 47-50 and R. Goodin 'Equal Rationality and Initial Endowments', p 126-127.

employer does not violate the proviso when she makes the most of her advantageous bargaining position. The poor person is again in a naturally coercive situation that the proviso cannot rectify.

I think that Gauthier misses the point when he addresses this objection. He wonders why it is morally objectionable if a party makes the most of a threat advantage. He argues that the disturbing aspect is more that to exploit someone in distress does not fit some standard practice of social morality. If you are in a position to help me in the desert, it seems a social convention that you will do so without expecting me to pay you a ludicrous amount in return.[291] What does this have to do with his project? He only tries to rationally derive a social morality not to defend the standard one. Such an argument is inconsistent with his theory. Why should his proviso rectify some form of coercion and not others? Why should it rectify some injustice and leave others deliberately outstanding? We can only agree with Buchanan: no rectifications at all better fits the contractarian project than partial ones.

We have just seen that the proviso fails to rectify all past injustices. Unfortunately, it can also generate unfair situations. For example, the proviso permits one to worsen the situation of others, if doing so is necessary to prevent one's own situation from being worsened even slightly. The proviso can also allow one to use people in some unpleasant way as long as the net effect on them is positive. Finally, the proviso allows one to kill or steal from others if it is accepted that someone else would do so if one didn't. [292]In most of these unfortunate situations, the proviso poses problems. Gauthier only provides an elusive answer to address these objections: "the proviso is not intended to be the last word on morality, but only the first."[293] This argument itself fails when we turn to another objection regarding the initial position and property rights.

P. Danielson argues that Gauthier's revised Lockean proviso generates a more insidious form of injustice. If "the only feature relevant to accepting a starting point is its effect on the resulting bargain", then, he argues, personal rights only should be treated in the initial position whereas both property rights and public goods should be subject to

[291] Gauthier also argues that such situations should not be considered as simple bargaining situations. It is obvious that application of MRC in this type of situations will never result in a mutually advantageous situation.
[292] These examples were suggested by D. Hubin and M. Lambeth's 'Providing for rights'.
[293] 'Uniting Separate Persons', p 183

bargain. Property rights can adversely affect some bargainers. They would not make them worse off in the initial position but it could worsen the outcome of bargaining for them. He illustrates his argument with the following example:

"Jim's building a house and leave homeless Bob no worse off until private home ownership acts as a constraint on cooperative interaction, perhaps preventing a socialised housing policy. In the prospect of cooperative agreement, Bob is worse off."[294]

Property rights should not be part of the endowment required to define agents for social contracting since they put at risk the impartiality of the proviso.

Gauthier first addressed this issue in 1988. He admitted:

"It is true that its [the proviso's] effect is to permit state of nature interaction that leaves some persons ultimately worse off than they would have been had everything (except personal rights) been decided by agreement... But those persons whose bargaining position is weakened are nevertheless better off than at the no interaction point... State of nature interaction does not confer short term benefits at the risk of long term overall costs."[295]

Gauthier then endeavours to demonstrate that the exclusion of property rights from initial endowments would affect the stability of the bargaining outcome.

What to make of such a reply? Gauthier admits that the proviso's rectificatory process is only partial but he claims that it does not affect the rationality of the proviso since it does not affect the stability of the future agreement. The comparison point is not what is best in the absolute for the bargainers but what is best in comparison to the no-interaction point. The choice of the no-interaction point as a comparison standard is not only counter-intuitive, it can also be unfair. If it can be unfair, how can it secure the stability of the bargaining outcome? Once again, Buchanan's argument makes sense. Any rectificatory device of the pre-bargaining position can only be partial and therefore necessarily adversely affects the bargaining outcome. Any manipulation of the existing pre-bargaining position is anti-contractarian by nature.

In 1993, Gauthier generalises his inclusion of property rights in the initial endowment: The exclusive use of particular resources is allowed "in so far – and only in so far as –

[294] The quote continues as follows: "Some criticisms of private provision of schooling or medical services often make a similar point. They focus not on the immediate detraction from the public services but on the shift in the initial position for bargaining that may block improvements to the public services." 'The Visible Hand of Morality', p 366

[295] 'Moral Artifice', p 411

this benefits the user without *net* cost to those no longer entitled to treat these resources as in common use."[296] How does one measure this net cost? Does one take into account post bargaining losses? These questions being left unattended, we can consider Danielson's objection as still valid.

The proviso is probably one of the core conceptions that Gauthier is the least willing to let go. Although, he is aware of its weaknesses, he believes that the proviso is an indispensable basis to his theory.

"The formulation of the proviso in *Morals by Agreement* is at best only a first approximation. But a more adequate formulation would leave unanswered the strong challenges that may be brought against its relevance for both agreement and compliance. Occasionally I find myself tempted to discharge the proviso from its contractarian employment ... But then I find myself once more convinced that it is exactly that power of the proviso – to convert as it were, a Hobbesian state of nature into a Lockean one – that is needed in a full contractarian moral theory."[297]

[296] 'Uniting Separate Persons', p 183
[297] 'Rational Constraint: Some Last Words', p 325

Conclusion: What is left of *Morals by Agreement*?

I believe that to understand Gauthier's achievement, we must remember what his goal was. We remember that he was dissatisfied with Hobbes' answer to the Fool. His concern was that, with the sovereign, Hobbes had given a political answer to the compliance problem intrinsic to any social contract theory. He believed that it was possible to give a moral solution instead. He was convinced that it was possible for individuals to internalise rationally derived constraints. In other words, his ambition was to provide a contractarian moral theory.

In order to do that, he used modern tools, relied on a subjective account of values, worked within the rational choice theory framework, used game theory, developed a bargaining theory and padded the baseline with a Lockean proviso. He has obtained a theory but this chapter has taken us through its flaws.

- The market as a morally free zone is a marginal core conception; it is not indispensable to the coherence of his theory. However, we can consider that he no longer defends it. Indeed, we saw that the market was proven to be morally sustained.

- One can say that Gauthier's bargaining theory is now abandoned. It rested on the assumption of equal rationality and its outcome was the Minimum Relative Concession principle. Gauthier has admitted that the assumption of equal rationality is flawed and that MRC at best needs modification. Gauthier does not show any interest in defending his bargaining theory anymore. In line with the above argument, I believe that his bargaining theory was either incompatible with the rest of his theory or was not a bargaining theory in the agreed sense.

- The conception of constrained maximisation suffered a major blow from the arguments of E.F. McClennen. However, this conception, so central to his moral theory, is very much alive. Gauthier has taken the full measure of the potential of McClennen's concept of resoluteness and the core idea of his moral theory will be re-activated within this broader framework.

• Gauthier does not want to let go of the Lockean proviso. Buchanan argued that the proviso was a rectificatory process incompatible with any contractarian theory. By nature, rectifications are likely to be marginal or partial and therefore unfair. Gauthier is aware of the objection but maintains that the proviso is rationally required to ensure the compliance to the agreement. He believes that without this proviso, his moral theory would collapse.

• As far as the Archimedean point is concerned, J. Hampton claimed that Gauthier's Archimedean point becomes more Rawlsian and less Hobbesian towards the end of chapter VIII of *Morals by Agreement.*[298] Her argument was based on the fact that the ideal chooser acts as a proto-individual who selects a scheme of cooperation not according to *who he is* but according to *who he could be* in any of these schemes. She insisted that such a turn in Gauthier's Archimedean standpoint reflected a change in his methodology. Gauthier's focus moved from the individuals to society. Individuals are structured by the social system in which they live and the moral frame has to be chosen before the individuals make any choices. We saw in section 1 that this Rawlsian turn had been emphasised in Gauthier's defence of the hypothetical contract. We also saw that this Rawlsian turn was incompatible with the rest of his theory.

If Gauthier accepts that *Morals by Agreement* as **a** contractarian moral theory is flawed, he is determined to demonstrate that **the** idea of a contractarian moral theory is the only way forward. From this point of view, one can say that *Morals by Agreement* as a project is very much alive. Gauthier is still convinced that it is possible to give a moral answer to the Fool. We need now to turn to his most recent work to see how he will defend the idea of contractarian moral theory without reference to a bargaining theory, to mutual unconcern or equal rationality.

[298] 'Can we agree on Morals?' especially, pp 344-352

CHAPTER IV: BARGAINING AND UTILITY MAXIMISATION

Now that we have reviewed most of the debates that have taken place around *Morals by Agreement*, I would like to suggest my own criticism in continuation of the interpretation provided in chapter II. Although, we have seen that most of the core conceptions are now either suspended or modified, I believe that the argument developed below against *Morals by Agreement* still applies to Gauthier's new attempt to defend the idea of a contractarian moral theory.

Gauthier's central idea is that it is possible and even intrinsic to human nature to change rationality once in social interactions. Reason is the bridge between natural or market interactions and social or cooperative interactions. It is through reason that we discover our need to change our rational dispositions. Reason helps us to overcome market failures and allow us to benefit from co-operation.

"Conceptions of rationality are not fixed in human nature, but rather the products of human socialization ... the capacity to make the choice of a conception of rationality is itself a necessary part of full rationality".[299]

This idea is repeated in *Morals by Agreement*:

"Reason, which increases the costs of natural interaction among human beings, offers not only a remedy for the ills it creates, but also the prospect of new benefits achieved mutually through co-operation."[300]

Or again:

"At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection."[301]

The change of rationality occurs once the pursuit of mutual advantage supersedes the pursuit of individual gain. In other words, there is an evolution or a change of rationality

[299] 'Reason and Maximisation', p 430-31
[300] *Morals by Agreement*, p 113
[301] *Morals by Agreement*, p 183

when individuals give priority to the pursuit of mutual benefit over pure individual gain. They constrain their natural disposition to solely pursue individual gain in order to achieve cooperation. The change is therefore from 'natural' to 'artificial' rationality *or from straightforward to constrained maximisation.* The rationality that prevails in market and natural interactions is supplanted by the rationality of cooperation. The rationality that sustains natural impartiality is taken over by the rationality that sustains artificial impartiality.

The question is now to know when the change of rationality occurs: before bargaining or after? Is this change of rationality a pre-condition to the agreement or a mere by-product of it?

Read:

"Cooperative interaction is not itself bargaining... Bargaining gives rise to cooperative interaction but is itself non-cooperative. This distinction is of great importance in subsequent discussion, for as we shall see in co-operating persons must at times constrain their utility maximizing behaviour, but in bargaining itself persons accept no such constraint. The constraints required by co-operation are arrived at through bargaining, but are not part of the bargaining process."[302]

Or again:

"Bargaining is a straightforwardly maximizing activity leading to an agreement on a joint strategy. Constraint enters in cooperative interaction, which requires adherence to this strategy even though the outcome is not in general equilibrium."[303]

We saw in chapter II that to consider the bargainers as straightforward maximisers poses three problems.

The first problem concerns the type of straightforward maximisation in question. We called it 'sophisticated' rationality and we noticed disconcerting assumptions as well as an unusual disregard for the bargaining power of the bargainers. A straightforward maximiser usually tries to maximise his chances to obtain a bigger share of the surplus using all the possible tools available such as threats or bluffs. We noticed that Gauthier's bargainers were unusually 'reasonable', avoiding putting the bargaining process at risk by fear of no co-operation.

---

[302] *Morals By Agreement,* p129
[303] *Morals By Agreement,* p151

Secondly, there is a chronology problem. Gauthier does not date the change of rationality. Instead, he distinguishes between two forms of rationality namely internal and external rationality. Internal rationality is the rationality of making the contract and corresponds to straightforward utility maximisation. External rationality is the rationality of keeping the contract and corresponds to constrained maximisation. External rationality applies both to relationships in the pre and the post-bargaining positions. Indeed, in order to establish the initial bargaining position, the future co-operators have to constrain their utility maximisation to guarantee the fairness of the process. It seems irrational to expect individuals to be constrained maximisers when they establish the initial bargaining position, straightforward maximisers when they bargain and constrained maximisers again when they co-operate.

The third problem is corollary to the second one and concerns the link between bargaining and the concept of translucency. It seems that this latter concept rests on a *pre-disposition* to comply. Read:

"The disposition to decide whether or not to adhere to one's covenants or agreements by appealing to directly utility-maximizing considerations, is itself disadvantageous, if known, or sufficiently suspected, because it excludes one from participating with those who suspect one's disposition, in those co-operative arrangements in which the benefits to be realised require each to forego utility-maximization."[304]

Or again:

"Only those disposed to keep their agreements are rationally acceptable as parties to agreements. Constrained maximisers are able to make beneficial agreements with their fellows that the straightforward cannot, not because the latter would be unwilling to agree, but because they would not be admitted as parties to agreement given their disposition to violation."[305]

I claim that it is necessary to re-establish the chronology for the coherence of the theory. Individuals must first change their rationality and acquire a disposition to obtain and maintain co-operation based on the pursuit of mutual interest rather than of individual gain. They can then establish the initial bargaining position, bargain and co-operate using the same rationality i.e. constrained utility maximisation. *I therefore claim that Gauthier's bargainers must be constrained maximisers.*

---

[304] *Morals by Agreement* p 162
[305] *Morals by Agreement* p 173

However one must not be confused: if the *change of rationality* must pre-exist the agreement, its *content* is fixed by it. If the *rational disposition* to cooperate and the *constraint on choices* have to pre-exist bargaining, the actual *constraint on behaviour* can only take place after agreement. Behaviour can only be constrained by a known and established principle.

The purpose of this chapter is to demonstrate that Gauthier's bargainers are constrained maximisers and to evaluate the consequences of such a claim on Gauthier's theory. After emphasising the incoherence of the theory as interpreted in chapter II, we will demonstrate that the logic of *Morals by Agreement* could be recovered if Gauthier's bargainers are constrained maximisers. However, if we accept this claim, it weakens Gauthier's achievement since he assumes what he meant to derive.

Section 1: Gauthier's bargainers are constrained utility maximisers

Gauthier describes the bargaining as a two-stage process: 1) each party advances a claim; 2) each party offers a concession by withdrawing some portion of his original claim and proposing an alternative outcome. We saw in chapter II that their starting point is the 'baseline'. They each come to the bargaining table with their own situation. It is this initial situation and the first stage of the bargaining process that is at stake in this section.

Bargaining conditions and constrained maximisation.

"In defending constrained maximisation we have implicitly reinterpreted the utility-maximizing conception of practical rationality. The received interpretation... identifies rationality with utility maximisation at the level of particular choices. A choice is rational if and only if it maximizes the actor's expected utility. We identify rationality with utility maximisation at the level of dispositions to choose. A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition."[306]

If constrained rationality is a disposition to choose, if it is a disposition to give priority to co-operation over pursuit of pure individual gain, I claim that the bargainers must possess this rational disposition in Gauthier's theory.

The difference is as follows: a straightforward maximiser devises an individual strategy that will enable him to strike an advantageous bargain and he will use all the bargaining power available to do so. A constrained maximiser is already in the joint strategy mode. He considers it more beneficial to co-operate than to manage solo. He *has internalised* the change of rationality from the pursuit of individual gain to the pursuit of mutual benefit. He is therefore disposed to constrain his utility maximisation in bargaining in order to better benefit from co-operation.

Gauthier imposes two constraints on his bargainers in the first stage of bargaining, two 'pre-conditions' on making claims:

---

[306] *Morals by Agreement*, p 182-183

"In deciding how much is possible [to claim], each is *constrained* by the recognition that he must neither drive others away from the bargaining table, nor be excluded by them. Hence each person's claim is limited by the overall cooperative surplus and more specifically by the portion of the surplus that it is possible for him to receive."[307]

"Each person may not always claim all of the cooperative surplus that he might receive but only that part of the surplus to the production of which he would contribute. Each person's claim is *bounded* by the extent of his participation in co-operative interaction."[308]

Is it really the way it works in bargaining between straightforward maximisers? It seems that bargainers would make the maximum claim *compatible with their bargaining power* not one proportional to their contribution. As we shall now see, Gauthier seems to ignore the role of threat advantages and unequal bargaining power in bargaining.

Would a straightforward maximiser abide by Gauthier's bargaining conditions?

At the end of his chapter on bargaining, Gauthier specifies 'a few' constraints on the bargainers:

"In ordinary bargaining persons may conceal significant features of their circumstances, or the full range of their options, may misrepresent their preferences, or the strengths of their preferences. But we suppose each person to be fully informed... In ordinary bargaining persons may bluff ... But here there is no place for bluffing; not only is each person fully informed but he is a rational-utility maximiser who knows his fellows to be also rational utility maximisers. In ordinary bargaining persons may make threats but among fully rational persons threats are useless; no one will believe anyone who claims that he will act in a non-utility maximising way should others not comply with his threat...Our bargainers have no psychological strengths to exploit or psychological weaknesses to be exploited."[309]

I believe that what Gauthier calls 'ordinary bargaining' should read 'bargaining between straightforward maximisers'. What he calls 'full rationality' should read 'constrained rationality'. I almost wonder if 'fully informed' should not read 'translucent' here. How would we 'know' the exact preferences, strengths, thoughts and dispositions of our fellow bargainers otherwise? It is true that if we are all rational utility maximisers, then we can simulate each others' reasoning. But it does not say anything about each bargainers' ability to conceal or distort his preferences or thoughts to the others, or his inability to spot the others' manipulations.

---

[307] *Morals by Agreement* p 134
[308] *Morals by Agreement* p 134
[309] *Morals by Agreement* p 156

What Gauthier assumes here is that, for the sake of obtaining co-operation, his bargainers are prepared to surrender the tools that would improve their bargaining power. If we can concede that it is rational to pursue co-operation in order to benefit from a co-operative surplus not available otherwise, it does not mean that it is irrational for a straightforward maximiser to improve his baseline in order to gain a bigger share of the co-operative surplus than he would be entitled to otherwise. The pursuit of mutual advantage does not preclude individuals from exploiting the advantages that their baseline provides them with. It is along these lines that Barry develops his argument against Gauthier's rationality in bargaining.

For example, Gauthier's main argument against threats is that it would be irrational to make a threat that can only be defective to carry out. Not only, if I am rational, I would not make a threat, but also even if I was irrational enough to make a threat, my fellow bargainers in assuming that I am a rational utility maximiser would not believe that I would carry it out.

What about the concept of 'threat advantage', common to Nash and Harsanyi? As explained by Barry[310], to threaten somebody is to attempt to affect his behaviour by stating a conditional intention. The conditional intention here is to do something unless the other performs and not to do it if he does. We remember Luke and Matthew story from chapter II. Matthew has the threat advantage since, unlike Luke, he prefers cacophony over silence. If Luke refuses to let him play solo by playing the piano, Matthew can bring about Luke's worst outcome by playing the trumpet at the same time. By using such a strategy, Matthew can improve his bargaining power when negotiating with Luke. The optimal threat is what puts one in the best position to strike an advantageous bargain.

Gauthier is aware of these strategies and he waves them away but claiming that

"they play a purely hypothetical role... since [the bargainers] do not actually choose them, but merely appeal to them to determine the costs that each could impose on the other in a strict competition for bargaining advantage. Maximally effective threat strategies would not be chosen by [the bargainers] were they to find themselves unable to co-operate; the threat point bears no particular relationship to the non-cooperative

[310] *Theories of Justice*, p 69

outcome. But if [the bargainers] would not choose these strategies, then they cannot credibly threaten with them."[311]

In this quote, we can see that Gauthier does not adequately distinguish between credible and non-credible threats. This poses the problem of specifying the non-agreement point.

As Barry points out, the non-agreement point plays a hypothetical role in Gauthier's theory just as much as in Nash's. So the first part of Gauthier's argument is irrelevant.

Now the next question is to know whether it is rational or not to *make a threat* with the intention to carry it out if the other party does not perform. Let us take the case of Matthew and Luke. It might be sufficient for Matthew to *state* that he prefers cacophony to silence for Luke to understand that he is in a weaker situation. Luke is then aware that if he wants co-operation and therefore a chance to play solo on some evenings, he has to concede more solo evenings to Matthew or else Matthew will bring about his worst outcome, namely cacophony. So it is rational to make a threat when one has an *obvious* threat advantage.

Let us go further: is it rational to *carry out* a threat when to make it is not sufficient to strike an advantage in bargaining. If, on the one hand, as it is the case in most real life situations, the relationship between the parties is on going, one could claim that it is rational to carry out a threat in order to be credible in future interactions. For a few nights, Matthew could play the trumpet while Luke is practising his piano until Luke comes in search of a compromise. It can be rational to pursue a worse outcome in the short run in order to bring about a better one in the long run. If, on the other hand the bargainers have a one off interaction, a threat is counterproductive and therefore irrational to carry out.

We are now touching the core of Gauthier's argument. What is to be avoided above all are deadlocks or no agreement. Whatever could lead to no agreement is labelled as irrational under the assumption that co-operation is necessarily better than the status-quo. ***A*** share of the co-operative surplus is necessarily better than ***no*** share at all. Once

[311] *Morals By Agreement*, p 200

again, we will follow Barry's argument against this logic. Indeed he argues that such a logic undermines completely the strategic aspect of bargaining:

"If we rule out any such strategic play as irrational and insist on non strategic maximisation as the only form of rational action, we automatically eliminate the rationale for any bargaining solution. For what drives every bargaining solution is the notion that rational actors will refuse any offer that they consider being inadequate even though it represents an improvement over the non-agreement point."[312]

A straightforward maximiser will always try to get away with less concession and more benefits.

The question now is why Gauthier is so adamant that threat behaviours are 'irrational'? Here comes his last argument: "In a community of rational persons, threat behaviour would be proscribed. Unlike cooperation, threat behaviour does not promote mutual advantage."[313] The same argument could easily apply to bluff and other standard bargaining tools. As claimed originally, I believe that the rationality Gauthier refers to here is already the rationality of mutual advantage and of morality. Bargainers are not to use these tools if they have internalised the benefit of mutual advantage. They must be prepared to surrender some of their personal advantage to better benefit from the future fruits of co-operation.

A straightforward maximiser, even pursuing mutual benefit and aiming at co-operation, would still try to improve his baseline in order to claim a bigger share of the co-operative surplus. Bluff and threats are only a few of the tools he would use in order to maximise his utility. The conditions imposed on Gauthier's bargainers forbid such practises. No straightforward maximiser would abide by his conditions, these are conditions for a constrained maximiser.

How was the confusion between straightforward and constrained maximisation possible?

From the above it seems easier to understand what has happened. Three of Gauthier's assumptions make the confusion possible. The assumptions of full information and equal rationality have vastly contributed to the distortion of Gauthier's bargaining

[312] *Theories of Justice*, p 60
[313] *Morals by Agreement*, p 186

theory. These two assumptions have been discussed at length in the literature and a review of their impact was provided above. I believe that a third more insidious and less noticeable assumption is equally damaging: the assumption of equal need and want for co-operation. Indeed, Gauthier's bargainers behave as if they all equally need and wish to achieve co-operation. They constrain their claims, they fear to put in jeopardy the 'bargaining process' in the apparent belief that the others need and will co-operation as much as they do.

As suggested above, these three assumptions together contribute to equip all the bargainers with a roughly equal bargaining power and an equal willingness to comply with the agreement reached. In allocating equal bargaining power to his bargainers, Gauthier has removed all the strategic aspects that usually characterise bargaining between straightforward maximisers.

More importantly, these three assumptions considerably increase the benefit of co-operation in comparison to solo behaviours. At equal bargaining power, the pursuit of mutual advantage almost merges with the pursuit of individual gain. If I know that my fellow bargainers are fully informed, equally rational and equally motivated to reach co-operation, it becomes my direct best interest to pursue mutual advantage.

The threshold between pursuit of mutual advantage and pursuit of individual gain becomes insignificant. There is a co-operative surplus that comes into existence only through co-operation and bargainers have to aim at co-operation in order to benefit from a surplus that would not be available otherwise. The choice faced by a rational utility maximiser is then between the pursuit of co-operation that gives access to a benefit not available otherwise and the status quo. Presented this way, it is obvious that straightforward maximisation dictates the pursuit of mutual advantage as an integral part of individual gain. Under these assumptions, straightforward and constrained maximisation merge into one undistinguished form of rationality.

Section 2: A new reading of Gauthier's theory

I expressed in chapter II my reservations regarding Gauthier's distinction between internal and external rationality. It seemed artificial to assume that individuals can simultaneously be straightforward maximisers while bargaining and constrained maximisers otherwise.

I can possibly accept that individuals are straightforward maximisers in the state of nature and during bargaining. Once in social interactions, they internalise the need for constraints on their behaviour and become constrained maximisers. Such an interpretation would have made sense if it was not for the proviso. The proviso introduces a chronological incoherence. Indeed, while still in the state of nature, Gauthier's prospective bargainers have first to constrain their utility maximisation according to the proviso's requirements in order to establish their respective initial bargaining positions. They then revert to straightforward maximisation to bargain. The outcome of bargaining leads them back to constrained maximisation.

I argue that if Gauthier believes that it is intrinsic to human nature to change rationality, this change of rationality has to precede bargaining. While in the state of nature, individuals become aware of the need for co-operation and they give priority to mutual advantage over pure personal interest. They then establish the original baseline applying the proviso, they bargain and they live in social interactions as constrained maximisers.

The two schemes on the next page illustrate my argument. Scheme 2 shows the new reading of Gauthier's theory suggested above. It shows that in natural and market interactions, individuals are straightforward maximisers, impartiality is naturally sustained and the factor endowments are given. Through reason, and in the face of externalities, individuals come to realise the need for co-operation. They accept, in principle, to constrain their utility maximisation in the hope of benefiting from mutual advantage and cooperation in the future. In this frame of mind, they agree on a baseline complying with the proviso requirements; then they bargain, choose the Minimum Relative Principle and re-establish an artificial form of impartiality.

Gauthier's theory has gained in coherence but does it still hold the same interest?

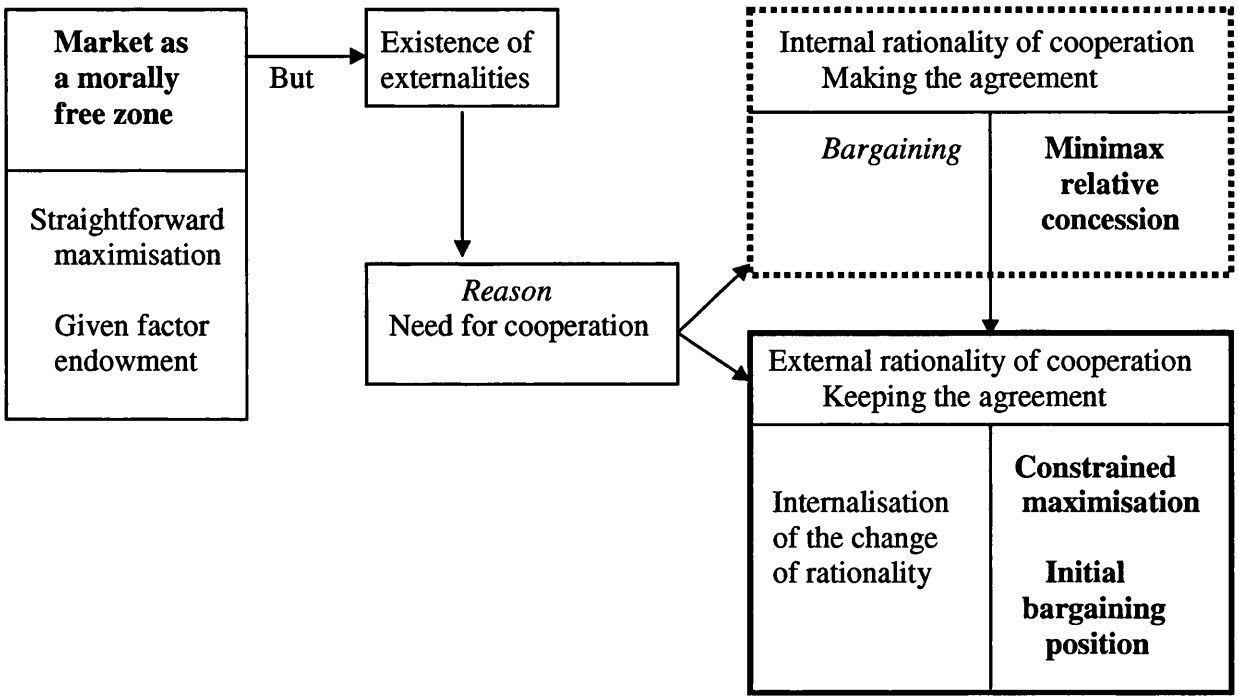
Constrained rationality is the rationality of justice and morality. "Our concern is to show that if co-operation results from rational agreement, the constraint it imposes is just".[314] If the rationality referred to in the above quote is already constrained i.e. the rationality of justice then Gauthier has not proven much. In other words, justice and morality pre-exist the theory. They are not just rationally derived.

