1-1-1989

The "inescapable" prisoner's dilemma.

Ishtiyaque H. Haji
University of Massachusetts Amherst

THE "INESCAPABLE" PRISONER'S DILEMMA

A Dissertation Presented

by

ISHTIYAQUE H. HAJI

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY

May 1989

Philosophy

THE "INESCAPABLE" PRISONER'S DILEMMA

A Dissertation Presented

by

ISHTIYAQUE H. HAJI

Approved as to style and content by:

Fred Feldm

Fred Feldman, Chairperson of Committee

Jerome B. King

Jerome King, Member

Gareth B. Matthews

Gareth B. Matthews, Member

John Robison

John Robison, Member

John Robison

John Robison, Department Head
Philosophy

ACKNOWLEDGMENTS

I am deeply indebted to Fred Feldman: Throughout the time I have been working on this thesis (indeed, throughout the time I've been at this department), I have benefitted enormously from the wit, encouragement, great support, bountiful suggestions, and penetrating criticisms of my teacher and friend.

I must also express my gratitude to Gareth Matthews who has exerted a powerful and very salutary effect on my philosophical development.

Lastly, I owe many thanks to Robert Sleigh and Bruce Aune for all that I have learned from these venerable philosophers.

ABSTRACT

THE "INESCAPABLE" PRISONER'S DILEMMA

MAY 1989

ISHTIYAQUE H. HAJI, B.A., SIMON FRASER UNIVERSITY

M.A., SIMON FRASER UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Fred Feldman

Do the requirements of morality and those of rational self-interest dictate performance of the same acts in every particular situation? In this thesis I examine and evaluate various proposed answers to this age-old philosophical question. I focus on a particular kind of situation in which the two sorts of requirement seem to be at odds with one another. These are situations of contract-keeping that are prisoner's dilemma-like. In such situations, if you are moral, then it appears that you should comply with an agreement to do the "cooperative thing." If you are rational, then it seems that you will do best for yourself if you refrain from cooperating. Reformulating the central question of this essay, is it rational in some sense of 'rational' to do the cooperative thing in, and so to "escape" a prisoner's dilemma?

I begin with Hobbes and submit he would answer in the negative. In analysing Hobbes' position, I critically discuss Jean Hampton's and Gregory Kavka's views on Hobbes on state-of-nature cooperation. I then consider more recent replies paying particular attention to David Gauthier's. I argue that his defense of an affirmative reply - the desirable reply - is flawed. I arrive at a similar verdict about Edward McClennen's opinion. Finally, I advance my own conclusion - there may be situations in which people must act in a way that is either immoral or irrational.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Two Problems

Clarifying the nature of the relationship between morality and rational self-interest is a perennial philosophical issue, one that has been debated by philosophers from Plato to the present. Of the numerous problems that arise in its discussion, there are two that I find especially intriguing. The gist of the first is captured by the question of Hobbes' Fool and the ethical sceptic: Do morality and rational self-interest require performance of the same acts in every situation? Of course, the sophisticated sceptic might readily concede that morality is superior to immorality, as a general policy, from the viewpoint of rational self-interest. But what about those special cases where an immoral act seems advantageous and where duty fails to have "the visage of a sweetie or a cutie?"[1] Does it not, in these situations, pay to be immoral? If reason construed as the maximization of one's self-interest is accepted as unproblematic, an affirmative response to this question, the sceptic may initially insist, has the consequence that a rational morality is a chimmera - that morality, as Thrasymacus seems to have believed, is simply a fool's game. But here, perhaps, the sceptic has spoken too quickly. That his judgment might be rash becomes somewhat more evident if we turn to the second problem:

While accepting reason, the sceptic shuns the constraints imposed by morality on the pursuit of individual interest. Morality, he believes, lacks a "validation," or a "justification," or a "rational foundation." The issue of what we mean when we say that a moral principle is justified is highly contentious. According to one prominent tradition - the contractarian one - noteworthy patrons of which are Hobbes and the Hobbesian contemporary, David Gauthier, moral principles are "generated" as a result of agreement among rational persons. They are justified for us because they impose demands which are, or would be, rational for each of us to accept. In other words, the validation of a morality, as the contractarian sees it, is to proceed by demonstrating that rational self-interested individuals, the moral sceptic included, would voluntarily choose to be moral. Assume this conception of moral justification sound and assume a set of moral principles so justified. Then the sceptic cannot relegate morality to the refuse heaps comprised of such dubious principles as those of astrology and voodoo. These assumptions, however, would do little if anything, I think, to alleviate the first problem - whether the dictates of duty are coincident in all circumstances with those of rational self-interest. Indeed, assuming the contractarian project a success - assuming a set of moral principles justified for everyone, may lead to an even more formidable third problem: If the prescriptions of morality and those of rationality are not concurrent under all conditions, and if both duty and interest provide us with reasonable grounds for action, then in cases of conflict does duty take precedence over, or give way to, rational self-interest?

In this thesis I focus primarily on the first problem, secondarily on the second, and not at all on the third. In discussing the first, I chiefly confine attention to situations of contract-keeping that are prisoner's dilemma-like. In these situations the requirements of morality and those of rationality often seem at loggerheads. In relation to them the focal question is this: Is the appearance of conflict between duty and interest merely that - sheer appearance, or is the conflict real? I'm inclined to side with Hobbes' Fool. Others are more optimistic. Before I can assess their positions and defend mine, the terms in which the problem is formulated must be clarified. In particular, I must explain the conception of rationality as "straightforward maximization." I must then say something about prisoner's dilemmas. The next two sections deal with these issues. I will then be in a position to show that while it appears morality requires that one keep one's contract to do the "cooperative thing" in a prisoner's dilemma, rationality seems to sanction violation.

1.2 Straightforward Maximization (SM)

Assume that at each moment of choice a person has several alternatives among which to choose. Assume, further, that for each alternative there is an outcome. The outcome is what would occur if the alternative were performed.[2] Assume that each outcome has a value for the agent. Her choice among her "timewise identical" alternatives is to be dictated by the value-for-her of the outcomes that would result if they were performed.

There are different conceptions about the value of outcomes for agents. I will distinguish two:[3] Let the utility of an outcome, o, of some action, a, for some agent, s, be the value of o for s. On one view, the value of an outcome is an objective measure of how beneficial or harmful that outcome would be for the agent. One and the same outcome may be objectively beneficial to a certain degree, and objectively harmful to a certain degree, for an agent. Assume that the net objective value of an outcome for an agent can be ascertained. On this approach, the agent-utility (or s-utility) of an action, a, for agent, s, is simply the net objective value of the outcome that would result were s to do a.

On a second approach, an outcome that has value has it relative to a person purely in virtue of the fact that the person has a preference for the outcome. Some theorists believe that only a certain class of preferences is relevant to a determination of an agent's values. Gauthier, for example, supposes that only considered preferences are germane. Considered preferences, he tells us, are those "that would pass the test of reflection and experience." (M by A, 31, 21-46) On this second view, the s-utility of action, a, for agent, s, can be conceptualized as a measure of s's preference for the outcome that would result if s were to do a.

Assume either one of these axiological approaches. Then on one variant, SM is the theory that for any act, a, and any agent, s, a is rational for s if and only if none of its alternatives has a higher s-utility than it has.

There is a more complicated variant of SM: Assume that for each alternative there is a set of outcomes that could result if it were performed. The complex variant requires that an agent weigh not only the utilities of outcomes realizable in action but also the probabilities on her evidence that they will occur. It requires that a rational agent endeavor to maximize the expected agent-utility of her actions. This parameter is to be understood in the following way: Take some course of action, a, to be performed by some agent, s. Consider all the posible outcomes of a that could result if it were performed. Of these, restrict attention to the subset of outcomes that would affect the welfare of s. For each of these, take the agent-utility of the outcome, U(oi), and the probability of the outcome given the action, P(oi,a), and multiply U(oi) and P(oi,a). Find the sum of the products. The sum is the expected agent-utility of a. On this variant, SM permits an agent to perform an action if and only if none of its alternatives has a higher expected agent-utility (or "expected utility" as I shall abridge) than it has.

1.3 Prisoner's Dilemmas (PDs)

The prisoner's dilemma is named after a story about two prisoners. Here's an illustrative tale: Suppose Butch and Sundance have been arrested for voyeurism. There is enough evidence to convict them on the charge for this felony, but the DA is after bigger game. He suspects they have robbed a bank together, and he believes he can get them to confess to it. Assume both felons are interested solely in spending the least time in prison and that the DA knows this. The felons are held in different cells and cannot communicate with one another. The DA approaches each with this proposition: "I'm going to offer your partner the same deal, so listen carefully. If one of you confesses to the bank robbery but the other does not, Confessor gets one year, and Sealed Lips gets thirty. If you both confess, I'll see to it that each of you lugs a ball and a chain around for a decade. If neither of you confesses, then each of you will enjoy the hospitality of the slammer for two years - Sam Maquerrie's wife has sworn under oath that you Peeping Toms violated the rights of Betty Loo Hot Lips. I'll be back in a while - for your confession."

The situation of the two prisoners can be represented by this matrix:

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	10,10	1,30
Remains silent	30,1	2,2

Figure 1.1 The Prisoner's Dilemma

Each "box" depicts an outcome of a pair of actions, one performed by each prisoner. The numbers indicate years-in-prison with Sundance's jail terms listed first.

If each is to act as a straightforward maximizer (SFM), confessing is the best policy - no matter what the other does each does best if he confesses. In the terminology of game theory, the act of confessing dominates for each player. The outcome of mutual confession, however, is not optimal, an outcome being optimal if and only if there is no alternative outcome that both gives some person a greater payoff and no person a lesser payoff. Each prefers mutual silence, an optimal outcome, to mutual confession. But this mutually preferred outcome lies beyond the reach of the SM-rational felons.

It is difficult to state precisely just what conditions must be satisfied in order for this situation to count as a PD. Following Professor Feldman, let's assume a minimal set of conditions:[4]

(i) The actions of the interacting agents must be independent in the sense that "no matter what choice either makes, he would still make that choice no matter what choice the other makes."[5] This "counterfactual independence condition" rules out, for instance, the possibility that Butch can somehow force Sundance to remain silent and then confess himself. Were the actions of the two literally interdependent in this way, it would not be true as it is in an authentic PD, that although confessing is individually rational, it is collectively irrational.

(ii) There are no hidden choices. Each can either confess or remain silent. It is not open for either, for example, to bribe the DA or to break out of prison.

(iii) There are no hidden payoffs. We assume that the payoffs shown exhaust what each felon stands to gain or to loose from each of the four possible outcomes. We assume, then, that each cares solely to minimize his prison-term. It isn't the case, for example, that Butch swayed by feelings of comradeship, is willing to sacrifice himself by remaining silent if Sundance confesses. If he were to be so moved, the payoff '1,30' in the upper right would be misleading.

The matrix representing the preferences of each felon for each outcome is this: (Higher numbers indicate lower preferences.)

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	3,3	1,4
Remains silent	4,1	2,2

Figure 1.2 The Prisoner's Dilemma Preference Matrix

(iv) The last condition is simply that the relevant persons' preferences for outcomes, resulting from sets of combined possible actions, one performed by each person, are ordered as in the above matrix.

Let's stipulate, perhaps redundantly, that a prisoner's dilemma is any two- or more-person situation that satisfies conditions (i), (ii), (iii), and (iv).[6]

Butch and Sundance, in a PD, both end up confessing as the DA correctly predicted. If only somehow they could curtail the pursuit of their own advantage and refuse to confess, each would be better off. It might be thought they could do this by cooperation: Suppose prior to the bank robbery the two make an agreement with each other to the effect that should they be captured each would not confess, since each is aware of the DA's tactic to get felons to confess, and is aware that the outcome of mutual silence is better for each than the outcome of mutual confession. But this pre-crime compact will not help matters. Merely having made such an agreement on prudential grounds does not provide any reason for a SFM to comply with it when the time comes for doing so: Although optimal, the "agreement outcome" is not in equilibrium. An <u>equilibrium</u> <u>outcome</u> is the product of a set of actions, one for each interacting person, such that for each such person, there is no alternative action that this person would prefer, the actions of all the others being fixed.

If Butch does his part - if he complies with the agreement - Sundance can do better by defecting. Since both are SFMs and since positions are symmetric, it seems that compliance with the pre-crime agreement would be rationally unjustified.

1.4 A Possible Conflict Between Duty and Interest

The inability of SFMs to adhere to mutually advantageous agreements requiring self-sacrifice in situations like the PD can be used to illustrate what seems to be a conflict between morality and SM: Assume morality requires that one adhere to a prudentially undertaken agreement even if adherence compromises one's own interests. Morality, we assume with Gauthier, imposes constraints on the direct pursuit of self-interest. If Butch and Sundance were moral, they would adhere to their pre-crime agreement if caught, but as SM-rational agents, they could not. It seems that whereas in such a situation morality constrains behavior in the direction of optimality, rational self-interest leads to a suboptimal outcome. Morality and SM, it appears, are therefore incompatible.

Call the outcome that would result if Butch and Sundance were each to keep silent the "cooperative outcome." To do the "cooperative thing" is for each felon to remain silent. The focal problem of this thesis can now be formulated in this way: Is it rational, in some sense of 'rational,' for each party in a PD to do the cooperative thing? Do the prescriptions of morality and those of rationality coincide in PDs? Yet alternatively, can rational agents "escape" the PD by doing the cooperative thing? If not, does the dilemma vindicate the charge of Hobbes' Fool?

The rest of this chapter is synoptic. I summarize what I undertake to show in each of the ensuing chapters.

1.5 Hobbes on Rational Cooperation

Hobbes seems to have been among the first to recognize the significance of PDs to the issues of whether there is harmony between the dictates of morality and those of self-interest, and to whether morality can be conceptualized as a "product" of rational agreement. Thus Professor Kavka writes:

> Hobbes's very point about the state of nature is that it has this multiparty prisoner's dilemma structure and hence must be abandoned. For so long as individuals remain in that state and act rationally, they will inevitably produce worse outcomes for themselves than they could obtain under other conditions. (K,113)

Gauthier traces the roots of his response to the Fool's contention that injustice may sometimes pay to Hobbes. (M by A, Chapter VI, section 1.3)

In consequence, it is only proper to begin with a consideration of what this venerable English philosopher had to say on state-of-nature cooperation. I devote Chapters 2 and 3 to this task.

In Chapter 2 I do two things. (a) First, I show that a careful examination of his views on state-of-nature cooperation reveals that it is controversial, contrary to popular assumption, whether Hobbes himself believed that natural-state individuals really are in a multiparty prisoner's dilemma.

Suppose, in light of this uncertainty, we overlook - in a judicious manner - certain passages in Leviathan and in other relevant Hobbesian texts. Then it is possible to interpret Hobbes as arguing for the point not infrequently attributed to him: The state of nature is one big PD. Hobbesian individuals in that multiparty dilemmatic situation would there reap suboptimal outcomes. Their very rationality prevents them from adhering to the interest-constraining dictates of morality - in particular, to the third law of nature, a moral principle requiring compliance with covenants rationally made - and so prevents them from escaping the dilemma. Hobbes may then be taken to be advocating a "political" solution to their predicament: Institute a sovereign who sees to it by his might that it is no longer advantageous for each in the state of nature to violate the constraints of duty.

(b) Second, I suggest that this political solution fails: Sovereign institution itself seems to involve PD-like problems that Hobbesian individuals are unable to overcome. It should be noted that even if this solution did succeed, by invoking it Hobbes appears to abandon any hope of reconciling duty and self-interest. The covenants enforced by the sovereign that enable rational agents to "escape" PDs are no longer interest-constraining. "Cooperation" is straightforwardly rational.

In discussing (a) and (b), I will comment on some of Kavka's views on Hobbes on state-of-nature cooperation.[7]

Jean Hampton takes issue with what I propose in Chapter 2.[8] She believes that although Hobbesian individuals are psychologically incapable of instituting a sovereign, the inauguration of the sovereign does not involve any PD-like problems. In fact, she argues that it is SM-rational for Hobbesian individuals in the state of nature to subjugate themselves to a sovereign. I believe she is mistaken on both counts. I show this in Chapter 3. In doing so, I will explain why her views on Hobbes on natural-state cooperation are not cogent.

1.6 Gauthier on Rational Cooperation

Chapters 4, 5, 6, and 7 are devoted to a critical examination of Morals By Agreement. In this work, David Gauthier launches a full scale defense of a contractarian theory of ethics - his aim is none other than to show that morality can be "derived" from rational agreement. If successful, this ambitious undertaking would accomplish a number of desirable goals: It would, for example, provide a response to the moral sceptic's demand for a non-moral justification for being moral: Moral principles, the contractarian might volunteer, are a requirement of practical rationality. As Gauthier says, moral principles are "a subset of rational principles for choice," so that "To choose rationally, one must choose morally." (M by A, 4) It would, in addition, rise to Hobbes' Fool's challenge that injustice may sometimes "stand with that reason which dictateth to every man his own good." Gauthier's brand of contractarianism attempts to defuse this challenge in a particularly engaging fashion: Gauthier attempts to show that SFMs can escape the PD without the brandishing sword of the sovereign. He first argues that SFMs will bargain with each other on the basis of the principle of minimax relative concession (MMRC). This principle, Gauthier tells us, is a moral principle. It's a principle of distributive justice. MMRC requires SFMs to do the interest-constraining cooperative thing in a PD. He then argues that it is rational for SFMs to change their very conception of rationality. Given a choice between straightforward maximization and constrained maximization (CM), it is SM-rational for SFMs to become constrained maximizers. CM, however, requires compliance with agreements that it is SM-rational to make. Since it is SM-rational to agree to cooperate in a PD, it is CM-rational to comply with such an interest-constraining agreement as well. The PD, Gauthier concludes, is escapable after all.