In chapter II[315], I showed that the rationality of morality is based on 'mutual advantage' rather than on pursuit of individual gain. According to my interpretation of Gauthier's theory, impartiality prevails in natural and market interactions. Partiality occurs when there is conflict between the rationality required to pursue one's individual gain and the one required to bring about mutual advantage. The reconciliation between these two types of rationalities then requires impartial constraints on individuals' utility maximization. These constraints define the boundaries of morality. Morality arises from the emergence of partiality and it implements an artificial form of impartiality by giving priority to mutual advantage over pure individual gain.

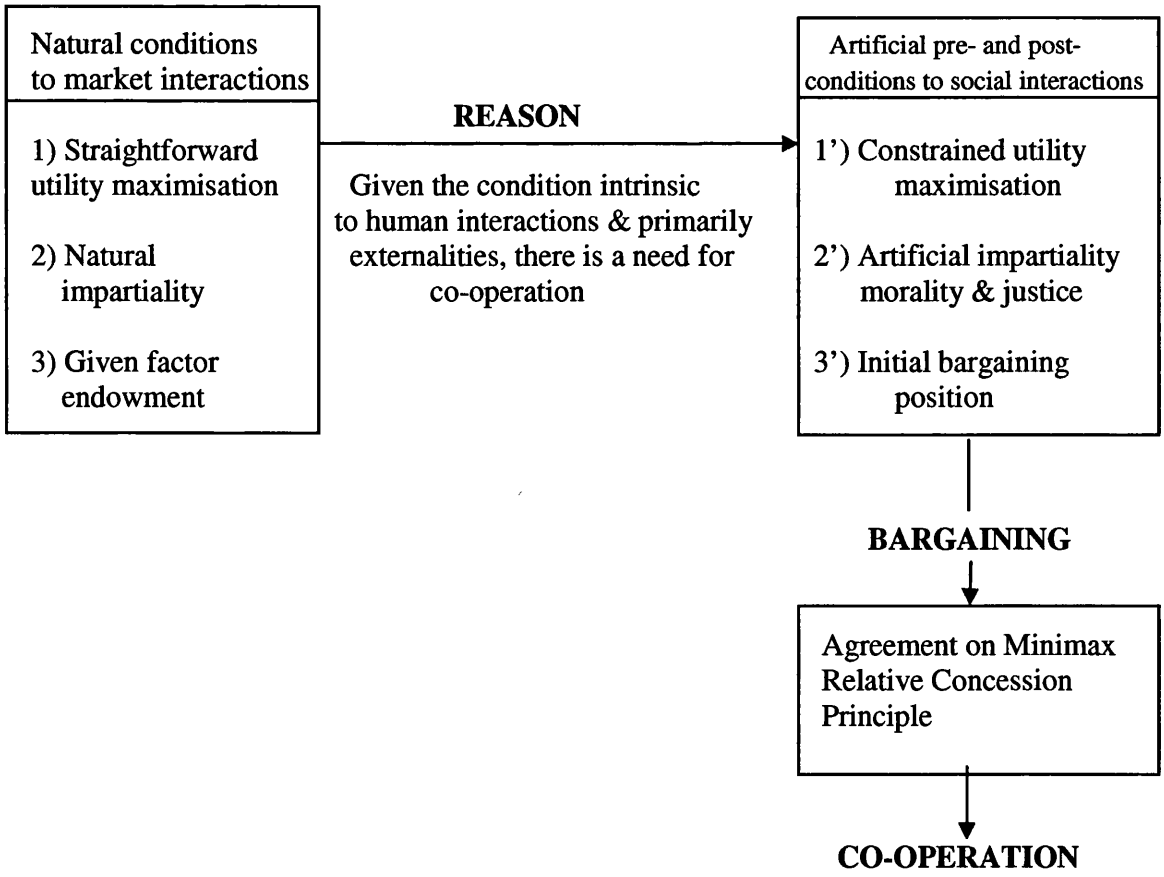
According to our above argument, individuals accept to constrain their utility maximisation in the state of nature in order to benefit from mutually advantageous cooperative activity in the future. The rationality of mutual advantage supersedes the straightforward rationality that prevails in the state of nature. Prior to bargaining, the parties have to internalise the need to constrain their behaviour. In other words they have acquired the artificial virtue of morality. It comes as no surprise if the outcome of the bargain carries through their virtuous mood.

[314] *Morals By Agreement,* p151
[315] See section 2, core conception 1.

Scheme 1: Initial reading of Gauthier's theory (See chapter II)



Scheme 2: A new reading of Gauthier's theory



Conclusion

We opened chapter III with Gauthier's moral theory and commented on the restricted field covered by it. We can now see why it is so. Gauthier understands morality as a constraint on the pursuit of individual utility. This constraint is rationally derived or rather the outcome of a change of rational disposition. Individuals in the state of nature come to realise the failure of the pursuit of individual interest in maximising their utility and rationally acquire a disposition to pursue mutual advantage instead. Morality appears when individuals are prepared to constrain the pursuit of pure individual gain to better pursue mutual benefit, i.e. when they internalise the change of rational disposition. The field of Gauthier's moral theory is therefore limited to the field of mutual advantage.

Whatever is not mutually advantageous is out of the scope of his moral theory. That leaves out the long list of individual values that make better persons. More worrying is that *whoever* does not contribute anything to mutual advantage has no place in the bargaining process and can only be referred to as a negative externality. Children, elderly people or some of the handicapped persons contribute nothing to the co-operative surplus but need instead schools, hospitals, old age and special homes. If we were mere rational machines, this exclusion would not be so shocking but since, as human beings, we do have affective bonds, such morality feels counter intuitive. Can we still label it 'morality'?

In 1988, at a time Gauthier was still defending *Morals by Agreement*, he wrote:

"we must show,..., that the constraint captures at least something of our intuitive or pre-analytic understanding of morality. Failure here would not deprive our argument of interest and significance, but it would disqualify it as a rational reconstruction of morality."[316]

It seems from the above that Gauthier failed in this project.

In this chapter, we demonstrated that morality, understood as constraints on individual utility maximisation was an input as much as an output of the agreement . We shall demonstrate in the rest of part II that this criticism carries through Gauthier's latest articles.

[316] 'Morality, Rational Choice, and Semantic Representation', p 177

CHAPTER V: SHALL WE BE RESOLUTE?

As we saw in chapter III, constrained maximisation remains the core idea of *Morals by Agreement*. However, further to McClennen's criticism, Gauthier realised that this conception was too narrow and did not take into account preference changes. Gauthier remained faithful to the core idea of morality as self-imposed constraints but, from 1993, he started to work within the framework of McClennen's resoluteness.

"I remain convinced that this [constrained maximisation] remains the most fruitful idea in *Morals by Agreement*... A constrained maximiser, as I shall use the term in a way that generalises somewhat from my use in *Morals by Agreement*, is someone who takes her reasons for acting, not only directly from the utilities of possible outcomes she may bring about, but also from her plans and commitments."[317]

Exploiting Bratman's research on plans and intentions as well as McClennen concept of resoluteness, he replaced the constrained maximiser by the resolute planner. Through this change, Gauthier also took a new turn in his approach. We remember that the idea of *Morals by Agreement* was born when he was first introduced to the *Prisoner's dilemma*[318]. He then realised that it was the modern way to express the Fool's challenge and he tackled the huge task of providing a complete coherent theory to address the multi-persons prisoner's dilemma that a society is.

But in the 1990's, Gauthier came to realise that the Prisoner's dilemma was the wrong starting point. A contractarian moral theory or any theory of rational co-operation must solve an *assurance game* instead. In 1998 he wrote: I claim that

'each does best to be accepted by her fellows as a sincere adherent to these arrangements [principles, practices and institutions that would realise mutual advantage and would obtain the agreement of everyone] and that the best way to be accepted is to be such a person. And from the standpoint of such a person – a *just* person – adherence to just arrangements has the character not of a prisoner's dilemma but an assurance game.'[319]

---

[317] 'Uniting Separate Persons', p 185-186
[318] See the preface of *Morals by Agreement.*
[319] 'Mutual Advantage and Impartiality', p 128

The distinction is obviously crucial since an assurance game involves dynamic choice, intentions and plans rather than mere strategic and static rational calculus as in the prisoner's dilemma.

Gauthier amended McClennen's concept of resoluteness to adapt it to his own theory. In this chapter, I will first review McClennen's concept prior to detailing Gauthier's amendments to it. I will then demonstrate that whether at *intra* or *inter*-personal level, sophistication is always a better strategy than resoluteness.

Section 1: Resoluteness and plans

The contradiction of standard rationality: an assurance game

We saw in chapter III[320] how McClennen uses the Ulysses example to highlight the difference between, myopic, sophisticated and resolute Ulysses. This example represents the case of *intra*-personal dynamic consistency. We turn now to an assurance game representing a case of *inter*-personal dynamic choice. We shall underline the contradictions of the standard concept of rationality *as described and interpreted by McClennen and Gauthier.*

David Gauthier suggests the following situation[321]. My crops will be ready for harvesting before yours. It will be better for both of us if we harvest together both yours and my crops but you will help me to harvest my crops only if you expect that I will help you to harvest yours later. I am rational and if you help me then when the time comes for me to help you in return, I will deliberate as follows: if I help you I am still better off than if I had harvested on my own but I could be even better still if I don't help you. Knowing that I am rational, you can guess my strategy as just described. Since you are rational too, you will not help me. We reach then a rather contradictory conclusion: our respective rational calculus brings about the worse outcome for both of us i.e. we both harvest on our own and fail to benefit from cooperation.

This 'interactive game' could be described as follows:

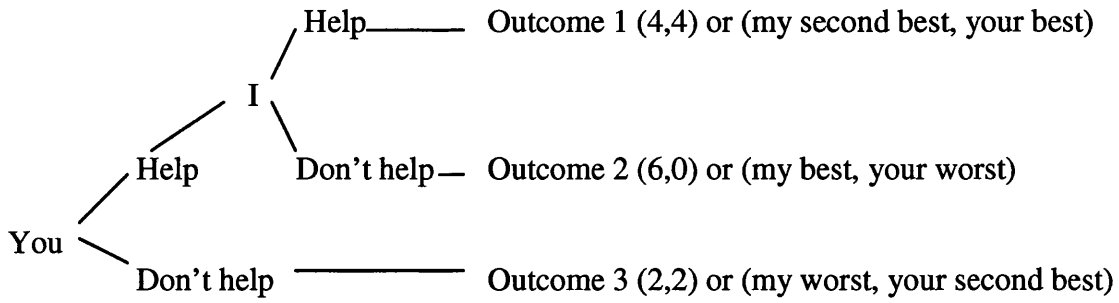


Figure 1: the assurance game

[320] See section 3.
[321] 'Assure and Threaten', p 692

The conditions of the game suggested are ideal:

- There is a Pareto-efficient outcome (we both harvest both crops together)
- We are both fully 'rational'
- We both know that we are both rational
- We both know the strategy structure of each other
- We have common knowledge of preferences with regards to the outcome

In order to ensure that the focus is exclusively on choosing the appropriate mode of rational deliberation, Gauthier also specifies that

- I am not concerned about my reputation.
- It is not an iterated game so I am not worried about what will happen next year.
- There is no pre-existing moral standard about promise keeping.

We can identify in the above story an incremental assessment of the situation. The situation is assessed at each point down the decision tree. The rational individual considers the outcomes *still* available at each point in time she has to make a decision. To understand the issue at stake let us review the situation from both parties' point of view both at t when my crops are ready for harvesting and then at t+1 when yours are ready.

As far as you are concerned, at t, the possible outcomes are either you help me and I don't help you and you gain 0, you don't help me and you gain 2 or you help me and I help you and you gain 4. Therefore, your preferred outcome would be that we cooperate. However you will not help me if you don't expect me to return the favour. From my point of view, the outcomes are you don't help me and I gain 2, you help me and I help you and I gain 4, you help and I don't help and I gain 6. At this stage of the reasoning, my preferred outcome is that you cooperate and I don't. The problem is that you will not cooperate if you are not convinced that I will cooperate when the time comes. If you don't believe that I will help you, you will stay at home while I harvest my crops. This brings about my worst outcome.

Since, according to the orthodox view developed above, I cannot secure my best outcome (you help, I don't), I must aim at my second best outcome (you help, I help). If I want to achieve this outcome I must convince you that I will help you.

The problem is then as follows: in an incremental reasoning, once you have already helped me, it is still *possible* for me not to help you. I have already benefited from your help and it is *now* in my best interest not to help you. How can I then sincerely intend to help you at date t and decide not to help you at date t+1? Are we facing a case of inconsistency? If we do, how can I consider myself rational and inconsistent?

One could object that the situation changes between t and t+1 and therefore the assessment of the outcomes available has to change too. At date t, the fact that you help me is a variable data to incorporate as such in my decision-making process, whereas at t+1 it is a fact (or fixed data). At date t, I had to include in my reasoning the fact that to gain your help I had to convince you and therefore believe myself that I would help in return. At date t+1, you have helped me and, when reassessing the situation, I realise that it is in my best interest not to help you.

This objection does not hold for two reasons:

• First according to the given characteristics of the story, we are both fully rational and there is a common knowledge of our respective rationality. We both have the same form of rationality and it will not take you long to sketch the incremental reasoning just described. Being no fool, you obviously will not help me.

• Second how is it possible for me to sincerely intend to help you at t, when I already know at t that at t+1, my rational calculus will lead me not to help you? A belief is a mental disposition, and I cannot acquire such a disposition when I know that the statement that is the object of my belief is not true. If I cannot believe that I will help you, I cannot sincerely assure you that I will and you will have a good rest while I harvest my crops alone.

McClennen and Gauthier argue that, intuitively, the solution to the problem underlined dwells in my intentions. At t, either I intend to reach my best outcome (you cooperate

and I don't) and I am heading to disaster or I genuinely intend to secure my second best outcome (we both cooperate). My sincerity in giving you assurance is entirely dependent on my intentions and the success of my strategy is entirely dependent on my sincerity. Following our intuition still, we understand that our intentions are the sole link between present and future deliberations. Therefore, the notion of intention introduces the notion of continuity in rational deliberation. One of the weaknesses of rational deliberation as described above is the disconnection between the rational calculus at t and the rational calculus at t+1. The rational agent deliberates at t+1 *as if* the situation was new and was not part of strategy decided upon at t. The assessment of rationality cannot be disconnected from the ultimate goal.

McClennen's solution.

The concept of resoluteness brings an interesting solution to the contradictions of incremental and separable rationality as underlined above. The reasoning starts at the *intra*personal level and is then extended to *inter*personal situations.

At an intrapersonal level, *on the standard theory*, a rational agent has to be dynamically consistent. In other words, once a rational agent has chosen his objectives and the plans that should enable him to attain them, he is not supposed to change these plans until completion unless he acquires new information that calls upon him to alter them. If he settles upon a plan, and then subsequently abandons it, even though no new information comes into play, then he is considered to be dynamically inconsistent. At the same time, the rational agent is supposed to look to what would bring about the best possible outcome as judged *at the time of assessment* and his choice of plans is meant to maximise his preferences for the outcomes.

This version of rationality allows for various adaptations:

- A plan which is likely to be later abandoned should not be considered as *feasible*. This weakness can be addressed by the introduction of a new degree of *sophistication*. An agent who, when planning at t, can foresee what is the best possible outcome but anticipates that the plan chosen to attain this objective is likely to be abandoned at t+i

can put constraints or rewards on the path to completion to ensure compliance with the original plan rather than treating the plan in question as not feasible.

- Most importantly according to this vision of rationality, it is not part of a rational thought process to ensure the coordination of past and present choices. At each point down the decision tree, the agent can reassess his situation *as if* he was at the beginning of a new decision tree; his past decisions matter only to the extent that they lead him to the present point. In other words, it is not expected from a rational agent to be able to follow through an *inter*temporal plan, or at least not solely for the reasons set at the time of planning.

From the above, we understand that the concept of rationality rests on two principles:

- The principle of *separability* states that an agent is entitled, at each point in time, to deliberate and decide about what to do next on the basis of whatever there and then he judges will maximise his present preference with respect to outcomes *still* available[322].

- The principle of *feasibility* states that a plan is feasible if and only if it is consequentially acceptable from a separable perspective at each successive step at which that plan specifies a choice to be made[323].

A plan is assessed as feasible or not feasible by application of backward reasoning but plans are made by looking forward. Therefore, present choices are made according to projected subsequent choices. Dynamic consistency is ensured by constraining present choices to expected future ones.

At this point of the reasoning, McClennen suggests reversing the logic of consistency by constraining future choices to present ones. The agent makes plan at t and constrains his future choices to the plan until completion. To be *resolute* is to be dynamically consistent in this way and is a pure act of will; the constraint on future choices is a mere *mental constraint.*

[322] 'Pragmatic Rationality and Rules', p 229
[323] 'Pragmatic Rationality and Rules', p 230

McClennen does not claim that to be resolute is always rational. Unforeseen circumstances can arise, of course, under which it will be rational to abandon the plan. He only claims that in certain circumstances to be resolute is a necessary condition of rational sequential choice. However, for very long term planning the concept of resoluteness could be too rigid[324].

The concept of resoluteness involves a two level deliberative approach where the first level is dedicated to choosing the best outcome available and planning future choices in order to reach the goal set and the second level is for choosing and acting upon the plan made at the first level. The concept of resoluteness ensures coordination over time for an individual and between interacting agents at an interpersonal level.

To be resolute at interpersonal level means that once you have agreed on a sequence of choices with other agents, you are disposed to comply with the plan agreed upon wherever you have to choose[325]. In the game described above[326], it is now easy to understand the power of the concept of resoluteness in solving the contradictions underlined. It is once again more 'efficient' or 'productive' to be resolute than to be merely rational.

Let us review our harvesting problem with our new binoculars. At t, your best outcome is for us to cooperate. As far as I am concerned, the situation is the same. I want you to help me to harvest my crops. In this context, it is in my best interest to convince you that I will help you to harvest yours when the time comes. If I want to convince you that I will help I have to plan at t that I will help you at t+1. At t+1, I will not reassess the situation anew since nothing has changed, I will just carry out my original plan and help you. It is because I am resolute and because I genuinely intend to help you that I can sincerely assure you and therefore convince you at t that I will help you if you help me. So you will help me.

---

[324] 'Pragmatic Rationality and Rules', p236

[325] McClennen acknowledges that unforeseen circumstances can arise, of course, under which it will be rational to abandon the plan.

[326] The conditions are: all the players are fully rational and there is common knowledge of the rationality of the players, of the strategy structure of the game and the preferences that each has with respect to outcomes. 'Pragmatic Rationality and Rules', p 243-244.

Most alternative solutions to the assurance game involve third party surveillance and enforcement systems. McClennen claims that voluntary compliance to plans as described in his theory is less costly and therefore more efficient.

In what follows, I will present Gauthier's defence of the superior rationality of the resolute mode of deliberation over the sophisticated one at both *intra* and *inter*-personal levels.[327] In both cases, I will defend the opposite argument and attempt to demonstrate the superiority of sophistication over resoluteness. [328]

[327] Intra-personal dynamic choice means individual making plans about future choices for himself and involving himself only. Inter-personal dynamic choice means individuals making plans about their future choices while interacting with each other.

[328] The assurance game described above is only one of the situations that have inspired Gauthier. However, in theorising about plans and commitments, Gauthier has also attempted to address Kavka's toxin puzzle and a classical threat game. Kavka's toxin puzzle goes as follows. A billionaire offers to pay one million dollars tomorrow morning if at midnight tonight I have convinced her that I *intend* to drink tomorrow afternoon a vial toxin that will make me violently ill for a day but will not be life threatening or have any lasting effect. Once she has put the money on my account in the morning, I can decide not to drink the toxin in the afternoon. If I don't intend to drink the toxin, can I convince her that I will? If I really intend to drink it, will I drink it when the time comes?
The threat game is also well known. Bob makes a threat to Bill. If Bill complies, Bob has much to gain but if Bill resists the threat, it will be costly to Bob to carry out his threat. A lot depends obviously on Bill's assessment of Bob's intention to carry out his threat. Is it rational to make such a threat? And if it is, is it rational to carry it out?
In willing to embrace all these cases under the same theory, Gauthier runs the risk of confusing goals. In particular the toxin puzzle is based solely on intention. Unlike in the co-operation case, compliance with the intention is inconsequential to the story. The threat game as well should remain out of the scope of his analysis since a moral theory is unlikely to spring from such a start. Threat games belong to strategic bargaining not to co-operation and moral practice. We shall leave these cases aside in this section.

Section 2: Rational agency & choice between modes of deliberation -Intra-personal dynamic choices

Gauthier's criterion

To give Ulysses a break, I would like to take another example (more meaningful to me!). We met with Michelle in chapter II and we remember her preference for chocolate cakes over fruit salads. Michelle cannot resist the smell of cakes when she passes her local bakery to go to work. Aware of her extra weight and knowing her weakness, she has two strategies available. She can be *sophisticated* and take a longer route that does not pass by the bakery. But Michelle has a very tight schedule in the morning and the time wasted to take the longer route is costly to her. Alternatively, she can be *resolute*, realise that in her best interest she must not eat cakes and decides (or plans) that she will pass by the bakery without stopping. To be resolute will save her the cost in time that she would encounter would she choose to be sophisticated.

When planning, an agent first chooses a goal or preferred outcome. Often a plan involves a sequence of choices. If, when planning, an agent can foresee that he will experience a temporary change of preference at one of the choice node, the plan is called not feasible. To be sophisticated is to be aware of this unwanted change of preference and to put in place an enforcement mechanism that will force the agent to comply with his original plan. Unfortunately such a strategy usually has a cost that to be resolute can spare.

McClennen suggests the following criterion for deciding when resolute choice is rationally superior to sophisticated choice:
'what forms a natural condition for the self being resolute is precisely its becoming aware that there are benefits to both the present self and the future self that will be foregone if the self cannot act resolutely.'[329]

---

[329] 'Rationality and Dynamic Choice: Foundational Explorations', p 212. To understand the difference, we can take another example. To avoid eating cakes *at home*, Michelle can either be sophisticated and have none in the house or be resolute and have cakes in the house but to decide to eat none. To have a cake-free home is costless to Michelle. It puts the full family on a diet (which cannot be that bad) and it is actually cheaper. Following McClennen's criterion in this case, sophisticated choice is rationally superior to resolute choice. There is no benefit to gain in being resolute rather than sophisticated.

I believe that this criterion is not sufficient and I am not the only one. Gauthier also defends a more stringent criterion.

His argument is primarily against the separability principle. Resoluteness does not provide the adequate rational mode of deliberation for agents anticipating preference shifts or changes in prospects. [330] An agent makes plans for herself based on her present concerns. However, at the time of choice, her concern might have changed and her past interests should not have any practical significance on her present life. Rational deliberation should allow for some preference shifts.

Unfortunately, resoluteness as a mode of deliberation lacks of flexibility and fails to adapt to change of concerns since it entails a violation of the separability principle. According to McClennen, the preference for the overall outcome becomes endogenous and outweighs immediate preferences at each node on the decision tree. So in Michelle's case, when adopting the resolute mode of deliberation, her preference for being slim outweighs her preference for cakes when she passes by the bakery. But does it?

Gauthier argues that McClennen's above criterion for deciding between modes of deliberation is not satisfactory and he endeavours to provide an alternative one.

Gauthier distinguishes between *proximate* (or immediate) preferences from *vanishing point* (or overall) preferences that an agent acknowledges when choice is not imminent.[331] Michelle has a vanishing point preference for being slim, however when passing by the bakery, she has a proximate preference for cakes. Agents rationally base their choices on vanishing point preferences. If Michelle still has a vanishing-point preference for being slim when passing by the bakery, rationality dictates that she chooses not to buy cakes. Indeed, if she surrenders to her immediate preference and buys a cake, she would do less well in achieving her overall preference for being slim. Given her preference at all time for being slim, her vanishing point preference is a more rational basis for deliberation than her proximate preference for cakes.

---

[330] 'Resolute Choice and Rational Deliberation', p 17
[331] 'Resolute Choice and Rational Deliberation', p 20

But what if, when passing by the bakery, Michelle judges that the benefits of realising her desire for cakes far outweigh the cost of a lifetime sacrifice to her shape? Surely deliberation cannot be rational if it dictates actions contrary to one's preferences at the time of choice. Gauthier argues that resoluteness stops being a rational mode of deliberation when the *strength* of proximate preferences is over a certain threshold that he calls the threshold of *immediacy*. Indeed, if the strength of Michelle's proximate preference when she passes the bakery is such that she will prefer to buy a cake she will do less well overall. The resolute mode of deliberation will not allow her to realise her overall preference for being slim. When the strength of her proximate preference is above a certain threshold, she is better off being sophisticated and taking a longer route than being resolute. So what is the threshold?

According to Gauthier, a preference *below the threshold* is such that at the time of choice, an agent would choose not to act on her proximate preference if the alternative were to act on all of her proximate preferences of equal strength. A preference *above the threshold* is such that at the time of choice, she would choose to act on her proximate preference if the alternative were to act on all of her proximate preferences of equal strength.[332] The threshold of immediacy now provides a comparative test for modes of deliberation:

"what survives as a candidate for rationality is the mode of deliberation that validates acting on proximate preferences if their strength exceeds the threshold of immediacy and validates acting on vanishing point preferences if the strength of her proximate preferences falls below that threshold."[333]

Resoluteness involves acting on the basis of vanishing point preferences. Therefore Gauthier argues that it is a rational mode of deliberation only when the strength of proximate preferences is below the threshold of immediacy.

How should we understand Gauthier's criterion for deciding between modes of deliberation? Let us apply it to Michelle.

Michelle's preference is *below the threshold* of immediacy if at the time of choice she prefers to eat a cake today but not to eat cakes on any other days, even if she was to experience on all these other days a similar desire for cakes as she does today. In other

[332] 'Resolute Choice and Rational Deliberation', p 21-22.
[333] 'Resolute Choice and Rational Deliberation', p 22

words, at the time of choice, she still prefers to be slim than to eat cakes ... except on this very occasion. On the contrary, Michelle's preference is *above the threshold* of immediacy if she chooses to eat a cake today and considers that she will do so whenever she feels, as she does today, that life is too short to be made of sacrifices.

How practical is this criterion? I really wonder. Indeed, when the proximate preference is below Gauthier's threshold of immediacy, one can interpret it as mere *ad hoc* rationalisation of temptation. Gauthier claims that it is then irrational to base deliberation on such proximate preferences.[334] However, when proximate preferences are above Gauthier's threshold one can wonder if the vanishing point preference has not changed. If, at each decision point, Michelle considers that life is too short to sacrifice her desire for cakes to her shape, it is possibly because she actually overall prefers enjoying cakes more than she prefers being slim. If our interpretation is correct, Gauthier's criterion only reveals a shift of vanishing point preference but does not deal with the classic case of weakness of will or *rational temptation.*[335]

It is all very nice to tell Michelle that it is irrational to base her deliberation on her proximate preferences when they only reflect her temptation for cakes. The reality is that wherever temptation exists, agents run the *risk* of succumbing to it. My argument is that a criterion for choosing between rational modes of deliberation would have to take into account this *risk.* I want to argue now that the *risk* of *rational temptation* actually can make resoluteness rationally inferior to sophisticated choice.

Defence of sophistication over resoluteness

One should assess the rationality of a mode of deliberation in terms of its ability to lead to the overall goal set. When deliberating, an agent chooses the goal she wants to achieve and indeed makes plans that will enable her to reach that goal. If she has to choose rationally between the resolute and the sophisticated modes of deliberation, one can describe her rational calculus as follows.

---

[334] 'Resolute Choice and Rational Deliberation', p 21

[335] The expression 'rational temptation' was borrowed from C. Finkelstein's 'Rational Temptation'. A rational temptation is a 'temptation an agent faces due to a temporary shift in her preference ordering' p 57.

Table1

	Advantages	Disadvantages
Sophistication	Secure achievement of most preferred outcome	Add cost of enforcement Introduce rigidity into plan
Resoluteness	Avoid cost of enforcement Allow for reconsiderations	Run a *risk* of not achieving most preferred outcome

In order to choose between these two modes of deliberation, it is essential to evaluate properly the risk involved in resoluteness. Let us attempt to offer an evaluation of this risk.

Who needs to make a plan at intra-personal level? Of course, we plan our day, our week-end or our holidays; we can also plan our career or our family life. We can have intentions about all these various aspects of our life but all these plans depend very much on context, opportunity or other criterion. The plans that we set for ourselves are usually the ones that enables us to fight a weakness, a recurrent temptation. We need resolutions and planning when we know that our will power is not strong enough to fight the temptations that are on the way to our most preferred goal.

For example, a student knows that, when the time comes to study, he usually prefers watching a good film, reading a good book or going out with the lads. He knows he cannot rely on his will power never mind how much he wants to be successful in his studies. A student who has no difficulty to motivate herself to study does not need any plan or any resolutions. When she has work to do or exams to take, she studies, full stop. On occasions, she probably also has to fight distractions (a good film or a party) but such 'sacrifices' hardly counts as counter-preferential choices since her preference for successful studies still strongly dominates her preference for the envisaged film or party.

To expect someone to choose counter-preferentially at the time of choice is irrational argues Finkelstein. Michelle has a strong preference for cakes when she passes by the bakery and her preference for cakes temporarily outweighs her preference for being slim. Finkelstein argues that people will always choose what they most prefer at the

time of action. She then demonstrates that resolute plans are either *infeasible* because imposing counter-preferential choices or *unnecessary* if proximate and vanishing point preferences are compatible. Her demonstration is simple: the resolute choice theorists would not be able to defend counter-preferential deliberation outside a plan. If counter-preferential behaviour cannot be ruled out by the theory of rationality outside the context of plan-execution, then the fact that an agent does better in terms of his vanishing point preferences would not be sufficient to make his adoption of a resolute plan rational.[336]

We need resolute planning when a plan is rejected as non-feasible by the standard account of rationality. There is no need for an enforcement system and therefore for the competing resolute planning when all the choices involved in the plan are compatible with all the agent's preferences. Resolute planning necessarily involves counter-preferential choice(s). Finkelstein argues that such plans are not feasible whether resolute or not. My argument is similar. I claim that resolute planning is either unnecessary or less rational than sophistication.

Let us consider several situations. In a first case, the vanishing point preference dominates the proximate preferences that are likely to spring out at decision points along the plan. The serious student would not mind watching a film rather than working at times. She does experience a mild shift of preference on occasions but her preference for successful studies dominates her temporary temptations. She can *foresee* that she will resist temptation at each decision point. Her plan is feasible according to the standard account of rationality since no counter-preferential choice is involved: at each decision point, she still prefers successful studies to a good film or a party. She is fully *aware of her motivation* to succeed and it dominates her choices. Since no counter-preferential choice is involved, resolute planning is *unnecessary*.

In a second situation, the agent is unsure about the strength of his vanishing point preference especially in comparison to the strength of his future proximate preferences. He *cannot foresee* what he will do, and he takes a *risk* of not reaching his most preferred goal. We need to prove that taking a risk is always more costly to the agent than enforcement measures. After all, there are various degrees of risk some more

[336] See her 'Rational Temptation'.

manageable than others. What if the risk to succumb seems slim enough *to the agent* and he finds it more rational to take the risk than to put in place enforcement measures?

The agent's *own evaluation* of the risk is the only evaluation that matters in this debate. His evaluation depends on two factors: his attitude to risk and the importance of the preferred outcome in his overall priorities. These two factors are obviously connected. If the outcome chosen is important to the agent, it is clear that he will be less prepared to take risk with it than if it is a very lower priority for him.

Now if the outcome of the plan is not so important to him than one can wonder why he is wasting time making plans and deliberating about them. One could also argue that if the outcome of the plan is a low priority for the agent then the strength of his vanishing point preference is mild. If it is mild, most proximate preferences would outweigh it. In this case, the agent would not act counter-preferentially if he acted upon his proximate preferences. We are back to feasible plan within the standard rationality framework.

From the above, we gather that individuals usually plan where there is an important issue at stake. Individuals have to choose between being resolute and being sophisticated in situations where they usually have a weakness to fight that threatens the achievement of their most preferred outcome. If they had no temptations on the way to their preferred outcome, they could rely on the standard account of rationality. If the issue is important to them so then they will not be willing to take a risk and fail to achieve their desired goal but if they have a weakness, the risk to succumb to temptation on the way is high. Therefore individuals would choose sophistication over resoluteness as a rational mode of deliberation.

If Gauthier has some reservations about resoluteness at intra-personal level, he believes that it is a valid mode of deliberation at an inter-personal level. I agree with him that the concept of resoluteness does apply very differently at an inter-personal level than it does at intra-personal level. However, I want to argue that resoluteness has no chance to take root without some pre-moral standards. Until these standards are in place sophistication remains the only rational mode of deliberation.

Section 3: Rational agency and choice between modes of deliberation - Inter-personal dynamic choices

Gauthier's concept of resoluteness

As a general comment, we can say that Gauthier's solution to the assurance game is largely inspired by McClennen's. However his objections to the concept of resoluteness are similar to the ones raised above in Michelle's case. We can sum them up in two points[337].

- Firstly, he disagrees with McClennen's rejection of the separability principle. Gauthier considers that agents are entitled to reconsider their plans at each node point. Reconsideration does not necessarily mean modification of the original plan. It only means checking that the original outcome-orientated plan is still the best plan available considering the agent's preferred outcome. We shall come back to his reconsideration criteria below.

- Secondly, he finds that McClennen concedes two much to the orthodox theory of rationality in supposing that directly maximising considerations are to brought to bear on each particular choice. Gauthier rejects the concept of endogenous or context dependent preferences and assumes that agents can make counter-preferential choices. Immediate preferences bear on current choices only if they can contribute to bring about the preferred outcome.

Gauthier's agents first choose their preferred outcome. They then make a plan that, if followed, will bring about the chosen outcome. The agent forms an intention to carry out the plan. At each decision point, he will check that the original plan-suggested choice is still the best choice available in the current situation. If the agent still expects to do better in executing his intention than if he had not formed it, then he has an adequate reason to execute it. The agent will then rationally disregard any immediate preferences that are not compatible with the original plan. If the agent no longer

[337] Gauthier has made these two points in several articles. For reference we can suggest looking at 'Commitment and Choice: An essay on the Rationality of Plans' §5.2, p 240-242

expects to do better than if he had not formed the intention, he is entitled to reconsider his original plan.[338] Let us now review our harvesting example.

Gauthier's suggestion to solve our initial harvesting problem is as follows.[339] My aim is that my life goes as well as possible. At t, I understand that to harvest my crops with you is part of a plan that makes my life going as well as possible. But if I don't help you, you don't help me. Therefore to help you is also part of a plan that makes my life going as well as possible. To gain your help I have to promise you that I will help you in return when the time comes to harvest your crops. Once I have benefited from your help on the basis of my promise to help you, I make outcome 2 (in figure 1) intentionally incompatible with my promise. At t+1, I have benefited from your help on the basis of my assurance. Not to help you is incompatible with my prior acceptance of your help. Therefore at t+1 I am left with only one possible outcome according to figure 1, namely to help you. It is the sincerity of my intentions to help you that convinced you to help me.

Like McClennen, he believes that:

- One should assess the rationality of a deliberative process in its ability to lead to the overall goal set. In reaching the goal set, I might have to perform 'irrational' actions i.e. actions which do not fit in the standard rational calculus (to help you in our example).
- From t already I can see that to help you is only my second best outcome but the best achievable outcome. At t+1 the assessment of the situation should be no different.
- The solution to the problem dwells in my intentions. The only way I can convince you at t that I will help you at t+1 is to genuinely intend to do so. In order to be able to give you this assurance sincerely I have to put mental constraints on my future choice.

I would like to challenge the last point defending two alternative solutions to the assurance game highlighting the danger involved in relying on intentions and sincere assurances. Gauthier and McClennen defend resoluteness *almost* as if there was only one supremely rational mode of deliberation, that we all acknowledge its superiority and adopt it. In any of their reasoning, whatever is not supremely rational is

[338] See 'Rethinking the Toxin Puzzle', p 48-49
[339] See 'Assure and Threaten', p 692-697

disregarded.[340] A. Morton claims: 'If we had one kind of ideal rationality we would be able to bind ourselves to cooperative actions... We are not ideally rational. We waver in our resolutions, and we do not see problems until they are upon us.' [341] Here, I want to argue that there are alternative rational options available. I do not claim that we will all deliberate as described below but for the sake of my argument, I just need to prove that some amongst us can use one of the following modes of deliberation. As before, I only need to prove the existence of a risk of non-compliance. Their mere existence threatens the supreme rationality of resoluteness.

Alternative 1: bluffing

It is possible to bluff. It is even more rational to be a successful bluffer than to be a genuine assurance giver. In short, if I successfully bluff you, I can secure my best outcome (you help, I don't) rather than my second best outcome (You help, I help). Bluffing might come more naturally to some agents than to others but the reality is that some amongst us are excellent bluffers.

We must remember the condition of the game: it is not an iterated game, there is no issue of reputation involved and I do not feel obliged by a pre-existing morality to keep my words. Based on pure rational ground, it makes perfect sense to bluff convincingly rather than to give a sincere assurance.

Gauthier or McClennen would object that I would not *be able* to convince you that I will help you later if I do not *genuinely intend* to do so. If such a claim was true it would mean that we are transparent. We are not. Even Gauthier had to assume that we are at best translucent. The same way some agents can disguise their dispositions, some can disguise their intentions. Bluffing exists and given the game structure, an agent has sufficient rational motivation to use her bluffing skills if she has any.[342]

---

[340] We must be careful not to claim that Gauthier and McClennen assume that agents have only one form of rationality. Both of them are perfectly aware of the various forms of rationality available from one agent to an other (See for example Gauthier's 'Public reason'). My only concern is that, at times, they seem to consider that what is not resolute or 'constrained' reflects a case of irrationality.

[341] See 'Psychology for Cooperators', p 167

[342] This argument is probably the main argument I hold against Gauthier's solution to the toxin puzzle as argued in his 'Rethinking the Toxin Puzzle'.

The main argument that makes resoluteness a rationally superior mode of deliberation is that it saves on the cost of an enforcement mechanism. We looked at the limits of this argument at the intra-personal level with Michelle. At the inter-personal level this argument meets a new challenge. Resoluteness is a mere mental constraint imposed by rational deliberation. In a one-off game, the only way one can 'check' the resoluteness of others' is *to be convinced* by their intentions.[343] It does not take much imagination to see that there is here an obvious room for error that a good bluffer would endeavour to exploit.

In our harvesting game, if I convince you that I have the intention to help you after you have helped me, you will *assume* that I am resolute and you will not feel the need for an enforcement mechanism whether I am resolute or just a good bluffer.

Alternative 2: myopic choice

This alternative is based on Gauthier's concept of resoluteness. His agents can reconsider their plan at each decision point and can make counter-preferential choices. Let us review once more our harvesting example.

At t I can see that it is in my best interest to help you if I want to gain your help. To obtain it I am *prepared* to help you in return. I adopt a resolute mode of deliberation. I want to avoid the cost of an enforcement mechanism (for example, I leave a provisional penalty to a third party: if I help you, the third party will give me back my money; if I don't help you , he will give it to you) and I form the intention to return your help. You are convinced by my sincere intentions and we decide not to have recourse to a third party. At t+1 however, we have harvested my crop and I no longer see the benefit I could gain from helping you. The question is now: does my original intention give me a reason to help you? Does my intention provide me with a *rational obligation* to return your help? I believe we can answer 'no' to both questions. Let us see why.

[343] Obviously in an iterated game, agents can also refer to past history of the game.

In a very good article[344], J. Broome argues that intentions are not reasons. His main argument is rooted in M. Bratman's famous 'bootstrapping objection': reasons cannot be created out of nothing. If something is not a reason, it does not become one because an agent takes it to be one. According to Broome, an intention is at best a *normative requirement.* If at the time of action, an agent has not changed her mind and her intention has not been *'repudiated'*, then it normatively requires her to do what she intended.

I believe that Gauthier would not dispute Broome's argument. He is the first one to claim that the weight of an intention on further reconsiderations depends on its strength. Not every intention is strong or firm enough to support an assurance[345]. He is careful never to claim that intentions *are* reasons. Gauthier only refuses to divorce current intentions from future actions. Unlike Bratman, he rejects the contrast between deliberation about what one intends now and deliberation about what one will do later. He argues that when deliberating about whether to give a sincere assurance or not, one cannot *simultaneously* intend to reciprocate later and therefore give a sincere assurance *and* decide not to reciprocate. Gauthier assumes that we deliberate only *once* about what to do, at the beginning of a planning process. In deliberating, one decides on the outcome to go for and on the set of actions required to achieve it. Despite their temporal disparity, *these future actions are part of the same deliberation process* and must therefore be consistent. Within a plan, one does not deliberate anew, one merely reconsiders.

At *t+1*, when I have to decide whether I will return your help or not, Gauthier claims that I must review the situation *within the deliberation framework I had at t,* when I made my plan. At t, I could foresee that in following my plan I would only achieve my second best outcome in sincerely assuring you that I will return your help. If the situation at t+1 is as I expected it to be then I have no ground to reconsider *although continuing on this plan would no longer bring about the best outcome available from my new standpoint.*[346]

---

[344] See 'Are Intentions Reasons?'
[345] See 'Intention and Deliberation', p 45
[346] See 'Intention and Deliberation', p 48-49

Do rational individuals really deliberate this way? If I was to deliberate *sequentially* at each decision point, I would be called myopic. But better be myopic than blind! In effect, agents are required to close their eyes to outcomes technically still available to them and better than what they could achieve were they to pursue their plan. Gauthier would argue that I was able to secure my second best outcome *because* I had the intention to return your help. Had I not made the plan to return your help, I would not have been able to give you a sincere assurance and therefore I would not have benefited from your help. My plan enabled me to secure my second best instead of my worst at a time my first choice was not available. So now I must be faithful to my plan. This reasoning sounds more like a moral than a rational argument: I must be faithful to a plan that served me well although now it is no longer rational to keep it.

If I agree that an intention can support a sincere assurance and therefore lead an agent to keep her assurance, I also believe that an intention does not generate an obligation. Therefore it cannot secure compliance with the action planned. Some agents do reason sequentially. At t, it is possible to realise the benefit I could gain from returning your help and genuinely intend to do so at t+1. Yet, it is also possible to deliberate anew at t+1 rather than looking backward on past deliberations and decide not to help you. The problem is that I was able to give you a sincere assurance and you had no way to *know* whether I was resolute or merely myopic. Again, we are not transparent: you have no means of measuring the strength of my intention or assess whether I will carry out my plan. My argument is that this degree of uncertainty must be taken into account when choosing between rational modes of deliberation.

Intentions are not reasons and a fortiori, they do not create any rational obligations whether with oneself or with others. An agent is not bound by his plan, his intention or his assurance. Gauthier or McClennen would argue that she has a rational 'mental bound' to her plan. But this mental bound is as strong as the agent can or want to make it. An intention does not necessarily travel well. An agent can 'suffer' from a temporary shift of preference between the time at which she formed the intention to act later and the time of acting upon her intention. The strength of her 'mental bound' can be outweighed by the strength of her proximate preference. In such a case, she has no *obligation* to stick to her plan or keep her assurance. In the meantime, she was able to give a sincere assurance and was convincing enough to be believed.

Defence of sophistication against resoluteness.

Our argument is very similar to the one presented at intra-personal level. Our main issue is that Gauthier and McClennen focus essentially on the agent *giving* the assurance. They argue that if he is sincere and adopts a resolute plan then he can convince his partner in the game of his resoluteness and together they can avoid paying for enforcement measures.

The alternatives presented above highlight one aspect: agents are not transparent. The agent giving the assurance can be genuinely resolute but the agent *receiving* the assurance has no way to *know for sure* whether the assurance received is genuine or not. If he accepts not to put in place enforcement measures, he runs a *risk* of being deceived or let down. We are back to the rational calculus described in table 1. Our demonstration of the superiority of sophistication over resoluteness follows the same path.

In short, the bigger the risk of being deceived or let down, the less likely agents will be willing to rely on the resoluteness of their partner. The game being one off, freed of pre-moral standards and without any issue of reputation being at stake, the risk taken by the agent receiving the assurance is rather substantial and will justify the cost of an enforcement system.

I believe, like Gauthier and McClennen that individuals are able to be resolute and that resoluteness *could be made* more rational than sophistication. However, resoluteness is not a natural mode of deliberation. The natural mode of deliberation is described in the standard account of rationality. It is what Gauthier once called straightforward maximisation. Resoluteness is an artificial or social mode of deliberation. Unlike Gauthier and McClennen, I do not believe that it will come naturally to straightforward maximisers precisely because of the risk taken by the ones who will have to perform first in any assurance situations. As long as resoluteness is not widely generalised, sophistication will remain the only rational mode of deliberation. Let me explain.

If we want resoluteness to get started in a society, agents must *first* change focus from the pursuit of individual benefit to the pursuit of mutual advantage. Contrary to what Gauthier and McClennen claim, *resoluteness is not a stand alone rational concept.* We are back to the comment we made about Gauthier's bargainers in chapter IV. What Gauthier calls morality is an artificial virtue developed by individuals in social interactions and that is supported by a change of focus from pure individual gain to mutual benefit. For resoluteness to be a more rational mode of deliberation than sophistication, this change has to *pre-exist* the assurance game or any social interaction. If agents have internalised this shift of interest, then they have acquired the social virtue called morality. Only then can the risk of defection (or deception) be significantly reduced and enforcement measures avoided. Moral standards have to pre-exist the assurance game for resoluteness to be a rational mode of deliberation. This is incompatible with the assumptions of the game as described above.

Gauthier is aware that at a social scale, morality and obligation have to pre-exist compliance with institutions and practice rules. Once again he finds inspiration in Hobbes. We remember that through *natural reason,* agents derived the laws of nature. Gauthier argues that Hobbes' laws of nature create an obligation to comply with the sovereign's civil laws.

CHAPTER VI: A RE-INTERPRETATION OF HOBBES' CONTRACTARIAN THEORY AND PUBLIC REASON

In chapter I we saw that according to Gauthier, Hobbes distinguishes natural and conventional reason. *Natural reason* enables men to achieve their primary end i.e. self-preservation and leads them to lay down their right of nature. But as they lay down their right of nature they also lay down this very natural reason as their aim switches from self-preservation to peace i.e. from individual to mutual advantage. A new conventional reason then supersedes natural reason. Since right reason is individually devised, it is no longer a reliable tool to harmonise all individual behaviours towards a mutual advantage. Therefore, Hobbes needs to introduce an arbitrator (the sovereign) to decide on the common right reason, the *conventional reason*. Each has grounds to accept it as long as it is common knowledge that most persons accept it and expect others to adhere to it.

Gauthier claimed that Hobbes's morality arises with the laws of nature that establish a set of conventions constraining each man's maximizing activity and distinguishing right from wrong. Gauthier suggested rephrasing the second law of nature as follows:

'As long as each person appeals solely to his natural reason, there can be no security to any man of living out the time that nature ordinarily allows. Thus a man must be willing when others are so too, as far as he shall think it necessary for peace, to lay down natural reason, and be contented with a standard of reason which allows him so much liberty against other men, as he would allow other men against himself'[347].

In the 1990's Gauthier's focus evolved and he developed two new themes in his interpretation of Hobbes. The first one was about Hobbes contractarian theory of law; he attempted to prove that the laws of nature, as commands of reason, do create an obligation to obey the sovereign. The second theme was a development of the conventional reason which he then named public reason. This little detour through Gauthier's interpretation of Hobbes will enlighten his latest research on a contractarian theory.

[347] 'Thomas Hobbes : Moral Theorist', p557

I repeat what I said before. Gauthier is a Hobbes scholar, I am not. In this chapter, wherever I refer to Hobbes, it is from Gauthier's writings. Whether his interpretation of the Master is correct and conventional or not is irrelevant for the purpose of this chapter. My only reason to develop on Gauthier's interpretation of Hobbes is to highlight the link between his understanding of the classic philosopher and his research on his own contractarian theory. Hobbes is an obvious source of inspiration in his research; his work on Hobbes is an integral part of his philosophy.

Section 1: Hobbes' theory of law.

In the state of nature, each agent is free and has a right to do whatever is necessary to further her own good which Hobbes assimilates to her self-preservation. Persons find themselves in competition and the right of nature of one opposes the right of others. Through reason, agents realise that in order to find security they all have to agree to give up their natural right. The laws of nature are the mere outcome of each individual's prudential reasoning. They are not laws as such and certainly not edicts of a lawgiver. However, the exercise of these laws effectively curbs individual rights of nature. Gauthier claims that 'in requiring us to curb the exercise of natural right, the laws of nature bring a moral order into existence.'[348] I would like to stop a moment on this quote and attempt to understand it.

As we saw before, in Gauthier's world, morality is an artificial virtue.

- It is *artificial* firstly because it emerges within social interactions and not in the state of nature. It is also artificial because it does not come naturally to individuals, it is derived through instrumental reason. It is supported by an artificial form of rationality where the pursuit of individual interest is supplanted by the pursuit of mutual advantage. Morality is the artificial impartiality that prevails in social interactions.

- Morality is a *virtue* because it applies to human beings as rational agents. Morality becomes a virtue once individuals have *internalised* the change of rationality or the change of focus required. Morality is 'an impartial constraint on the direct pursuit of individual utility'.[349] It is a *self-imposed* constraint on behaviour.

So how are we to understand the sentence 'the laws of nature bring a moral order into existence'? I believe that the only way to read it is as follows. The emergence of the laws of nature creates moral standards, i.e. they dictate what individuals *should* do if they are rational and they want to secure their self-preservation. The laws of nature do not create the actual virtue but only the normative standards of the virtue. The

[348] 'Thomas Hobbes and the Contractarian Theory of Law', p 20
[349] *Morals by Agreement*, p 95

acquisition of the virtue is a personal 'effort'. It is an effort because it requires fighting the human (or 'natural') rationality and replacing it by a 'social' one. In Gauthier's interpretation of Hobbes, it requires the acceptance of public reason over private judgement or natural reason. I believe it would be a massive mistake to assume that instrumental reason on its own is sufficient to tame human natural rationality. The fact that the change of rationality is a prudential edict does not mean that agents will actually perform the change required. The prudential edict creates the standards for morality but it does not create the virtue. The virtue appears only once the prudential command has been *internalised* by the agents.

Let us illustrate our argument with an example. Most of the rules of the road could be the outcome of instrumental reasoning. It is in our best interest not to use a mobile phone while driving, not to speed, or to stop at a red light. These rules are also in our social, mutual interest: when these rules are kept, roads are a much safer place for all of us, drivers, passengers and pedestrians. If we were in the hypothetical situation to decide whether we would choose these rules or not, we can safely assume that most of us would derive and agree on them. It does not mean that we keep them. It has rapidly become a necessity to make these common sense rules into laws and to enforce compliance to them. Every day people are caught by enforcement cameras or the police. We each have a good reason not to keep such or such rule (I am late and the roads are free) and we think that it will go unnoticed. We give priority to our local momentary needs or reasons over the mutual interest. Is it a case of irrationality? Or does it only mean that we have retained our 'natural reason' (or human natural rationality) and we have not yet switched to the social one?

Gauthier is fully aware that the link between the dictate of instrumental reason and the emergence of morality is an effort from each individual on his natural calculus and behaviour. '*In requiring us to curb the exercise of natural right*, the laws of nature bring a moral order into existence.' However, the inherent weakness of the natural laws, as mere prudential edicts means that they need to be backed by a powerful sovereign. Hence Gauthier continues:

'the laws of nature are in themselves insufficient. Only insofar as their rational authority is backed by the coercive power of the sovereign, can human beings be expected to conform to them.'[350]

We need now to turn to the creation of the sovereign and the link between natural and civil laws in order to understand the role of the moral standards created by the laws of nature in Hobbes' contractarian theory of laws.

Through reason, agents realise that in order to find security they have to all agree to give up their natural right to govern themselves to an individual or group, namely to the sovereign. Hobbes' sovereign is an artificial person, constituted by the authorisation of his subjects. At this point Gauthier notes the ambiguity of the sovereign. On the one hand the sovereign acts as the universal agent of his subjects which implies subordination to his principal. On the other hand, in giving up their right to the sovereign, persons have agreed to subordinate themselves to him.[351] The key point to reconcile these two contradicting claims is to remember that, for Hobbes, there is no obligation on any man which does not arise 'from some act of his own'.[352]. The sovereign's words and actions are not his own by they represent the words and actions of his subjects. Therefore, 'in owning the sovereign's actions, each person is bound by them as if they were her own.'[353] Gauthier argues that, underlying the authorisation of the sovereign, the agents' intention involves giving up the right to govern themselves and so the assumption of an obligation to obey.

'A command is a law if it is issued to those previously obligated to obey the issuer. For Hobbes, this obligation of obedience is assumed in the act by which those to whom the command is issued have authorised the actions of the person or group who issues it.'[354]

Gauthier insists that for Hobbes there *is no obligation to obey a powerless sovereign.* The purpose of the subjects in giving up their right to the sovereign is to secure peace and their self-preservation. No power to the sovereign means no security and no peace and therefore there is no rationale in authorising it. Firstly, an obligation to obey goes hand in hand with the power of the sovereign to protect and secure peace. It would be irrational to entrust the sovereign with my security if he does not have the power to

[350] 'Thomas Hobbes and the Contractarian Theory of Law', p 20
[351] 'Thomas Hobbes and the Contractarian Theory of Law', p 11
[352] *Leviathan*, chap 21, p 561
[353] 'Thomas Hobbes and the Contractarian Theory of Law', p 11
[354] 'Thomas Hobbes and the Contractarian Theory of Law', p 12.

protect me. Secondly, if the subjects are to be under an obligation to obey, the sovereign must have the power to enforce his commands. The sovereign laws would be vain if they were not enforced. The expectation of punishment does enter the individual rational calculus: a person who breaks the laws must expect to be punished so that the cost of breaking the law outweighs the benefit she sought to gain.

Now Gauthier details the connection between natural and civil laws. As we saw, he considers the natural laws as theorems of reason.[355] On the one hand, the laws of nature are part of the civil laws since the sovereign gives the laws of nature determinate content. The sovereign is responsible for interpreting the laws of nature and for giving them their content. On the other hand, the civil laws are part of the laws of nature since the laws of nature dictate obedience to the civil laws. Gauthier is now able to distinguish between the *legal obligations* generated by the civil laws from the *moral obligation* generated by the natural laws. 'One has a moral obligation to carry out one's legal obligations.'[356] Obedience to civil laws depends on a prior obligation. This prior obligation was acquired through instrumental reasoning and as such was a voluntary act. The natural laws are not proper laws until the sovereign gives them determinate content with the civil laws and enforces compliance to them. Reciprocally, the civil laws are not laws unless they are based on the prior obligation generated by the laws of nature.

Section 2: Public reason

The second key idea in Gauthier's interpretation is that in giving up their right of nature, agents also give up their individual natural reason. There can be only one reason applying to all and it has to be the sovereign's. The sovereign's judgement of right and wrong supplants private judgements and provides a standard of public reason which is expressed in civil laws. In appointing the sovereign using our natural reason, we have given our approbation to public reason. 'Public reason supplants private reason but is founded on it ... It is itself a rational creation.'[357]

---

[355] For a full demonstration of his claim see his 'Hobbes: The Laws of Nature'. In this article published in 2001, he demonstrates that the laws of nature are primarily rational precepts and only secondarily commands of God and commands of the sovereign.
[356] 'Thomas Hobbes and the Contractarian Theory of Law', p 21
[357] 'Thomas Hobbes and the Contractarian Theory of Law', p 13

Why is it a rational creation? Gauthier explains: Hobbes is aware of the fallibility of human rationality. In a way there is no right reason until men agree to accept the reason of one man or group of men as right reason. The method of rational deliberation is then transformed:

'The individual mode of deliberation, in which each person judges for herself what she has reason to do, is supplanted by a collective mode, in which one person judges what all have a reason to do.'[358]

Once individuals have agreed the Judge or Arbitrator's reason as right reason, then rationality is exercised by him. An agent who then acts on the basis of his own judgement would exhibit a deficient rationality. The correct rationality would be to act in conformity with right reason.