Professor Gauthier's contractarianism raises many intriguing issues. I concentrate on these:

(1) The "Bargaining Problem." Assume cooperation makes possible a surplus of goods that would not be forthcoming if the prospective cooperators were to act independently. The generation of such a surplus creates a problem of distribution: how is the surplus to be alloted among those who produce it? More specifically, on what principle will SFMs agree to apportion the cooperative surplus? Gauthier argues that there is a unique principle that governs the relevant distribution, the principle of minimax relative concession, or its twin principle, maximin relative benefit. In Chapter 4, I question this claim. I argue that SFMs will not necessarily minimax in many bargaining situations. They may well reach agreement on

some other basis. So there is no unique solution to the bargaining problem.

Bargaining theory presupposes that what one brings to the bargaining table - one's prebargaining endowment - is settled or fixed. A "Lockian Proviso," Gauthier argues, defines a system of property rights that constrains one's initial endowment. Since each agent is entitled to her initial endowment as a matter of right, only the surplus, if any, generated by cooperation is subject to distribution. I mention this crucial aspect of Gauthier's theory only to set it aside. The difficult and highly interesting problems it raises will not be discussed in this thesis.

(2) The "Compliance Problem." Suppose the bargaining problem solved and solved in the direction Gauthier recommends. It would then be rational according to the principle of MMRC to agree to do the cooperative thing in a PD. There is, however, still a problem of compliance: Suppose Butch and Sundance have made a rational pre-crime agreement to remain silent if captured and questioned. Why should they comply when the time comes to perform? After all, the outcome of mutual silence is not in equilibrium. If one complies the other will do best by reneging. Gauthier's answer is that it is rational for them to comply - not SM-rational, of course, but "PR-rational." PR is the theory that an act is rational if and only if it

"expresses" a rational disposition of choice. "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." (M by A, 183) Gauthier limits choice to two decision-making strategies, SM and CM. He argues that CM and not SM is the rational disposition. Any SFM, he claims, confronted with a choice between SM and CM will (under specified conditions) choose the latter. CM requires compliance with prudentially rational agreements. Assuming PR is unproblematic the compliance problem, Gauthier thinks, is solved.

I believe the "choice argument" is not sound: It is false that SM is not utility maximizing. I endeavor to show this in Chapter 5.

(3) Assessing Theories of Rationality. I do, however, concede that it is CM-rational (under certain conditions) to abide by agreements that it is SM-rational to make. But this fact should not by itself impress a SFM. With the failure of the choice argument, some rationale is needed to establish the superiority and so the preferability of CM to SM. In an earlier work, "Reason and Maximization," Gauthier proposes a "self-support" criterion of adequacy for theories of practical rationality. His claim is that although CM is self-supporting, SM is not.

CM is therefore a better theory than SM. The criterion and variants of it are developed, discussed, and rejected in Chapter 6.

(4) The Justification of Principles of Practical Reason. I said earlier that the contractarian attempts to answer the moral sceptic by showing that reason requires acknowledgment of at least some moral constraints. This way of answering the sceptic assumes that although there is a problem about the justification of morality, there is no anlaogous problem about a theory of practical rationality. Gauthier's project is particularly attractive because of its implicit rejection of this assumption. The project offers a unified scheme for the justification of both a morality and a theory of practical rationality. That scheme I call "abstract choice theory." In Chapter 7 I argue against the rational credentials of choice theory.

Gauthier's contractarianism, I conclude, cannot deflate the Fool's challenge.

1.7 McClennen on Rational Cooperation

In the penultimate chapter, I examine Edward McClennen's view that "resolute" agents are rationally able to secure cooperative outcomes that are optimal in PD-like cases.[9] This view, I believe, is once again questionable: resolute choosers fare no better than SFMs in the relevant dilemmatic situations.

I conclude that there are cases in which persons must act in a way that is either immoral or irrational. The challenge of Hobbes' Fool has yet to be answered.

Notes

1. In Morals By Agreement (1986), Oxford: Clarendon Press, p. 1, Gauthier informs us:

> The unphilosophical poet Ogden Nash grasped the assumptions underlying our moral language more clearly than the philosopher Hume when he wrote:
>
> > 'O Duty!
> > Why hast thou not the visage of a sweetie or
> > a cutie?'
>
> We may lament duty's stern visage but we may not deny it. For it is only as we believe that some appeals do, alas, override interest or advantage that morality becomes our concern.

2. In strategic contexts in which an outcome for an agent depends in part on the actions chosen by other persons, this assumption does not hold. The account of SM presented in this section is standard for parametric contexts. But it should, nevertheless, facilitate an understanding of the reasoning of straightforward maximizers in the strategic context of the prisoner's dilemma.

3. See Chapter 7, section 1.4 of this thesis, for an elaboration of these two theories of value.

4. Fred Feldman, "On The Advantages of Cooperativeness," forthcoming in Midwest Studies in Philosophy, section 3.

5. Fred Feldman, "On The Advantages Of Cooperativeness," section 3.

6. Professor John Robison has indicated that whether or not two individuals like Butch and Sundance in the real world are in a PD partly seems to depend on how we characterize their alternatives. If each has the choice of either confessing or remaining silent, then (providing the relevant conditions just adumbrated are satisfied) it seems that the two felons are indeed in a PD. But after consulting with Butch, the DA could equally well have apprised Sundance that he, Sundance, could either match whatever Butch did (Butch can either confess or remain silent, Sundance is informed), or he could fail to match

the action of his partner. With Sundance's alternatives characterized in this way, the felons don't appear to be in a PD, as matrix 1.3 illustrates. One might perhaps object

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	1,1	9,9
Remains silent	0,10	10,0

Figure 1.3 Robison's Matrix

that whereas confessing is a *real* action - it's something Butch can do, matching isn't. But there's an easy reply here: Why not suppose that to match Butch's actions, Sundance must place a token in a basket. Placing a token in a basket seems to be as real an action as confessing - it's certainly something Butch can do.

What may be deep and interesting implications of Professor Robison's point await further study.

7. Gregory Kavka, *Hobbesian Moral and Political Theory* (1986), Princeton, New Jersey: Princeton University Press. Page references to this work are given in this way: (K, page number).

8. Jean Hampton, *Hobbes and the Social Contract Tradition* (1986), Cambridge: Cambridge University Press.

9. See Edward F. McClennen, "Constrained Maximization and Resolute Choice," forthcoming in *Social Philosophy and Policy*; Edward F. McClennen, "Dynamic Choice and Consistency," forthcoming; and Edward F. McClennen, "Prisoner's Dilemma and Resolute Choice," in *Paradoxes of Rationality and Cooperation* (1985), eds., R. Campbell and L. Sowden, Vancouver: The University of British Columbia Press, 94-104.

CHAPTER 2

THE SYMMETRY ENIGMA

2.1 The Symmetry Problem

Are the requirements of morality and those of rational self-interest coincident in every particular situation? Hobbes confronted this question - the central question of this work - amongst other places, in Leviathan.[1] In discussing it he devoted serious attention to situations of covenant-keeping of an highly interesting nature: The situations, it appears, can be depicted by the game theoretic matrix of the PD. Since we are concerned with precisely these sorts of cases, it is well worth considering what Hobbes has to say about them. More specifically, my principal aim in this chapter is to attempt a clarification of his views on the rationality and morality of contract-keeping in these dilemmatic situations. Having declared this ambitious aim, let me immediately admonish the reader that the chapter will not prove very illuminating in this respect. However, I will have some things to say about what Hobbes' views on the relevant issue could not have been, or perhaps, should not have been.

I begin with a brief summary of the "logical form" of Hobbes' laws of nature which constitute the moral principles of his ethical system. This will help in apprehending his third law "that men perform their covenants made." (MW,3,15,130) A covenant of mutual trust is an agreement in which both parties are required to discharge their covenantal obligation, in sequence, at some time after the contract is made. (MW,3,14,120-121,124) The brief explication of the third law should enable at least a rudimentary understanding of Hobbes' views on the morality of contract-keeping. I then present passages in which Hobbes seems to affirm that it is not rational in the state of nature for covenant-parties who have to perform first - "first-party members" - to keep their agreements, although it is rational for "second-party members" to do so if first-parties have already performed. These views on the rationality of agreement-keeping give rise to the "symmetry" problem: precisely what is the asymmetry between the situations of first-parties and second-parties in the state of nature that brings it about that whereas it is irrational for first-parties to cooperate, it is rational for second-parties to do so?

The significance of the symmetry problem is striking: On the one hand, assume as has often been done, that persons in the state of nature are in a multiparty PD. If we now also assume that these individuals are rational, Hobbes can be read as undertaking the fascinating project of demonstrating how such persons could escape their predicament. One might be sceptical of his prospects of success. If Hobbes' rational egoists really are in an authentic PD, it appears they would have to remain there: cooperation on a voluntary basis seems impossible.[2] It would take no less than an "extra-natural state" savior who saw to it that state-of-nature denizens would in fact maximize their advantage by "cooperating," if these persons were to have any chance to leave the state of nature. Hobbes' sovereign, not being an "extra-natural" being, could not assume this role. For the sovereign, if he were to be inaugurated, would be so by a social contract. State-of-nature inhabitants would covenant to set up their savior whom they would select from within their own ranks. Matters are here further complicated in this respect: If the social contract is "interest-constraining" and so significant from the point of view of rational cooperation, then it appears that it would be beyond the reach of rational egoists. If, alternatively, it requires no restraint whatsoever on maximizing activity, then it is questionable whether effecting such a contract could in any way further the ends of natural-state individuals, and also questionable whether such individuals are in a real PD in the first place.

Suppose on the other hand that state-of-nature dwellers are in some sort of situation - possibly some variety of iterated PD - in which cooperation is rational.

Then Hobbes' views on cooperation become, I think, far less engaging. On this alternative, Hobbes could profitably be interpreted as recommending how less-than-rational persons like "real world" people could be brought to see, maybe with the help of a benevolent sovereign, that it is in fact in their long-term interests to cooperate, cooperation being conducive to "commodious living."

I suggest that the correct interpretation of Hobbes' project in Leviathan and the prospects of its success depends on a resolution of the symmetry enigma. Are state-of-nature habitants really in a PD as Hobbes' remarks on first-parties suggest, or are they in a situation in which cooperation is rational, as his comments on second-parties intimate? I consider a view on this issue that has recently been defended by Gregory Kavka. I argue that it is not cogent.

2.2 Hobbes on the Morality of Contract-keeping

Persons in Hobbes' state of nature have as their primary end their own self-preservation and their individual well-being.[3] The ought-principles - the laws of nature, Hobbes tells us, prescribe actions that these persons must perform as means to the fulfilment of their primary goal.

> A LAW OF NATURE,....., is 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 taketh away the means of preserving the same; and to omit that, by which he thinketh it may be best preserved. (MW,3,14,116-117)

These precepts are, in fact,

> conclusions or theorems [of correct reasoning] concerning what conduceth to the conservation and defence of [men]. (MW,3,14,147)

What may be called the "logical form" of these laws, as both Kavka[4] and Hampton[5] indicate, is this:

LF: One ought to do some action, a, provided others are doing so as well.

There is a question, of course, of how the qualifying clause of these laws is to be construed. Who are the relevant others? Is one freed from one's duty to do what is prescribed by a law's principal clause just in case a few of these others, or a substantial number of them, are not conforming? Furthermore, is one so freed when the appropriate number of relevant others are violating merely the law in question or a number of other laws, or both? For our purposes, we can safely ignore these complications.

Textual support for LF is provided, among other things, by Hobbes' statement of the first three laws, in particular the second, and a passage in which he distinguishes between laws holding "_in_ _foro_ _interno_" and those holding "_in_ _foro_ _externo_."

The first law tells us that

> 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. (MW,3,14,117)

It would be to no avail to seek peace unilaterally. Hobbes may consequently be taken to be advocating that persons seek peace by putting an end to the state of war only on condition that enough others are willing to do so as well.

The second law specifies what must be done if peace is to be attained. Each person must

> be willing, when others are so too,.....for peace, and defence of himself.....to lay down [his] right to all things; and to be contented with so much liberty against other men, as he would allow other men against himself. (MW,3,14.118)

The right in question is the "right of nature" or

> the liberty [i.e. the moral permission] each man hath, 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 judgment and reason, he shall conceive to be the aptest means thereunto. (MW,3,14,116)

So the second law avers that the moral permission attributed by Hobbes to each person in the state of nature to do whatever he believes is necessary to preserve and to enhance his life, is to be relinquished provided others are willing to do so as well. Unilateral capitulation of this right would simply result in the relevant person's making "himself a prey to others."

But how is universal restraint to be attained? One way is by agreement. Persons are to covenant to constrain their pursuit of individual self-interest; they are to agree, Hobbes might say, to be moral. Entering into this sort of agreement, however, even if necessary for being moral is clearly not sufficient: To be moral requires that it not only be rational for persons to make the interest-constraining agreement but that it also be rational for them to keep that agreement. For these reasons, Hobbes introduces what he needs - the third law which enjoins compliance with one's valid covenants.

On the face of it, the two-part structure of the third law is not obvious. That it has this logical form becomes fairly evident on considering an important passage in *Leviathan*. The passage strongly intimates that every law of nature exemplifies the logical form in question. It also suggests that each law requires concurrence with the prescriptions of its principal clause, or binds "*in foro externo*," if and only if others are concurring as well.

> The laws of nature oblige *in foro interno*; that is to say, they bind to a desire they should take place: but *in foro externo*; that is, to the putting them in act, not always. For he that

> should be modest, and tractable, and perform all he promises, in such time, and place, where no man else should do so, should but make himself a prey to others, and procure his own certain ruin, contrary to the ground of all laws of nature, which tend to nature's preservation. And again, he that having sufficient security, that others shall observe the same laws towards him, observes them not himself, seeketh not peace but war; and consequently the destruction of his nature by violence. (MW,3,15,145)

The third law, in light of this passage, is to be understood as requiring that one adhere to one's rationally made agreements if and only if the relevant others are doing so as well. In the event that they are not, then non-compliance is morally sanctioned.

It should be noted, although the issue is controversial, that the third law holds even in the state of nature. At least that's what I believe Hobbes believed. Thus, for instance, he claims that the "laws of nature are immutable and eternal," (MW,3,15,145) implying that their injunctions hold in every situation, the state of nature included. Furthermore, there is a passage that seems to lend unequivocal support to the view I am attributing to Hobbes:

> Covenants entered into by fear, in the condition of mere nature, are obligatory. For example, if I covenant to pay a ransom, or service for my life, to an enemy; I am bound by it: for it is a contract, wherein one receiveth the benefit of life; the other is to receive money, or service for it; and consequently, where no other law, as in the condition of mere nature, forbiddeth the performance, the covenant is valid. (MW,3,14,126-127)

Finally, in his reply to the Fool soon to be considered, Hobbes again seems to embrace what is here being ascribed to him.[6]

2.3 Hobbesian Stooges in a Prisoner's Dilemma

It has recently occurred to many that persons in Hobbes' state of nature are in a situation that can be represented by a PD-like matrix. As I indicated earlier, Kavka seems to think that this is so. Jean Hampton believes that

> in the state of nature people will.....tend to mistakenly treat PD games as one-time occurrences rather than as members of a series. (H,80-81)

To support their views, such authors interpret Hobbes as giving a description of the state of nature and its inhabitants that makes it possible for us to understand the situation in which the latter find themselves as a PD. One such description is this:[7]

"Hobbesian individuals," have four features that are particularly important for our purposes: (i) They are rational egoists, unrelentingly seeking to maximize the satisfaction of their desires.[8] (ii) As noted, they have as their predominant desire their self-preservation and their individual well-being. (iii) Each has desires that conflict with the desires of others so that the satisfaction of one person's desires frequently interferes with or precludes the satisfaction of another person's desires. This results primarily from the short supply of natural resources required to satisfy their needs and wants.[9] (iv) They are roughly equal in their physical and intellectual capacities so that no one individual is advantaged over any other in his quest for resources required to ensure his well-being.[10]

In attempting to satisfy their needs, Hobbesian individuals exert their power. An individual's power is "his present means; to obtain some future apparent good." (MW,3,10,74) In so doing they inevitably come into conflict over the appropriation of the same goods as goods are in limited supply. Since there is no common power to adjudicate conflicts in the state of nature, Hobbesian individuals are left to their own wits to retain the goods already appropriated and to acquire more of the same to ensure continuance of their well-being. Each person is well-aware that every other is in the same competitive situation. It is therefore probably in an individual's interest to make a preemptive strike. In Hobbes' words:

> [T]here is no way for any man to secure himself, so reasonable, as anticipation; that is, by force or wiles, to master the persons of all men he can. (MW,3,13,111)

If we now limit ourselves to any pair of such individuals, the dilemmatic nature of the situation in which they are becomes more-or-less evident. Each either attacks or fails to do so. If both resist attack and lie low, neither gains nor loses. If both attack, neither vanquishes the other but both suffer small losses. If one attacks while the other lies low, the aggressor makes large gains but the other seriously jeopardizes her well-being. Assuming symmetric payoffs, the game-theoretic matrix depicting their situation is something like this.