In effect, it cannot be so difficult. Although, the sovereign remains the sole interpreter of the natural laws and no person may put her private reason in place of the public reason, each may and must use his private reason to determine what the public reason is. Gauthier argues:

'each person than has rational access to the chain that leads from authorisation to law. By grounding public reason in each person's private reason, Hobbes provides both a moral foundation for civil law, and a rational foundation for morality.'[359]

I would like to stop here just to make a couple of comments about this interpretation.

• Firstly, I believe that this explanation gives the full measure of the artificiality of this mode of deliberation. Natural reason is our human reason, the one we follow, listening to our interior voice. Through instrumental reason, we come to realise that we have to harmonise reasons amongst us, social persons. This harmonisation involves surrendering our rationality to follow someone else's. It is asking a lot from an intrinsically rational being whose primary feature is to reason.

• Secondly, we can now see better the connection between Gauthier's own research on a contractarian theory and his interpretation of Hobbes. As we saw in the previous chapter, Gauthier argues that *rationality is to choose the correct mode of deliberation.*

---

[358] 'Public Reason', p 25
[359] 'Thomas Hobbes and the Contractarian Theory of Law', p 24-25

We see now that right or public reason is the mode of deliberation that rational individuals would choose. One more step and we get: public reason, i.e. a common standard of deliberation and judgement, w*ould be* the one chosen by *any* rational individual reflecting on the most rational mode of deliberation from a *hypothetical* perspective.

Gauthier continues. Public reason is not only the means to resolve the controversies between the various private judgements, it is also meant to resolve the conflicts of interests between individuals. Even if reason was not fallible, we each pursue our own good (i.e. what is rational for us to pursue). Different agents might rationally reach different decisions about what is good and that, in itself, can create conflict and controversies. The Arbitrator, in harmonising private judgements also decides on the good and on what constitute moral actions. Each individual's reason is directed to his own good whereas the sovereign's reason is directed to 'the impartial resolution of controversies among individuals'.[360]

At this stage of the interpretation Gauthier steps in the debate and suggest restricting the scope of public reason: 'I accept what I take to be Hobbes' idea that public reason justifiably supplants the private reason of the individual, but I want to insist that he does so only within limits.'[361] The reason why Hobbes has to give full power to the sovereign is because individuals, once they have set up the Arbitrator, do retain their natural mode of deliberation.[362] They will never accept the Arbitrator's reason in place of their own and they will carry on deliberating about the laws, led by their individual judgement of the good. Public reason will never be brought into being.

Instead Gauthier makes the following suggestion.

'The core of public reason is to be found in the transformation it effects in the deliberation of those subject to it... Public reason provides the citizen with a common standard for determining right and wrong, good and evil'.[363]

---

[360] 'Public Reason', p 33
[361] 'Public Reason', p 35
[362] We come back to a comment that Gauthier made before about Hobbes i.e. that Hobbes gives a political rather than a moral answer to the Fool.
[363] 'Public Reason', p 36

Gauthier suggests that agents agree to authorise a public person to judge and will in their names. Agents through their authorisation agree to treat as right reason the public person's judgement within the scope so defined. Once the public person is constituted, citizens can pursue their own good as long as it does not clash with right reason. In order to do so agents must have *internalised* the common standards established by public reason. Public reason becomes a 'construction from the individual reasons of the member of society.'[364]

Morality arises once agents have internalised these common standards. We shall ask ourselves later whether, based on pure instrumental reason, one can actually internalise someone else's reason. My claim is that one cannot. It takes a lot more than mere reasoning to change our intrinsic mode of deliberation especially to such an artificial one. Morality cannot be the outcome just of instrumental reasoning. But let us finish with Gauthier's interpretation of Hobbes and the problem of punishment.

Public reason takes normative precedence within its scope of authorisation over what would otherwise be individually rational. However, normative precedence does not guarantee motivational efficacy, hence the need for enforcement to ensure that the normative precedence of public reason is motivationally effective. The problem of a contractarian theory is that it must accommodate punishment within the limits set by individual rationality when no rational person would directly will his own punishment. Punishment can only be portrayed as a necessary evil: the benefit expected from securing the others' compliance outweighs the cost of being caught in case of one's own violation. [365]

As we can easily suspect Gauthier cannot be entirely satisfied with such a 'weak' rational calculus. He argues that if an agent found it reasonable to authorise public reason, than surely it is *less costly* and therefore more rational to voluntarily comply with its requirements.[366] He sees agents as more sociable than Hobbes. He hopes 'to use the language of rational agreement to articulate the basis and character of the institutions that enable the secure establishment of such co-operative and sociable

[364] 'Public Reason', footnote 7, p 25
[365] See 'Thomas Hobbes and the Contractarian Theory of Law', p 32 - 33
[366] See 'Public Reason', p 41

interaction.'[367] Again he remains very evasive about the concrete reality of such a rational agreement. As I will argue later, rationality is not sufficient to establish what he calls 'civic friendship'.[368]

[367] 'Thomas Hobbes and the Contractarian Theory of Law', p 34
[368] See 'Thomas Hobbes and the Contractarian Theory of Law'

Conclusion

Before we turn to Gauthier's latest contractarian theory, I believe we should recapitulate what he wants to borrow from Hobbes: firstly, the concept of prior moral obligation contained in the natural laws and secondly, the concept of right reason as a common standard of deliberation and decision, judgement and evaluation. Gauthier identifies five main normative features of Hobbes' rational contractarian theory of law, skimmed from Hobbes own details. He suggests that a plausible normative theory of law would have to build on the following general Hobbism.

1) Law depends on a prior normative relationship between lawgiver and those to whom law applies; the subject has an obligation to obey the lawgiver.
2) The obligation to obey the lawgiver must issue from the rational agreement of those obligated.
3) The extent of their obligation is determined by their intention in authorising the lawgiver. In other words the scope of law is restricted and Gauthier specifies: it is restricted to the regulation of interactions for their mutual advantage.
4) In authorising a lawgiver, the members of society agree to treat the law as the expression of public reason. Each puts aside her own private judgment and accepts the judgment of the laws.
5) The obligation to obey the lawgiver depends on his power both to ensure mutual advantage and enforce his edicts.[369]

His primary focus is on the obligation created by the laws of nature and the authorisation of the sovereign. The obligation emerges with the agreement of all. Moral standards arise with the laws of nature and create an obligation to conform to public reason. The backbone of *Morals by Agreement* and of his contractarian theory is all contained in this general Hobbism. Gauthier concludes:

'the contractarian idea is that reasonable conduct is to be evaluated ... in terms of the pursuit of one's own interests constrained by the requirements of fair mutual advantage. So understood, reasonableness involves the internalisation by each member of society of

[369] 'Thomas Hobbes and the Contractarian Theory of Law', p 15-16

the same standard of decision-making that, for the contractarian, governs legitimate legislative and judicial activity.'[370]

It is now time to turn to his latest contractarian moral theory. Gauthier has written several articles on the topic but none develop a complete and overall theory. However based on *Morals by Agreement*, his latest work on resoluteness and his new development on Hobbes' interpretation, we should be able to draw a fairly accurate picture of what Gauthier would consider as an adequate contractarian theory.

[370] 'Thomas Hobbes and the Contractarian Theory of Law', p 26

CHAPTER VII: GAUTHIER'S LATEST CONTRACTARIAN MORAL THEORY

This morning I opened my mail to find an unpleasant surprise: I was caught speeding at 43 mph in a town centre where the speed limit is 30 mph. The upset was double: firstly because I had a £60 fine and secondly because I usually consider it irresponsible to speed in town streets. Had I been at a bargaining table where the speed limits were to be decided, I most probably would have lobbied for a 30mph speed limit in towns. So if I firmly believe that speed should be limited to 30 mph in town centres and I am caught speeding at 43 mph, am I being irrational? We shall attempt to answer this question in this chapter.

Introduction

I believe that Gauthier's latest work is characterised by a complete revolution in his approach to contractarianism. However some features remain.

Common features of his contractarianism

Firstly, his contractarianism is *hypothetical.*[371] Gauthier does not describe what happened or should happen in establishing a social contract. His approach is normative rather than descriptive. However, there is a difference in his approaches. In *Morals by Agreement*, Gauthier started from a state of nature in which agents were using the orthodox rationality referred to in both economic and game theories as individual utility maximisation. Interactions did not yield an optimal outcome and agents came to realise that they had to constrain their pursuit of individual interest. The state of nature represented natural interactions and was opposed to society. In his most recent writings,

[371] Gauthier has developed that aspect of his theory in numerous (if not most of his) articles. Some are referred to in chapter III. We can also mention 'Public Reason' or 'Political Contractarianism'.

Gauthier provides agents in the state of nature with tools to step back and evaluate their existing practices and institutions. His approach is still normative but the purpose is justificatory. Agents straight away adopt a mode of deliberation specific to social interactions. Equipped with this new mode of deliberation, agents should be able to reflect on their conditions of interaction based on purely rational grounds and in complete abstraction from any environmental influence. The integrity of the rational calculus should secure the agreement of *all* members of society who would attempt this normative exercise. This retrospective approval of a practice after rational deliberation is called the 'contractarian test'[372].

Secondly, his contractarianism is based on *individual rational deliberation*. Gauthier remains convinced that the key to morality is a change of rationality. Based on that change, Gauthier demonstrates how agents rationally derive moral standards understood as constraints on their behaviour. Even if Gauthier's contractarianism takes a political turn in its most recent version, he remains faithful to the idea that agents, in changing their mode of deliberation, internalise the constraints required by the moral standards so derived and become moral. Roughly, he assumes that if someone rationally agrees with a rule (i.e. the rule passes the 'contractarian test') then the person has a sufficient reason for conforming to it. Even better, if the rule passes the contractarian test, then each agent is entitled to expect others to conform to it. Therefore, according to Gauthier, there are two theories within a contractarian moral theory: a theory of justice that characterises the common moral standards that should constrain the agents and a moral theory that deals with the internalisation of these common standards by the agents. Gauthier wrote: 'That some justification must be afforded in support of deliberative constraints to the constrained individual is at the core of what I call a contractarian moral theory.'[373] The only justification that can be appealed to relates to the deliberative rationality of the members of society.[374] Only once a constraint is so justified can the agent agree to it and comply with it.

[372] See for example 'Political contractarianism'.
[373] See 'Mutual advantage and Impartiality', p 130
[374] See 'Political Contractarianism', p 141

A new approach

Gauthier still wants to address the Fool's challenge and give him a moral answer but this time he wants to distance himself from the constraints imposed by economic rationality. Gauthier no longer tries to artificially rectify market failures. Instead, Gauthier refers to the market in order to emphasize the different mode of deliberation between this naturally mutually advantageous venture and society as 'a cooperative venture for mutual advantage'[375]. Talking about the role of politics he writes:

'One way of characterising this view of the role of politics is to represent it as making possible market success (by eliminating force and fraud) and remedying market failure (by eliminating free riding). But the reference to market success and market failure might misleadingly suggest the subordination of the political realm to the economic... The rationale of politics is to supplant or constrain strategic interaction...'[376]

Politics is no longer in the service of economics, it has a rationale and a purpose of its own. In effect, it means that Gauthier negates any connections between the rationality required in politics and the one required in economics. He is freed from all the constraints imposed by the theory of rational choice and his theory is no longer about solving the problem of strategic interactions in game theory. Forget the artificial distinction between internal and external rationality, between straightforward and constrained maximisation and all these attempts to remain connected to the economic rationale. Gauthier starts with a clear page. As we saw above, he has resourced himself in Hobbes and he has a set of new philosophical tools to tackle his new challenge.

This change of approach has a huge importance in the coherence of his contractarianism. We remember that Gauthier is convinced that it is intrinsic to human nature to be able to change their rationality. In chapter IV, I wondered when that change was supposed to take place in the process described in *Morals by Agreement.* I noted that the change was supposed to take place after bargaining and I emphasised the incoherence caused by this timing. In assuming a complete new rationale to politics, Gauthier solves this incoherence de facto. Politics is about a completely new mode of deliberation that applies all along the process.

---

[375] Gauthier regularly borrows this description from Rawls' *A Theory of Justice*, p 4
[376] 'Constituting Democracy', p 316

In chapter IV I suggested that, in *Morals by Agreement,* morality was precisely about this change of rationality. What has happened to Gauthier's project to derive morality from pure rational grounds? Where does morality fit in this new picture? I intend to answer these questions and address the dilemma posed by my speeding ticket in this chapter.

In order to do so I will attempt to draw a picture of Gauthier's latest contractarian theory. I insist on the fact that what follows is a *reconstruction* of his theory derived from his most recent articles and based on my understanding of his approach. Since *Morals by Agreement,* Gauthier has never developed a new complete theory. He has merely commented upon contractarianism here and there, continuing to refer to *Morals by Agreement* for the broad lines.

Section 1: Strategic interactions versus deliberative rationality

Overview

For the first time Gauthier *opposes* the rationale of the market to the rationale that prevails in the political area. The market is the area of strategic rationality whereas politics is the field of deliberative rationality. Deliberative rationality takes over when the perfectly competitive market fails to bring about natural harmony.

I would like to emphasise the fact that for Gauthier, market rationale continue to cohabit with the rationale of politics. Political tools are only there to take over wherever the market fails.

'We do not want to replace the strategic interaction of free individuals, in those circumstances in which they are able to achieve optimal outcomes in which benefits are matched to costs, with the strategic interaction of an unrestricted majoritarian politics, which can by no means be expected to yield either optimal or fair outcomes. In effect then, we want to achieve a constitutional balance between two modes of strategic activity – market and government.'[377]

The field of government is probably the one of public reason i.e. 'on those matters and in those respects that significantly affect the interactions of the citizens and the public good available to them.'[378] Gauthier distinguishes between the two forms of rationale and he expects them to exist side by side.

The implementation of an artificial harmony in the fields not covered by the market involves a double harmonisation exercise. The first level of harmonisation is through a centralised mode of decision-making.[379] The centralised decision-making institution, or government, will be responsible for structuring agents' interactions in such a way that its product, the social decision, will yield a fair and optimal outcome. The second level of harmonisation is through public reason as expressed in laws. Public reason is the expression of right reason and agents have to submit their private reason to the authority of public reason.

---

[377] 'Constituting Democracy', p 325
[378] 'Public Reason', p 37
[379] 'Constituting Democracy', p 325

The functioning of this artificial harmony involves constraints on strategic interactions. Gauthier excludes market-like dynamic interactions between individual life-plans and chosen social norms. The framework of social interactions is chosen once and for all.[380] Social structures have to pre-exist the agents' individual life-plans. Or said otherwise, agents have to bend their individual life plans to the requirements imposed by the laws and institutions.

We remember that the purpose of the contractarian approach is justificatory. The contractarian test is to ensure that an existing rule would be accepted by deliberators. Gauthier insists that the only common norm of deliberation between all members of society is rationality. Religious or other pre-established norms are not suitable since they are not common to all the deliberators. Rational deliberation is the only common ground. Not only all deliberators are required to justify their social norms on the sole basis of rational deliberation but deliberators who would reject deliberative rationality *cannot* be included in society.[381]

This point is essential to Gauthier. Agents can make plans for their lives but they must do so within the existing social practices and institutions. Individual life-plans must bend to the social structure. If an agent has some non-rational normative requirements (such as religious requirements), he can include them in his life-plan as long as they fit in the social structure. The social structure should in no way be the outcome of non-rational deliberation or suffer from the influence of any non-rational norms. On the contrary, during deliberation, life-plans have to be reasoned and submitted to the critical considerations of the fellow bargainers for approval.[382]

We can see how the concept of resoluteness will fit in this picture. Agents must decide on what is in their overall best interest. Once they have chosen their most preferred outcome, they must constrain their choices and life plans accordingly. Deliberative rationality is about choosing the best outcome and aiming at it. It is the capacity to be resolute.

---

[380] Gauthier refers specifically to the Constitution that frames the functioning of social interactions in a democratic state. See 'Constituting Democracy' p 327
[381] See 'Political Contractarianism', p 142
[382] See 'Constituting Democracy' , p 321

Gauthier writes:

'[A]ny actual system must largely pre-exist the real person who enjoy the benefits and fall under its constraints.'[383] Or again '[V]alues, aims and plans are all subjects for deliberation and choice... The contractarian supposes that social practices are to be justified, not primarily by showing how they accommodate independently-given life-plans, but rather by showing how they permit and encourage the formation of life-plans for which they offer the prospect of satisfying fulfilment.'[384]

About rationality

As emphasised before, rationality is an individual feature. An institution, a rule or a practice is rational if it is considered as such by a rational individual assessing it. In differentiating between economic and political rationality, Gauthier has to define a new form of human rationality. If political rationality is no longer about the maximisation of individual utility as it is in the economic realm, what is it?

Gauthier explains: beings act for reasons. Reasons are representations of the world. Beings can have the capacity to represent some features and to be moved by the content of those representations. Rationality is the capacity to be moved by the *appropriate* representations. It is the ability to assess the appropriateness of the reasons. A rational agent determines and endorses the fit between her actions and her beliefs.[385] Rationality is therefore the normative extension of reason.

The most striking feature of this definition is its vagueness. How to evaluate the appropriateness of representations? One can give any content to this assessment and call it normative. Rationality so defined can only be fallible. Not only is rationality fallible, but it necessarily generates conflict of interests. We do not deliberate from the same standpoint. What is rationally good for you in your situation can be rationally bad for me in my circumstances. A controversy may arise between us about the state of affairs we each want to bring about. That is why we had to rely on a public agent's

[383] 'Political Contractarianism', p 133
[384] See 'Political Contractarianism', p 138
[385] See 'Public Reason', p 20

reason.[386] As we saw above, Hobbes and his interpreter Gauthier both agree that the only way to resolve the intrinsic weakness of rationality in social interactions is to appoint a public agent and to abide by her representations of the world.

Gauthier goes further. Agents do not only *abide* by the public agent's concept of rationality they also *internalise it.* In doing so they alter their mode of deliberation; they give a common content to the *appropriate* representations of the world, namely *the public agent's.* They deliberate according to her representations. 'The core of public reason is to be found in the transformation it effects in the deliberation of those subject to it.'[387] Agents assess the appropriateness of representations of the world not according to their own perception but according to the common and mutually agreed perception of the public agent.[388] Since they have endorsed her representations as their own by appointing her, Gauthier claims that this mode of deliberation is still individual. But is it really?

First of all, I find surprising to expect individuals to continue to use their private reason in market interactions and to expect them to use public reason in social ones. This dual rationality makes the internalisation of public reason more difficult and less probable.

More importantly, this transfer of individual rationality from each citizen to a public agent seems far-fetched. Can an individual actually internalise someone else's reason? Is it really possible for someone to picture accurately someone else's representations and deliberate according to them?

Gauthier is aware of this difficulty and he later comes back to it. If he remains convinced that the construction of social institutions relies importantly on the

---

[386] Reason here is a synonym of rationality. Gauthier wrote: 'Although some persons have insisted that rationality must yield some form of universality or unanimity, I am unconvinced by their arguments. No doubt that two agents, similarly placed ... will reach similar conclusions about what acts are rational. But this is uninteresting. The important question is whether the way in which two states of affair bear on a particular agent may rationally affect her decision as to which to bring about so that different agents might rationally reach different decisions. As long as this is possible, then even infallible reasoners may find themselves facing disputes and controversies. And nothing in my account of rationality would rule this out.' 'Public Reason', p 29

[387] 'Public Reason', p 36

[388] Citizens are allowed to reconsider their authorisation if the public agent betrays their expectations and does not fulfil the role he was allocated.

deliberative normativity of the individuals, he acknowledges that it cannot be the only resource required.

'[T]he idea that the only resources available for the construction of authority *must* come from individual norms of deliberation now seems to me to confuse the need to justify authority in terms of deliberative rationality with the endeavour to create authority through the transfer of deliberative rationality from the individual members of society to a social agent.'[389]

Beyond the problem of transfer of authority from individuals to a public agent, this concept of rational deliberation meets what we called in chapter III the circularity problem. If I deliberate about an existing rule using the representations of the public agent already in place, I merely assess the consistency of the rule with the public agent's right reason. C. Morris' comment is truer than ever:

'A choice would not be a fundamental evaluation of the institution, for its standards are not independent of the domain of evaluation. For a rational choice evaluation of a social practice or system to be fundamental the preferences are to be more than coherent and considered. They must also be independent of the practice or system. In the absence of such independence, all that can be shown by rational choice is that the preferences are, broadly speaking, consistent with the practice or system.' [390]

Description of deliberative rationality

In order to address some of these issues, let us make an attempt to construct Gauthier's concept of deliberative rationality within his own framework.

To do so we need to come back to Gauthier's interpretation of Hobbes's laws of nature and public reason. We remember that he suggested that each agent should be able to rationally derive the laws of nature. The laws of nature are considered as the moral standards that must constrain the agents' future strategic interactions. In order to give a unique content to these laws and to secure their enforcement, they appoint a public agent. This public person is considered as the interpreter and guardian of the laws of nature. She acquires her authority from the agreement of all on her appointment. Gauthier initially argued for a direct transfer of authority from individual reason to public reason in agents' representations. Gauthier now claims that deliberative

[389] 'Political Contractarianism', p 142
[390] C. Morris 'Relation Between Self-interest and Justice', p 140.

rationality alone is not sufficient to justify the public agent's authority. So let us suggest the following alternative.

Agents first derive rationally the common moral standards. They then deliberate according to the *representations provided by these moral standards*. They assess their existing rules and practices, as provided by their current public agent, according to these representations. In other words, they step into the shoes of the public agent as an interpreter of the laws of nature. In order to do so they must not deliberate about each rule or practice in separation from the others. For the assessment work to be consistent and harmonious, they must assess their social structure as a whole.

There are then two possible scenarios. In the first one, the rules, practices and institutions of their existing social structure match the representations provided by the moral standards. The public agent has done her interpretation work properly. The fact that alternative interpretations were available is irrelevant as long as the existing social organisation is deliberatively acceptable as an interpretation of the moral standards.[391] Within its scope, public reason can then take over individual reason in deliberation. The authority of public reason over individual reason is now justified.

In the second scenario, the rules, practices and institutions of their existing social structure does not match the representations provided by the moral standards. The public agent has not done her interpretation work properly. In this case, either the public agent should be replaced by a more capable agent or she should be advised to change the existing rules.

Issues raised by this description

How can the agents rationally derive the moral standards in the absence of any representations of the world? That is where the state of nature enters the picture. I shall argue that the change of rational deliberation takes place in the state of nature and that

[391] The possibility that other norms, practices and institutions might be equally reasonable to accept is no argument against an existing social order.' 'Political Contractarianism', p 140

it bears in itself the impartial constraints on individuals that bring about the moral standards.

Which representations do the agents need to internalise and when: the moral standards or the public reason? I believe that in Gauthier's theory, the agents internalise first the change of rational deliberation and then whatever change comes with it. However, I shall argue that the change of rational deliberation is artificial and that it cannot carry the internalisation of the moral standards.

Has Gauthier solved the circularity problem? I believe he has but he had to take a Rawlsian turn to do so. Firstly, if individuals have to assess their existing systems, they have to abstract from their environment and their circumstances in the pre-bargaining position. They can no longer be real people in their real life. Secondly, they are now required to decide not only on the way to distribute the co-operative surplus but also on which society they want to live in. They have to agree on social practices and institutions. In doing so, they also devise the concept of value. I would like to develop both these points now.

Gauthier argues: agents become 'real' people in their real life only once in social interactions, i.e. once they have chosen their institutions and practices after rational deliberation. '[A]ny actual system must largely pre-exist the real person who enjoy the benefits and fall under its constraints.'[392] So if the bargainers are not 'real' until they enter social interactions, who are they? Gauthier writes: 'We regard each bargainer as serving as an ideal representative of the particular person he will be in the social world to be shaped by the constitution on which all agree.'[393]

The bargainers are no longer real people, but they are 'proto-individuals'[394] who select a social framework not according to *who they are* but according to *who they will be* in any of these frameworks. Individuals are structured by the social system in which they live and the moral frame has to be chosen before the individuals make any choices. The

[392] 'Political Contractarianism', p 133

[393] 'Constituting Democracy', p 324

[394] This expression is borrowed from J. Hampton, 'Can we agree on Morals?' (especially, pp 344-352). Her comments was then about the ideal chooser.

deliberators have to leave aside the non-rational aspects of their life-plans and abstract from their moral and cultural environment.

Where does it leave the concept of value as a measure of coherent and considered preference? Is value still a measure of utility? Can we still say that value is relative and subjective?

'[V]alues, aims and plans are all subjects for deliberation and choice.'[395] Value is the outcome of deliberation, it is the fruit of this double harmonisation process. In pure strategic interactions as exemplified on the market, outcomes result from the independent decisions of individuals whereas in deliberative politics, outcomes result from the decision-making of a centralised institution that artificially harmonises interactions. In pure strategic interactions, individuals are guided by their own private reason whereas, in social interactions, individuals are guided by public reason as expressed in laws. As explained above, 'public reason provides the citizen with a common standard for determining right and wrong, good and evil'.[396] Their interactions are structured and constrained by it. However, Gauthier cautiously argues that the constraints are also rationally derived. Agents have internalised the requirements of right reason and they therefore auto-constrain their life-plans. Once in society, they continue to pursue their own good but their conception of the good is then framed by public reason.

Value is no longer solely the outcome of individual rationality. It has also to be submitted to general confrontation prior to being adopted by the agents. Value is then a measure of rational and mutually acceptable preferences. It can no longer be accepted as a measure of individual utility but rather as a measure of mutual utility. As such it becomes difficult to consider it as plainly subjective and relative. Indeed, value is not subject to the sole rational assessment of the valuer but it has also to be accepted by his fellow bargainers; it is not relative to his present circumstances but to his future situation once deliberation has taken place. The concept of value is the object of deliberation between bargainers. Agents can only internalise a concept of value that has gained mutual approval.

---

[395] See 'Political Contractarianism', p 138
[396] 'Public Reason', p 36

Section 2: Derivation of the common moral standards

This step is crucial since the full theory rests on it; Gauthier's claim that morality can be derived from rational agency depends on it. In what follows, we shall develop a similar argument to the one developed in chapter IV: Gauthier's agents change of rational mode in the state of nature. This change bears in itself the moral standards they arrive at. These standards are derived from 'rational' agency but the rationality referred to is already moral.

Gauthier's agents start from the state of nature and have to work their way to society. The state of nature is the realm of unconstrained strategic interactions; it represents the social vacuum. The moral standards are mutual advantage and equal respect. How do the agents derive these standards from their state of nature?

About 'civic friendship'

In one article Gauthier suggested relying on 'civic friendship' to go from the state of nature to society. [397] We remember that Gauthier had to leave aside the assumption of mutual unconcern. However, he remained convinced that a contractarian theory could not rest on affective bonds. In the scenario below he develops instead an artificial concept of mutual concern.

In the state of nature, agents are interacting reluctantly and see each other as a cost. However, they come to realise over time that others can also afford opportunities. Their view on each other evolves: from being tolerated, they become useful partners. This change of perception creates artificial bonds of convenience. As the areas of mutual benefit expand, they see each other as sharing in a way of life. From useful partners, others become 'public' or 'civic friends'. Their presence is then welcomed rather than merely accepted. A bond of mutual civic concern is born [398] and with it the standard of equal respect. Indeed, civic friends 'affirm each other's good in willingly making and honouring whatever commitments are needed to make their mutual activities successful

[397] The concept of 'civic friendship' is mentioned in 'Thomas Hobbes and the Contractarian Theory of Law' but is described only in 'Constituting Democracy'.
[398] See 'Constituting Democracy', p 317-318

from each partner's perspective.'[399] In doing so they treat each other's life –plans with equal respect.

I believe that such a descriptive scenario does not fit in Gauthier's otherwise normative and justificatory theory. Although I believe we must keep in mind the concept of civic friendship, I prefer to develop my argument on the path described in most of his other articles.

'A co-operative venture for mutual advantage.'