	Mo	
	Attacks	Lies low
Larry		
Attacks	1,1	10,0
Lies low	0,10	9,9

Figure 2.1 Two Stooges in the State of Nature

The numbers in the matrix represent expected utilities with Larry's utilities ranked first. (Numbers in forthcoming matrices in this chapter, unless otherwise specified, are to be similarly interpreted.) In such a situation, no matter what the other does, an individual does best if she attacks. Since positions are symmetric, the strategy of attack is dominant for either player. Consider, now, the entire state-of-nature population. With respect to it, universal restraint is better for all than universal agression, but each does best by attacking no matter how many others are lying low or doing otherwise. So persons in Hobbes' state of nature are in a multiparty PD, or so some have supposed.

I emphasize that I do not claim this description of the state of nature and its inhabitants is correct or even complete. But it is helpful in providing a framework to understand Hobbes' views on state-of-nature cooperation.

2.4 Hobbes on State-of-nature Cooperation

To attempt to clarify these views let's first suppose that two characters in the natural state, Larry an Mo, have made a defense covenant, which we assume is a covenant of mutual trust, that each will come to the aid of the other in case the person or property of that other is attacked by some third party. Recall that a covenant of mutual trust for Hobbes is an agreement calling for the parties to perform one after the other at some time after the agreement has been consummated.[11] Is it rational for either to keep his agreement? Hobbes first discusses whether it is so for the party who has to perform first:

> If a covenant be made, wherein neither of the parties perform presently, but trust one another;

> in the condition of mere nature, which is a condition of war of every man against every man, upon any reasonable suspicion, it is void: but if there be a common power set over them both, with right and force sufficient to compel performance, it is not void. For he that performeth 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; which in the condition of mere nature, where all men are equal, and judges of the justness of their own fears, cannot possibly be supposed. And therefore he which performeth first, does but betray himself to his enemy; contrary to the right, he can never abandon, of defending his life, and means of living. (MW,3,14,124-125)

Hobbes seems to be claiming that with no power to enforce agreements, and so with no assurance that the other party will comply, it would not be rational for first-party members in the state of nature to comply.

Assume, plausibly, that Larry and Mo each ranks his preferences for outcomes from most preferred to least in this way: (1) Unilateral violation; (2) mutual compliance; (3) mutual violation; (4) unilateral compliance. Then Larry and Mo are in a PD that can be represented by matrix 1. (See page 36 below. Numbers in parentheses depict preference strengths, higher numbers indicating stronger preferences. Numbers not enclosed within parentheses portray absolute values of outcomes for agents.) Call situations representable by a matrix of this kind in which outcomes are valued by either party in the manner shown, "matrix 1" situations. Since noncompliance is dominant for

	Mo	
Larry	Keeps the agreement	Breaks the agreement
keeps the agreement	6,6 (2,2)	-2,10 (4,1)
Breaks the agreement	10,-2 (1,4)	0,0 (3,3)

Figure 2.2 Matrix 1

either party in a matrix 1 situation, if Larry and Mo are in fact in this sort of situation in the state of nature, then Hobbes' conclusion about first-party's noncompliance makes eminent sense.

Hobbes enunciates his views on compliance with respect to second-parties in responding to the objection of the atheistic Fool.

> The fool hath said in his heart, there is no such thing as justice; and sometimes also with his tongue; seriously alleging, that every man's conservation, and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also to make, or not make; 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, and the observance of them justice: but he questioneth, whether injustice,.....may not sometimes stand with that reason, which dictateth to every man his own good; (MW,3,15,132)

I take the Fool to be objecting to a thesis about the third law he ascribes to Hobbes, that it is sometimes rational to keep one's covenants even when the expected costs of doing so exceed the expected gains, as in a matrix 1 situation. Hobbes replies in this fashion:

> For the question is not of promises mutual, where there is no security of performance on either side; as when there is no civil power erected over the parties promising; for such promises are no covenants: but either where one of the parties has performed already; or where there is a power to make him perform; there is the question whether it be against reason, that is, against the benefit of the other to perform, or not. And I say it is not against reason. For the manifestation whereof, we are to consider; first, that when a man doth a thing, which notwithstanding any thing can be foreseen, and reckoned on, tendeth to his own destruction, howsoever some accident which he could not expect, arriving may turn it to his benefit; yet such events do not make it reasonably or wisely done. Secondly, that in a condition of war, wherein every man to every man, for want of a common power to keep them all in awe, is an enemy, there is no man who can hope by his own strength, or wit, to defend himself from destruction, without the help of confederates; where every one expects the same defence by the confederation, that any one else does: and therefore he which declares he thinks it reason to deceive those that help him, can in reason expect no other means of safety, than what can be had from his own single power. He therefore that breaketh his covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any society, that unite themselves for peace and defence, but by the error of them that receive him; nor when he is received, be retained in it, without seeing the danger of their error; which errors a man cannot reasonably reckon upon as the means of his security: and therefore if he be left, or cast out of society, he perisheth; and if he live in society, it is by the errors of other men, which he could not foresee, nor reckon upon; and

> consequently against the reason of his preservation; and so, as all men that contribute not to his destruction, forbear him only out of ignorance of what is good for themselves. (MW,3,15,133-134)

Hobbes' idea seems to be this: A second-party could not expect violation of the agreement to be advantageous. He could not because a consequence of violation that is both highly probable and substantially costly is exclusion from civil society - a cooperative enterprise of great potential benefit. Since the alternative to entering civil society - remaining in a perpetual war of each against each - is so bleak, a second-party could ill afford to cultivate a reputation of untrustworthiness by breaking a state-of-nature covenant. For founders of the society to come, the passage suggests, would exclude from their joint enterprise such untrustworthy others.[12]

If this is a correct construal of Hobbes' reply to the Fool, then it seems that persons in the state of nature couldn't be in a matrix 1 situation. Matrix 1 fails to take into account the serious consequence of violation just described. When the expected costs of this consequence are tallied, perhaps the correct matrix representing the situation in which individuals in the state of nature really are is the one illustrated in Figure 2.3 on the next page. In a "matrix 2" situation, compliance is dominant. Hobbes would then be right in his insistence on the

	Mo	
	Keeps the agreement	Breaks the agreement
Larry		
keeps the agreement	6,6	-2,5
Breaks the agreement	5,-2	-5,-5

Figure 2.3 Matrix 2

rationality of second-party compliance.

Hobbes' reply to the Fool is curious. It's so because as construed it isn't <u>really</u> a reply: Cooperation in a matrix 2 situation is rational. The Fool needn't deny this. But surely we may charitably interpret him as demanding to know why cooperation is not against reason in matrix 1 sorts of cases. Perhaps Hobbes introduces the Fool as an expedient to clarify his views on cooperation: He introduces the Fool, has the Fool attribute to him a thesis about cooperation, and then gives a response which shows he does not ultimately endorse the thesis in question. Reason cannot prescribe keeping the kinds of agreement that would enable persons to escape a PD. Hobbes, I believe, clearly saw this.

Be that as it may, we now have a vexing problem. If state-of-nature individuals really are in a matrix 2 situation, cooperation is rational for either party, and it is rational for each no matter what the other is doing. The rationality of the second-party's cooperating does not, as Hobbes seems to believe, depend on whether the first party has performed. Hobbes might well be right, given the views of morality he endorses, to hold that it would be morally permissible for a second-party not to perform if, for some reason, the relevant first-party did not. In such a case, the requirement of the qualifying clause of the third law would presumably not be satisfied, thus freeing the second member from her duty. But it could not be rational for a second-party to fail to cooperate, any more than it could be for a first-party to do so. Notice, further, another potential worry if Hobbesian individuals are in a matrix 2 situation. Suppose the explanation of warfare in the state of nature requires that it not be rational for Hobbesian individuals to cooperate in that situation. If cooperation is rational why war with one another? Then once again it would seem that Hobbesian individuals couldn't be in a matrix 2 situation. Hobbes, though, is pretty emphatic that cooperation is rational for second-party members in cases where first-party members have already performed, even though it is not rational for first-party members to have performed.

2.5 Unsuccessful Attempts to Avoid Symmetry

What's to be done? Are we required to ascribe to Hobbes the position that sometimes parties in the state of nature are in matrix 1 situations whereas at other times they are in matrix 2 situations? This would only engender further problems. For one thing, suppose we accept what I think is Hobbes' explanation of the irrationality of second-parties failing to cooperate - they risk nonadmittance into the society to come. If nonadmittance is really a cost, and a substantial one at that, then it is difficult to see how persons in the state of nature could be in a matrix 1 situation. Such a matrix incorrectly reflects the alleged costs and benefits. For another thing, even if there were matrix 1 situations in the state of nature, Hobbes would have to maintain that it would be rational for second-parties in _these_ situations to fail to cooperate, despite irrational first-parties having done the cooperative thing. He does not say so.

A second, and perhaps more promising possibility, is this: It's clear, or so it may be thought, that you cannot represent the costs and benefits for each party as being the same as those of each other. It should, in consequence, be pretty obvious that one symmetrical matrix will not suffice to depict the payoffs that first-party and second-party players may reasonably expect. Rather, there are two decision problems and so two matrices that need to be considered separately: One matrix is required to focus on the choice question: going first/violating first; a second different matrix is required to focus on the choice question: going second when someone has gone first/violating at this stage. Maybe the first matrix will be a PD, and the second won't be. In this way, it might be possible to account for Hobbes' views on state-of-nature cooperation.

Let's explore this possibility more closely, beginning with the choice problem: going first/violating first. Larry rears horses and his neighbor, Mo, farms cattle. Suppose they have made a defense pact agreement; each has pledged to aid the other in case that other is attacked by a third party. One fine morning, as each is brewing coffee on his porch, each spots a band of notorious cattle bandits fast approaching Mo's ranch. Their intention is clear - a raid on Mo's establishment is imminent. What should Larry do? Should he come to the aid of Mo as called upon by the defense pact? Let's distinguish a few possible scenarios.

First, assume each knows that the other is rational. Assume, in addition, that each is transparent to the other: each is aware not only of the rational disposition of the other but also of how that other will act on each occasion of choice. Assume, also, that "the force of transparency extends through time:" Suppose Larry is first-party player. Each is so transparent to the other that prior to performing, Larry then knows, or at least has very good reason to believe, that having performed, Mo will keep his part of the agreement. If he performs, he now has assurance that Mo will perform in the future. Each player's true colors reveal themselves across time. Assume, further, that transparency is irreversible. Under these conditions Mo can count on Larry coming to his aid - it's straightforwardly rational for Larry to do so: Since each is transparent, neither can dissemble. Larry, for example, cannot enter the agreement with the intention of not later performing: Mo would know of Larry's intention, and would refuse to negotiate the agreement in the first place. Larry, in turn, would know this. Under conditions of transparency it seems that each would come to the aid of the other if and only if the other were to do the same. If Larry were to aid Mo, he could expect Mo to reciprocate. If he were to stay put, he could expect Mo to do the same. Transparency, it appears, forges a link between the actions of the two players. What each does seems at least partly to "determine" what the other does. With their actions interdependent in this way, it would be forthrightly rational for each to honor the terms of the defense pact. If all this is correct, then Larry's decision matrix in this case does not seem to be a PD.

The assumption of transparency is unrealistic. So let's next assume that Larry and Mo are "opaque." What should Larry do? If he adheres to the agreement, he can expect Mo to renege: It seems Mo would do better to renege than to reciprocate in the event that Larry's farm were raided. If Larry reneges, he can once again expect Mo to renege. It would be foolish for Mo to comply unilaterally. Whatever he does, Larry can expect Mo to renege. Knowing this, it seems it would be rational for Larry to renege as well. Perhaps in this second case, Larry is in a genuine PD, and that the game in which the two find themselves, a game that can only be completed in two consecutive moves - first player plays first/second player plays second, is itself PD-like, as diagram 1 on the next page illustrates.

Finally, consider a third case. Again, assume opaque agents. But this time assume also that forays of the kind described are commonplace in the state of nature, so that both can expect to find themselves in a similar situation many times over. Assume neither knows how many times over. Opacity, as the second case suggests, lends credence to the view that Larry is in a PD, and that each game consisting of two consecutive moves, is itself PD-like. The third case differs from the first two in that each game is only one in a series of such games. We're dealing, in other words, with some variety of iterated PD. Assume (as both

Larry: Moves First	Mo: Moves Second	Preferences for outcomes after Mo’s second move	
		Larry	Mo
Keeps the defense pact	Keeps the pact	5	5
	Breaks the pact	0	10
Breaks the defense pact	Keeps the pact	10	0
	Breaks the pact	1	1

The numbers represent expected utilities, higher numbers indicating stronger preferences.

Figure 2.4 Diagram 1

Kavka and Hampton do) that in these kinds of situation, the rational thing to do is to abide by the defense pact agreement. Then Larry will help Mo fight the bandits.

These results are not very encouraging, at least not from the perspective of the defender of Hobbes: First, it is only in the second case that it is rational for first-party Larry to renege. In the other two cases, the rational thing for Larry to do is to abide by the defense pact. Second, and more importantly, symmetry is all pervasive: Consider the second choice question. Assuming Larry has abided by the agreement, should Mo come to his aid, this time let's suppose, to fight off some horse thieves? In a one-shot transparent situation, the answer seems to be that he should, for reasons analogous to those already considered: Rational Larry would not have abided had he known or had he had good reason to believe that Mo would fail to abide in the future. Since, by assumption, Larry did abide he must have had good reason to believe that Mo would also abide. We can therefore conclude that Mo will abide as well.

Now consider an opaque, one-shot situation analogous to the situation in which players' preferences are represented by diagram 1. As the diagram clearly indicates, second-party Mo, just like first-party Larry, would do best for himself if he were to renege. Of course, if Larry did for some reason or other abide in this situation, then he failed to act rationally.

Finally, assume players are in an iterated dilemma, each game of which terminates only after two consecutive moves - Larry going first, Mo going second. In this case, I think, it would also be rational for second-party Mo to adhere to the defense pact.

The twin matrix manoeuvre won't save Hobbes.

A third possibility is to take Hobbes to be distinguishing between two states of nature, an ideal state with all agents fully rational, and a "real world" state with an admixture of agents, some rational, others irrational. If all agents in the ideal state are in a PD, then rational cooperation remains a dream. With respect to the real world state, the question is whether cooperation is rational given the presence of the two sorts of agents. If rational Spock were transported from rational man's heaven (hell, perhaps?) to the real world state, what should he do? I will shortly consider a real world situation of this type.[13] For now it suffices to note that if first-parties in a real world natural state are in a matrix 1 situation, and second-parties in a matrix 2 situation, then the presence of irrational parties makes no difference whatsoever to the rationality of cooperation in these situations. Suppose irrational first-party Larry complies. Hobbes advises second-party Mo to reciprocate on pain of being excluded from civil society. The old problem simply recurs: If founding members of the society to come don't take kindly to second-party violators, shouldn't they also frown upon first-party violators, thereby calling into question whether first-parties really are in a matrix 1 situation? Aren't untrustworthy first-parties just as bad as second party exploiters?

As I see it, then, an interpreter of Hobbes is forced to come to terms with the "symmetry" enigma: If persons in the state of nature are in some sort of PD, then Hobbes' views on first-party covenant-keeping seem reasonable. These persons, if in a PD, would do best for themselves by reneging. If, on the other hand, they are in a situation in which doing the "agreement required thing" is dominant, then his remarks on second-party covenant-keeping also seem well-founded. The trouble is that first-parties and second parties seem to be in the same sort of situation. Each seems to stand to gain or to loose whatever each other seems to stand to gain or to loose by cooperating. The payoffs to each appear to be symmetric. Yet, their positions in the state of nature must not be symmetric if Hobbes' views on the rationality of natural-state cooperation are to be believed. In light of the symmetry problem, how are these views of Hobbes' to be accounted for? Let's consider Professor Kavka's reply.