In most of his writings, Gauthier's deliberators starts with a vision of society as a co-operative venture for mutual advantage and work their way from the state of nature to such an ideal. Why should agents, deliberating from a state of nature, start from the idea of society as a cooperative venture for mutual advantage? It makes sense nowadays in view of the recent developments of modern moral and political philosophy. However, if individuals deliberate in abstraction from pre-conceived ideas of the social structure they want to put in place, how would they rationally derive this model? In the state of nature, the only representations that the agents have are their own. Their individual representations are full of their own best interest (and possibly the best interest of the persons they care for). Representations are strategic and unconstrained. How does the idea of society as a cooperative venture for mutual advantage spring from such representations?

The first step out of nature is the willingness to agree. Agreement is a voluntary act. Individuals must first realise that unconstrained strategic interactions do not yield an optimal outcome; they must be willing to escape the costs of disadvantageous interactions and benefit from co-operation. The existence of social structures is universally preferable to 'the vacuum that would exist in the absence of norms and practices.'[400]

Gauthier assumes that agents come to realise that strategic deliberation will take them nowhere and that they must project themselves in the society they want to live in. In

[399] 'Constituting Democracy', p 318
[400] 'Political Contractarianism', p 140

order to do so they change of rationale. They replace their individual representations by the representation of an ideal society. In deliberating, they no longer take into account their strategic advantages or weaknesses and they no longer pursue their sole best interest since these representations are doomed to failure. Instead, in deliberating they only take into account what would gain general agreement and therefore take them away from the state of nature. Their natural representation is replaced by an artificial one. They no longer reason from where they are but from where they want to be. They become the famous 'proto-individuals who select a social framework not according to *who they are* but according to *who they will be* in any of these frameworks.'[401] They project themselves in the future. They become resolute.

We can see how Gauthier has changed his strategy from *Morals by Agreement.* In this first version of his contractarian moral theory, he attempted a mimic of a bargaining process with real agents pursuing their best interest. He labelled it internal rationality. We demonstrated the artificiality of the process and brought about the underlying change of rationality that had really taken place. Here Gauthier no longer attempts to describe a bargaining process, he directly replace the rationale of the market by the rationale of politics. Strategic and dynamic interactions are replaced by deliberative rationality. However in changing the rational mode, Gauthier has replaced real individual with 'proto-individuals'. These proto individuals are not moved by the representation of their best interest but by an impartial representation of the society they want to live in. They must constrain the pursuit of their own best interest in order to take into account the best interest of others that are parties to their endeavour of exiting the state of nature. In doing so, they exhibit impartiality.

The moral standards flow naturally from this change of rational mode. Agents want to agree in order to exit the state of nature. Agreement has to be voluntary and no agent would voluntarily agree to a constraint that is not mutually advantageous. Mutual advantage should be understood as

'requiring that each normal member of society both find value and contribute value to others. Someone who did not find value in society would have no reason to agree to its conditions of interaction. Someone who did not contribute value to others would give them no reason to accept her within the scope of society's conditions of interaction.'[402]

[401] See above in the sub-section *strategic interaction versus deliberative rationality.*
[402] 'Political Contractarianism', p 135

Civic friendship can step in the discussion here. In order to pursue mutual advantage, agents constrain their life-plans. It means that among the life-plans that would fulfil their personal requirements (whether rational or not), they must choose one that can be acceptable to others i.e. that does not involve a net cost to society. Individuals are free to pursue the life plan of their choice as long as in doing so they are net contributors to society. As quoted above the now civic friends 'affirm each other's good in willingly making and honouring whatever commitments are needed to make their mutual activities successful from each partner's perspective.'[403] In doing so, they exhibit equal respect. Equal respect is clearly a by-product of mutual advantage and it secures impartiality.

Moral standards and Lockean proviso.

In imposing equal respect and mutual advantage to the agents deliberating about their social structures, Gauthier puts in place a forward-looking rationale that secures future compliance. Indeed, if my situation is treated with respect and if I do gain from the agreement then I will be motivated to comply. If the situation of others is treated with respect and they do gain from the agreement then it secures their compliance.

This rationale obviously belongs to the family of the Lockean proviso core conception. Gauthier suggests adapting the Lockean proviso and replacing it by the condition of equal respect that should ensure the ex-ante fairness of the institutions chosen. He writes:

'Bargaining must proceed from an appropriate base point if it is to be considered deliberative. In *Morals by Agreement* I introduce a modified Lockean proviso to ensure that the base point does not incorporate force, fraud, or free-riding. Here, however, the concern to manifest civic friendship through the expression of equal respect is intended to safeguard against these ills. The question, pressed on me especially by Bruce Ackerman, is whether this concern may not also rule out inequalities in the base point that would be permitted by the proviso. I think not. Mutual advantage and equal respect do not require that the parties begin from an equality of condition.'[404]

This last debate is of mild interest to me since it is marginally relevant to the present thesis. I will ignore it here.

---

[403] 'Constituting Democracy', p 318
[404] 'Constituting Democracy', see footnote 12, p 333

About the internalisation process

Gauthier defines the human's deliberative capacity as follows. It is

'the capacity to be motivated by representations or states of affairs as they are and as they might be given one's possible actions, and the further capacity to be able both to ask in any particular case whether it makes sense to be motivated to some act by some set of representations, and actually to be motivated in the light of the answer one gives to this question.'[405]

The process we have just described is a perfect illustration of this definition. Agents asked themselves whether it 'made sense' to be moved by their individual representations (the pursuit of best interest) – 'primary representations' hereafter. They realised it did not and they changed their set of representations (to the pursuit of mutual advantage) – 'secondary representations hereafter'. The deliberative capacity described by Gauthier assumes the corollary ability to be motivated to act upon this new set of representations. Such an assumption amounts to an internalisation of the change of representations. Indeed, if someone can be motivated to act by a set of representations it surely means that this set of representations is *sufficient* to generate her action. Or is it?

I personally believe that the internalisation of the moral standards would require more than just rational deliberation. It is part of human nature, not to say part of human psychology, to pursue individual interest. In order to change this representation to impose the artificial pursuit of mutual advantage, I would suggest the old Aristotelian remedy: education.

I agree that agents act upon representations of the world in order to bring about their chosen state of affair. However, I do not believe that they can act upon abstract representations. These representations must have been experienced or felt before they can be vivid enough to motivate the agents to act upon them. At the stage we describe, the secondary representations are a mere rational and artificial abstract. They have no concrete reality and are meaningless to them. Representations of the world must be perceived to effectively motivate actions. In our example, I believe that agents must first be inculcated the moral standards. The pursuit of mutual advantage and equal

[405] 'Political Contractarianism', p 133

respect must become like a second nature to each of them. Once they have enjoyed the benefits to be gained from keeping these standards, they can then, but only then, internalise them.

In fairness to Gauthier, I believe that he has an *intuition* of this problem of representations. Several elements betray his intuition.

Firstly, his one-off description of civic friendship is nothing else than another version of the solution suggested here. Indeed, through practice of co-operation, agents experience the benefits of mutual advantage and slowly modify their perception of others. They then learn to treat them with equal respect. As mentioned before, the problem of such a descriptive concept is that it does not fit in Gauthier's justificatory theory. The emergence of civic friendship would owe more to perceptions and emotions than to rationality and deliberation. Since civic friendship is an essential step to equal respect and therefore to impartiality, such an origin is not acceptable.

Secondly, we remember Gauthier's doubt about the resources required to the transfer of authority from individuals' reason to the public one. 'Society may be conceptualised as a construction by deliberatively rational individuals, but it need not be a construction using only the normative resources of deliberative rationality.'[406] Gauthier then falls short of suggestions for other resources but I believe that this doubt is another betrayal of his intuition.

Last but not least, he never claims that his deliberators internalise the moral standards. The internalisation process of his contractarian moral theory is always about the social structures not about the moral standards. I assume that it is because, at the stage of the reasoning, the moral standards don't have the form of specific rules or norms. They are mere rational constructions without concrete representations. Let us continue our reconstruction of his theory.

[406] 'Political Contractarianism', p 141

Section 3: Agreement on social structure.

His agents have now left the state of nature and they deliberate about a compatible social structure. They constrain their life plans according to mutual advantage and they consider others as civic friends that they treat with equal respect. Gauthier expects his deliberators to evaluate their social structure seeking mutual benefit and equal respect. In doing so, they must display good will: 'I understand good will as involving openness and good faith – the willingness to acknowledge the nature and strength of one's true concerns in the process of reaching agreement, and the subsequent willingness to adhere or keep whatever is agreed. Within the framework of agreement, each must exhibit a positive concern for impartiality and mutuality.'[407] The outcome of this deliberation must be an *agreement* on social structure. If they find that their existing social rules, practices and institutions are compatible with the moral standards just derived, then they will have a moral obligation to keep them and comply with them as if they had derived them themselves.

Gauthier imposes two conditions to his contractarian deliberators.

- Condition 1: Equal desire to agree. Agents want *equally* to contract. 'The pressure to reach agreement arises solely from its desirability, which is felt equally by the members of society, and not from any difference in capacity or temperament or position which might bear differentially on the members, and so benefit some at the expenses of others.'[408] This condition seems to be more a corollary to the standard of equal respect than a stand-alone condition.
- Condition 2: Full information. This condition also prevailed in *Morals by Agreement*. Bargainers must have full information, they cannot bluff, threaten or exploit weaknesses. We have already commented upon the incredible demand of this assumption. However, it makes more sense in this new version of *Morals by Agreement* than it did in the original one. If agents seek mutual benefit and equal respect then: firstly they must know the situation of others to ensure that they do gain from the agreement without imposing a net cost to the others; secondly, as civic friends, they must not exploit differential advantage.

[407] 'Constituting Democracy', p 324
[408] 'Constituting Democracy', p 320

Since each agent retains his capacities, circumstances and constrained life plans, Gauthier insists that this stage should take the form of strategic bargaining between merely constrained deliberators. However, the combination of these two conditions amounts to the artificial equalisation of the bargaining powers of the deliberators. As argued in chapter IV[409], this artificial equalisation removes any form of strategic bargaining from the deliberation process taking place. In treating others with equal respect, Gauthier's bargainers are required to display impartiality; they impose constraints on their behaviour and exhibit morality. Gauthier's pre-bargaining position is no less morally loaded than Rawls'.

It is clear that deliberators must assess the social structure as a whole. They cannot evaluate each rule or norm in isolation. In this global assessment, some rules will be advantageous to some agents and disadvantageous to others. Due to their difference of circumstances or capacities, agents cannot agree on all rules and practices. However, for mutual agreement to emerge, the agents must gain from some rules when they loose from others in such a way that their overall benefit is proportionally equal to the benefit of the other parties to the agreement.

Gauthier writes.

'We should not expect the outcome of the social contract, or of any other agreement, to be the best possible for each party. What matters is that there be social structures that are clearly universally preferable to the vacuum that would exist in the absence of norms and practices, so that each party as a strong incentive to agree, and that the actual social structure not only in this way be mutually beneficial, but also that it ensures that the proportion between benefits and the contribution that each chooses to make be roughly the same for all. The terms of the cooperative venture are the same for all insofar as they provide for each to benefit in proportion to his contribution to the same degree as his fellows.'[410]

This balancing act between costs and benefits in agreement raise a new issue. Indeed, the advantage of this contractarian theory is to be rationally derived. I have a reason to comply with a rule if I have rationally agreed to it. Will I have a reason to comply with a rule that is rationally disadvantageous to me but that I accepted only in order to benefit from another one? How can I expect others to conform to a rule disadvantageous to them when I know that they agreed to it only to gain from another rule? That is where the concept of resoluteness takes its full dimension.

---

[409] See Barry's comments in Chapter IV.
[410] 'Political Contractarianism', p 140

Section 4: Compliance with agreement

What have we achieved so far? Agents have derived two moral standards and they have evaluated their social structure in the light of these standards If the structure has not passed the contractarian test, then they must supplant it by a compatible one. If the structure has passed the contractarian test, then they have acquired a moral obligation to comply with it. The rules and norms of their social framework have then the force of law. More importantly, and this is Gauthier's moral answer to the Fool, the deliberative rationality of the agents in agreement has secured their internalisation of these social rules and practices. This is a key concept in Gauthier's contractarianism. This is the moral theory within his contractarian theory. What was once called constrained maximisation is now called resoluteness but the concept remains at the heart of Gauthier's contractarian theory.

About resoluteness

We can now see how resoluteness fits in the big picture. Our agents first realise that pursuing mutual advantage and seeking equal respect will take them out of the state of nature and will enable them to benefit from co-operation. They then assess their social structure in the light of these two standards. If their set of social norms and institutions pass the contractarian test, then it means that the social structure under scrutiny represents a good interpretation of the standards sought. Therefore, in complying with all the requirements of their social system, agents are secured to achieve their overall best interest i.e. benefiting from a co-operation that would not be available otherwise.

Deliberative rationality involves complying with all rules and norms of the social system, including the ones that are not directly advantageous. It is rationally justified in the bigger scheme of each agent's life-plan. If, in social interactions, agents come across a rule that is disadvantageous to them but that they had a good reason to accept in their overall best interest at the time of agreement then they have a sufficient reason to conform to it now.

At the end of chapter V, I expressed my reservation about resoluteness as a mode of deliberation. I claimed that for resoluteness to be a more rational mode of deliberation

than sophistication, agents had *first* to change of focus from the pursuit of individual benefit to the pursuit of mutual advantage. I argued that resoluteness was not a stand alone rational concept. We can now see why and how. Resoluteness is part of a deliberative process that begins with the derivation of moral standards; it is *the* part that deals with the keeping of these standards.

Has Gauthier established a moral theory? As repeated many times, morality is human virtue. The deliberative and resolute agents can be considered as moral agents if and only if they have internalised their social rules and practices. But have they? Have we solved our representation problem? To address this question, I suggest coming back to my speeding ticket.

About the internalisation of the social norms and practices

As explained at the beginning of this chapter, I have been caught speeding at 43 mph in a town centre where the speed limit is 30mph. Had I been at a negotiation table where speed limits were to be decided, I would have fought for a 30mph in town. *My* reason for wanting a low speed limit in towns is that there are a lot more pedestrians and a lot more occasions to stop then on an intercity road. In keeping a low speed limit, we can stop more quickly and avoid more accidents or at least avoid more fatal accidents. So what did happen to this rational deliberation process on that day?

The facts were as follows: it was a Sunday morning at 8.00 am, the streets were empty and I was on my way to an appointment with a doctor after being unwell during the night. I was well enough to drive and to think but... I was not thinking and I did not know that there was a speed camera on that road. Let me explain.

In most of the actions we perform, we act in the immediacy of the action. In the immediacy of the action, the only representations we appeal to are our natural ones i.e. our best interest ones. The streets were empty so I did not *feel* I was taking a risk. At that time of the day and on that day of the week, I did not *expect* to meet anybody and the speed limit rule did not *strike* me as obvious. Yet the rule still made sense since there *could* have been a couple of pedestrians – for example an adult after a night of alcohol, or someone who, like me, had assumed that at that time on that day the streets

would be empty. *On the spur of the moment, I reassessed the situation, not according to what was mutually advantageous but according to my most immediate interest*: I wanted to arrive at the doctor as soon as possible. Had I realised that there was a speed camera, I would have *surely* slowed down... *firstly* in order to avoid a speeding ticket and *secondly* because it would have probably reminded me of the rule and of its purpose.

I believe that this example illustrate the weakness of Gauthier's argument. It also gives some indications of the remedies. I agree with Gauthier that if we were to reflect (or rationally deliberate) on appropriate social interactions, we would be able to derive some common sense rules that are mutually advantageous and make sense if we want to live well together. However, I disagree with Gauthier *if* he thinks that rational deliberation is sufficient to secure the internalisation of the rules so arrived. In other words, I disagree with Gauthier when he claims: that the deliberative capacity includes the 'further capacity to be able both to ask in any particular case whether it makes sense to be motivated to some act by some set of representations, and actually to be motivated in the light of the answer one gives to this question.'[411]

As noted above, I believe that the primary remedy is education. The pursuit of mutual advantage is a rational but artificial representation of the world. It does not come naturally to agents in their everyday interactions because it usually does not match exactly their *immediate* best interest. By definition, what is mutually advantageous is in each agent's *overall* best interest but it will usually involve an immediate cost. Therefore, I am convinced that to link the deliberative capacity to the action, the motivation needs to be fed to the agents from an early age. In their everyday gestures and within the scope of social interactions, children must be taught to think mutual advantage rather than immediate best interest.

Let us take a simple example. We can all agree that clean and tidy streets are a lot more pleasant for everybody than filthy ones. However, many people dispose of their litters in the street. Most of the time, it is because they can't find a bin when they need it and it bothers them to keep their litter until they can find one. Sometimes, they cannot even

[411] 'Political Contractarianism', p 133

be bothered looking for a bin. If they were to reflect on their gesture, they would probably agree that the practice of disposing of their litter in a bin makes sense and is mutually advantageous. But, when acting, they deliberate according to their most immediate interest not according to what is mutually advantageous. Their most immediate interest is to get rid of their litter straightway without suffering the inconvenience of keeping it until they can find a bin.

Children should be taught to dispose of their litters in bins and they should be encouraged to think for themselves of the reason behind such a practice. This way, when they are adults, they naturally wait until they find a bin to dispose of their litter. They don't evaluate the situation in the immediacy of the action when they can succumb to the temptation of immediate interest. They would act upon an inculcated practice that makes sense when they reflect on it. Only then can we consider that they have internalised public rules.

As more children are taught this way, more people keep the rules and practices The advantage of mutually advantageous rules don't only make sense in abstract deliberation but agents can actually experience their benefits. It becomes easier and easier to keep these rules. It will come more naturally to everybody. As a consequence, if the pursuit of mutual advantage was like a second nature, we would hardly need an enforcement system. The enforcement system would be for the rarer agents who would not keep the rules or practices. For all the others, it would also act as a constant reminder of their commitment to public reason. Punishment is the secondary remedy to the weakness of Gauthier's moral theory.

About punishment

As we saw above[412], Gauthier sees a potential conflict between motivational efficacy and normative precedence of public reason over individual reason, hence the need for an enforcement system. The public person disposes of the power to enforce its edicts. Gauthier justifies this by the now famous contractarian test. Would someone,

[412] For what follows see 'Public Reason', pp 41-42. We touched on the subject of punishment at the end of the previous section.

deliberating from a state of nature from which such power was absent, agree to it? Gauthier claims that anybody would.

But Gauthier argues: 'The judgement that an agent ought to be compelled or required to act in some way does not entail the judgement that the agent ought to act in that way voluntarily.'[413] Laws and public reason appeal for the former type of judgement. If agents find it reasonable to act according to the requirements, then, considering the threat of punishment, they will find it costly not to do so voluntarily. Gauthier concludes:

'the only rationale that should suffice for a general power to enforce certain edicts would also establish the normative priority of those edicts in the agent's unforced deliberation and address simply the gap between normative priority and motivational efficacy.'[414]

As before, I do not believe that deliberative rationality is sufficient to short-cut the need of an enforcement system. As agents internalise public reason, society can reduce the costs of enforcement measures. However punishment is also part of the educational process. Children should be inculcated and explained the rules and practices they have to comply with. The fear of punishment is a constant reminder of public reason.

It took me a few months to write this chapter. I have had plenty of time to reflect on my speeding ticket. When I see the speeding cameras, I slow down. Only then do I remember their purpose. It still does not come to me naturally but I am a lot more aware of it. Not only does the prospect of punishment modify the immediate rational calculus to reconcile it with the outcome of deliberative rationality, but it also educates the agents and secures their compliance. Only from a generalised compliance, can the concept of resoluteness get started. As agents experience the benefits of complying with public reason, they internalise the constraints it comes with.

I insist that the fear of punishment is only a complement to education. If agents continue to think best interest and to despise mutual advantage, the cost of the enforcement system would be disproportionate to the advantage it is meant to secure.

---

[413] 'Public Reason', p 41
[414] 'Public Reason', pp 41 - 42

Conclusion

'For we suppose that the capacity to make such choices [among dispositions to choose] is itself an essential part of human rationality. We could imagine beings so wired that only straightforward maximisation would be a psychologically possible mode of choice in strategic contexts. Hobbes may have thought that human beings were so wired that we were straightforwardly-maximising machines. But if he thought this he was surely mistaken.'[415]

Gauthier is convinced that we can change of mode of deliberation. He is convinced that this change can not only secure impartial bargaining but can also provide a moral answer to the Fool. In internalising this change of rationality, agents become moral.

In this chapter we detailed the rational change he suggests and we replaced it in a reconstruction of his *Morals by Agreement*. The outcome is a new contractarian moral theory which has taken a serious Rawlsian turn. It has gained in coherence, it flows more than *Morals by Agreement* but it carries the same weaknesses. Firstly, the change of rational mode occurs before bargaining and bears in itself the morality it is supposed to derive. Secondly, Gauthier's agents are supposed to internalise a change of rational deliberation based on pure rational grounds.

I am like Hobbes. I believe that we are and remain wired 'straightforwardly-maximising machines'. The change of rational mode suggested by Gauthier is artificial and incompatible with human nature. If I agree that people can deliberate rationally about their social interactions, I am not convinced firstly that they can agree and secondly that they can rationally internalise morality. Indeed, agents would retain their natural rationality in bargaining and play on any strategic advantage or weakness to tailor the outcome to their advantage. Agents would also retain their natural rationality in interactions and refuse most personal costs even if it is in the interest of mutual advantage. I believe that only education and fear of punishment can secure the realisation of Gauthier's contractarian theory.

[415] *Morals by Agreement*, p 183

CONCLUSION

The purpose of this thesis was to assess David Gauthier's success in giving a moral answer to the Fool. Gauthier believes that it is possible to become moral by mere rational calculus. This belief is based on the assumption that agents are able to change mode of rational deliberation once in social interactions: 'At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection.'[416] His contractarian moral theory is primarily rooted in this belief and its supporting assumption.

In the first and the fourth chapters of this thesis, I emphasized the interactions between Gauthier's interpretation of Hobbes' *Leviathan* and his own contractarian moral theory. In particular I attempted to show how the idea of a dual rationality sprang from Hobbes's distinction between natural and public reason. Gauthier took further this distinction and placed it at the heart of his own contractarian theory. Gauthier's interpretation of Hobbes' classic evolved over the years and we highlighted the parallel between the evolution of his interpretation and the development of his own research.

The first achieved answer to the Fool was given in *Morals by Agreement.* We detailed it in chapter II. Let us review briefly this first answer and our criticism of it.

Morality is defined as 'an impartial constraint on the direct pursuit of individual utility'[417]. In order to take root, cooperation requires some bending of individual interest and therefore some constraints on individuals' behaviour. Agents internalise the need for these constraints once they realise the benefits they can gain from cooperation over individually devised strategies. The internalisation of these impartial constraints comes with a change of mode of rational deliberation, i.e. a switch from the pursuit of

[416] *Morals by Agreement*, p 183
[417] *Morals by Agreement*, p 95

pure self interest to the pursuit of mutual advantage. Morality is the artificial virtue supported by this change of mode of rational deliberation.

Gauthier defines rationality as utility maximisation and justifies at length the reason for this definition. He then demonstrates that, when pursuing mutual advantage, agents continue to maximise their utility but in a constrained way. The rationality that prevails in natural interactions is described as *straightforward utility maximisation* whereas, the rationality that prevails in social interactions is labelled *constrained utility maximisation.* Gauthier also distinguishes between internal and external rationality of the agreement. The internal rationality is about making the agreement and requires straightforward maximisation. It is the rationality that prevails during bargaining. The external rationality is about keeping the agreement and requires constrained maximisation. Gauthier claims that compliance to the agreement is dependent on the fairness of the process, which in turn depends on a fair baseline. In the state of nature, prior to any bargaining, agents must first erase the effects of past force, fraud, free-riding or parasitism. That is what he calls his revised Lockean proviso. External rationality therefore includes not only compliance to the agreement once in social interactions but also the establishment of the pre-bargaining baseline. We saw in chapter IV that this second distinction between internal and external rationality is rather problematic. Let us briefly summarise why.

It is obvious that if the making of the agreement was based on constrained maximisation, the agreement would already be morally loaded. Indeed constrained maximisation is the instrumental rationality that supports morality. However, we demonstrated that Gauthier's bargainers were constrained maximisers: firstly because no straightforward maximiser would accept Gauthier's bargaining conditions and secondly because it seems irrational to expect the same agents to be constrained maximisers when establishing the baseline, straightforward maximisers when bargaining and constrained maximisers again once in social interactions. Either the agents have internalised the need to constrain their utility maximisation to benefit from cooperation or they haven't. The theory has to keep some chronology to be coherent.

We saw that this first version of Gauthier's moral contractarian moral theory had been the object of heavy criticisms. I reviewed most of them in chapter III. Gauthier had to

surrender his assumptions of equal rationality and mutual unconcern. He had to put in brackets his defence of the market as a morally free zone and he had to suspend his minimax relative concession principle until further research. However, the core of his contractarian moral theory - i.e. the change of rationality that we highlighted above – was revived within a broader framework. Indeed, McClennen's concept of resoluteness came to the rescue of Gauthier's constrained maximisation. Gauthier very rapidly took the measure of the tremendous potential of resoluteness and has explored it ever since. In chapter V we reviewed the advantages and the shortcomings of McClennen's concept. We also attempted to reconstruct what would be Gauthier's latest contractarian moral theory

- using *Morals by Agreement* as the backbone;
- removing all the acknowledged failures of this original version and
- taking into account all of Gauthier's latest publications on the topic.

Let us once again briefly recapitulate our findings.

In the newest version of Gauthier's contractarian moral theory, both definitions of morality and rationality change but his purpose remains to derive morality from rationality. Gauthier's social contract remains hypothetical and justificatory but his framework changes considerably.

Gauthier distinguishes between the rationale of the market – *strategic rationality* - and the rationale that prevails in politics – *deliberative rationality*. Both rationales cohabit in his new system and his focus is obviously on the latter. Deliberative rationality takes over when the market fails to bring about natural harmony between conflicting interests. This distinction means that rationality in politics is no longer connected to utility maximisation. Deliberative rationality is defined as the capacity to be moved by the representations provided by public reason. Public reason is the reason of a public agent appointed by the people to interpret the moral standards that should prevail in social interactions. In a hypothetical contractarian theory, the public reason is represented by the existing sets of public rules. Gauthier's agents rationally derive the moral standards required to exit a hypothetical state of nature (described as a vacuum of rules and laws). They then assess the conformity of their existing set of public rules (or public reason) with the moral standards so derived. If their existing set of rules matches the moral standards, then they are *justified* in bending their individual life-plans to these

rules. If their existing set of rules are not in conformity with the moral standards, then the set of rules needs changing.

Morality is the impartial constraint that individuals impose on their respective life plans in accordance with public reason once public reason has been rationally justified by the procedure described above. In the state of nature, individuals are moved by their individual representations (the pursuit of their best interest). Their deliberative capacities enable them to be moved by a new set of representations once this set has been validated by instrumental reason. This new set of representation encompasses the pursuit of mutual advantage. An agent has become moral once he has deliberated about his existing set of rules in the light of rationally derived moral standards, he has accepted its rationality and he has internalised it. The moral agent self-constrains his life plans in accordance with public reason.

I raised two issues. The first one was similar to the one made about *Morals by Agreement*: the change of mode of rational deliberation from strategic to deliberative has to pre-exist the derivation of the moral standards but this change carries all the requirements of impartiality that characterise morality. The second issue raised was about the internalisation of public reason. I wondered whether agents could internalise a set of rules based on sole rational justification. Although I cannot substantiate my claim without further research, I explained why I believe that education and punishment would have to play an indispensable complementary role in the internalisation process.

When I first read *Morals by Agreement*, I felt that I was reading an enduring masterpiece. Although I am now aware of its shortcomings, I remain convinced that the future of political philosophy lies in it. Like Gauthier and McClennen, I believe that it is possible to develop a contractarian moral theory. They have done most of the work. If I was to pursue my research in this field I would probably explore a more Aristotelian path and would abandon the project to connect morality solely to rationality.

I would probably start like them by highlighting the benefits that we could all gain from cooperating. I would insist on the overall rationality of cooperation for each individual. I would describe cooperation as politically moral or politically 'right'. A contractarian

moral theory deals only with the restricted field of political morality i.e. the set of rules that makes social life possible.