2.6 Kavka on Symmetry

Kavka commences his chapter on cooperation by introducing the notion of a defensive cooperative. A defensive cooperative

> consists of an explicit agreement, among some group of individuals in the state of nature, that each will come to the aid of any other in the group

> whose person or property is attacked or threatened. (K,127)

Each individual member of such a group must grapple with a problem of compliance. Is it rational to act in accord with the terms of the agreement when called upon to do so? In discussing this problem, Kavka tells us that we can represent a typical state-of-nature agreement by a PD matrix analogous to matrix 1. (K,138) Since in such a situation violation is dominant, it would not be rational for first-party members to comply. Indeed, he claims

> Hobbes apparently accepts this reasoning, agreeing (with the Fool) that the first party should not perform. We may say that Hobbes approves of _defensive_ violations of state-of-nature agreements, that is, violations motivated by the desire to avoid being taken advantage of, to avoid becoming a unilateral complier. (K,139)

What about compliance on the part of second-party members given that "some naive first party" (K,139) has performed? Should he "_offensively_ violate," thereby securing his maximum payoff? The Fool, Kavka tells us, thinks so, but Hobbes refuses to endorse offensive violations. (K,139) He refuses, according to Kavka, for reasons similar to those I suggested earlier:

> The Fool, looking only at the immediate payoffs available, sees the situation depicted in [matrix 1]. He therefore proclaims breach.....as the most reasonable response.....Hobbes, however, views the second party's response as having long-term effects on that party's prospects of future cooperation

> with others. Specifically,.....that a particularly vital sort of cooperation, inclusion in defense confederations will be denied known offensive violators of agreements, since other members will not trust them to keep their promises of aid.....The danger of being excluded from future cooperative defense arrangements here.....transforms a situation in which noncooperation is a dominant strategy into one in which cooperation is dominant. (K,140)

If this is the correct explanation of Hobbes' views, Kavka plainly faces the symmetry problem:

> A[n].....objection to Hobbes's reply to the Fool, raised by Jean Hampton, is more serious.....[I]f Hobbes is right that it is rational for second parties to keep state-of-nature agreements, then it must be rational for first parties to keep them as well. The argument is that the latter, following the logic of Hobbes's argument, will not expect rational second parties to cheat them, and that even if they are cheated, first parties can expect to gain more than compensating advantages by proving themselves trustworthy partners in future defense cooperatives or commonwealths.[14] (K,147-148)

Although the problem to which this passage alerts us is not quite the symmetry problem, it's very close. Let's consider what Kavka has to say about this objection. First, he proposes that

> The clear response to the first part of Hampton's objection, suggested by passages in chapter 14 of Leviathan, is that first parties to state-of-nature contracts cannot count on second parties acting rationally. It is in the long-term interest of second parties to comply with state-of-nature agreements, as the reply to the Fool shows, but their short-term interests (and strong passions, like greed) dictate a different course.

> Given the frequent tendency for people to be carried away by short-term interests, it is not safe to be a first-party complier with a state-of-nature agreement because of the substantial risk of irrational noncompliance by the second party. (K,148,153)

Hampton objects that if it is rational for second-parties to perform, first-parties would know this, and would therefore expect compliance from these parties. But since each does better if everyone complies, first-parties would themselves do best by complying. Hampton should insist that mutual compliance has the result that there is probably no conflict in the state of nature. Kavka rejoins that although it is rational for a second-party to comply if a first-party has done so, most second-parties because of shortsightedness will not comply. First-parties, cognizant of the irrationality of many or perhaps even most second-parties cannot, in consequence, expect second-parties to comply. So first-parties will refrain from complying as well.

Kavka's response, it seems to me, is not very compelling: Suppose most Hobbesian individuals in the state of nature are shortsighted. Then I think we can assume this of both first-parties and second-parties. It would be arbitrary to suppose that only second-parties are shortsighted and that only firsts are rationally perfect. Now let's distinguish two cases. (i) Suppose Hobbesians can identify who's rational and who isn't. Kavka admits, or would admit, that it is rational for a rational second-party to perform if a first-party has done so. A rational first-party, knowing this, would expect a rational second-party to perform. So if he is interacting with a rational second, he would comply. In this instance, both would comply. There would then probably be no conflict between these two. But a rational first cannot expect an irrational second to comply. So when interacting with such an agent, a rational first would fail to comply. The irrational second-party would probably not comply either. So neither would comply and neither would get out of the state of nature.

(ii) Consider the second case. Again, assume a "real world" natural state, with an admixture of agents, some rational, others irrational. Suppose Hobbesians cannot tell who's rational and who isn't. Suppose it is generally known that in this natural state there is a "frequent tendency for people to be carried away by short-term interests and immediate passions." So even if such irrational parties were in an iterated PD in which cooperation is rational, they would be swayed by their short-term interests and would fail to cooperate in each round. Suppose, since there are many irrational people in this real world state, that the probability that one's prospective interactive partners are irrational is high.

Suppose, finally, that if a person abides by a defense pact agreement while the others do not, the outcome for the person is bleak. Call such a state of nature an "adventurous" one. In an adventurous state compliance is irrational - I think - for a first-party. Now suppose an irrational first has performed. What should the relevant second-party do? It would be rational for a second-party to comply if Kavka's claim that a high cost of violation on the part of second-parties, exclusion from defense confederations, is correct. But would such confederations exist in an adventurous state of nature? That's a moot issue: Kavka informs us that many such confederacies are formed by tacit agreement or explicit agreement. (K, 126) If such confederacies exist, then, it must be rational to make and honor "defense confederacy agreements." (See pages 55 and 56 below.) But even if it were rational to do these things, since many in an adventurous state of nature are irrational, these irrational persons would fail to appreciate the advantages to be had by upholding such defense covenants. So it's likely that few, if any, defense confederacies would exist. Exclusion from such confederacies could then not be a substantial cost second-parties in an adventurous natural state could anticipate were they to renege.[15] So in this second case, even rational second-parties would not comply.

Kavka's second response is far more engaging because it meets the symmetry problem head-on. He suggests that first-party and second-party members suffer differential costs when they violate their covenants.

> [D]efense group members are likely to react *differently* to first-party and second-party violators of state-of-nature agreements. In particular, the defender of Hobbes's view must believe that defense-group members will regard first-party violators as more reliable and desirable partners than second-party violators. (K,149-150)

Kavka then adduces what he takes to be some "plausible grounds for this belief." Assume he is right about this. It is still not quite clear what matrix (or matrices) correctly depicts the situation he has in mind. It probably wouldn't, for instance, be matrix 3.

	Mo	
Larry	Keeps the agreement	Breaks the agreement
keeps the agreement	10,10	-2,8
Breaks the agreement	8,-2	5,5

Figure 2.5 Matrix 3

In a matrix 3 situation neither party, if each is a rational egoist, could decide on the rational course of action without further information. Supply the required information and the matrix no longer remains the same. If no such information is available, each might elect to maximin - to do whatever will result in the best worst outcome. If they did maximin, both in this situation would do the noncooperative thing. None of this, though, shows that Kavka's "differential cost" reply is problematic. But I think it is:

The term 'defense group,' Kavka tells us, is "intended to encompass both commonwealths and state-of-nature defensive groupings." (K,149) In the present context, the term presumably refers to the latter. He further informs us that defense groups in the state of nature may originate in sundry ways - "by family ties, conquest and submission, tacit agreement, explicit agreement, or some combination thereof." (K,126) Defense alliances formed by conquest and submission, we are told, are probably unreliable, unstable (K,108-119-120) and arguably their existence in the state of nature is not "even highly likely." (K,141) Family ties may prove tenous, and even if they are not, we may simply treat each family as an individual and inquire whether cooperation amongst such individuals is rational. Ignoring the last category, this leaves the two of interest to us - defense groups formed by tacit or explicit agreement. Among rational individuals, however, no such "defensive cooperative groups" could exist unless it were rational both to enter into and keep a "defensive cooperative group agreement." After all, that's how such groups are formed.

Now reconsider Kavka's differential cost reply. As Kavka is himself aware, an objection to this kind of reply is that it assumes the viability of defensive cooperative groups in the state of nature. (K,141) What has not been stressed enough, if at all, is this: If the reply assumes the existence of such defensive cooperative groups, it must also assume, as has just been explained, the rationality of keeping defensive cooperative group agreements: If it is irrational to keep such defense covenants, rational agents could not expect rational parties to keep them; no such covenants would be rationally honored, so no defense confederacies that depend upon the rationality of keeping such covenants would exist. Clearly, however, the reply is not entitled to this assumption since one of its objectives is to show that agreement-keeping, at least by second-parties, is rational in the state of nature. In addition, in a real world adventurous natural state as we saw, it seems irrational - irrational for any party - to make and to keep a defense pact agreement.

In the interests of accuracy, let me add that Kavka considers this kind of objection - the objection concerning the viability of defensive cooperative groups - in a slightly different context, namely, in his exposition that second-party violators are really in a matrix 2 rather than in a matrix 1 sort of situation. Recall that he there appealed to the idea that a serious cost of violation is exclusion from defense confederations in the state of nature. (See pages 49 and 50 above.) However, the same kind of objection, as we have seen, can be directed against the differential cost reply.

Kavka has a response. Again the response is meant to parry the objection raised in connection with accounting for the rationality of second-party adherences to state-of-nature agreements. But presumably, Kavka would offer a similar response as far as present matters are concerned. So let me first cite the passage containing Kavka's response to the former objection and then indicate how it might be adapted to explain the differential costs incurred by first-party and second-party violators of state-of-nature agreements. The relevant passage is this:

> [S]uppose one.....doubts that state-of-nature defense pacts are viable and can provide substantial security benefits. In that case one may simply read 'society' in.....the reply to the Fool as standing for civil society, or the commonwealth. Then Hobbes may be viewed as pointing out that founders, or preserving members, of a commonwealth will not accept unreliable

> parties, such as offensive violators of agreements, as members. Second-party state-of-nature agreement violators are thus not simply risking future membership in shaky state-of-nature defense cooperatives, they are risking their chances of permanent escape from the state of nature via the only effective mechanism thereof, membership in a commonwealth. (K,141)

Adapted to present concerns, rather than claim "defense group members are likely to react differently to first-party and second-party violators of state-of-nature agreements," Kavka might claim "founding members of the society to come are likely to react differently to first-party and second party violators."

But once again the response, I believe, is questionable. Assume all persons in the state of nature are rational. It appears that even these fully rational agents should be sceptical about their prospects of entering civil society. If this is true - if Hobbesians are unable to leave the state of nature, then these natives cannot really be "founders, or preserving members, of a commonwealth" to come. Exclusion by any state-of-nature founder from any future commonwealth could not then be any threat to anyone: Suppose escaping the state of nature, as Hobbes believes, does require the instituiton of a sovereign. Doing this, in turn, minimally seems to require that it be rational for each party both to make and to keep an agreement to surrender his right of nature to the sovereign and to obey the sovereign's commands. But the sort of agreement (or agreements) in question here seems analogous to defensive cooperative group agreements. Both kinds of agreement are mutually advantageous though requiring self-sacrifice. Although it is collectively rational to keep such agreements, it is not individually rational to do so. Such agreements are "interest-constraining." So just as a defensive cooperative group couldn't exist unless it were rational for Hobbesian individuals to make and keep a defensive cooperative group agreement, so it seems a sovereign could not be instituted and, in consequence, a commonwealth couldn't come into existence, unless it were rational for Hobbesian individuals to make and to keep a "sovereign-instituting" agreement.[16]

2.7 More Kavka on Symmetry

It's worth, I think, briefly diverting at this point to note the following: Ignoring certain passages in Leviathan, suppose Hobbesian individuals really are in a matrix 1 PD in the state of nature. Then we can further suppose that these individuals, being rational, realize that the goods they require to satisfy their needs are in short supply in the absence of cooperation, but would be in plentiful supply if they could only restrain their pursuit of self-interest and enter into fruitful cooperative ventures. A a possible way to attain universal restraint is by the making and the keeping of an "interest-constraining" agreement. But if the requisite sort of agreement really is interest-constraining, then the very rationality of these rational egoists would inhibit them from keeping such an agreement. Perhaps Hobbes saw this, and in light of seeing this, sought a "political" solution: Institute a sovereign that penalizes breaches of agreements so that it is no longer advantageous to violate one's covenants. The sovereign's job, in effect, would be to "transform" the original matrix 1 situation into something like a matrix 2 situation. This is one way of understanding what Hobbes is up to in Leviathan. On this understanding, the symmetry problem vanishes: Each state-of-nature individual is in the same matrix 1 PD.

The discussion in the paragraph preceeding this diversion, however, seems to cast doubt on the tenability of even this political solution: If the very institution of a sovereign necessitates the making and the keeping of an interest-constraining agreement, then rational egoists will be unable to do what is required by the "solution" to escape the natural state.

Not to be defeated yet, Kavka anticipates and responds to this sort of worry.

> A final problem for Hobbesian social contract theory.....centers on the question of why the

> parties should comply with the social contract once it is made, why they should obey the orders of the government created by it. After all, the social contract is a state-of-nature agreement, and Hobbes says you should not keep these (as a first party) since the other party or parties may not follow suit.....
>
> The general nature of our solution to the first party compliance problem [is this:] The social contract is different from other state-of-nature agreements in that it promises, if successful, to remove the parties from the state of nature. Thus, each has a tremendous amount to gain by its success. This means that the risk of being a unilateral complier is worth running if it attaches to a reasonable chance of mutual compliance. Further, since others also obviously have much to gain by the effectiveness of government, one runs little risk of being a lone complier - others will be only too glad to make the arrangement work once you have set the example.....[Y]ou are almost certainly better off under [the government's] protection than you would be returning to the state of nature by refusing to comply. (K,243-244)

With respect to the sovereign-instituting agreement or the social contract, either first-parties are in a matrix 1 PD, or they are not. The passage suggests that it is individually rational for first-parties to comply with such an agreement if there is a "reasonable chance of mutual compliance." But in a matrix 1 situation even if each has a "tremendous amount to gain by its success," keeping an interest-constraining agreement is not rational. It would seem that in connection with the social contract, first-parties are not in a matrix 1 PD. The situation of these first-parties is not analogous to the situation in which first-parties of other state-of-nature covenants find themselves. So let's suppose that in relation to the social contract no symmetry problem exists, first-parties just like second-parties are in fact in a situation - perhaps a matrix 2 situation - in which compliance is rational. Then there are two problems:

First, the agreement in question - the social contract - despite strong appearances to the contrary, is not interest-constraining. Such an agreement, it seems, imposes no restraint on an individual's exercising her unlimited right of nature. As David Gauthier explains

> The role of so-called 'moral' conventions [would] then be not to constrain our behavior, but rather to enable us to coordinate that behavior to maximal advantage, effecting, like the perfectly competitive market, the harmony of non-tuisms.[17]

The second problem is even more serious: The social contract requires at least that each agree to surrender his right of nature to, and to authorize all the actions of, a sovereign-elect. (See, for example, K,180-181) Kavka suggests that the reason why compliance by first-parties (and presumably also by second-parties) with the social contract is rational is that it increases their chances of membership in a viable commonwealth:

> We may see this solution as an extension of Hobbes's reply to the Fool. There it was suggested that compliance with a state-of-nature contract is rational if it increases your chances of later membership in a viable commonwealth. But surely first compliance with the orders of a government

> newly founded by agreement increases your chances of membership in a viable commonwealth. All problems of commonwealth formation have been solved at this point, save possibly one: firmly establishing general expectations of obedience to the government. (K,244)

I do not here take Kavka to be responding to the unfounded concern about whether there is a compliance problem in connection with the edicts of a fully "empowered" sovereign: Suppose the sovereign is somehow established. Now he's in charge, and he tells me to pay taxes. If I'm rational, I'll pay - because he can inflict very serious penalties on me. Rather, Kavka seems to be responding to the legitimate worry about whether setting up the sovereign - abiding by the agreement that tells you to do whatever is required to establish the sovereign - is rational. But what Kavka suggests about why persons in the state of nature have reason to believe they can escape that state by entering a viable commonwealth seems misguided: Kavka apparently assumes that first-parties and second-parties will incur severe costs in the event that they fail to abide by the social contract, costs associated with non-admittance into the commonwealth of the future. This assumption presupposes that such a commonwealth is viable. This, in turn, assumes that even if Hobbesian individuals are able to solve the problem of who among them is to become sovereign,[18] there is no compliance problem in connection with, for example, agreeing to surrender one's right of nature to the sovereign-elect; it is simply straightforwardly rational to do so. This is a strong and controversial assumption, one I believe Kavka has not sustained. It is at least clearly an assumption that cannot be defended by supposing that compliance by first-parties and second-parties with the social contract is rational, since it improves their chances of membership in a viable commonwealth. It cannot be so defended because this reason implicitly assumes that compliance is rational: no commonwealth is viable, because no sovereign would exist, if it were not rational to surrender one's right to, or to authorize all the actions of, a sovereign-elect.

To recapitulate briefly: Kavka suggests that the symmetry problem is to be solved by noting that first-parties and second-parties incur differential costs when they renege on their state-of-nature agreements. This is because founding members of the society to come will react differently to first-party and to second-party violators. It is obvious that these founding members could be none other than Hobbesian individuals themselves. If these individuals, or at least a portion of them, are to react differently to these violations, they must have reason to believe that a commonwealth of the future is in fact viable. They must, for instance, have reason to believe that the inauguration of a sovereign involves no PD-like problems. Kavka argues that there is no compliance worry in relation to the social contract. He suggests that it is forthrightly rational for first-parties, and indeed for second-parties, to abide by the social contract. Doing so increases your chances of membership, he suggests, in a viable commonwealth. But this reason for compliance assumes that compliance with the social contract is rational: With no such compliance, no sovereign and so no commonwealth could come into existence. The differential cost reply, then, at least as I understand it, ultimately assumes something that is crucial to what it is meant to establish - that compliance with natural-state agreements, on the part of second-parties, is rational.