Then I would highlight the concessions on the immediate best interest that cooperation would impose on each of us. I would then be in a position to demonstrate the artificiality of the mode of deliberation that cooperation requires in day-to-day life. Using examples of conflict between individual and mutual advantage, I would highlight the necessity to remedy the fallibility of human rationality.

My remedy would obviously be education. I would develop an educational model destined at 'shaping' human reason so the cooperative mode of deliberation becomes like a 'second nature' to most agents. Children would be trained *first* to have the 'right' behaviour. Only once they would have experienced the benefits of such behaviour, would they be explained its rationale.

The agents so educated will have *properly* internalised the rules required by cooperation. These rules will be obvious and natural to them. Most importantly they will consider these rules as part of their well being and obvious to incorporate in their life-plans. They will no longer take them as self-imposed constraints.

Punishment would then be a mere safety net for the agents who would not comply with these rules. Punishment is a delicate issue and it is difficult to say more about it without proper research.

This very broad outline of an alternative theory obviously owes a great deal to Gauthier's pilgrim work in the field. Despite all my criticisms, I remain an admirer of his achievement in political philosophy.

APPENDIX: SETTING THE CONTEXT

This appendix is an introduction to the themes covered in *Morals by Agreement.* A vast range of concepts and theories is introduced as neutrally as possible, from the minimax theorem, the Nash solution, the Arrow's theorems or Harsanyi's general theory of rational behaviour to Rawls's theory of justice or Buchanan's theory of consent. Most, if not all, of these subjects have generated an important literature. We shall not present, discuss or criticise any of it here. We shall read them with Gauthier's binoculars later.

However, we need to note one common point between all these theories: they all make the same basic assumption namely that individuals are rational and as such they are utility maximisers. The famous rationality postulates are at the core of modern moral and political philosophy. This assumption has become so obvious that it is the natural starting point of any modern political philosopher.

This appendix provides a selection of information relevant to the good understanding of the subsequent ones. Its purpose is two-fold: firstly, it is designed to present the context in which Gauthier worked and carried out his research. Most of the material presented here influenced his writings; secondly, it aims at providing the indispensable background knowledge needed to assess Gauthier's contribution to moral and political philosophy.

Many subsections are borrowed from various authors whose names and publications are clearly given as and when. Those familiar with the materials presented can skip this appendix without any loss.

As far as game theory's role in social science is concerned, we can (very!) schematically distinguish between before and after Harsanyi's *Rational Behaviour and Bargaining Equilibrium.* The state of game theory and its potential use in social sciences before Harsanyi's classic was more or less as described by Luce and Raiffa in

1957[418]. In 1976, when the Nobel prize published his book, he considerably enlarged the debate. His work has been discussed and disagreed with at length in the literature but it remains an uncontested reference in the field. The first section is a presentation of game theory and the theory of social choice. The second section is a presentation of Harsanyi's innovations in both fields. The third section is a presentation of the two most recent versions of modern contractarianism, namely Rawls' theory of justice and Buchanan's theory of consent.

[418] *Games and Decisions Introduction and Critical Survey.*

Section 1: Game theory and social choice theory

In 1944 and then 1947, Von Neumann and Morgenstern published a book[419] that launched a new field of research in social sciences, namely the use of game theory in social situations. In 1957, Luce and Raiffa provided a critical survey on the topic that is still referred to due to its clarity and use to understand the philosophical issues behind the mathematical formulas. This book also offers a broader picture of the subject since it also incorporates two chapters on utility theory and a chapter on welfarism and Arrow's impossibility theorem. The table below provide their classification to be later compared to Harsanyi's. In this section, we shall review each decision-making theory separately.

[419] *Theory of Games and Economic Behaviour.*

LUCE AND RAIFFA'S CLASSIFICATION

	Decision-making theories		
Decision maker	Individual		Group
Theory	**a) Utility theory**	**b) Game theory**	**c) Social choice theory**
Number of players	One	Two or more	Large number
Conditions	Certainty or Risk &/or Uncertainty	Another form of uncertainty	

Definitions:

Individual vs group: 'Any decision maker (single human being or organisation) which can be thought of as having a unitary interest motivating its decisions can be treated as an individual in the theory. Any collection of such individuals having conflicting interests which must be resolved, either in open conflict or by compromise, will be considered to be a group'.[420]

Condition of certainty: when each action is known to lead invariably to a specific outcome

Condition of risk: when each action leads to one of a set of possible outcomes, each outcome occurring with a probability known to the decision-maker.

Condition of uncertainty: When each action has as its consequence a set of possible specific outcomes but where the probabilities of these outcomes are unknown.[421]

Comments

Luce and Raiffa insist on the fact that any index could have been chosen instead of the ambiguous and often misused index of utility. However, due to convention they will use the concept of utility to refer to the measurement of preferences. They also insist on the fact that utility theory is not a part of game theory and it can stand apart from it. Like social choice theory, they inserted it in the picture to provide a better overview of the connections between all the decision-making theories.

[420] *Games and Decisions* p13
[421] *Games and Decisions* p13

a) Utility theory

The concept of utility has been extensively described and commented in the literature. We shall just bear in mind here that utility is used in decision-making theories as an index reflecting preference-orderings. For mathematical convenience, we attribute a numerical value to a preference but one must be careful not to confuse this numerical value with a reflection of the satisfaction level. Preferences (and the corresponding utilities with them) have to satisfy the condition of transitivity (if A is preferred to B and B to C then A is preferred to C). We shall distinguish below the three conditions of decision-making as identified in the above table, namely conditions of certainty, risk and uncertainty.

Under *conditions of certainty*, each action is known to lead invariably to a specific outcome. Decision making under certainty can be summed up as follows: given a set of possible acts, one has to choose one (or all) of those that maximize the utility index. If a person's preferences satisfy certain consistency and continuity axioms, then his preferences can be represented by a well-defined, continuous and *ordinal* utility function. Rational behaviour is then equivalent to maximizing this utility function.

Under *conditions of risk*, each action leads to one of a set of possible outcomes, each outcome occurring with a probability known to the decision-maker. In the case of decision-making under risk, Bayesian decision theory then identifies rational behaviour with maximizing one's *expected utility* i.e. the mathematical expectation of one's *cardinal* utility function.

If a person is able to express preferences between every possible pair of gambles, where the gambles are taken over some basic set of alternatives, then one can introduce expected utility associations to the basic alternatives. If a person 'is guided solely by the utility expected value, he is acting in accord with his true tastes, provided only that there is an element of consistency in his tastes'[422]. The consistency of his tastes is described through six axioms[423].

Given a finite set of basic alternatives (or prizes) that we denote $A_1, \ldots, A_n$, each lottery assigns to each alternative a known probability. The set of probabilities is noted $p_1, \ldots, p_n$ with $p_i \geq 0$ and $\Sigma p_i = 1$.

For any two lotteries $L = (p_1A_1, p_2A_2, \ldots, p_nA_n)$ and $L' = (p'_1A_1, p'_2A_2, \ldots, p'_nA_n)$, if they satisfy the six consistency axioms, then
- there are numbers u_i associated with each Ai such that the magnitude of the expected value of L and L' are respectively $p_1 u_1 + p_2 u_2 + \ldots + p_n un$ and $p'_1 u_1 + p'_2 u_2 + \ldots + p'_n u_n$
- these magnitudes reflect the preference between the lotteries in the sense that $u(L) > u(L^*)$ if and only if $L > L^*$.

In 1954, Savage generalised Von Neumann Morgenstern utility function to conditions of uncertainty in which probabilities were not given exogenously, but derived subjectively from the preferences themselves. Savage named the probabilities so arrived 'personal probabilities'.

b) Game theory

So far we have only reviewed the case of an individual choosing in situation where his environment and the behaviour of others are given. But choices are more often made when other agents' behaviour is a variable. Game theory deals with the various issues raised by agents' interactions and strategies.

---

[422] *Games and Decisions* p 21
[423] The six axioms are: ordering of alternatives, reduction of compound lotteries, continuity, substitutability, transitivity and monotonicity.

Players in the game are to be characterised by three assumptions:

i) Each player has preferences over the outcomes which meet the axioms of expected utility theory

ii) Each player is fully aware of the rules of the game and the utility functions of each of the players

iii) Of two alternatives that give rise to outcomes, a player will choose the one which yields the more preferred outcome, or more precisely, in terms of the utility function, he will attempt to maximize expected utility[424].

Assumption ii) is obviously rather problematic since it is unlikely that in real situations, the actors will have the knowledge assumed. Discussion on this problem is beyond the scope of the present review.

Assumption iii) is referred to as the postulate of rational behaviour. We will see later the central role of this assumption in Harsanyi's work.

A player's preferences can be represented by a linear utility function U_i (or *payoff* function) whose values depend upon the strategy choices of all the players. A game consists of a set of n players, n sets of pure strategies s_i and n linear payoff functions U_i, one for each player.

A vast category of games (potential social situations) is encompassed under the generic term of game theory. We shall first review the variety of game situations prior to entering any further details about some of them.

First of all we have to distinguish the game situation by the numbers of players. Before Harsanyi, most of the research was on two-players games, although some solutions had been found for n-players games where $n>2$. When there are more than two players, new extra-game-theoretic concepts are introduced such as side payments, coalition structures, limitations on collusion, correlated strategies, transferable utilities and interpersonal comparisons of utilities. For the purpose of understanding Gauthier, we

[424] *Games and Decisions* p 49-50

don't need to review the technical (lack of?) solutions for n-person games (where n>2). We will limit our presentation to two-person games.

The second distinction regards the type of situations.

- Games can be strictly competitive or non strictly competitive. In the first case, it means that whatever one player gains is lost by the other player. There are also called zero-sum games because when you add up the utilities of each player to the game, the total makes zero. In the case of non-strictly competitive games, the players can gain from cooperation and or bargaining since their utilities can be convergent.

- Games can also be cooperative or non-cooperative. Can the players communicate and agree on a joint strategy that can be beneficial to both? If yes, then there is room for cooperation. Cooperative games are a lot more difficult to handle since players can have recourse to threats and assurances during bargaining or violate after bargaining the agreement reached during bargaining. Some psychological aspects have to be taken into account that are not necessarily easy to formulate mathematically.

The table below provide a summary of the solutions to the various game situations just mentioned for two-person games. Each solution will be broadly presented but without proofs or technical details.

Two-person games			
Non cooperative		Cooperative (arbitration schemes)	
Strictly competitive (zero-sum)	Non strictly competitive	Strictly competitive (zero-sum)	Non strictly competitive
Minimax theorem Equilibrim pairs with pure or mixed strategies.	Nash solution equilibrium pair(s) of mixed strategies	1) Von Neumann Morgenstern (set of) solutions 2) Nash bargaining problem	1)Shapley value 2)Nash extended bargaining pb

Non cooperative - strictly competitive games: Maximin (or minimax) theorem

Strictly competitive games are also called zero sum games. In such games, whatever is gained by one player is lost by the other. Each player attempt to maximize his security level i.e. maximize his gain and minimize his loss. An equilibrium pair is obtained when the maximum gain of one equals the minimum loss of the other. Von Neumann & Morgenstern proved that under certain conditions such an equilibrium exists. This theorem is known as the *maximin theorem.*

Non cooperative – Non strictly competitive games: Nash 'solution'

There are two players and each has a set of pure strategies. Both players have knowledge of all possible outcomes and of each other's preferences. They also have a preference ordering over these outcomes. Player 1 has the set $\alpha = \{\alpha_1, \ldots, \alpha_i, \ldots, \alpha_m\}$ and player 2 has the set $\beta = \{\beta_1, \ldots, \beta_j, \ldots, \beta_n\}$ of pure strategies.
For each strategy choice α_i, β_j from respectively players 1 and 2, there is a certain outcome O_{ij} with utility a_i for player 1 and b_j for player 2.

A player's mixed (randomised) strategy is obtained by applying a probability distribution to his set of pure strategies. A mixed strategy set for player 1 is $\alpha' = x\alpha = \{x_1\alpha_1, .., x_i\alpha_i, \ldots, x_m\alpha_m\}$ with all probabilities $x_i \geq 0$ and $x_1 + \ldots + x_m = 1$ and for player 2 is $\beta' = y\beta \ \{y_1\beta_1, \ldots, y_j\beta_j, \ldots, y_n\beta_n\}$ with all $y_j \geq 0$ and $y_1 + \ldots + y_n = 1$.

Two equilibrium pairs $(x\alpha, y\beta)$ and $(x'\alpha, y'\beta)$ are *equivalent* if the returns to each player are the same i.e. $U1 (x\alpha, y\beta) = U1 (x'\alpha, y'\beta)$ and $U2 (x\alpha, y\beta) = U2 (x'\alpha, y'\beta)$.
They are *interchangeable* if $(x\alpha, y'\beta)$ and $(x'\alpha, y\beta)$ are also in equilibrium.
A strategy pair $(x\alpha, y\beta)$ *jointly dominates* strategy pair $(x'\alpha, y'\beta)$ if each player prefers $(x\alpha, y\beta)$ i.e. if $U1 (x\alpha, y\beta) > U1 (x'\alpha, y'\beta)$ and $U2 (x\alpha, y\beta) > U2 (x'\alpha, y'\beta)$.

Nash proved in 1951 that

- every non cooperative game with finite set of pure strategies has at least one mixed strategy equilibrium pair.
- A non cooperative game is solvable if all equilibrium pairs are interchangeable.
- The solution to the game is its set of equilibrium pairs

'A Nash solvable game need not have equivalent equilibrium pairs, so Nash was led to define the upper value for a player as the most he can get from some equilibrium pair and the lower value as the least he can possibly get'[425].

Cooperative games – Nash's bargaining problem.

Von Neumann and Morgenstern identified a solution set as follow:

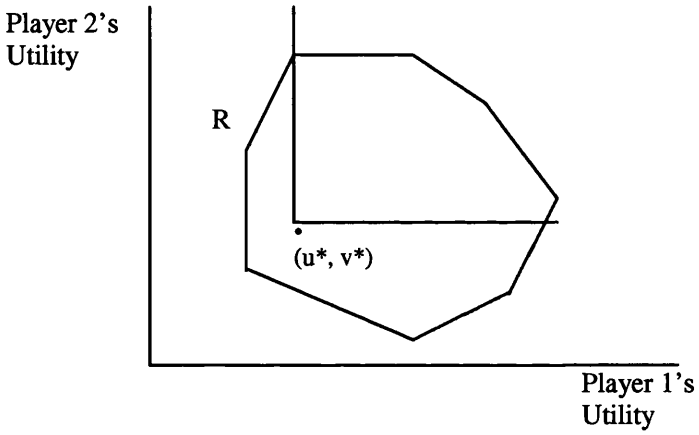
- It belongs to the set of possible payoffs.
- It includes all the jointly undominated outcomes that yield each player at least as much as he can secure independently by playing his maximin strategy.
- The solution set hereby obtained is called the *negotiation set.*
  The set of all undominated outcomes is known as the *Pareto optimal set.*

The problem is now to find within the negotiation set an optimal outcome. That is where communication between players comes into the picture.

Luce and Raiffa seemed convinced that cooperative games were dependent on an arbiter. Their argument in favour of an arbiter is rather simple: in negotiation, the players can fail to reach an agreement because they are not prepared to make concessions. However in most cases, it is more advantageous to all (and therefore rational) to reach an agreement than not. An impartial arbiter can more coolly assess the situation. Luce and Raiffa note: 'We may suppose that the arbiter sincerely envisages his mission to be "fairness" to both players; however, there are not, as yet, any simple and obvious criteria of "fairness", so, in effect, he is being asked to express a part of his ethical standards when resolving the game'[426]. They therefore see all the suggested solutions to cooperative games as arbitration models with different views on fairness. As we shall see later, both Harsanyi and Gauthier disagree with this approach, although on completely different grounds.

[425] *Games and decision* p106-7
[426] *Games and decision* p121

In 1950, Nash tried to give a formal definition and find a solution to the bargaining problem between employer and labour union[427]. The game can be described as follows: two players come to the market with each a bundle of good to exchange. Each trade T brings about a different outcome. Their preferences over randomised outcomes are consistent and therefore can be mirrored by a numerical utility index. To each trade T corresponds a pair of utilities (u, v) with u representing player 1's utility and v representing player 2's. Amongst all possible trades, there is one corresponding to no trade: T* (u*, v*). Let us call R the region of all possible trades and randomisations between trades. R is bounded, convex and closed (see figure below).



Each bargain is represented by a region R and a *status quo* point (u*, v*) and denoted [R, (u*, v*)]. Any trade off would be a point above and to the right of (u*, v*) (the existence of such points is assumed). The problem is to identify a unique payoff (u°, v°) that is "fair" to both players. In other words we look for a function F that operates on [R, (u*, v*)] to give a unique (u°, v°).

After formulating the problem in such a way, Nash suggested the following solution.

i) Change the origin of measurement of utility for each player so that the point (u*, v*) is transformed into (0, 0), and let the resulting transformation of R be denoted by R'.

ii) In the region R' find the unique point (u°', v°') such that u°', v°' is the maximum of all products uv, where (u, v) is in R', i.e.,

a) (u°', v°') is a point of R', u°'>0, v°'>0

b) u°'v°' $\geq$ uv for all (u, v) belonging to R' such that u $\geq$ 0 and v $\geq$ 0.

[427] The following paragraphs are borrowed from Luce and Raiffa's presentation of Nash's bargaining problem. (see *Games and Decisions*, p 124-137)

The point $(u^{\circ\prime}, v^{\circ\prime})$ is the Nash solution to the bargaining game [R', (0, 0)]. The solution to [R, (u*, v*)] is obtained by inverting the utility transformations on $(u^{\circ\prime}, v^{\circ\prime})$. this point can be characterized by the unique point (u°, v°) of R such that
$(u^{\circ} - u^*)(v^{\circ} - v^*) \geq (u - u^*)(v - v^*)$, for all (u,v) belonging to R and such that $u \geq u^*$ and $v \geq v^*$.

Nash's function is the only function that satisfies four indispensable rationality properties that a solution should fulfil namely: invariance with respect to utility transformation, Pareto optimality, independence of irrelevant alternatives and symmetry (independence of labels of players).

It is important to note that Luce and Raiffa understand Nash's solution as an arbitration model and therefore concentrate their criticism on the 'fairness' of the solution obtained.

Nash attempted to extend his model to non-strictly competitive games. In short, each player adopts a mixed strategy as a threat. The pair of threats establishes a payoff that, in turn, acts as the status quo point for further bargaining; and the bargaining problem is solved as above. Each player's purpose is therefore to select a threat strategy so as to influence the status quo in the most favourable manner. The solution hereby obtained belongs to the negotiation set and is Pareto optimal but it fails to be unique.

c) Social choice theory and welfarism

It is not the place here to give an historic of social choice theory and welfarism but minimum background information is appropriate to understand the issues at stake.

Through two different approaches of welfare theory, we shall review two problems:

- The nature of the welfare function (cardinal or ordinal) and the problem of aggregation of utilities or rankings.
- The role of ethics and morality in social choice theory.

Welfare Economics[428]

Although welfare economics is no longer really in fashion, it is worth emphasizing some of its salient features. As a definition Winch offers: 'Welfare economics is the study of the well being of the members of a society as a group, in so far as it is affected by the decisions and actions of its members and agencies concerning economic variables.'[429]

Welfare economics takes the objectives of society as given and work out the appropriate policies as means of achieving them using testable and tested economic tools. One of the main tool used by classical economic theory is the utility theory combined with the theory of rational behaviour that we reviewed above. In short, the preference of an individual can be represented by her utility function; if rational her objective will be to maximize this utility function.

In the case of an individual (or single household) the utility function Ui is composed of a bundle of goods (X1, ...Xg) the acquisition of which to maximize under the constraint Σ Pi Xi – I = O where Pi is the price of good Xi as set by its supply and demand on its competitive market and I is the income of the individual exogenously derived from the individual's work factors sold.

In the case of an entrepreneur, the utility function Ui is composed of bundle of goods (X1, ...Xg) the production of which to maximize under the constraint Σ Pi Xi – (I' + π) = O where I' and π are respectively the total cost and the profit of the firm.

Analogously, society as a whole has to maximize its welfare function dependent positively on its components utility levels. Winch notes that most of the conventional theories of welfare economics rest on the Paretian's objective: if one person is better

[428] This subsection owes much to David Winch's *Analytical Welfare Economics*, which is often referred to in Gauthier's *Morals by Agreement*.
[429] *Analytical Welfare Economics*, p 13

off, and no one is worse off, welfare is increased. As a consequence, welfare is an increasing function of individuals' utilities: $W = W(U1, \ldots, Un)$ with $\delta W / \delta Ui > 0$.

Under certain assumptions and conditions, welfare economics is then able to identify *the* Pareto optimum. In particular, the assumption of perfect competition can lead to the *identity of the equilibrium*, obtained when all forces in the model have worked themselves out, *to the optimum*, the desirable situation as defined by the theory. The theory also provides some solutions to the problems posed by the removal or modification of some of the basic assumptions or conditions.

The analogy between the maximization of individuals' utility function and the maximization of the social welfare raises **three major ethical issues.**

1) Objectives and social values

The role of welfare economics is not to decide what the objectives of society 'ought to' be, it is rather to take this objectives as given. The role of setting the objectives of society belongs to moral and political philosophy.

However, Winch insists: 'all economics is concerned with the making of choices, and rational choice necessitates the stipulation of an objective. The nature of the objective depends entirely upon the value judgements of the person stipulating it.'[430] When classical economics assume that individuals maximize their utilities, there is no implication that such behaviour is good or bad.

'To assume that individuals attempt to maximize their own objectives makes the outcome of the economic system dependent on the individuals' value judgements, but it says nothing about the desirability of those values. Concepts of morality, by contrast, stipulate social values that override individual choice.'[431]

The welfare function is dependent on the objectives of the policy makers and these objectives themselves depend on the assumption of certain value judgements. We saw that by convention the Paretian criteria had been chosen in welfare economics but the

[430] *Analytical Welfare Economics*, p25
[431] *Analytical Welfare Economics*, p26

reality of policy making is a lot more complex than what the theory assumes. Would welfare economics be sufficiently equipped to deal with other value judgements?

2) Pareto optimum and distributive justice.

Do we consider the initial factors (work or production) as given? If we do so, is the optimum derived from the theory fair and equitable? Winch discusses the distributive justice of the optimum obtained in condition of perfect market competition and concludes:

'either the competitive equilibrium achieves distributional equity as a matter of definition, because equity existed in the initial pattern of factor ownership, and therefore exists in the results that ensue therefrom; or equity is considered to be an attribute of the distribution of income or utility, rather than factor ownership, in which case any effort to achieve distributional equity in a changing world undermines the foundations on which the efficiency of the competitive system was based.'[432]

3) The existence of a 'collective' rationality.

The analogy of utility with welfare maximization is based on two related assumptions.

- The first assumption is that such a welfare function does exist.
- The second assumption is that there is such thing as a collective rationality. It is assumed that, the same way individuals maximize their utility function as a mean to achieve their individual goals, society maximizes its welfare function as a mean to achieve common goals.

Do such common goals exist? This issue raises the problem of aggregation of individual wills and preferences. It is now time to turn to Arrow's impossibility theorem.

[432] *Analytical Welfare Economics*, p99

***Arrow's impossibility theorems*[433].**

How do we aggregate the preference rankings of more than two individuals? It has been proven that, due mainly to intransitivity, simple majority rule is not an option. So what else? Arrow in 1951[434] proved that, given five apparently innocuous conditions of aggregation, it was impossible to obtain a social ranking of preferences from individual ones. We shall provide below a broad picture of his social choice theory.

Arrow is an ordinalist concerned with collective social choice and two connected problems: 1) the mechanisms for such a social choice and 2) the consistency of various value judgements as to the mode of social choice. Arrow excludes the concepts of cardinal utility and interpersonal comparisons of utility.

Two definitions

Social state: 'a complete description of the amount of each type of commodity in the hands of each individual, the amount of labour to be supplied by each individual, the amount of each productive resource invested in each type of productive activity, and the amount of various types of collective activity'.[435]

Social welfare function: 'a process or rule which, for each set of individual orderings R_1, ...R_i,..., R_n for alternative social states (one ordering for each individual i) states a corresponding social ordering of all alternative social states, R.'[436]

Two axioms about the social ordering

Given several individuals (voters) and a minimum of three alternative social states x, y and z, we shall write 'x is preferred or indifferent to y' with the symbol xRy.

---

[433] In what follows I have chosen to present Arrow's original theorems although Sen's version of these theorems is more widely accepted nowadays. Gauthier refers to (and question) both version in *Morals by Agreement* (See pp 123-125)
[434] *Social Choice and Individual Values*
[435] *Social Choice and Individual Values* p17
[436] *Social Choice and Individual Values* p23

Axiom I: (Connectivity) For all x and y, either xRy or yRx.
Axiom II: (Transitivity) For all x, y and z, xRy and yRz imply xRz.

Three Assumptions:

- There is a basic set of alternative social states that can be presented to voters.
- Each individual in the community has a definite ordering of all conceivable social states, in terms of their desirability to him.
- Individuals are rational i.e. their social orderings Ri satisfy axioms I and II

Five conditions about the welfare function

Condition 1: there are at least two individuals ordering a minimum of three alternatives. The social welfare function is defined for all possible individuals' orderings.

Condition 2: (Positive association of social and individual values). If one alternative social state rises or remains still in the ordering of every individual without any other change in those orderings, then it rises or remains the same in the social ordering.

Condition 3: (Independence of irrelevant alternatives). It requires that the social ordering of any two outcomes shall not be influenced by the presence of any other alternatives in the set to be ordered.

Condition 4: (Citizens' sovereignty). No social ordering can be imposed on society. Said differently, the social ordering is dependent on, and derived from individual orderings.

Condition 5: (Non dictatorship). The social ordering shall not coincide with the ordering of any particular individual regardless of the ordering of others.

One ethical note

Arrow distinguishes between the voters ordering of social states 'according to the direct consumption of the individual and the ordering when the individual adds his general

standard of equity. We may refer to the former ordering as reflecting the tastes of the individual and the latter as reflecting his values'[437] ....The individual orderings which enter as arguments into the social welfare function as defined here refer to the values of the individuals rather than to their tastes.'[438] This embryonic dichotomy will become relevant once we present Harsanyi's ethical theory in the next section.

Two theorems

The two theorems are known as the possibility theorem for two alternatives and the general possibility theorem (when there at least three alternatives). Both theorems state that no welfare function can satisfy the five above conditions simultaneously.[439]

[437] *Social Choice and Individual Values* p18
[438] *Social Choice and Individual Values* p23
[439] See *Social Choice and Individual Values* p 48 and 59

Section 2: Harsanyi and the general theory of rational behaviour

Harsanyi has extensively written and developed on his theory. His main publications are a collection of essays written between 1953 and 1976[440] and his famous *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations* respectively published in 1976 and 1977. In both publications, he attempts to demonstrate the supremacy of Bayesian decision theory in what he calls the theory of rational behaviour whether on parametric or strategic situations of choice.

He notes that in social sciences most of the games are two persons nonzero-sum games or n-persons games but that these games don't have real solutions. In this context Harsanyi offers a theory that suggest a solution for these games within a more unified framework. His ambition is to treat all the independent decision-making theories reviewed above under the large umbrella of a general theory of rational behaviour and to provide a systematic solution to each type of situation. In order to achieve this goal, he defines the rational-behaviour concept by means of a few additional and more powerful rationality postulates. Harsanyi's purpose is also to make game theory a readily available tool at the service of social sciences and the analysis of social behaviour.

Harsanyi's classification (shown on the table below) provides relevant background information to understanding his approach. One can note two essentials departures from Luce and Raiffa's classification that reflects the innovations made by Harsanyi.

1) He removes the partitioning between the various theories. The independent decision-making theories are now all part of the same general theory of rational behaviour

2) The moral dimension belongs solely to social choice theory. Morality is not a concern of the social scientist when dealing with utility and games theories.