2.8 Conclusions

A variety of states of nature are possible. Some may contain defensive groups that are stable. In such states of nature, Kavka's differential cost reply may well resolve the symmetry problem. In other states of nature no such groups might exist. In yet others with extant defensive groups, the groups might be unstable and short-lived. In real world states of nature where defensive groups can only come into existence through the making and keeping of defensive group agreements or agreements relevantly similar to them, Kavka's differential cost reply won't solve the symmetry problem. What is perhaps even more important is that in such states of nature, it appears that Hobbes' reply to the Fool that it would be rational for second-party members to keep their state-of nature agreements isn't very compelling either: Suppose both first-party and second-party violators, as I argued, have little or no reason to believe that they will be able to leave the state of nature and enter the great Leviathan. Then exclusion from state-of-nature defensive groups or from civil society cannot really be taken to be a cost of reneging on a state-of-nature agreement. If this in turn is true, then parties in the state of nature will probably not find themselves in a matrix 2 situation. They may be doomed to a matrix 1 PD. Interestingly enough, there are worries even if they do find themselves in a matrix 2 situation in which exclusion from civil society is a real cost. First, it may then well be misleading, as Farrell[19] and Gauthier[20] have emphasized, to read Hobbes as attempting to "ground," morality on the basis of self-interest: The kinds of agreement that it is rational to make and keep in matrix 2 situations, as we saw in section 2.4, do not require constraining one's pursuit of self-interest. It may then be charged that such agreements do not require persons to comply with moral requirements since <u>moral</u> requirements mandate a restraint on maximizing activity. Second, we have to reconsider the matrix for going first. If it is rational to go second, then the matrix for going first probably does not depict a PD situation, and there probably is no conflict in the state of nature.

Maybe a reconstrual of Hobbes' account of state-of-nature cooperation will make possible a resolution of the symmetry problem. Although she does not directly address this problem, Jean Hampton offers us an account of natural-state cooperation that is distinctly different from what has been considered so far. She also has many interesting things to say about sovereign institution. It is to these views that I now turn.

Notes

1. Thomas Hobbes, Leviathan (volume 3) in The English Works of Thomas Hobbes, edited by W. Molesworth, London: John Bohn, 1839. Forthcoming citations are as follows: (MW, volume number, chapter number, page number(s)). References to other writings of Hobbes are to appropriate volumes in the same work by Molesworth.

2. David Braybrooke's paper, "The Insoluble Problem of the Social Contract," in Paradoxes of Rationality and Cooperation eds., R. Campbell and L. Sowden (1985), Vancouver: The University of British Columbia Press, 277-306, has a nice discussion of this issue.

3. See, for example, (MW,3,13,116).

4. Gregory S. Kavka, Hobbesian Moral And Political Theory (1986), Princeton, New Jersey: Princeton University Press, especially pp. 309-314 and 338-349. References to this work are given in this manner: (K, page number).

5. Jean Hampton, Hobbes and the Social Contract Tradition (1986), Cambridge: Cambridge University Press, pp. 89-92. References to this work are given in this way: (H, page number).

6. Controversy over the present issue is generated by the view expressed by Hobbes in passages such as this:

> But because covenants of mutual trust, where there is a fear of not performance on either part, as hath been said in the former chapter, are invalid; though the original of justice be the making of covenants; yet injustice actually there can be none, till the cause of such fear be taken away; which while men are in the natural condition of war, cannot be done. Therefore before the names of just, and unjust can have place, there must be 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; and to make good that propriety, which by mutual contract men acquire, in recompense of the universal right they abandon: and such power there is none before the erection of a commonwealth. (MW,3,15,131)

For thoughts on how what Hobbes says in this passage may be reconciled with what he says in the passage cited in the text, see Kavka, Hobbesian Moral And Political Theory, pp. 350-352 and Daniel M. Farrell, "Reason and Right in Hobbes' Leviathan," History of Philosophy Quarterly 1 (1984), 297-314, especially pp. 305-306.

7. The description that follows is suggested by Hobbes in Leviathan, Chapter 13.

8. See, for example, (MW,3,5,30), (MW,2,2,16), (MW,3,15,133).

9. See (MW,3,13,111).

10. See (MW,3,13,110)

11. Hobbes explains that to lay down a right is to place oneself under an obligation. (MW,3,14,118-119) A contract is "a mutual transfering of Right." (MW,3,14,20) A covenant is a type of contract in which "one of the contractors, may deliver the thing contracted for on his part, and leave the other to perform his part at some determinate time after" (MW,3,14,121) or in which "both parts may contract now, to perform hereafter." (MW,3,14,121) Kavka tells us that if both parties are to perform later and in sequence, we have a covenant of mutual trust. (Kavka, Hobbesian Moral And Political Theory, p. 304)

12. This construal of Hobbes' reply to the Fool has a serious shortcoming that I discuss in section 2.6 below.

13. Both Hampton and Kavka seem concerned with this type of situation.

14. A footnote at the end of the first sentence in this passage in the original says this:

> Jean Hampton, 'Hobbes, Contract, and the Wisdom of Fools' (unpublished paper presented at the University of Colorado, Boulder, 1979). See also her 'Hobbes's State of War,' Topoi 4 (March 1985): 47-60. David Zimmerman has raised a similar objection in a letter to me.

15. There is a second problem with assuming the existence of defensive cooperatives in an adventurous state of nature that I discuss below.

16. Jean Hampton argues that it is straightforwardly rational for each person in the state of nature to subjugate himself to a sovereign. I will discuss this interesting argument in the next chapter.

17. David Gauthier, "Thomas Hobbes: Moral Theorist," *The Journal of Philosophy* 76 (1979), 547-561, p. 556.

18. Gauthier's "Thomas Hobbes: Moral Theorist," p. 556.
Kavka, in fact, believes that the choice of a sovereign by rational parties in the state of nature is an impure coordination problem:

> It is primarily a coordination problem because, given the likely miseries of the state of nature (an active war of all individuals or small groups), it matters much more to each that there be a sovereign than who in particular it is. It is an impure coordination problem, because various individuals would expect to fare better under different sovereigns (e.g., each party might most prefer that he himself were sovereign). [Kavka, *Hobbesian Moral And Political Theory*, p. 185]

An impure coordination problem, Kavka says, is a problem of the sort displayed in matrix 4.

	Mo does	
Larry does	A	B
A	5,4	0,0
B	0,0	4,5

Figure 2.6 Matrix 4

Kavka seems to be suggesting that it is straightforwardly rational to subjugate yourself to a sovereign - no matter who in particular it is - than to remain in the misery-laden state of nature. Why he thinks this is so is not clear to me. Hampton has something to contribute to this issue. See footnote 17 above.

19. See Braybrooke's "The Insoluble Problem of the Social Contract," p. 309.

20. David Gauthier, "Thomas Hobbes: Moral Theorist," pp. 555-556.

CHAPTER 3

HAMPTON ON HOBBES ON STATE-OF-NATURE COOPERATION AND SOVEREIGN INSTITUTION

1.1 Introduction

In Hobbes and the Social Contract Tradition,[1] Jean Hampton contends that Hobbes' argument for absolute sovereignty - the argument that self-interested people in the state of nature would be able to institute a sovereign - fails. She tells us that

> Hobbes's argument does not fail because he cannot establish the rationality of creating an absolute sovereign, nonetheless it fails because he cannot establish, given his psychology, that men and women are able to do what is required to create a ruler satisfying his definition of an absolute sovereign. (H,197)

Hampton, as the passage reveals, does not believe the "sovereignty" argument fails for the reason that what she identifies as its first premise - that it is more in each person's interest to be subjugated to a sovereign rather than to remain in the state of nature (H,148,186) - fails. Rather, she believes it does not succeed for other reasons. She acknowledges that if the problems involved in sovereign institution required that persons be able to escape a PD - that they be able to do the cooperative thing in such a dilemmatic situation - Hobbesian individuals, being the SFMs that they are, would be unable to overcome them.

Suppose, for example, the institution of the sovereign requires each person in the state of nature to make an agreement to surrender his unlimited right of nature to the sovereign. Suppose, as I proposed in the last chapter, the agreement in question is "interest-constraining:" Each prefers mutual adherence to mutual violation, but each would do best if he were to violate unilaterally. Then although it would be collectively rational to abide, it would not be individually rational to do so. Hobbesian individuals who were fully rational would therefore, if the social contract were interest-constraining, be unable to escape the natural state and to enter civil society. But Hampton believes that although the inauguration of the sovereign requires that Hobbesian individuals be capable of resolving complex problems, none of these is a PD. (H,136,138,157,148) I think, however, Hampton's argument to the contrary notwithstanding, that even the first premise of the argument for sovereign inauguration is problematic. I believe that it is problematic because sovereign institution <u>does</u> in the end succumb to the very kind of PD-like problems Hampton claims it evades. To see that this is so, one first needs to understand her views on Hobbes on natural-state cooperation.

Accordingly, in this chapter I begin with a summary of Hampton's account of Hobbes on state-of-nature cooperation.

Although highly interesting, this account - I think - is afflicted with a serious difficulty. I then show how these views on cooperation undermine her reasoning in support of what she takes to be the first premise in Hobbes' sovereignty argument.

1.2 Hampton's Shortsightedness Account of Conflict

Hampton (like Kavka) is concerned, among other things, to give an account of the state of nature that entitles Hobbes to his conclusion that in this situation each individual is at war with every other.[2] In adducing her explanation of warfare, she expounds what she takes to be Hobbes' views on rational cooperation in the natural state.

Hampton begins by suggesting that individuals in the state of nature are not in a single play PD. Rather, they are in an iterated PD (IPD). Roughly, an IPD is a complex game, consisting of a sequence of other games, each game in the sequence being PD-like. (H,75) Second, Hampton tells us that given certain assumptions, it is rational for Hobbesian state-of-nature individuals to cooperate in an IPD.

> Even if one of the parties behaves irrationally by breaking his contractual promise in the first game (or successive games), [the] "iterated PD game".....counsels that the long-run benefits accruing from faithful contract keeping will prompt the other party to continue to keep his part of the bargain for a time in order to try to "teach" the breaching party to choose the promise-

> keeping act. The idea is to make the breaching party realize that it is in his best interest to reward rather than punish his partner's cooperative act, because otherwise he will be forcing his partner to renege in subsequent games, and a pattern of contractual breaches will be established that will deprive both of them of the benefits of future bargains. (H,75-76)

Assume so far so good.[3] Hampton then introduces a novel idea. She claims that because of shortsightedness many in the state of nature will mistakenly take themselves to be in a single-play PD rather than in an iterated one.

> The account [of conflict] would contend that many people fail to appreciate the long-term benefits of cooperation and opt instead for the short-term benefits of noncooperation, and the rest are legitimately fearful enough of this shortsightedness afflicting their partners to doubt that cooperation would have any educative effects. This worry could then force even a farsighted person to take a single-play orientation, with the result that the uncooperative action would dominate. (H,81)

We can better appreciate the shortsightedness account of conflict by briefly summarizing two other accounts of warfare that Hampton develops, and by understanding her reasons for rejecting them.

Hampton tells us that Chapter 15 of Leviathan suggests what she calls the "passions" account of conflict.[4] According to this account, various natural psssions of Hobbesian individuals like partiality, pride, revenge, and the passion for glory are responsible for warfare. (T,48-49) On this account, although cooperation among persons in the state of nature is generally rational, (T,48) people will fail to cooperate. They will fail to do so because some of their natural passions like the ones mentioned will

> disrupt many people's reasoning and cause them to behave irrationally, while the rest [will] fear this disruption and [will] (rationally) refuse to cooperate in order to avoid being exploited. (T,48)

This account of conflict, Hampton argues, engenders difficulties for Hobbes. On the one hand, if these passions are not widespread among Hobbesian individuals - if the actions of these persons are not often disrupted by them, then these passions in combination with the predominant desire for self-preservation each Hobbesian individual has, will at most generate only moderate amounts of conflict. The state of nature will then not be a state of war of each against each. In consequence, no sovereign will be needed to rescue these individuals from a disastrous plight. (T,49-50) Suppose, on the other hand, that these passions are sufficiently widespread and deep-seated to generate total war. Then the passions account seems to conflict with Hobbes' psychological postulate that the desire for self-preservation is the predominant desire of Hobbesian individuals. In addition, the frequent disruption of people's cooperative activities by these passions would make the creation and institution of a sovereign a near impossibility. (T,49-50)

An appreciation of these shortcomings of the passions account, Hampton proposes, may have moved Hobbes to present a very different explanation of conflict in Chapter 13 of Leviathan. According to this "rationality" account, conflict in the state of nature is a function of the rational pursuit of self-preservation by each individual:[5] Hobbesian individuals, who are roughly equal in their powers and capacities, compete for scarce goods in order to acquire more of the same, and to ensure that they continue to live in a "commodious" way. Each is aware that all the others are in the same competitive situation. It is therefore probably in an individual's interest to strike preemptively. In fact the account, as Hampton indicates, suggests that each person in the state of nature is in a PD: Each prefers mutual forbearance to mutual invasion, but each does best by invading unilaterally. Since invasion is dominant, each ends up invading. The result of invasion for purposes of seizing goods coupled with invasion for the sake of glory, is a condition of total war. (T,50-51)

Hampton explains that the rationality account, just like the passions account, suffers serious problems. First, if Hobbesian individuals in the natural state are really in a PD, they could not cooperate to institute a sovereign. Second, she believes that this account does not seem to be true. For persons in the state of nature are probably in an IPD in which cooperation is rational and not in a one-shot PD. (T,52)

The shortsightedness account of conflict, Hampton urges, seems to have "all the advantages of the passions and the rationality accounts, but.....none of either account's disadvantages." (T,52) The account takes seriously the idea that Hobbesian individuals are in an IPD situation. This fits well with Hobbes' psychology, since being in an IPD links cooperation as a means of achieving peace, with the predominant desire for self-preservation each Hobbesian individual has. (T,52) The account also allows for sovereign institution: If Hobbesian persons are in an IPD in which cooperation is rational, then it is not impossible for them to cooperate for purposes of inaugurating a sovereign. (T,52) Yet at the same time, Hampton insists, since the numerous shortsighted persons in the state of nature treat PDs as one-time occurrences rather than as members of a series, "we would be acknowledging the soundness of the iterated PD argument for cooperation but still endorsing, in the main, Hobbes's Chapter 13 account of conflict." (T,53)

Hampton summarizes the "shortsightedness" account in this way:

> 1. The iterated PD game argument establishes that although it is not rational to cooperate in single-play games, many and perhaps even most cooperative situations in the state of nature are multi-play PD game situations in which it is in one’s long-term best interest to cooperate.
> 2. The complexities of life in the state of nature are such that many people will reason, mistakenly, that it is rational not to cooperate in iterated PD game situations. This fallacious reasoning will be common, but not ubiquitous, in the state of nature.
> 3. The fear that one’s partner is too shortsighted to appreciate the long-term benefits of cooperation, or the fear that one’s partner will believe that one is shortsighted, will lead one (a) not to cooperate in high-risk cooperative ventures (as dictated either by the maximin rule or by an expected-utility calculation),[6] and (b) not to cooperate in many (although not all) medium and low-risk cooperative situations, as dictated by an expected-utility calculation (where the probability that one’s partner will behave uncooperatively is generally high).
> 4. The desire for glory, understood as the desire to have the power and ability to get one’s own way, is a powerful but subsidiary cause of conflict; and insofar as it encourages the belief that one is superior to one’s fellows, it leads one to overestimate one’s chances of winning a conflict, and this encourages the conclusion among people in this state that (as defined in 3(a) or 3(b)) it is rational not to cooperate. (H,88-89)

1.3 Two Natural-State Situations

I want to indicate a problem with this interpretation by asking whether or not on it cooperation is rational in the state of nature.[7]

Item (1) in Hampton’s summary tells us that there are cooperative situations in the state of nature that are IPDs in which cooperation is rational. Item (2) tells us that in many, and maybe even in most, such cooperative situations individuals will fail to cooperate because of mistaken reasoning. Item (3), however, suggests that in many of the cooperative situations in which persons reason incorrectly, the rational thing to do, "as dictated by an expected utility calculation" or a maximin rule is to fail to cooperate. But if these cooperative situations are IPD situations as (1) tells us, then since it is allegedly rational to cooperate in such situations, (see item (1)) rationality cannot also prescribe that persons in them do the noncooperative thing. The problem can be seen more clearly if we distinguish between two different sorts of situation in which Hobbesian individuals in the natural state might find themselves. In one of these, what I call a "C1" situation, cooperation is rational. In the other - a "C2" situation - cooperation is not rational.

C1 Situations

Assume that each person in the state of nature is rational. These persons have true beliefs and reason correctly. Assume secondly, that they are in a kind of iterated PD. Here's an illustrative scenario borrowed from Chapter 2: Larry and Mo are two inhabitants of the state of nature. Larry raises cattle and his neighbor Mo rears horses. Cattle and horses are highly coveted by the numerous bandits who roam the plains. Each of our ranchers can expect his establishment to be marauded. Aware of this, each makes a defense pact with the other. The pact requires that each come to the aid of the other in case the person or property of that other is attacked by a third party.

Suppose Larry is attacked. Should Mo honor the pact and come to Larry's aid? From a short-term perspective, Mo might do best by reneging: If he adheres to the defense pact agreement, he can expect Larry to renege. After all, in the short run Larry would probably do better if he were to renege than if he were to help Mo, in the event that Mo's farm were raided. If Mo now violates the pact, he can once again expect Larry to renege in the future. From a short term perspective, then, it may seem to Mo that he is in a genuine PD, and that the game in which the two find themselves, a game that completes in two consecutive moves - first player plays (or aids) first/second player plays second, is itself PD-like, as diagram 1 on the next page confirms.

But suppose raids of this kind are frequent so that each can expect to find himself in "raid" situations many times over, but neither knows how many times over. Then from a long-term perspective, each will probably do best by upholding the defense pact. Since each is rational and neither suffers from shortsightedness, each in this IPD situation will cooperate.

Larry: Moves First	Mo: Moves Second	Preferences for outcomes after Mo's second move	
		Larry	Mo
Keeps the defense pact	Keeps the pact	5	5
	Breaks the pact	0	10
Breaks the defense pact	Keeps the pact	10	0
	Breaks the pact	1	1

The numbers represent expected utilities, higher numbers indicating stronger preferences.

Figure 3.1 Diagram 1

C2 Situations

Assume, firstly, that most but not all persons in the state of nature are shortsighted. Shortsighted people reason badly. They may do so in two distinct ways. First, their reasoning may be correct, but nevertheless they reason badly because their reasoning appeals to beliefs that are false. (H,82-83) For instance, these persons might be in some kind of matrix 2 situation in which cooperation is rational. Recall that in such a situation cooperation is dominant for each. But they might believe that they were in a single-shot matrix 1 PD situation. The rational thing to do in a matrix 1 situation, they would reason correctly, is to fail to cooperate. Since in this imaginary case, they would (mistakenly) take themselves to be in a single-round PD, they would not cooperate. Here's another example, more relevant to a discussion of Hampton's views: The persons might be in some kind of situation in which if they were rational, smart, and farsighted, it would be an IPD in which cooperation is rational. But as it stands, due to their shortsightedness, it isn't. Once again, suppose they believe that they are in a single-shot matrix 1 PD. Given their mistaken beliefs, they correctly reason that in this (hypothetical) case, they should refrain from cooperating. Second, their reasoning itself might be faulty. (H,82) They may not, for instance, be astute enough to ascertain that in an IPD of a certain type, the rational thing to do is to cooperate. They may be incapable of any "reasoning that requires any sophisticated long-term reasoning ability." (H, 149)[8] Assume, secondly, that each shortsighted person in the state of nature does in fact believe that he is in a one-time matrix 1 PD and will, in consequence, do the noncooperative thing. Assume, thirdly, that each person in the natural state knows that most persons in this state are shortsighted and that a few are not shortsighted. Assume, fourthly, that these Hobbesian individuals cannot determine whether their prospective interactive partners are either shortsighted or non-shortsighted.

Suppose I'm a non-shortsighted person in this type of state of nature and it's Thursday. Suppose, in addition, that it's now time to decide whether or not to keep a defensive pact agreement that I made on Monday. What should I do? I know I'm in a "C2" situation that abounds with shortsighted people. Suppose I also know that shortsightedness, as Hampton cautions, "is difficult, if not impossible to cure." (H,85) Now I know that shortsighted persons believe they are in a matrix 1 situation and will do the noncooperative thing come what may. Since there are many shortsighted people in C2, the probability that my partners are shortsighted is pretty high. Furthermore, I know that if I cooperate while the others do not, the outcome for me will be either pretty calamitous or at least fairly damaging. This, together with the aforementioned items of relevant information I have about C2, should be sufficient to convince me that cooperation in a C2 situation is not rational. Assume that when it comes to deciding whether or not to cooperate, each non-shortsighted person reasons in this way. The outcome will be mutual non-cooperation. If this in turn is true, then we know that a C2 situation cannot be a C1 situation or an IPD of Hampton's variety; in the latter cooperation is rational, in the former, it is not.

We can allow that this might be true: Maybe if there were no shortsighted people, the situation would not be of the C2 type. Perhaps it then would be an IPD analogous to a C1 situation. It is possible that if everyone were farsighted, the rational thing to do would be to cooperate. But even if this is true, it should make no difference to whether cooperation is rational in a real C2 situation in which many but not all are shortsighted: What's rational to do depends in part on what the others will do. If most are shortsighted, it would not be rational for me to behave as if they were farsighted.

1.4 A Problem with Hampton's Account of Cooperation

Reverting to Hampton's views on cooperation, some passages suggest that she believes Hobbesian individuals to be in a situation relevantly similar to a C2 situation. For instance, in the (full) passage cited on page 75 above, Hampton appears to believe that noncooperation will be dominant for each party. A second germane passage is this:

> Hobbes can make a fairly good case for the claim that the rationality of cooperation is sufficiently

> difficult to understand that the number of people who will reason badly and who thus will fail to cooperate is high enough to make cooperation generally too risky. (H,88)

A third relevant passage occurs later in the book:

> the iterated PD game argument is supposed to show the long-term rationality of performing the collectively rational act in those prisoner's dilemmas that are part of an indefinite series; but as we discussed in chapter 3, in most situations too many Hobbesian people are likely to be shortsighted to make it rational for even farsighted people to trust that their partners will be true to their commitments. (H,134)

Other passages, however, indicate that Hampton believes the appropriate situation is relevantly like a C1 situation. Item (1) in her summary is a nice example.[9] It should be stressed that Hampton clearly wants to ascribe to Hobbes the view that in some manner cooperation in the state of nature is rational: She believes, as I intimated earlier, that were it not rational, Hobbesian individuals would be unable to institute a sovereign and escape their predicament. (H,78-79)[10]

The summary, though, strongly suggests Hampton wants to have it both ways - that on the one hand, cooperation really is rational in many cooperative situations in the state of nature but on the other hand, since such situations are replete with shortsighted people, cooperation in them is not rational. This would, of course, be a mistake. In C2-like situations, cooperation is not rational whereas in C1-like situations, it is.[11] The problem here, perhaps, is this: In some situations like matrix 1 (or matrix 2) situations, what it is rational for a party to do does not depend on certain information about the relevant others. One needn't know, for example, that the other is incorrectly perceiving the situation as a matrix 2 situation and will thus probably do the cooperative thing, in order to know that the rational thing to do in a matrix 1 situation is to fail to cooperate. Other situations like matrix 3 situations are not of this sort. In such situations, depending on what the second party does, one would do best by adjusting one's actions accordingly. Perhaps the state of nature is similar in this respect to a matrix 3 situation: If, as Hampton believes, it is known that many in the state of nature are

	Mo	
Larry	Keeps the agreement	Breaks the agreement
keeps the agreement	10,10	-2,8
Breaks the agreement	8,-2	5,5

Figure 3.2 Matrix 3

shortsighted and will do the noncooperative thing, this bit of information, as cases 1 and 2 intimate, should considerably influence one's decision as to the rational course of action.

It is possible that Hampton does conceptualize matters in a somewhat similar fashion to what has just been suggested. She may be inclined to the view that knowledge that many in the state of nature are shortsighted "transforms" a situation that is initially an IPD into one that is analogous to C2. But even such a view wouldn't enable her to have her cake and to eat it as well. For either such a "transformation" occurs or it does not. If it does not, then contrary to summary item (3), an expected utility calculation should not prescribe noncooperation in most cooperative situations. If it does, then pace summary item (1), most cooperative situations in the state of nature will not be IPDs.

Finally, Hampton could say that it's wrong to think of the state of nature as one big cooperative situation in which cooperation either is or isn't rational. Rather, think of the natural state as abounding with many cooperative situations, most of which are of the C2 variety, but some of which are of the C1 variety. (She could not here reverse the order, supposing an abundance of C1 situations. For then in most cooperative situations cooperation would be rational. The state of nature would, as a result, fail to exhibit the kind of conflict Hampton believes it does.)

The problem with this suggestion is that it is inconsistent with summary item (1). Contrary to what that item tells us, the suggestion now under consideration proposes that most cooperative situations in the natural state are of the C2 variety in which cooperation is not rational.

In light of the foregoing, let's tentatively conclude that Hampton's shortsightedness account of conflict supposes Hobbesian individuals in the state of nature to be in either a C2-like situation or in a C1-like situation. I now want to show that this conclusion undermines what Hampton believes is the first premise in Hobbes' argument for absolute sovereignty.

1.5 Hampton on Sovereign-institution

The premise in question is this:

> 1. It is in each person's interest to be subjugated to a sovereign rather than to remain in the state of nature. (H,186)

Hampton argues for this premise in the following passage:

> If we consider how Hobbes describes the state of nature and what he says about the reasons people have for instituting a sovereign, his remarks

indicate that it is not best characterized by the matrix in Figure 6.6, but by the matrix in Figure 6.7. [These matrices are reproduced on page 91 below.] Actually, this matrix represents an idealized version of what I will eventually argue is the real deliberation of the parties regarding the institution of the sovereign, because it helps us to clarify what the preferences of the people are for remaining in the state of nature versus surrendering their rights to some person or assembly. In the matrices of Figures 6.6 and 6.7 we are supposing, for simplicity's sake, that there are three people in the state of nature, that one of them person Z, has already been selected by some process as potential sovereign, and that the other two people, X and Y, are deliberating whether or not to surrender their rights to all things to Z. Suppose that I am individual X and you are individual Y. Would our preferences match those in Figures 6.6 or 6.7? In the PD matrix of Figure 6.6, I reason that I would be better off in a partial state of war (where you have surrendered your rights to Z but where I have not) than I would be in either a complete state of war or a complete state of peace. However, this does not seem to be the preference I would actually have if I were deliberating whether or not to surrender to Z; instead, it would seem that I would believe, as the matrix in Figure 6.7 indicates, that I would be worse off in this partial state of war than I would be in the total state of war. In the latter state there would exist only individuals (i.e. Y and Z) of strength and abilities roughly equal to my own, who might be deterred from attacking me because they would be uncertain of having sufficient strength to overcome me, or who could be repelled successfully by me, given their attack, if I had a slight advantage in strength. But if you should surrender your right to Z and I did not, there would be a consolidation of powers in this small confederacy, making the group significantly stronger than any single individual like myself. This confederacy would therefore be likely to attempt, and be successful in, an attack against me. But this means I would perceive my life in such a partial state of war to be less secure than in a total state of war, so that I would prefer the latter to the former.

However, as the matrix indicates, I would regard being a member of this confederacy in a partial state of war as preferable to being a lone

> individual in the state of total war. If I surrendered my right to all things to Z, I would gain additional security because I could rely on the support of Z if I were attacked, and the two of us together would fare better in any preemptive strike against another individual like you because of the strength of our numbers. However, I would best prefer the situation in which both you and I would authorize the same person or assembly of persons as sovereign. My security would be greatest when there were no other individuals or groups who were still at war with me, and this would occur when everyone in the state of nature had authorized the same person as sovereign. (H,148-149)

	X	
Y	Surrender to Z	Do not surrender to Z
Surrender to Z	2,2	4,1
Do not surrender to Z	1,4	3,3

Figure 3.3 Hampton's Figure 6.6

	X	
Y	Surrender to Z	Do not surrender to Z
Surrender to Z	1,1	2,4
Do not surrender to Z	4,2	3,3

Figure 3.4 Hampton's Figure 6.7

The numbers in Figures 6.6 and 6.7 represent preference-strengths, lower numbers indicating stronger preferences.

Suppose Hampton is right that the matrix in her Figure 6.7 (or as I shall say "matrix 6.7") correctly portrays persons' preferences for sovereign institution. Then it is straightforwardly rational for Hobbesian individuals to institute a sovereign. Her inauguration, contrary to the discussion in the last chapter, involves no PD-like problems. I believe, though, that Hampton's argument for premise 1 leaves something to be desired.

Assume, as Hampton does, that the question of who will be sovereign is settled. Assume, also, that Hobbesian individuals are psychologically able to do what is required to create a sovereign. In deciding whether or not it is rational to institute a sovereign, each individual in the state of nature, I think, faces at least two decision problems. It is a failure to distinguish these two problems that is the ultimate source of the difficulty for Hampton's argument:

Let individuals x, y, and z be state-of-nature individuals. Suppose I am individual x and I am deliberating whether to surrender my right to all things to the sovereign-elect, z. Hampton seems to attribute to me the following kind of reasoning: "Suppose individual y surrenders his rights to z. Then y and z will unite in a defense confederacy. The confederacy will be stronger than any lone individual like me. I will then in all likelihood be attacked, because doing so will further the interests of y and z, now united in a strong team. (Assume that if their attack were successful, they would split the spoils 50-50.)

I now have the choice of either surrendering to z, or refraining from doing so. If I fail to surrender, I will be assailed. I will be unable to repel the invaders and my life will be hanging on a thread. If I surrender, I will be spared. Furthermore, I will almost certainly be able to enjoy a share of the fruits of further raids on lone individuals by the then enlarged group consisting of x, y, and z. So if y surrenders, I do best by surrendering.

Suppose, on the other hand, y does not surrender. Then if I don't surrender, everyone remains in the state of nature. That would be bad. If I surrender, then z and I will unite in a defense confederacy. I could rely on the support of z if I were attacked, and the two of us together would fare better in any preemptive strike against another individual like y. So if y fails to surrender, I do best by surrrendering.

Hence no matter what y does, I do best by surrendering. So I should surrender."

Notice, however, what this line of reasoning presupposes. Hampton's argument that matrix 6.7 correctly represents the preferences of individuals in the natural state "for being in such a PD-prone state versus being in a commonwealth (or something in between)" (H,150), presupposes an affirmative answer to a different decision question: whether it is rational for an Hobbesian individual in the state of nature to cooperate with another such individual in order to form a defense confederacy.[12] It presupposes, in other words, an answer to the very kind of question discussed in section (1) above and in the preceeding chapter - whether cooperation in the state of nature is rational.

1.6 Prior Results Summarized

In the last chapter I arrived at some general conclusions on the rationality of the making and the keeping of defense pact agreements in the state of nature.[13] I argued that in a transparent one-shot sequential play PD-like situation, it is rational for both first-party and second-party members to uphold a defense pact. Hampton, however, eschews considerations of transparency (H,218-219). So this conclusion would fail to impress her. In an opaque, single round PD-like game, I argued that it is not rational for either first-party or second-party Hobbesian individuals to adhere to a defense pact. Again, I think this conclusion would not perturb

Hampton since she seems to believe that Hobbesian individuals in the state of nature are in an IPD. The third conclusion I assumed is that Hobbesian individuals in an IPD will find it rational to abide by a defense pact agreement. But again, this conclusion may not be very germane as far as Hampton is concerned, since she believes shortsightedness to be widespread in the state of nature. I assumed, in endorsing the third conclusion, that all individuals in the IPD were rational.

So let's reconsider Hampton's views on natural-state cooperation and see how they bear on her argument in support of premise 1.

1.7 Why Hampton's Account of Sovereign-institution Fails

I concluded section (1) by proposing that on Hampton's views, people in the state of nature are either primarily in C1 sorts of situation or they are principally in C2 kinds of situation. Let's examine each possibility.

Assume that raids or preemptive strikes in the state of nature are common, so that individuals like x and z can expect frequently to find themselves in raid situations in which they must defend themselves against attack by third parties. Well aware of this, neighbors x and z ponder whether to form a defense confederacy. It is not unreasonable to suppose that a situation of this sort is an IPD, or a sort of situation analogous to a C1 situation.

In a situation of this type cooperation is rational. In any case, we can assume as Hampton seems to, that it is. If x, y, and z are in fact in this kind of situation, then providing each is rational, it is in the self-interest of each to make and to keep a defense pact agreement. Hampton, as we saw, believes many in the state of nature to be shortsighted. But let's initially suppose, to simplify matters, that x, y, and z are fully rational, farsighted individuals. Then it would seem that each would make a defense pact treaty and would abide by that treaty.

Support that Hampton takes x, y, and z to be in this sort of situation, or at least in a situation in which cooperation is rational, is found in what she has to say about the notion of surrendering one's rights to someone else:

> An individual authorizes another as sovereign [or surrenders one's rights to all things to the sovereign (H,174)] by
> 1. participating with the other inhabitants of the state of nature in a process in which one of them is selected as sovereign (e.g., in a peaceful election process or in violent competition for leadership in which one confederacy emerges as dominant, followed by an explicit or tacit agreement that this individual is sovereign) and
> 2. obeying the punishment commands (a) of only this individual (b) to refrain from interfering in the punishment of another and (c) to actively assist in the punishment of another, insofar as these commands are (or have been made by the sovereign to be) individually rational.....
>
> [A]ll of these actions involved in authorizing the sovereign can be performed by Hobbesian people.

> (H186-187) [This is because it is straightforwardly rational for each to perform each of these actions. (H,187-188,207)]

So, for example, if x "surrenders his rights to all things" to z, x undertakes an action or a series of actions, each of which is SM-rational. If it is rational for x (and indeed for y) to form a defense confederacy with z, then Hampton's argument in support of premise 1 may seem pretty strong.

However, there's a problem: If it is straightforwardly rational for x and y, in fact for any individual, to unite in a defense confederacy with others, then the state of nature, I think, would fail to exhibit substantial warfare.[14] There would then be no need for an absolute sovereign. To put the point slightly differently, the assumption that it is rational for individuals in the natural state to make and to keep defense pact agreements seems necessary to sustain Hampton's argument for premise 1. At the same time, this very assumption seems to undermine the shortsightedness account of conflict.