[440] *Essays on Ethics, Social Behaviour, and Scientific Explanation*

HARSANYI'S CLASSIFICATION

	General Theory of Rational Behaviour			
	Theory of Rational Behaviour			Theory of Moral Value Judgements
	Theory of individual interests			Theory of *common* interests
	a) Utility theory		**b) Game Theory**	**c) Ethical Theory**
	Classical Economic Theory	Modern Decision Theory		
Situation (under)	Certainty	Risk &/or Uncertainty	Another form of uncertainty	
Maximization	Utility	Expected Utility	Individual Utilities	Average Utility Level
Perspective	One individual		Several individuals interacting	Impartial Impersonal
Choice	Parametric		Strategic	Social (moral)

Definitions

Condition of certainty, risk and uncertainty : same definitions as above

Parametric choice: choice in which the actor takes his behaviour to be the sole variable in a fixed environment and regards himself as the sole centre of actions.[441]

Strategic choice (case of interaction – Game situations): choice in which the actor takes his behaviour to be but one variable among others so that his choice must be responsive to his expectations of others' choices while their choices are similarly responsive to their expectations.[442]

Expected Utility: utility assigned to a lottery i.e. the sum of the products obtained by multiplying the utility of each of the prizes (or outcomes) of the lottery by its probability.

[441] *Morals by Agreement* p21
[442] *Morals by Agreement* p21

a) Utility (decision) theory

There is not much innovation in the way Harsanyi proves the existence of a utility function and uses the corresponding rationality axioms in situation of certainty or risk.

In the case of uncertainty, we saw earlier that Savage had established in 1954 the existence of 'personal probabilities' and a utility function based on seven consistency postulates. He had extended the concepts of expected utilities and Bayesian decision theory to situations of uncertainty by amending the (rationality) postulates. Harsanyi uses Anscombe and Aumann's axiomatisation which combines Savage and Von Neumann Morgensten'. The 'personal probabilities' are now called 'subjective probabilities'. As we shall now see, Harsanyi applies the tools of subjective probability and Bayesian decision theory to game theory.

Up to that point, Gauthier does not have much to object to Harsanyi's classification:

'The theory of rational choice identifies practical rationality with utility maximization. If the utility of a lottery is its expected utility, then in condition of risk and uncertainty practical rationality is identified with *expected utility-maximization*. This is the central thesis of Bayesian decision theory. Although the theory does not command universal assent, we shall not enter into its merit and defects here, but only assert (dogmatically!) that we know no satisfactory alternative.'[443]

b) Game theory[444]

Two central novelties in the use of game theory in social sciences.

Harsanyi's method leads him to unique solutions for each game situation.

i) *Extension of the Bayesian theory* of rational behaviour not only to utility theory under uncertainty but also to game theory by using subjective probabilities.

'Our theory of rational behaviour in game situations will be a direct generalisation of the Bayesian theory of rational behaviour under uncertainty. Following the Bayesian

[443] *Morals by Agreement*, p 44
[444] In this sub-section, we follow Harsanyi's own presentation of his theory step by step.

approach, we shall assume that any player i will express his expectations about the behaviour of another player j by assigning subjective probabilities to various alternative actions that player j may possibly take'[445].

However, each player does not decide arbitrarily of the subjective probabilities. Instead, these probabilities rest on two essential principles:

- The principle of mutually expected rationality.
The subjective probabilities that player i will assign to various possible actions by player j have to be consistent with the assumption that player j, like player i himself, will act in a rational manner in playing the game. Decisions about the subjective probabilities players should entertain play an important role in Harsanyi's theory since it enables him to define a determinate solution to each particular game.

- Zeuthen's principle
Most of the game theorists before Harsanyi relied on the pay-off dominance relations and intuitively assumed that they were all based on the principle that rational players will always prefer strategies yielding higher payoffs. Unfortunately this approach, as we saw before, did not provide a determinate solution in most situations. Harsanyi believes that the concept of payoff-dominance relations should be replaced with risk-dominance relations. This concept is based on Zeuthen's criterion: 'at any given stage of bargaining between two rational players, the next concession must always come from the party less willing to risk a conflict – if each party's willingness to risk a conflict is measured by the highest probability of conflict that he would be prepared to face rather than accept the terms proposed by the other party'[446].

ii) The use of bargaining models rather than arbitration models

Harsanyi clearly distinguishes between game theory that deals with players pursuing their respective self-interest and ethical theory that deals with the common interests of society as a whole. Morality comes into the picture of decision making only when dealing with social choice. Moral issues within game situations are incorporated in utility functions and should not pollute the decision-making process. Each player

[445] *Rational Behaviour*, p 11
[446] *Rational Behaviour* p 12

attributes positive utilities to what he values most. If each player worries only about maximization of his own (expected) utility function, the game theorist only concentrates on the rationality postulates that go with this maximization process. Gauthier completely disagrees with this dichotomy and prefers to reintroduce the concept of morality in game theory but not using arbitration schemes.

Therefore rather than arbitration models, Harsanyi prefers to refer to bargaining models. This method solves the cumbersome problem of the fairness of the solution(s). Another side effect of this method is that there is no need for interpersonal comparisons in game theory. As we shall see below, interpersonal comparisons according to Harsanyi become relevant only in the context of his ethical theory.

Two classes of rationality postulates

Class A: postulates of rational behaviour (i.e. criteria for players' strategies)

A1: *Maximin postulates.* In a game G unprofitable to you always use a maximin strategy

A2: *Best reply postulate.* In a game G profitable to you, always use a strategy σi representing a best reply to $\sigma\hat{\imath}$ of the n-1 other players.

A3: *Subjective bet reply postulate* (Bayesian expected utility maximization postulate). In a bargaining game when the other players bargaining strategies are unknown us, we have to rely on the subjective probabilities that we assign to the various combination of the other players' bargaining strategies. Hence, in a bargaining game, our bargaining strategy β is a subjective best reply to the other players' expected mean bargaining strategy combination.

A4*: Acceptance of higher payoffs postulate.* If, in a bargaining game, a player agrees to a joint strategy σ, then, if there exists a joint strategy σ^* yielding him a higher payoff than σ, he would be even more willing to agree to σ^*.

A5: *Postulate of indifference between strategies yielding equal payoffs* (equiprobability or centroid postulate).

Class B: postulates of rational expectation, (i.e. criteria for players' expectations about each other's strategies). The validity of the subjective probabilities is reliant on this second class of postulates.

B1: *Mutually expected rationality postulate.* Always expect the other players to follow the class A and B rationality postulates.
B2: *Symmetric expectations postulate.* Never expect the other players to choose a strategy (especially a more concessive strategy) that you would not choose should you be in his situation.
B3: *Expected independence of irrelevant variables postulate.* Don't expect the other players to establish their strategies on variables which relevance for bargaining behaviour cannot be established on the basis of the rationality postulates.

Two person cooperative games – The Nash solution.

Unlike Luce and Raiffa, Harsanyi interprets Nash's solution to two-person cooperative games as a bargaining model rather than as an arbitration model[447]. Harsanyi's theory, based on the Zeuthen principle and the strong rationality postulates leads to the Nash solution in the case of two persons cooperative games. This solution is called the Nash – Zeuthen – Harsanyi solution in Gauthier's *Morals by Agreement.*

Harsanyi demonstrates three important results:
Zeuthen's principle in bargaining leads to the Nash solution.
Harsanyi's strong rationality postulates (i.e. A3, A4, B1, B2 and B3) leads to similar bargaining behaviour as the Zeuthen's principle and brings about the Nash solution.

In the general case of cooperative games, Harsanyi demonstrates that Nash's extended solution of optimal threats is once again equivalent to the Zeuthen's bargaining behaviour. In substance: 'the solution is based on the concept of mutually optimal threat strategies. Intuitively speaking these represent the best possible compromise between trying to *maximize* the damage that one can cause to the opponent in a conflict situation and trying to *minimize* the cost of the conflict to oneself.'[448] Therefore, the

[447] *Games and Decision* p 145-6
[448] *Rational Behaviour and Bargaining Equilibrium*, p179

Nash – Zeuthen – Harsanyi's solution takes the form of a minimax (or maximin) formula.

c) Ethical theory

Harsanyi's version of welfarism is also a major innovation since he clearly places morality at the heart of his welfare function. Let us follow him.

General features

- Harsanyi's ambition is to provide a model of moral value judgement freed from any cultural or pre-conceived input and assessed exclusively on rational grounds. His model is meant to unable us to assess rationally the morality of a situation, system or society. Harsanyi does not provide a rationally derived moral system, he only provides us with rational tools to assess the morality of existing situations. Ethics is defined as being in the service of the *common* interests of society as a whole.

- Moral rules should comply with the principle of universality suggested by Kant, hence Harsanyi's preference for rule utilitarianism over act utilitarianism. In his opinion, a moral rule is a basic general rule which is valid in any given situation of the same sort and is therefore a 'behavioural rule that would maximize social utility if it were followed by everybody in all situations' it applies to[449].

- The *principle of consumers' sovereignty* states that the consumer is free to choose and his preferences are not judged or morally assessed. The interests of each individual must be defined fundamentally in terms of his own personal preferences and not in terms of what somebody else thinks is right for him. However, Harsanyi allows some major exceptions to this principle through the true and the antisocial preferences (see below).

[449] *Rational Behaviour*, p41

- Harsanyi considers that an agent biased by the pursuit of his personal interests is not capable of a moral value judgment. A moral point of view has to come from an impartial observer i.e. an outside observer whose personal interests are not involved, or at least from an agent consciously and willingly ignoring his personal interests to assess the morality of a situation or a system. Morality emerges from and is guaranteed by impartiality and impersonality.

- One can notice some similarities with the concept of veil of ignorance later developed by Rawls[450]: an individual i's choice amongst alternative social situations would satisfy the requirement of impartiality and impersonality,
'if he simply did not know in advance what his social position would be in each social situation'[451] or more specifically ...
'if he thought he would have an equal probability of being put in the place of any among the n individual members of society'[452]. This latest point is known as the *equiprobability postulate*. This postulate enables an individual to abnegate his own position.

- According to the *similarity postulate,* 'once proper allowances have been made for the empirically given differences in taste, education etc.., between me and another person, then it is reasonable for me to assume that our basic psychological reactions to any given alternative will be otherwise much the same'[453]. Because it is not open to any direct empirical test, the similarity postulate is to be considered as an a priori postulate. According to this postulate, it is possible to make interpersonal comparisons.[454]

Model of moral value judgement

A, B, C... are social situations;
Ui is the von Neumann-Morgenstern (VNM) cardinal utility function of individual i (i=1, ...,n);

[450] Although Harsanyi was the first one to create the concept of the original position.
[451] 'Morality and the Theory of Rational Behaviour', p 49
[452] 'Morality and the Theory of Rational Behaviour', p 50
[453] *Rational Behaviour*, p50
[454] This postulate is obviously highly controversial. A discussion on it would be out of scope here. We know that Gauthier completely rejects it.

Now, let

1) Ai denote i's personal objective position in social situation A (with regards to income, wealth, consumption level, health...) and
2) P_i denote his subjective attitude (including his personal preferences).
3) R_i denote a vector consisting of all objective causal variables needed to explain the subjective attitudes expressed in Pi.

When individual i tries to make an interpersonal utility comparison between the utility levels U_i (A_i) and U_j (B_j), it amounts to deciding whether he would prefer to be in the objective position A_i with his own subjective attitudes P_i ([A_i, P_i]) or rather to be in the objective position B_j with j's subjective attitudes P_j ([B_j, P_j]). An individual's preferences among extended (hypothetical) alternatives such as [A_i, P_i] and [B_j, P_j] are called extended preferences.

Let $V_i = Vi\ [A_j, P_j]$ be the extended utility function[455] for individual i over all extended alternatives [A_j, P_j] with j = 1...n, then:

(1) $U_j\ (A_j) = U_j\ (A) = V_i[A_j, P_j]$
(2) $W_i\ (A) = (1/n)\ \Sigma\ V_i\ [A_j, P_j] = (1/n)\ \Sigma\ U_j\ (A)$[456]

In other words, a rational individual i attempting to force upon himself a moral attitude would choose to maximize his welfare function which is nothing else than the social average utility level.
In his model, Harsanyi identifies several types of preferences.

• The personal (or manifest) preferences that guide our everyday behaviour and are expressed in our *utility functions* U_i.

[455] The utility function representing the extended preferences.
[456] When individual i is constructing his social welfare function Wi, he is required to make interpersonal utility comparisons by comparing the utility units of the n individuals utility functions U1...Un using the appropriate conversion ratio qj to homogenise them. As a result his social welfare function becomes: Wi (A) = (1/n) Σ qj Uj (A)

• The moral / social (or conditional) preferences are the preferences that would guide our thinking would we force a moral attitude upon ourselves. They are expressed in our *social welfare functions* W_i.

• The extended preferences are the expression of what individual i would prefer if she had to decide about j's preferences placing herself in j's objective position (income, wealth, health, social position) and assessing j's subjective attitude (as expressed by Uj). Extended preferences are imaginary and are embodied in the *extended-utility function* V_i.

• The true preferences are the preferences we would have if we had all the relevant information, we always reasoned with the greatest possible care and we were in a state of mind most conducive to rational choice. Our rational wants are those consistent with our true preferences whereas irrational wants are those that fail the test. This has repercussions for i's assessment of U_j: 'instead of using j's actual utility function Uj which represents j's actual manifest preferences, i may wish to use a *corrected utility function* U'_j, representing i's own estimate of what j's preferences would be if j had better information about the relevant facts'[457].

• The antisocial preferences are preferences that in Harsanyi's opinion should be excluded since utilitarians' 'basis of all moral commitments to other people is a general goodwill and human sympathy'[458]. If j's preferences conflict with i's own fundamental value judgments, i would be justified in using a *censored utility function* U''_j instead of U_j in order to disregard j's antisocial preferences.

We have now reviewed the necessary background to the theory of rational choice as well as to the game and social choice theories. It is now time to turn to field of contractarianism.

[457] 'Morality and the Theory of Rational Behaviour', p 61
[458] *Rational Behaviour*, p56

Section 3: Two modern versions of contractarianism

It is now time to present two versions of contractarianism, contemporary to *Morals by Agreement*, chronologically Rawls's theory of justice and Buchanan's theory of consent. Rawls is probably more known but both these theories have inspired Gauthier and are direct challengers to his *Morals by Agreement*. Buchanan, like Gauthier, follows the Hobbesian contractarian tradition, whereas Rawls is more Kantian in his approach.

A) Rawls' theory of justice

A Theory of Justice was published in 1971. It is a classic of moral and political philosophy. The main features of the theory are usually well known and it is not our purpose to summarize it all here. We shall only raise a few relevant issues. After providing the basic features of the theory, we shall focus on the Kantian background, the conception of justice obtained and the concepts of rationality displayed in the theory. Through each of these issues, we should have a better grasp of Rawls's project and a broader picture of his achievement. It is important to note that, when writing this book, Rawls was familiar with Harsanyi's early works on the general theory of rational behaviour and he extensively compares his approach to Harsanyi's.

Basic features of the theory

a) The method

Rawls adopts a two step theory: 1) from practical reasoning, the agents derive their concept of the right (i.e. the principles of justice for the basic structure of society as well as for individuals); 2) Then within the framework of the right so derived, they each choose their individual good (i.e. their individual long term rational life plan).

The first question to answer is therefore: what principles of justice *would* rational agents choose if they were placed in a specified *original position.* Since to each original position chosen corresponds a different set of principles of justice the next logical question is: how to choose the original position? Rawls suggests adjusting the original position and therefore the principles of justice to our *'considered judgements'.* As we grow, we develop a sense of justice and we acquire a skill of judging what is just and what is unjust. We have a desire to act in accordance with this sense of justice and we expect others to feel the same. The principles rationally derived should match the moral capacity so acquired. If they don't, either one needs to modify the original position and the corresponding principles of justice or one needs to modify one's considered judgment until one feels happy that the principles obtained satisfy a finely tuned sense of justice. Judgements and principles are then in *reflective equilibrium.* Rawls seems to work his way backward: from our considered judgements he chooses the principles of justice one should arrive at and then deduces the corresponding original position. We shall follow his logic here.

b) The principles of justice

There are of two sorts of principles of justice: some apply to the basic structure of society (primarily institutions) while others apply directly to individuals. We shall only provide the *social* principles of justice here.

First principle: each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberties for all.

Second principle: social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.

Priority rules: the principles of justice are to be ranked in lexical order and therefore (a) [priority of liberty] the first principle has priority over the second one; (b) [priority of justice over efficiency and welfare] the second principle of justice has priority over efficiency and maximizing of the sum of advantages; fair opportunity is prior to the

difference principle.[459] Rawls regards such a ranking as analogous to a sequence of constrained maximum principles.

Rawls demonstrates that anybody in the original position would choose the above maximin principles of justice. Why choose maximin principles and not the expected utility or the average utility principles? Rawls provides an interesting argument. We shall only recall the following points: 1) in the original position the agents lack information or known probabilities to calculate any expected utilities. The maximin is therefore their safest option; 2) the principles must be such that all the parties will keep to them once in society, even the worst off amongst themselves; the agreement would not be stable otherwise; 3) the priority rules guarantees that the agents treat each other as ends rather than as means and each endeavours self respect.

c) The original position

The circumstances of justice are as follows: individuals are roughly similar in physical and mental powers so no one can dominate the others, there is a condition of moderate scarcity so cooperation is necessary but fruitful ventures can flourish; individuals have a common interest in cooperation despite conflicting claims on the resources available; individuals knowledge are incomplete and their power of reasoning and memory are naturally limited.

Individuals are placed behind a veil of ignorance when choosing the principles of justice. This means that they don't know their place in society, their conception of the good (i.e. their individual long term rational life plan), the special features of their psychology or the particular circumstances of their own society. However, they do know the general facts about human society; they understand political and economic affairs and they know the laws of human psychology[460].

[459] *A Theory of Justice* p266

[460] As a side comment, one can note two common points between Harsanyi and Rawls: 1) morality goes with impartiality; 2) They assume the existence of basic human psychological laws.

A Kantian background

Rawls is Kantian and he is the first one to acknowledge and emphasize his affiliation[461]. Here are the main features Rawls claims to borrow from Kant.

1) Rawls is a contractarian: 'My aim is to present a conception of justice which generalises and carries to a higher level of abstraction the familiar theory of the social contract as found, say, in Locke, Rousseau and Kant.'[462] Like Kant's, Rawls's social contract is *hypothetical*: 'Kant is clear that the original agreement is hypothetical'[463]. The contract is not historically grounded. Rather, it is the result of what rational agents *would* choose if placed in a certain original position. The original position has nothing to do with a general assembly taking place at a given time. Rather, it must be interpreted 'so that one can at any time adopt its perspective. It must make no difference when one takes up this view point, or who does.'[464]

2) Being contractarian, Rawls rejects any form of teleological theory. However, he admits that rational agents must pursue some individual *good.* Each person's good is determined by what is for her the most rational long-term plan, given favourable circumstances; a rational plan is one that cannot be improved upon. Each individual good must be framed within the right. The *concept of right* acts as a *constraint* upon individual plans as well as on the principles of the basic structure of society or any other ethical principles The concept of right is anterior to the one of the good. The priority of the right over the good is a central feature of Rawls's justice as fairness but it is also, according to Rawls 'a central feature of Kant's ethics.'[465]

3) The *concept of right* itself seems to be borrowed from Kant. A conception of right is a set of principles that fulfil five conditions: 1) they are *general* in form; 2) they are *universal* in application (i.e. 'they must hold for everyone in virtue of their being moral persons ... [they] are to be chosen in view of the consequences of everyone's

---

[461] See *A Theory of Justice* §40, pp221-227 and 'Kantian Constructivism in Moral Theory'. From his first chapter, he waves away the Hobbesian heritage with a simple comment: 'For all his greatness, Hobbes's *Leviathan* raises special problem'. (*A Theory of Justice*, p 10 n4). That is it about Hobbes!
[462] *A Theory of Justice*, p10
[463] *A Theory of Justice*, p11 n5
[464] *A Theory of Justice*, p120
[465] *A Theory of Justice*, p 28 n16.

complying with them'[466]); 3) They are publicly known ('The publicity condition is clearly implicit in Kant's doctrine of the categorical imperative insofar as it requires us to act in accordance with principles that one would be willing as a rational being to enact as law for a kingdom of ends.'[467]); 4) They must impose an ordering on conflicting claims (Completeness and transitivity of the ordering); 5) Reasoning successfully from them is conclusive (i.e. they must pass the test of practical reasoning).

4) Rawls assumes that individuals are to choose the social principles of justice from behind a veil of ignorance. They don't know how the various alternatives will affect them and they are obliged to evaluate the principles from a general point of view. According to Rawls, the formulation of the veil of ignorance is implicit in Kant's doctrine of the categorical imperative:

'Kant held, I believe that a person is acting autonomously when the principles of his action are chosen by him as the most adequate possible expression of his nature as a free and equal rational being. The principles he acts upon are not adopted because of his social position or natural endowments, or in view of the particular kind of society in which he lives or the specific things that he happens to want. To act on such principles is to act heteronomously. Now the veil of ignorance deprives the persons in the original position of the knowledge that would enable them to choose heteronomous principles.'[468]

5) In the original position, individuals are assumed to be *mutually disinterested.* In other words, they try to advance their system of ends as far as possible without seeking to confer benefits or impose injuries on one another. They are not moved by affection, rancour, envy or vanity. Again this assumption parallels Kant's notion of autonomy. Any of these feelings would pollute individuals' rationality when establishing the principles of justice. It is part of their liberty to be freed from their interest in others, whether this interest is positive or negative. Men have common as well conflicting interests but they must allow for freedom in the choice of a system of final ends when deciding on the principles of justice.

[466] *A Theory of Justice*, p114
[467] *A Theory of Justice*, p115
[468] *A Theory of Justice*, p222

The concept of justice obtained

a) Narrowing down of the field of justice to distributive justice.

Rawls understands *justice as fairness*. According to his concept of justice, society is interpreted as a 'cooperative venture for mutual advantage'.[469] Rawls is cautious not to call his theory a moral theory. He restricts the field of morality to the one of justice putting aside straight away the concepts of value or moral worth[470]. Within justice he tackles primarily social and individual justice (in this order) leaving aside the laws of nations. Within social justice, he narrows down his theory to the field of distributive justice. Whatever is to be distributed is the fruit of cooperation therefore the distribution process is an integral part of the cooperation process. In procedural justice,

'the correctness of the distribution is founded on the justice of the scheme of cooperation from which it arises and on answering the claims of individuals engaged in it. A distribution cannot be judged in isolation from the system of which it is the outcome or from what individuals have done in good faith in the light of established expectations.'[471]

He distinguishes his conception of distributive justice from the utilitarian conception of allocative justice.

b) An egalitarian conception of justice

Rawls insists that the difference in initial endowments between individuals is neither just nor unjust but is a source of injustice if taken as such in society. Therefore he suggests using the 'principle of redress'[472]: undeserved inequalities generated by the difference in initial endowments should call for redress. For example, greater resources should be spent on the less intelligent or less socially advantaged children rather than on the most intelligent or the ones with a better social background. The principle of redress is a natural corollary of the principles of justice. It does correct the arbitrariness of the natural assets distribution or of the initial position in society.

---

[469] *A Theory of Justice*, p4

[470] 'Justice as fairness is not a complete contract theory. For it is clear that the contractarian idea can be extended to the choice of more or less an entire ethical system, that is, a system including principles for all the virtues and not only for justice.' *A Theory of Justice*, p 15. The next step would be the concept of 'right as fairness'.

[471] *A Theory of Justice*, p76

[472] *A Theory of Justice* p 86 -88

The concepts of rationality displayed in the theory

Rawls seems to distinguish between two concepts of rationality that bring about two different concepts: the concept of right and the concept of individual good.

a) Social rational prudence and the concept of right

Free, equal and rational individuals placed in the original position should be able to derive a concept of right (i.e. principles of justice). If the approach is essentially Kantian, the rationality in question belongs to the prudential category. Rawls acknowledges his debt to Gauthier for this concept:

'Just as an individual balances present and future gains against present and future losses, so a society may balance satisfactions and dissatisfactions between different individuals... The principle of choice for an association of men is interpreted as an extension of the principle of choice for one man. Social justice is the principle of rational prudence applied to an aggregative conception of the welfare of the group.' [473]
Rawls substitutes for an ethical judgment a judgment of rational prudence.

Although both for Rawls and Gauthier, rational prudence is at the root of mutual advantage and of compliance with the social contract, they don't understand rational prudence in the same manner. Gauthier understands it as *individual*: prudent agents rationally bring about a social order. Rawls understands it as *social*: rational prudence applies directly to the social order and frames individual goods, hence the anteriority of the concept of right over individual goods. In Rawls theory, when the agents choose the principles of justice, they don't know their individual long-term plans and therefore cannot apply rational prudence to their individual circumstances. When choosing, they merely assume that they will prefer more primary social goods rather than less. The primary social goods are rights, liberties, opportunities, income and wealth.

b) Individual rationality and the concept of good

A person's good is determined by what is for her the most rational long-term plan of life given reasonably favourable circumstances. A rational plan is a plan that cannot be

[473] *A Theory of Justice* p 21

improved upon. It is *constrained* by the principles of justice that prevail in the society she belongs to. Individual rationality is the standard rationality used in social theory: 'a rational person is thought to have a coherent set of preferences between the options open to him. He ranks these options according to how well they further his purposes; he follows the plan which satisfy more of his desires rather than less, and which has the greater chance of being successfully executed.'[474]The only exception is to the standard theory is that Rawls' agents do not suffer from envy and they are mutually disinterested. This feature has already been discussed above. Rawls seems to assume that individuals each have a life plan that they keep in the long term. This is an implicit assumption that we find as well in Gauthier.

B) Buchanan's limits of liberty[475].

Buchanan is an economist who wrote mainly in the 1970's. His intrusion into the field of political philosophy was motivated by his frustration at the increasing role of the state in every day life. His ambition was to find the right balance between on the one hand the anarchy's useless waste of resources in defence and predation and on the other hand the Leviathan's potentially abusive restriction on individual liberty. He was in favour of an enforcement system, but within the limits of individuals' consent.

More originally, he wanted to rethink the 'original position', or the position from which individuals come to contract. Economic theory takes as 'given' individuals' initial endowment. Buchanan wanted to look behind the curtain and reflects on this initial endowment, on any pre- and contractual stages. He noted that any exchange depends upon mutually agreed and defined rights and laws. How we came to agree and define these rights was the question Buchanan wanted to address. Starting from the beginning, he wanted to know: how was the initial distribution established, what was the threshold for viable (stable) agreement to take place, what did we agree on, who was to enforce the social contract and within which boundaries, how to deal with the special case of public goods and what should be the dynamic of any constitutional contract. In order to

[474] *A Theory of Justice* p 124
[475] *The limits of liberty: Between Anarchy and Leviathan*

answer these questions, he distinguished between three periods in the making of the social contract:

- The stage of natural distribution is based on the assumption that individuals are unequal.
- The stage of constitutional contract is the stage at which the laws, the property rights and the state as enforcement system are created
- The stage of post constitutional contract is the stage at which trading can commence

- in private goods: being the traditional field of modern economic theory, it is well known and therefore not discussed in the book.
- in public goods: being often neglected in traditional analysis, it is an integral part of his theory.

Natural distribution

We are all different in style, beauty, intelligence, personality, talents and we don't all come from the same environment. Buchanan starts from the rather revolutionary assumption that we are by our very nature different and therefore unequal. Our inequalities are rooted in our differences in preferences or tastes, capacities, skills or talents and environmental settings. These three types of differences characterize the natural distribution.

At the beginning, in the state of anarchistic jungle, unequal individuals have to survive. In order to do so, they have to waste some of their resources in predation and defence. Buchanan demonstrates that even in this pre-contractual stage, individuals do reach an equilibrium since each comes to know the limits of what she can take and protect from others. This equilibrium is stable since there is no incentive for any person to modify his behaviour in the absence of external shocks. As such, chaos does not accurately describe this anarchistic equilibrium since there exists a certain degree of predictability.

Can we move from this equilibrium to a Pareto better position? A way to improve our position would be to save on predation and defence resources. The question is then to

find out from which threshold we could *all* gain from, and therefore agree on, a change of behaviour (no need to attack or defend). Agreement can only emerge from a mutually advantageous state of affair. Precisely because we are unequal, some amongst us might prefer the anarchistic equilibrium. Let us call 'direct-production position' the position in which we each keep for our own use what we produce given our natural preferences, capacities and environmental settings without wasting resources in predation and defence. Buchanan demonstrates that the anarchistic equilibrium is not necessarily Pareto inferior to the 'direct-production position'.