Hampton may object that I have assumed that all persons in the state of nature are rational and that none is shortsighted. But she might protest that it is this very characteristic of Hobbesian persons that is essential to understanding why the natural state abounds with conflict. She stresses that "shortsightedness makes warfare inevitable." (H,148)

So let's consider the second possibility. Assume that most individuals in the state of nature are shortsighted. These persons, being shortsighted, would fail to cooperate. They would fail to form defense confederacies with others. They would fail to do so because shortsighted persons, when deliberating whether or not to enter into a defense pact agreement, would take themselves to be in a one-shot PD in which cooperation is irrational. A non-shortsighted person in this C2 state-of-nature situation would also fail to cooperate: Suppose Hobbesian individuals x, y, and z are in this situation. Suppose x is rational whereas the other two are shortsighted. On Monday x makes a defense pact agreement with z. It's now Thursday and x must decide whether to act in accord with the agreement. x knows that the probability that the other two are shortsighted is high. He knows that shortsightedness is almost impossible to cure. He knows that shortsighted persons believe they are in a one-shot PD and will not do the cooperative thing. x rightly concludes that it would be irrational to abide by the defense pact. Cooperation in this C2 sort of situation is not rational. Since it is not rational for x to adhere to the defense pact, x will find it irrational to subjugate himself to z. Hampton argues to the contrary. But her argument assumes that it is rational for x and z to form a defense confederacy. This is not true in a C2 situation. Her argument therefore, when the relevant situation in which x, y, and z are is a C2 situation, is not persuasive.[15]

Perhaps what is needed to solve the problem of the Hobbesian social contract is a reconceptualization of Hobbes' views on rational cooperation. That is precisely what David Gauthier gives us.

1.8 Prelude to Gauthier's Views

Gauthier's views are discussed in some detail in the next four chapters. As a prelude to this discussion, I think it fitting to conclude this chapter with these provoking remarks of his on Hobbes:

> But 'this specious reasoning is neverthelesse false.' Hobbes has another, and better, reply to the Foole, in his account of right reason.....Not only morality, but rationality as well, must come within [the] ambit [of conventionalism].....The Foole, in appealing to natural reason in support of injustice, falls into inconsistency, through his failure to appreciate the tight conceptual connection between right and reason which is necessary to Hobbes's thought. The right of nature expresses right reason. If one lays down some portion of that right, then one also renounces the rationality that was the basis of the right laid down. If one lays down some portion of one's right to do whatever seems conducive to one's preservation and well-being, so that one may find peace, then one renounces preservation as the standard of reason, in favor of peace. The Foole appeals to that reason which dictates to every man his own good - to natural reason, so that he may

show injustice to be rational. But injustice is a violation of covenant, and, in covenanting, in laying down one’s right, one has renounced natural reason as the court of appeal, in favor of a reason that dictates to every man what all agree is good.[16]

Notes

1. See Jean Hampton, Hobbes and the Social Contract Tradition (1986), Cambridge: Cambridge University Press. Page references to this book are as follows: (H, page number).

2. See Hobbes and the Social Contract Tradition, chapters 2 and 3.

3. The issues of just what an iterated PD is, and whether cooperation in such a PD is rational, are controversial. See, for example, David Braybrooke, "The Insoluble Problem of the Social Contract," in Paradoxes of Rationality and Cooperation eds., R. Campbell and L. Sowden (1985), Vancouver: The University of British Columbia Press, 277-306; Peter Danielson, "The Moral and Ethical Significance of TIT FOR TAT," Dialogue 25 (1986), 449-470; Gregory S. Kavka, Hobbesian Moral and Political Theory (1986), Princeton, New Jersey: Princeton University Press, pp. 129-136, and J. H. Sobel, "Utility Maximixers in Iterated Prisoner's Dilemmas," Dialogue 15 (1976), 38-53. In this paper Sobel argues that cooperation in an iterated PD is not rational.

4. Hampton presents the passions account in Hobbes and the Social Contract Tradition, pp. 63-74, and in "Hobbes's State of War," Topoi 4 (1985), 47-60. Page references to this paper are given in this way: (T, page number).

5. The rationality account is discussed by Hampton in Hobbes and the Social Contract Tradition, pp. 58-63; 74-79; and in "Hobbes's State of War." It is basically the same account that I developed in section 2.3 in Chapter 2.

6. In a high risk situation,

> one can suffer crippling losses if the other party reneges and takes advantage of one's cooperation. [Hampton, Hobbes and the Social Contract, p. 81. See, also, p. 71 of this same work, and Hampton's "Hobbes's State of War," p. 53.]

7. Let's ignore the complication of distinguishing, as Hobbes seems to do, between the rationality of cooperation with respect to first-party members and with respect to second-party members in sequential play PD-like games.

Hampton does not directly address the symmetry problem. However, it's abundantly clear that she attributes to Hobbes certain views on the rationality of cooperation in the state of nature. Her position on the symmetry enigma may then simply be this: If her interpretation of Hobbes' account of cooperation is correct, then the problem fails to arise. In other words, the symmetry problem is to be evaded by reconstruing Hobbes' account of rational cooperation.

8. There is a nice discussion of the sources of shortsightedness in the state of nature in Hampton's "Hobbes's State of War," pp. 54-56.

9. It should be clear that Hampton does not intend 'cooperative situation' to refer solely to a situation where the rational thing to do is to cooperate. So, for instance, in summary item (3), she claims that it is not rational to cooperate in many medium-risk and low-risk cooperative situations.

10. See also, for example, Hobbes and the Social Contract Tradition, pp. 78-79.

11. The tendency to conflate the two situations seems apparent in the following passage, this time in "Hobbes's State of War," p. 58:

> According to the shortsightedness account of conflict, although cooperation is rational, the prevalence of shortsighted and glory-prone people will mean one usually cannot risk cooperation oneself because one cannot be sure enough in most cooperative situations that one's partner will be cooperative (or learn to cooperate).

12. Braybrooke is sceptical about this. If x and z are equally resourceful in the state of nature, then

> why should either have less to fear from utterly subjecting himself to the other than from entering into a contract which the other would be free to violate? [David Braybrooke, "The Insoluble Problem of the Social Contract," p. 287]

13. See Chapter 2, section 2.5, this thesis.

14. It's possible that a state of nature will emerge with multiple defense groups. It might then be tempting to explain the existence of conflict in terms of inter-defense-group warfare. But this suggestion doesn't fare too well either: We need only take each group as an individual and ask whether cooperation among such individuals is rational. On the possibility now being entertained, cooperation among natural-state individuals is rational. So "group individuals" would cooperate to fend off third parties, again with the result that there would be insubstantial warfare in this sort of state of nature.

15. Richard J. Arneson, in correspondance, has suggested the following sort of reply available to Hampton in defense of the claim that the existence of natural-state conflict is consistent with the eventual ratification of the social contract:

It isn't true that if it's rational for individuals to unite in a defense confederacy with others, the state of nature would not exhibit much conflict:

> Given shortsightedness, many individuals would not form defense confederacies, and these individuals would be prone to war against each other and they would also be tempting prey for such defense confederacies as do form. If a single defense confederacy becomes doiminant, and subordinates all individuals over a large territory, then we have sovereignty by conquest. If this does not occur, we have the state of war, hence the strong motive to form sovereignty by institution. Hampton raises the question of whether it would be rational to submit to a sovereign when one knows that many persons are shortsighted. Her answer is that even shortsighted people can see their way to cooperation in this instance, because the payoffs favor cooperation (submission) even in the very short term.

Since cooperation is rational in the state of nature evisaged by Arneson, perhaps we're in a C1-like situation. In this sort of situation, raids and preemptive strikes are common (see page 95 above). Arneson suggests that even shortsighted people in such a situation "see their way to" submission. It is better for these Hobbesians to submit to a sovereign than to keep warring. Suppose all this is true. Why, then, wouldn't these very same shortsighted persons unite in defense confederacies given the frequency of raids and preemptive strikes? Surely, if anything, it is more in their immediate short-term interest to defend themselves against raiders than to inaugurate a sovereign. If these shortsighted persons can "see their way to" submission, they should also, barring any plausible explanation to the contrary, be able to "see their way to" forming defense confederacies. If this, in turn, is true, I don't see, *pace* Arneson, how *many* shortsighted persons in this state of nature would "not form defense confederacies" and "would be prone to war against each other."

16. See David Gauthier, "Thomas Hobbes: Moral Theorist," *The Journal of Philsophy* 76 (1979), 547-561, pp. 556-557.

CHAPTER 4

RATIONAL BARGAINING

1.1 Gauthier on the Prisoner's Dilemma

The matrix representing Butch's and Sundance's PD is this:

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	10,10 (a)	1,30 (b)
Remains silent	30,1 (c)	2,2 (d)

Figure 4.1 The Felons' Dilemma

The numbers represent years with Sundance's prison-terms shown first. Think of (a), (b), (c), and (d) as "agreement outcomes," or outcomes that would result if each felon were to agree with the other to act in a certain fashion, and were then to act in that fashion at the appropriate time. (d) is the "cooperative outcome."

The parts of Gauthier's argument germane to showing that the dilemma is escapable - that in the example, Butch and Sundance could rationally cooperate and secure optimal outcome (d) - are essentially two: First, Gauthier argues that as SFMs, agents in the position of Butch and Sundance, engaged in genuine bargaining over the agreement outcomes, will settle on (d), the outcome selected by the principle of minimax relative concession. This principle of distributive justice, according to Gauthier, expresses the principle of utility maximization in the context of bargaining. (M by A, 145,151) Second, rational bargainers can assure themselves that their would-be partners do not cheat when the time comes to act on the terms of a rational agreement, because their "constrained rationality" enjoins compliance with agreements that are rationally made.

In this chapter I discuss rational bargaining. In the next, I deal with the issue of compliance.

1.2 The Bargaining Issue

Bargaining over "possible PD agreement outcomes" like (a), (b), (c), and (d) may seem to pose no special problem: the cooperative outcome, (d), seems to be salient. It seems to be the one that rational agents should seek to attain. So let's for the moment leave Butch and Sundance and talk about others. We'll then return to our two felons.

Gazing at the starry skies one night, Tom and Dan - two SFMs - are promised by VOICE that if they were rationally to agree on how to divide $100 between them, each would find the appropriate sum under his pillow the next day. To reach agreement would be to consent to act in accord with a strategy. A strategy - roughly speaking - consists in a set of actions, one for each maximizer. Since there are many possible splits, Tom and Dan have available to them many different strategies. In conformity with what principle, if any, would it be rational to choose among these strategies? If we suppose that in settling on an option some compromise will be required - neither on the face of it, if rational, will agree to a $100/$0 or $0/$100 split - then each will have to bargain with the other to make certain concessions.

David Gauthier proposes that SFMs in such bargaining situations will settle on a compromise in which the largest concession anyone makes is smaller than it would be in any other compromise available: They will choose agreement outcomes on the basis of the principle of minimax relative concession (MMRC).[1]

In this chapter, I begin with a summary of Gauthier's theory of rational bargaining.[2] This is a theory that purports to specify the terms of rational agreement. It sets aside the issue of whether compliance with an agreement rationally entered into is itself rational. I then argue that SFMs need not bargain on the basis of the principle Gauthier recommends. They would be no less rational, in certain bargaining situations, to strike agreement in some other way. This questions Gauthier's assumption that there is a uniquely attractive principle of rational bargaining.

1.3 Gauthier's Theory of Rational Bargaining

Assume, plausibly, that cooperation often makes possible a surplus of goods that would be unavailable if those who cooperate to produce this surplus were to refuse to interact, and were instead to act independently. Bargaining theory is addressed to how this surplus is to be apportioned among those who contribute to its production.

Gauthier tells us that in bargaining it is natural to think of each person as beginning from a base point - a prebargaining payoff that is not called into question by the bargaining situation. The prebargaining payoff may initially be identified with what each person could expect to garner in the absence of any cooperative interaction.[3] Each person then makes a claim reflecting her desire to gain as much as possible from agreement, subject to the constraint not to drive others away and not herself to be excluded from the bargaining table. Given this constraint and the desire to maximize payoffs, each person's claim is the largest slice of the pie he could get consistent with each other bargainer getting his prebargaining payoff. In the example of our star-gazers, each has a prebargain payoff of $0 and each will claim $100. A rational bargain gives each individual no more than his claim and no less than his base point. Incompatible claims will require bargainers to make concessions if agreement is to be reached. Since each wishes to gain as much as possible, each will endeavor to minimize her concessions.

> Now the magnitude of concession is established not with reference to some absolute scale of utility, but rather with reference to the particular bargaining situation; concession is a measure of the proportion between the part of one's claim that one abandons, and the entire claim, or gain over one's base-point payoff, that one originally advances. Since the bargainers are equally and fully rational, the maximum concession - the greatest proportion of his or her original claim that any bargainer gives up - must be minimized. Since all benefit from reaching agreement, some set of concessions is rational for all to accept, but a particular set is rational for all only if any alternative would require a concession at least as great as the maximum in the given set.
>
> In bargaining, therefore, rational persons will act on a principle of minimax concession - the greatest or maximum concession must be a minimum.[4]

Let's introduce some abbreviations.

(1) UA(Oi) is the value of outcome Oi for some agent A.

(2) UAmin is A's minimum cooperative utility. It is the minimum utility she must be guaranteed by any agreement outcome if she is to choose any such outcome.[5] UAmin is equal to what A could expect to acquire from her own efforts in the absence of agreement.

(3) UAmax is A's maximum cooperative utility or A's claim.

Suppose the total gains that cooperation may bring to a two-person group of bargainers consisting of members A and B, gains over and above the base points of A and B, is some quantity k. (If k is some good, assume that bargainers' utilities are linear with respect to their share of this good.) k is the *cooperative surplus*. When there are only two bargainers, as in the case of A and B, A's claim will be the whole cooperative surplus. When there are more than two bargainers, one's claim will be that portion of the cooperative surplus to which he contributes. At least for two-person bargaining situations, it appears that your claim is equivalent to the maximum amount of utility you could hope to obtain from cooperation compatible with your partner receiving her minimum cooperative utility.[6]

(4) RA(Oi)adv is the relative advantage of Oi to A. If Oi is any agreement outcome, then RA(Oi)adv = (UA(Oi)-UAmin)/(UAmax-UAmin). The numerator of this ratio is the difference between the utility a person is going to obtain from cooperation and that person's minimum cooperative utility.[7]

(5) RA(Oi)conc is the relative concession of Oi to A. If Oi is any agreement outcome, then RA(Oi)conc = (UAmax-UA(Oi))/(UAmax-UAmin). The numerator of this ratio is the difference between a person's maximum cooperative utility and the utility she is going to receive from agreement. This difference is, therefore, simply the amount of utility

a person forgoes in not getting her maximum.[8] For any agent, a, and agreement outcome oi, the sum of Ra(oi)adv and Ra(oi)conc is one.[9]

We can now finally formulate Gauthier's principle of rational bargaining, the principle of minimax relative concession.

MMRC: Given a range of agreement outcomes each of which requires concessions by some or all bargainers if it is to be selected, an agreement outcome should be selected only if the greatest or maximum relative concession it requires is no greater than the maximum relative concession required by every other agreement outcome.[10]

Think of MMRC in this way: An agent's actual concession is the difference between her claim and what she gets from the outcome eventually agreed to. Her maximal concession is the difference between her claim and her prebargaining payoff. Her relative concession is simply a ratio of her actual concession to her maximal concession, or what she actually concedes in bargaining relative to the most she could concede in bargaining. For each possible outcome of bargaining, this parameter can be calculated for each agent. We can somewhat unrealistically think of bargainers acting on MMRC as doing the following: First, they inspect all possible agreement outcomes. Second, they demarcate the largest relative concession required in each outcome. Call this set of largest relative consessions "LRC." Each outcome has an LRC member. Third, bargainers select the outcome with the smallest LRC member.

Intuitively, Gauthier's elegant formal machinery is designed to capture this idea: Assume a two-person bargaining situation. Assume in addition that there is a single good, produced in fixed quantity and divisible in any way among the cooperators.[11] Suppose that by acting independently, each agent can produce k units of the good. Suppose that by pooling their resources and cooperating, they can generate 4k units of this good. The cooperative surplus, as we noted, is the difference between what can be produced by individual activity and what can be produced by joint activity - in terms of our example, this amounts to 4k-(k+k) = 2k units of the good. Each cooperator, Gauthier reasons, is entitled to her prebargaining payoff, k, and to a share of the cooperative surplus proportional to the amount of that surplus that he makes possible. In a two-person cooperative venture, if there is a cooperative surplus, each party is equally responsible for its production. So in such a case, each cooperator is entitled to an equal share of the surplus (2k/2 = k units).[12]

A down-to-earth example should further help illuminate Gauthier's intuitions about bargaining: Fred has $200 to invest and Bruce has $300. Investments under $500 earn 5%

and those of $500 and over earn 10%. Their only options are to invest separately, or to pool their funds and to invest jointly. Investing separately, Bruce can expect a return of $15, and Fred a return of $10. If they pool their money, they can expect a return of $50 on the collective fund. Joint investment generates a cooperative surplus equivalent to $50-($15+$10) = $25. Without the cooperation of the other, no such surplus would be possible. Each contributes equally to its generation, and so each, Gauthier proposes, is entitled to an equivalent share. Bruce will end up with a total of [$300 (his principal) + $15 (his prebargaining payoff) +$12.50 (his share of the cooperative surplus)] = 327.50; and Fred with [$200 +$10 + $12.50] = 222.50. The example highlights what appears to be Gauthier's conviction that a principle of rational cooperation must specify, in an acceptable fashion, how the contributions each makes in a cooperative endeavor are related to the production of the cooperative surplus.[14]

Let's now revert to Butch and Sundance. Assume their utility matrix is identical to the one in Figure 4.2 on page 114 below. Numbers portray expected utilities so that more is better. What each could expect to secure in the absence of cooperation is one utile. In other words, Umin for each is equivalent to this quantity. Umax for each =

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	1,1 (a)	10,0 (b)
Remains silent	0,10 (c)	9,9 (d)

Figure 4.2 The Felons' Utility Matrix

18-1 = 17. We can now calculate the relative concession each would make if each "selected" one of the outcomes (a), (b), (c), or (d). The formula for relative concession is this: (Bargainer's claim - UBargainer (Oi)/(Ubargainer-max - UBargainer-min). The relative concession of outcome (a) to each, then, is (17-1)/(17-1) = 1. Similarly, Reach(b)conc = 0.44; Reach(c)conc = 1.1; Reach(d)conc = 0.5. We may represent each outcome as a set of relative concessions: {(a) or (1,1), (b) or (0.44,1.1), (c) or ((1.1,0.44), (d) or (0.5,0.5)}. By inspection, MMRC prescribes (d) - it recommends that each do the cooperative thing.