Let us imagine a society made of two individuals A and B. A can produce more goods than B with less efforts. B will not agree to change his behaviour unless A is prepared to transfer some initial quantity of goods or endowments to B prior to negotiating any change of behaviour agreement. The temptation for B to attack A will be too strong for any such agreement to be stable. Once the transfer of good or endowment has taken place, a new anarchistic equilibrium is reached. As phrased by Buchanan: 'Positive rights may be established, once the initial transfer has taken place to bring the two parties into a setting where the direct-production assignment is, in fact, Pareto-superior to anarchistic equilibrium'.

Constitutional contract: the theory of law.

Once the natural distribution is in the anarchistic equilibrium described above, agreement can take place. Two questions come to mind immediately.

- Are the agents behind a 'veil of ignorance'? For Buchanan, *when agreeing on a constitution, individuals do know their position*, their initial endowment, preferences, capacities and environmental settings; they can actually 'use' the agreement as a means to adjust some of the known inequalities they suffer from. The constitutional contract takes more the form of a *collective bargain*. Each individual makes an economic calculus of what is best for him in terms of utility maximizing. This approach is obviously a lot closer to Gauthier's than to Rawls' or Harsanyi's.

• What about justice in this process? Buchanan considers justice as an economic calculus in disguise and as a hypocritical tool. When talking about constitutional changes over generations, here is the way he refers to 'justice': 'In those situations where individuals may have rational economic reasons for accepting some reassignments of rights, where genuine constitutional change may be possible, the public discussion may be conducted in the rhetoric of "justice". Even the advocates of structural change may not be fully aware of the rational or utility-maximizing motivation that lies at the base of their proposals... Arguments which may find their origins in rational economic calculation, ... when presented under the disguise of justice tend to attract support.'[476]

The constitution will be an agreement between all parties about:

1) Behavioural changes or disarmament contracts
2) Property rights as established in the natural distribution
3) Conditions of enforcement
4) Rules regarding public goods.

1&2) *Disarmament contract and property rights.*

The first two sets of laws are rather straightforward. I use all my own endowment to produce my own goods without wasting any resources on predation or defence and whatever equilibrium we reached in the anarchistic stage puts the limits of what is mine and what is yours. You don't attack me and I don't attack you. The disarmament contract and the property rights are directly derived from the natural position.

3) *The protective state and enforcement system*

'Straightforward utility maximization will lead each person to defect on his contractual obligation if he expects to be able to accomplish this unilaterally.'[477] The bigger the society, the more likely such calculus and the whole agreement quickly becomes void. Therefore, *when contracting*, the parties must enter into some sort of enforcement

[476] *The Limits of Liberty*, p80-81.
[477] *The Limits of Liberty*, p64

agreement. Individuals write the rules that are to be kept by the enforcing agent (usually referred to as the state). It is not the role of the *protective state* to write or interpret the rules or to make value judgements. Its role is purely scientific, mechanical: was the law violated yes or no and if yes what punishment is applicable for such violation.

The establishment of the state as enforcement system is part of the constitutional contract. The rules and the precise role of the enforcing agent are agreed upon as part of the social contract. The reason for this is two-fold. Firstly, as straightforward utility maximisers, it is more rational to defect than to comply when others do comply. However, the viability of the contract depends upon the general knowledge of mutual compliance. The enforcing agent is the guarantor of post contractual compliance.

Secondly, as human beings, we are weak by nature and even if we know what is best for ourselves in the long term and decide upon a plan of actions, we might be tempted to behave adversely to our best interest at the time of action. Buchanan gives the example of Crusoe who, aware of his interest to wake up early and work when the sun was low, built and set an alarm clock. Buchanan writes:

'Crusoe imposes rules on his own behaviour because he recognises his own imperfection in the face of possible temptation.... Crusoe constructs his alarm clock, an impersonal and external device designed to impose constraints on his own choice behaviour. He may, of course, also select internal rules or precepts which, once adopted, will be rigorously followed. But there remains an important difference between the two cases, one that has significance for the broader problems. With the alarm clock, Crusoe disturbs his dozing in advance. *He closes off one behavioural option that would continue to remain open under a voluntaristic rule.*'[478]

When we agree on an enforcement system at the time of contracting, we program in advance an *external* constraint that will force us to keep a beneficial agreement. This approach will obviously become highly relevant when we consider Gauthier's concept of constrained maximization.

Buchanan also raises several issues about the enforcing agent (or protective state) such as what constraints on the enforcing agent's behaviour or what limits to its power, how to reassess its role over generations. Unfortunately, these issues fall outside the scope of

[478] *The Limits of Liberty*, p 93

this presentation. Buchanan offers original analysis and suggestions that would deserve attention in another enquiry.

4) The productive state and rules regarding the public goods

The state will have the responsibility of creating, managing and monitoring public goods. As such, it becomes a *productive state.* However the dividing line between private and public sector of the economy is settled in the basic constitution Why is the problem of public goods sorted out at the time of constitutional contracting and not later? Buchanan's argumentation is worth a little detour since it will become highly relevant when we review Gauthier's concept of translucency.

The distinction between public and private goods is established as follows: private goods are rivals in consumption (if you use it, I cannot use it), they are usually divisible amongst persons, whereas public goods are usually available to all. The markets 'naturally' regulate the distribution of private goods. An abundant economic literature exists on the topic. Free riding problems arise once we turn to public goods. In short: 'The individual maximizes his utility by refraining from making an independent contribution toward the provision and the financing of the commonly shared good and service.'[479]

The dilemma posed by public goods can be summed up as follows: Pareto optimality or efficiency cannot be attained until all persons are brought into the trading agreement. Unfortunately, some might benefit less from public goods than others and might require more gains out of it. Such unequal treatment at the outset is contrary to any basic principle of political order. Therefore, Buchanan suggests the concept of *exclusion.* In order to respect the rule of unanimity and the choice of some to keep out of public goods cost sharing, it is possible to put in place an exclusion mechanism. Exclusion can be extremely costly and resource wasteful to those in the sharing group but might never be observed since the certainty of being excluded from the enjoyment of the subsequent benefits could motivate all persons to join in the basic contract. In short, each individual has to decide at contracting time whether she wants to participate in public

[479] *The limits of liberty*, p37

goods financing or not. If she does, she is allocated a right to enjoy the benefits of the public good, if she does not, she is deprived of such a right and she is excluded from the community.

Buchanan notes two important points:

- 'The outcomes that defines the amount of publicly provided goods and services and the means of sharing their costs are themselves contracts, and as such, these, too, require enforcement. This creates a necessary interface between the productive and the protective state.'[480]

- 'The necessary condition is only that public goods exchange, conceived as games, be positive sum for all participants. There is no necessity that aggregate payoffs be maximized.'[481] However, ideally, the state is a productive process when it 'enables the community of persons to increase their overall levels of economic well-being, to shift toward the efficiency frontier.'[482]

Post constitutional contract

Now that all the rights have been allocated, trading can commence both in private and public goods under the conditions and rules specified in the constitutional contract.

Buchanan then introduces a fabulous dynamic in his theory. All the above is very nice as long as the parties to the contract are the ones who have to keep the rules ex-post. What about the following generations who had no part in the original constitutional contract? They might be more inclined to defect unless the rules can be adapted. Buchanan describes the Status Quo as the existing current situation of a political order. Whether it falls within the description above given or not is irrelevant. The rules are as they are and have to be modified from there not from where they should be. Once again the dynamic part of Buchanan theory falls outside the scope of this presentation but would deserve attention elsewhere.

[480] *The limits of liberty*, p 98
[481] *The limits of liberty*, p 47
[482] *The limits of liberty*, p 97

BIBLIOGRAPHY

Aristotle (340's BCE) *Nicomachean Ethics*, in Classics of Moral and Political Theory, 3rd Edition, Indianapolis: Hackett Publishing Company, pp 195 –300

Arrow KJ (1951) *Social Choice and Individual Values*, New York: John Willey & Sons.

Axelrod Robert A. (1984) *The Evolution of Cooperation*, New York: Basic Books, Inc.
_____ (1997) *The Complexity of Cooperation*, Princeton University Press

Bacharach Michael (2000) 'Framing and Agency: A New Theory of Cooperation', Unpublished

Baier Annette (1988) 'Pilgrim's Progress', *Canadian Journal of Philosophy* 18, pp 315-330

Baier Kurt (1958) *The Moral Point of View: A Rational Basis of Ethics*, New York: Ithaca or Cornell University Press.
_____ (1988) 'Rationality Value and Preference', in *Social Philosophy and Policy* 5/2, pp 17- 46

Baillie James (2000) *Hume on Morality*, London: Routledge Philosophy Guidebooks

Barry Brian (1989) *Theories of Justice*, Volume 1 of *A Treatise on Social Justice*, Berkeley: University of California Press.

Binmore Ken (1993) 'Bargaining and Morality, in Rationality', in *Rationality, Justice and the Social Contract*, edited by D. Gauthier and R. Sugden, Hemel Hempstead: Harvester Wheatsheaf, pp 131- 157

Braithwaite R.B. (1955) *Theory of Games as a Tool for the Moral Philosopher*, Cambridge: University Press.

Braybrooke D. (1987) 'Social Contract Theory's Fanciest Flight', *Ethics* 97 (4)

Bratman Michael (1987) *Intention, Plans and Practical Reason*, Cambridge, MA: Harvard University Press

_____ (1996) 'Planning and Temptation', in *Mind and Morals* edited by L. May, M. Friedman and A. Clark, Cambridge, MA: MIT Press.

Broome John (1991) *Weighing Goods*, Oxford: Basil Blackwell

_____ (1999) *Ethics out of Economics*, Cambridge: Cambridge University Press.

_____ (2001) 'Are intentions reasons? And how should we cope with incommensurable values' in *Practical Rationality and Preference*, edited by Morris & Ripstein, Cambridge: Cambridge University Press, pp 98 – 120.

Buchanan James M. (1975) *The limits of Liberty: Between Anarchy and Leviathan*, Chicago: The University of Chicago Press

_____ (1988) 'The Gauthier Enterprise', in *Social Philosophy and Policy* 5/2, pp 75 – 95

Danielson Peter (1988) 'The Visible Hand of Morality', *Canadian Journal of Philosophy* 18, pp 357-84

Elster Jon, (1979) *Ulysses and the Sirens: Studies in rationality and irrationality*, Cambridge: Cambridge University Press

_____ (1983) *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge: Cambridge University Press

Finkelstein Claire (2001) 'Rational Temptation', *Practical Rationality and Preference*, edited by Morris & Ripstein, Cambridge: Cambridge University Press, pp 56-80

Fishkin James S (1988) 'Bargaining Justice and Justification: Towards Reconstruction', in *Social Philosophy and Policy*, pp 46-65

Freeman Samuel (1998) '*Contractarianism*', in *Routledge Encyclopaedia of Philosophy*, edited by E. Craig, London: Routledge.

Gaertner W. and Klemish-Ahlert M. (1991) 'Gauthier's approach to Distributive Justice and other Bargaining Solutions', in *Contractarianism and Rational Choice* edited by P. Vallentyne, Cambridge: Cambridge University Press, pp 162 -177

Gauthier David (1963) *Practical Reasoning: The Structure and Foundations of Prudential and Moral Arguments and their Exemplification in Discourse*, Oxford: Clarendon Press.

_____ (1965) 'Rule Utilitarianism and Randomization', *Analysis* 25, pp 68-69

_____ (1967) 'Progress and Happiness: A Utilitarian Reconsideration', *Ethics*, Volume 78 (1), pp 77-82

_____ (1967b) 'Morality and Advantage', in *Morality and Rational Self-Interest,* p 166-184. (First published in Philosophical Review 76, pp 460-75)

_____ (1967c) 'How decisions are caused', *Journal of Philosophy*, 64, pp147-151

_____ (1969b) *The Logic of Leviathan: The Moral and Political Theory Of Thomas Hobbes,* Oxford: Clarendon Press.

_____ (1970) *Morality and Rational Self-Interest*, edited by D. Gauthier, Englewood Cliffs, NJ: Prentice Hall.

_____ (1974a) 'Rational Cooperation' in *Nous* 8, pp 53-65

_____ (1974b) 'The Impossibility of Rational Egoism', *The Journal of Philosophy* (Volume 71, N 14) p 439- 456

_____ (1974c) 'Justice and Natural Endowment: toward a Critique of Rawl's Ideological Framework' in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 150-170

_____(1975a) 'Reason and Maximization' in *Canadian Journal of Philosophy* 4, pp 411- 433

_____ (1975b) 'Coordination', in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 275-97. (First published in *Dialogue* 14, pp 195 – 221)

_____(1977a) 'Social Contract as Ideology', in *Philosophy and Public Affairs* 6, pp 130 – 64

_____ (1977b) 'Why ought One obey Gd? Reflections on Hobbes and Locke', in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 24- 45

_____ (1978a) 'Economic Rationality and Moral Constraints', *Midwest Studies in Philosophy*, 3 pp 92-3

_____ (1978b) 'The Social Contract: Individual Decision or Collective Bargain?' In *Foundations and Applications of Decision Theory* edited by Hooker C.A., Leach J.J. and McClennen E.F. , Vol. II, pp 47 –67 or Philosophy and Public Affairs, 6 (1977) pp 63-5

_____ (1978c) 'Social Choice and Distributive Justice', in *Philosophia* 7 pp 249-50

_____ (1978d) 'Critical Notice of John Harsanyi's Essays', *Dialogue* Vol. 17, pp 696 - 706

_____ (1979a) 'David Hume: Contractarian', *Philosophical Review* 88 pp3–38

_____ (1979b) 'Bargaining our Way into Morality: a Do it Yourself Primer', *Philosophic Exchange* 2, No 5, pp 14-27.

_____ (1979c) 'Thomas Hobbes: Moral Theorist', *The Journal of Philosophy*, Volume 76 (10), pp 547-559

_____ (1982a) 'On the Refutation of Utilitarianism' in *The Limits of Utilitarianism* edited by Miller HB and Williams W.H. pp 144 – 163

_____ (1982b) 'Justified Inequality?', *Dialogue* 21, pp 431-443

_____ (1982c) 'Three Against Justice: The Fool, the Sensible Knave, and the Lydian Shepherd' in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 129-149

_____ (1982d) 'No Need for Morality: the Case of the Competitive Market', *Philosophic Exchange* 3, No3 pp 41-54.

_____ (1983) 'Critical Notice of Ulysses and the Sirens by Jon Elster', *Canadian Journal of Philosophy* 13, pp 133-140

_____ (1984a) 'Deterrence, Maximization and Rationality', *Ethics* 94 (N3), pp 474-95

_____ (1984b) 'The Incomplete Egoist' in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 234 - 273

_____ (1985) 'Justice as Social Choice', in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 171-186

_____ (1985a) 'Bargaining and Justice' in *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press, pp 187-206

_____ (1985b) 'The Unity of Reason A Subversive Reinterpretation of Kant', *Ethics*, Volume 96 (1), pp 74-88

_____ (1986) *Morals by Agreement*, Oxford: Oxford University Press

_____ (1987) 'Taming Leviathan, Critical Notice of Jean Hampton's Hobbes and the Social Contract Tradition'. *Philosophy of Public Affairs* 16, pp 280-298.

_____ (1988a) 'Hobbes' Social Contract', *Nous*, 22, pp 71 – 82

_____ (1988b) 'Morality Rational Choice and Semantic Representation: A Reply to my Critics' in *Social Philosophy and Policy*, 5 / 2 Gauthier's New Social Contract, pp173-221

_____ (1988c) 'Moral Artifice' in *Canadian Journal of Philosophy* Vol. 18, N2 pp385-418

_____ (1990a) *Moral Dealing: Contract, Ethics and Reason*, Ithaca, London: Cornell University Press

_____ (1990b) 'Economic Man and the Rational Reasoner', in *From Political Economy - and Back?* Edited by Nichols J.H. and Wright C., San Francisco: ICS Press, pp 105-131

_____ (1990c) 'Le Promeneur Solitaire: Rousseau and the Emergence of the Post-Social Self', *Social Philosophy and Policy* 8

_____ (1990d) 'Thomas Hobbes and the Contractarian Theory of Law', in *Canadian Philosophers: Celebrating twenty years of the Canadian Journal of Philosophy*, edited by D. Copp, Calgary: Calgary University Press, pp 5-34

_____ (1991a) 'Why Contractarianism?' in *Contractarianism and Rational Choice* edited by Vallentyne P., Cambridge: Cambridge University Press, pp 15 –30

_____ (1991b) 'Rational Constraint' in *Contractarianism and Rational Choice* edited by Vallentyne P. , Cambridge: Cambridge University Press, pp 323-331

_____ & Sugden Robert (1993) *Rationality Justice and The Social Contract, Themes from Morals By Agreement*, Hemel Hempstead: Harvester Wheatsheaf

_____ (1993) 'Between Hobbes and Rawls' in *Rationality, Justice and The Social Contract* ed. D. Gauthier and R. Sugden, Hemel Hempstead: Harvester pp24-40

_____ (1993) 'Uniting Separate Persons', in *Rationality, Justice and The Social Contract* edited by D. Gauthier and R. Sugden, Hemel Hempstead: Harvester pp 176-193

_____ (1994 July) 'Assure and Threaten', *Ethics* 104, p 690-721

_____ (1995a) 'Public Reason', *Social Philosophy and Policy* 12, pp19-42

_____ (1995b) 'Constituting Democracy', in *The Idea of Democracy* edited by D. Copp, J.E. Roemer, J. Hampton, Cambridge University Press

_____ (1996) 'Commitment and Choice: an Essay on the Rationality of Plans', in *Ethics, Rationality, and Economic Behaviour*, ed. Francesco Farina, Frank Hahn, and Stefano Vannucci, Oxford University Press.

_____ (1997a) 'Resolute Choice & Rational Deliberation: a Critique and a Defence', *Nous* 31 (1): pp 1-25

_____ (1997b) 'Political Contractarianism', *Journal of Philosophy*, volume 5, 2, pp 132-148.

_____ (1997c) 'Rationality and the Rational Aim' in *Reading Parfit* edited by J. Dancy, Blackwell Publishers, pp 24-42

_____ (1998a) 'Intention And Deliberation', in *Modelling Rationality, Morality and Evolution*, edited by P. Danielson, Oxford University Press, pp 41-54

_____ (1998b) 'Rethinking the Toxin Puzzle', In *Rational Commitment and Social Justice: Essays for Gregory Kafka* edited by J. Coleman and C. Morris, Cambridge: Cambridge University Press, pp 47-59

_____ (1998c) 'Mutual Advantage and Impartiality', in *Impartiality, Neutrality and Justice: Re-reading Brian Barry's Justice as Impartiality* edited by Kelly P.J., Trowbridge: Edinburgh University Press

_____ (2001) 'Hobbes: the Laws of Nature', *Pacific Philosophical Quarterly*, vol. 82, issue 3-4, pp 258-284

_____ (2004) *Rousseau: The Social and the Solitary*, Cambridge: Cambridge University Press.

Goodin R.E. (1993) 'Equal Rationality and Initial Endowments', in *Rationality, Justice and the Social Contract*, edited by D. Gauthier and R. Sugden, Hemel Hempstead: Harvester Wheatsheaf

Griffin James (1986) *Well Being: its Meaning Measurement and Moral Importance*, Oxford: Clarendon Press

Hampton Jean (1980) 'Contracts and Choices: Does Rawls Have a Social Contract Theory?' *The Journal of Philosophy*, 77, June, pp 315-338

_____ (1986) *Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press

_____ (1988) 'Can We Agree On Morals?' *Canadian Journal of Philosophy* 18, pp 331-56

_____ (1991) 'Two faces of Contractarian Thought' in *Contractarianism and Rational Choice* edited by P. Vallentyne, Cambridge: Cambridge University Press, pp 31 - 55

Hardin Russell (1988) 'Bargaining for Justice', in *Social Philosophy and Policy* 5/2, pp 65 - 75

Hare R.M. (1952) *The Language of Morals*, Oxford: Clarendon Press

Harman Gilbert (1977) *The Nature of Morality: An Introduction to Ethics*, New York: Oxford University Press

_____ (1983) 'Justice and Moral Bargaining', *Social Philosophy and Policy* 1 pp114-31

_____ (1988) 'Rationality in Agreement', in *Social Philosophy and Policy* 5/2, pp 1-17

Harsanyi, J (1976) *Essays on Ethics, Social Behaviour and Scientific Explanation*, Dordrecht, Holland; Boston: D. Reidl Pub. Co.

_____ (1977) 'Morality and the Theory of Rational Behaviour' in *Utilitarianism and Beyond* edited by AK Sen & B. Williams, Cambridge: Cambridge University Press, pp 39 - 62

_____ (1977) *Rational Behaviour & Bargaining Equilibrium in Games & Social Situations*, Cambridge: Cambridge University Press

Hobbes Thomas (1651) *Leviathan*, in Classics of Moral and Political Theory, 3rd Edition, Indianapolis: Hackett Publishing Company pp 491- 621.

Hubin D.C. and Lambeth M.B. (1991) 'Providing for Rights', in *Contractarianism and Rational Choice*, edited by P. Vallentyne, Cambridge: Cambridge University Press, pp 112 - 127

Hume David (1740) *Treatise of Human Nature* (selections), in *Classics of Moral and Political Theory*, 3rd Edition, Indianapolis: Hackett Publishing Company pp 692 –714.

Kant Immanuel (1785) *Grounding for the Metaphysics of Morals*, in *Classics of Moral and Political Theory*, 3rd Edition, Indianapolis: Hackett Publishing Company pp 833 – 875.

Kavka Gregory S. (1983) 'The Toxin Puzzle', *Analysis* 43, pp 33-36

_____ (1987) 'Morals By Agreement', *Mind* 96. pp 117-21

_____ (1995) 'Why Even Morally Perfect People Would Need Government?', in *Contemporary Political and Social Philosophy* edited by Paul E.F., Miller F. and Paul J., Cambridge: Cambridge University Press, pp 1- 18.

Kraus J. & Coleman L. (1987) 'Morality and the Theory of Rational Choice', *Ethics* 97 (4)

Locke John (1689) *Second Treatise of Government*, in *Classics of Moral and Political Theory*, 3rd Edition, Indianapolis: Hackett Publishing Company pp 624 – 689

Luce, R.D. and Raiffa, H. (1957) *Games and Decision*, New York: John Willey and Sons

Mackie John (1978) *Ethics: Inventing Right and Wrong*, Harmondsworth: Penguin

McClennen Edward F. (1988) 'Constrained Maximization and Resolute Choice', *Social Philosophy and Policy* 5, pp 95-118

_____ (1990) 'Foundational Explorations for a Normative Theory of Political Economy', in *Constitutional Political Economy*, 1, pp 67 - 89

_____ (1997) 'Pragmatic Rationality and Rules', *Philosophy & Public Affairs*, Summer 1997, Volume 26, Number 3.

_____ (2001) 'The Strategy of Cooperation' in *Practical Rationality and Preference*, edited by C. Morris & A. Ripstein, Cambridge: Cambridge University Press, pp 189-208.

_____ (2002) 'The Rationality of Being Guided by Rules', *The Oxford Handbook Of Rationality*, February 2004, pp 222-240

Mendola J. (1987) 'Gauthier's Morals by Agreement and Two kinds of Rationality', *Ethics* 97 (4)

Mill John Stuart (1836) 'On the Definition and Method of Political Economy', in *The Philosophy of Economics, an Anthology* edited by D.M. Haussman, 2nd Edition, Cambridge: Cambridge University Press pp 52 - 68

_____ (1859) *On Liberty,* in *Classics of Moral and Political Theory*, 3rd Edition, Indianapolis: Hackett Publishing Company pp 876 - 934

_____ (1863) *Utilitarianism*, in *Classics of Moral and Political Theory*, 3rd Edition, Indianapolis: Hackett Publishing Company pp 935 - 968

Morris Christopher W (1988) 'The Relation Between Self-Interest and Justice in Contractarian Ethics' in *Social Philosophy & Policy*, 5, pp 119-153

_____ & Ripstein A. (2001) *Practical Rationality & Preference: Essays for David Gauthier*, Cambridge: Cambridge University Press

Morton Adam (2001) 'Psychology for Cooperators', in *Practical Rationality and Preference,* edited by C. Morris & A. Ripstein, Cambridge: Cambridge University Press, pp 153 -172

Nagel Thomas (1970) *The Possibility of Altruism*, Oxford: Clarendon Press

Nash J. F., (1953) 'Two-Person Cooperative Games', *Econometrica*, 21, pp 128 - 140

Nozick Robert (1974) *Anarchy State and Utopia*, Oxford: Blackwell

_____ (1993) *The Nature Of Rationality* Princeton: Princeton University Press

Paul E.F., Miller F. & Paul J. & P. Ahrens (1988) *The new Social Contract: Essays On David Gauthier*, New York: Basil Blackwell

Parfit Derek (1984) *Reasons and Persons*, Oxford: Oxford University Press.

_____ (2001) 'Bombs, Coconuts or Rational Irrationality', in *Practical Rationality and Preference*, edited by C. Morris and A. Ripstein, Cambridge: Cambridge University Press, pp 81 -98

Plato (380's BCE) *Republic*, in Classics of Moral and Political Theory, 3rd Edition, Indianapolis: Hackett Publishing Company, pp 28-191.

Rawls John (1958) 'Justice as Fairness', *Philosophical Review*, 67, pp 164 - 194

_____(1971/99) *A Theory of Justice, Revised Edition*, Oxford: Oxford University Press

_____ (1980) 'Kantian Constructivism in Moral Theory', *Journal of Philosophy* 77

Rousseau Jean-Jacques (1762) *On the Social Contract*, in *Classics of Moral and Political Theory*, 3rd Edition, Indianapolis: Hackett Publishing Company pp 771 – 830

Rubinstein Ariel, (1982), 'Perfect Equilibrium in a Bargaining Model', *Econometrica*, Vol. 50, No1, pp 97- 110

Sayre-McCord Geoffrey (1991) 'Deception and Reason to be Moral', in *Contractarianism and Rational Choice*, edited by P. Vallentyne, Cambridge University press, pp 181 - 196

Scanlon Thomas M (1982) 'Utilitarianism and Contractualism' in *Utilitarianism and Beyond* edited by AK Sen & B. Williams, Cambridge: Cambridge University Press pp103-128

Schelling Thomas C. (1960) *The Strategy of Conflict*, Cambridge Massachusetts: Harvard University Press.

Sen Amartya (1977) 'Rational Fools', *Philosophy & Public Affairs* 6 317, 44

_____(1987) *On Ethics and Economics*, Oxford: Basil Blackwell

Smith Holly (1991) 'Deriving Morality from Rationality', in *Contractarianism and Rational Choice*, edited by P. Vallentyne, Cambridge University Press, pp 229 - 254

Social Philosophy and Policy, 5 / 2 (1988) *Gauthier's New Social Contract*

Sugden Robert (1986) *The Economics of Rights Cooperation and Welfare*, Oxford: Basil Blackwell

_____ (1993) *'Rationality and Impartiality: Is the Contractarian Enterprise Possible'*, in *Rationality, Justice and the Social Contract* edited by D. Gauthier and R. Sugden, Hemel Hempstead: Harvester Wheatsheaf

_____ (2000) 'Team Preferences', *Economics and Philosophy* 16, pp 175-204

Thomas Laurence (1988) 'Rationality and Affectivity: The Metaphysics of the Moral Self', in *Social Philosophy & Policy* 5/2, pp 154 - 173

Vallentyne Peter (1991) *Contractarianism and Rational Choice, Essays on David Gauthier's Morals by Agreement,* Cambridge: Cambridge University Press.

_____ (1991) 'Gauthier's Three Projects', in *Contractarianism and Rational Choice*, edited by P. Vallentyne, Cambridge: Cambridge University Press, pp 1- 13

_____ (1991) 'Contractarianism and the Assumption of Mutual Unconcern', in *Contractarianism and Rational Choice*, edited by P. Vallentyne, Cambridge: Cambridge University Press, pp 71 – 76

Weale Albert (1993) 'Justice, Social Union and Separateness of Person', in *Rationality, Justice and the Social Contract* edited by D. Gauthier and R. Sugden, Hemel Hempstead: Harvester Wheatsheaf

Williams G. (1998) 'The Problems of David Gauthier's Attempt to Derive Morality from Rationality', *Philosophical Notes* 51

Winch DM (1971) *Analytical Welfare Economics*, Harmondsworth: Middlesex