An alternative route to the same result is this: MMRC allocates to each his base point and the part of the cooperative surplus to which he contributes. The cooperative surplus = (9+9)-(1+1) = 16. Each contributes equally to the realization of this surplus. In consequence, MMRC awards to each 1 + (16/2) = 9 utiles.

I have assumed that if IS is a bargaining situation with n agents, an agreement outcome may equally well be represented as a set of n relative concessions, one for each agent of IS. Let S be a bargaining situation and let AS1,.....,ASi be the agreement strategies available to members of S. A _minimax concession in S_ is a relative concession that is the maximum concession in the set of relative concessions (or the agreement strategy) in which it occurs, and that is no greater than a maximum concession in each ASi that is a possible outcome of S.

1.4 _An Argument for MMRC_

Gauthier argues that rational persons like Tom and Dan, bargaining over monetary splits - or more generally, bargaining over agreement outcomes - will act on the basis of MMRC:

> If there is to be agreement, then someone must make a concession at least equal to the minimax. Now if it is not rational for me to make such a concession, then, since the policy which is rational for me is rational for everyone, it is not rational for any person to make such a concession, and there can be no rational agreement. But it is rational for me to enter into an agreement; hence it must be rational for me to make a minimax concession. Furthermore, since agreement can be reached without any person making a larger concession, and since it cannot be rational for me to make a greater concesssion than necessary, it

> cannot be rational for me to make a concession larger than the minimax. Hence it is rational for me to enter into any agreement requiring at most the minimax concession from me. Since everyone reasons similarly, bargaining among rational persons proceeds on the _principle of minimax concession_.[16]

Let S be a bargaining situation. Then presented somewhat more systematically, the argument can be summarized in this way:

1. If it is rational for any member of S to make an agreement (i.e. if it is rational for any member of S to select an agreement outcome), then it is rational for some member of S to make a concession at least equal to the minimax.[17]

2. It is rational for any member of S to make an agreement. (Any member of S stands to benefit from agreement.)

3. Therefore, it is rational for some member of S to make a concession at least equal to the minimax. (1,2)

4. a. Assume I am a member of S.

   b. If it is not rational for me to make a concession at least equal to the minimax in reaching agreement, it is not rational for any member of S to make such a concession.

Line (4) is justified by an appeal to the equal rationality of all bargainers.

5. Given (3) and (4b), it is rational for me to make a concession at least equal to the minimax in reaching agreement.

Notice (5) allows the possibility that it would be rational for me to make a concession larger than the minimax in selecting an agreement outcome. To block this possibility the argument continues in this way:

6. Rational persons would reject, and they would expect any other rational person to reject, a given relative concession if no one need make such a large relative concession.[18]

7. In reaching an agreement no one need make a concession larger than the minimax.[19]

8. If (6) and (7), then (9).

9. In reaching agreement rational persons would reject, and would expect any other rational person to reject, a given relative concession if it is larger than the minimax. (6,7,8)

10. If (9), then (11).

11. It would be irrational for any person and so irrational for me to make a concession larger than the minimax in reaching an agreement. (9,10)

(12), (13), and (14) complete the argument.

12. If it is rational for me to make a concession at least equal to the minimax but no larger in reaching an agreement, then it is rational to make an agreement requiring at most the minimax concession from me.

13. Therefore, it is rational to make an agreement requiring at most the minimax concession from me. (5,8,9)

14. Since all rational agents reason similarly, bargaining among such agents proceeds on the principle of MMRC.

1.5 MMRC's First Competitor

Premise (6) is crucial in establishing the irrationality of any bargainer making a concession larger than the minimax. But this premise imposes an arbitrary constraint on rational bargaining - at least that is what I now want to argue.

Rewritten, we can interpret (6) as a precept of rational bargaining:

B1: Rational persons ought (rationally) to reject, and they would expect any other rational person to reject, a given relative concession if no one need make such a large relative concession.

There are alternatives to B1: The total cooperative gain is the maximum total payoff that would be available to a group of potential bargainers were they to act on an agreement strategy. If Tom and Dan, for example, were to reach agreement and cooperate, the maximum total payoff available to them would be a $100.00. A maximin rule for bargaining tells an agent to compare the minimum values of each outcome and to settle for the outcome whose minimum is the maximum value for all the minimums. Now compare B1 with B2.

B2: Rational persons ought to reject, and they would expect any other rational person to reject, a given net utility (that is, a utility a bargainer stands to receive from an agreement outcome) if (i) no one need have such a small net utility, and (ii), distrubution of the total cooperative gains by the maximin procedure in (i) results in each bargainer receiving more than his prebargaining payoff. If (ii) is not satisfied, then bargainers should proceed in accordance with B1.

B1 and B2 generate different results in bargaining as the table on the next page illustrates. Abel's and Mabel's minimum cooperative utilities are 180 and 80 respectively as indicated by the payoff each would receive if each were to select O1. Assume that the total cooperative gain is $500.00. Abel's and Mabel's maximum cooperative utilities can now be calculated: UAmax = 500-80 = 420; UMmax = 500-180 = 320. Having established minimum and maximum cooperative utilities, relative advantages and relative

Table 4.1 The Abel/Mabel Case

Agreement Outcomes		O1	O2	O3	O4	O5	O6	O7
Relative Advantages	A	0	1	0	0.1	0.2	0.3	0.4
	M	0	0	1	0.9	0.8	0.7	0.6
Utilities	A	180	420	180	204	228	252	276
	M	80	80	320	296	272	248	224
Relative Concessions	A	1	0	1	0.9	0.8	0.7	0.6
	M	1	1	0	0.1	0.2	0.3	0.4

Agreement Outcomes		O8	O9	O10	O11	O12	O13
Relative Advantages	A	0.5	0.6	0.7	0.8	0.9	0.29
	M	0.5	0.4	0.3	0.2	0.1	0.71
Utilities	A	300	324	348	372	396	250
	M	200	176	152	128	104	250
Relative Concessions	A	0.5	0.4	0.3	0.2	0.1	0.71
	M	0.5	0.6	0.7	0.8	0.9	0.29

concessions can now easily be determined in conformity with definitions (4) and (5) given in section 1.3 above. By inspection B1 prescribes O8, and B2 O13. Were O13 not a possible outcome, B1 would once again prescribe O8, but B2 would now prescribe O6. Suppose we change the example by assuming that O1 yields Abel 260 units of utility, that is, Abel can get this without cooperating. If clause (ii) of B2 were omitted, then since the total gain is 500, and by assumption utility is fully transferable, B2 without clause (ii) would assign Abel 250 units - less than he would recieve by not cooperating. B2, then, without clause (ii) would be a non-starter as a bargaining principle; no rational person will employ a principle that might require him to accept less than he could get by not bargaining. So if O1 affords Abel 260 units, B2 requires the two to revert to Gauthier's minimax rule.

Since B1 and B2 generate non-equivalent results, why prefer the first to the second? A suggestion in the texts is that B1 enjoys a theoretical advantage over B2.

> Given the claims of the bargainers, what concessions is it rational for them to make? To answer this question we must first consider how concessions are to be measured. The absolute magnitude of a concession, in terms of utility, is of course the difference between the utility one would expect from the outcome initially claimed and the utility one would expect from the outcome proposed as a concession. But this magnitude offers no basis for relating the concessions of different bargainers, since the measure of individual utility does not permit interpersonal comparisons. However, we may introduce a measure of relative concession which does enable us to compare the concessions of different bargainers, and which thus gives us a basis for determining what concession each must rationally make. (M by A, 134-135)

> Hence, although we may not assume that the numerical utilities of different persons are comparable, we may assume the comparability of relative advantage.[20]

The suggestion, then, is that B1 enables a comparison of the concessions each bargainer makes without need to postulate an interpersonal utility scale, or a scale that indicates how much one person likes one item relative to how much other people like the same item. B2, on the other hand, needs to presuppose some such scale. Relative to B1, B2 is therefore at a disadvantage since it is probably not possible to make interpersonal utility comparisons.[21]

Such reasoning, however, ought not to convince a SFM to prefer B1 to B2: Gauthier informs us that in most cases acting in accordance with MMRC will result in bargainers selecting an agreement outcome that will require each to make equal relative concessions.[22] Now suppose there is no valid interpersonal scale of utility and rational bargainers Tom and Dan select an outcome, O*, that requires each to make the same relative concession. Assume, now, that had there been a valid interpersonal scale of utility, Dan would discover that by selecting O*, he has conceded in absolute terms - that is, in terms of the amount of utility conceded - twice as much as Tom. Given this possibility, Dan may reasonably conclude that acting on the principle of MMRC yields unfair results. More to the point, however, it seems that measures of relative concessions (or relative advantages) provide rational bargainers with no basis to decide whether each is better or worse off relative to the others _in the absence of assuming some valid interpersonal_ scale of utility: Given the assumptions of our conjecture, Dan would not know whether he were better or worse off than Tom in choosing O* if there were no way to compare the utilities afforded each by O*. It is no consolation to Dan to know that he has made the same relative concession (or equally, gained the same relative advantage) in selecting O* if their utility scales differ markedly. Similarly, Dan would be in no position to know whether he was doing better (worse) than Tom in selecting an agreement outcome that afforded him a larger (smaller) relative advantage than it did Tom, in the absence of a valid interpersonal scale of utility. If Dan's selecting an outcome that offers him a larger relative advantage than it does Tom is no guarantee that he is better off than Tom in so selecting, why should he be in the least inclined to select agreement outcomes on the basis of MMRC, or as I shall say, on the basis of minimaxing, if persons' utilities are not comparable? Matters could, of course, be rectified by postulating a scale that allows interpersonal comparisons of utility. In that case, though, B1 would be no more attractive than B2 since it would no longer enjoy the theoretical advantage it is suppposed to have over B2.

So it seems that B1 is an arbitrary constraint on rational bargaining. SFMs have been supplied with no cogent reason to prefer it to B2.

1.6 MMRC's Second Competitor

The arbitrariness of B1 is accentuated by recognizing that in addition to B2 it has other competitors. Let's introduce the notion of maximum shared gain: For each combination of strategies, take the total cooperative gain, C. Then divide C by the number of participants. Call the result "shared gain." A combination of strategies maximizes shared gain if no alternative combination has higher shared gain. Call the shared gain of such a combination of strategies "the maximum shared gain." So, for example, the shared gain of the combination of strategies, "CS1," that would result in O1 is (180+80)/2 = 130. CS1 fails to maximize shared gain since the shared gain of each other combination of strategies is 250. Now consider B3.

B3: Rational persons ought to reject, and they would expect any other rational person to reject, a given agreement outcome if each did not receive from that agreement outcome a utility equal to the maximum shared gain.

Were Abel and Mabel to act in accord with B3, they would reject all outcomes except O13, the outcome that affords each equal maximum shared gain.[23]

Since Abel is a maximizer, he might argue that he does worse by acting on B3 than he does by following B1. B1 would require him to reject the outcome prescribed by B3 and to accept the outcome that gives him 300 utiles (or dollars) and Mabel 200 utiles. Why should he make a larger relative concession than Mabel? Since Mabel is also a maximizer, she might retort that Abel places a premium on equalizing shared gains. Why is it any more rational to prefer equalizing relative concessions than it is to prefer equalizing shared gains? "But I brought to the table more than you. Why should I then settle for an equal split of the cooperative gains?" Abel might complain. Of course, Mabel should reply: "You ought to because if we didn't cooperate you'd be worse off. (Assume a universe with only two potential cooperators, Abel and Mabel.) By accepting 250 utiles you still do better than you would in the absence of cooperation."

Gauthier suggests that

> "The fundamental rationale for the principle of minimax relative concessions.....turns on an interpersonal comparison of the proportion of each person's potential gain that he must concede. However, were we to assume a measure of utility permitting interpersonal comparisons, and were we then to be tempted by some principle of equal gain, we should remind ourselves that any such temptation could be countered by a principle of equal loss, in relation to one's claim." (M by A, 139)

On Gauthier's suggestion, Abel should rejoin that rather than equalize total cooperative gains they should equalize losses in relation to their claims. But, assuming a measure of utility permitting interpersonal comparisons, to equalize such losses in relation to bargainers' claims is just to act in accordance with B1. In the Abel/Mabel case, for example, let x be the loss that is to be equalized in relation to their claims. The total cooperative gain for all combinations of strategies excluding the combination that would result in O1 is 500. Abel claims 500 less Mabel's minimum cooperative utility - that is, he claims 500-80 = 420. Mabel claims 500 less Abel's minimum cooperative utility or 500-180 = 320. If losses from claims are to be equalized, then (420-x)+(320-x) = 500. Since x = 120, losses are equalized when each accepts the outcome prescribed by B1. The point, of course, is that by countering that they ought to equalize losses in relation to potential gain, Abel has not offered a reason why it is any more rational to do so than it is to equalize cooperative gains.

In summary, it is not evident that the principle of minimax relative concession "expresses the principle of utility maximization in the context of bargaining." (M by A, 145, 151) There is nothing in the theory of SM, for example, that compels favoring B1 over B3. At least, I see no argument to show that it would be SM-irrational to adopt B3.

1.7 MMRC's Third Competitor

Shifting away from the argument in 1.4, I now want to make some general comments on Gauthier's theory of bargaining. In particular, in this section I want to question Gauthier's assumption that minimaxing is the only rational way to break a deadlock.

Since B1 is arbitrary in the sense explained, suppose Tom and Dan act on a diffferent precept, B4.

B4: A rational person ought to reject, and would expect any other rational person to reject, a given agreement outcome if that person need not obtain as small a utility as she would were she to select that outcome.

If anything, it seems B4 and not B1 is a corollary of SM. But this aside, if Tom and Dan subscribed to B4, they would deadlock. Tom would select the outcome that afforded him $100.00 and Dan nothing, whereas Dan would select the outcome that afforded him $100.00 and Tom nothing.

B4 may seem to be a non-starter. Suppose Tom and Dan are offered two unbreakable gold pieces, one worth $100.00 and one worth $101.00. Any decision they reach about who is to get which piece will be irrational according to B4.

But interestingly enough, B1 seems to suffer a similar defect: If we assume that the prebargaining payoff of each is $0, and there are only two possible outcomes - either Tom gets the coin worth $101.00 and Dan gets the other, or Dan gets the coin worth $100.00 and Tom gets the other, would B1 not countenance an analogous problem as B4? If one of B4 or B1 were true, then perhaps sometimes SFMs would be no more successful in making agreements than in escaping the PD.

Be that as it may, perplexed by their stalemated situation and cognizant of the fact that entering into an agreement is mutually advantageous, suppose star-gazers Tom and Dan decide to break the deadlock. Suppose, further, that Tom is aware of Dan's penchant for gambling and Dan is also aware of Tom's similar passion, both having frequented the same casinos in Monte Carlo. In true gambler's spirit Dan reasons with Tom: "Gauthier's a conservative. He thinks that deadlocked SFMs like us should always minimax. We're adventurous, why not flip a coin and winner take all? Besides, notice an interesting fact about our situation. If we minimax, the probability of each of us getting $50.00 is unity; if we coin-flip, the probability of one of us getting a $100.00 is 0.5, and nothing in case one of us lucks out, is also 0.5. The expected utility of minimaxing, assuming our utilities are linear with money values and assuming we either both coin-flip or both minimax, is therefore 50X1 = 50 and that of gambling is [(0.5X0)+(o.5X100)] = 50. So it is no more rational to coin-flip than to minimax. What say you we coin-flip?"

Suppose Tom and Dan do in fact coin-flip. Have they acted irrationally? I'm inclined to answer in the negative. Their case, I think, strongly suggests that what concessions it is rational to make partly depends on one's dispositions (other than the disposition to be SM-rational) and the dispositions of one's fellow cooperators. At least this much is true: SM does not preclude one from being disposed to being adventurous any more than it precludes one from being disposed to conservativism. In addition, if a batch of bargainers were disposed to gambling, then in certain situations I think it would be no less rational - SM-rational - to break a deadlock by coin-flipping than it would be by minimaxing.

1.8 A Final Problem with MMRC

I want to conclude by suggesting a final worry with Gauthier's theory that has to do with the size of the total cooperative gain. Recall, the total cooperative gain is the maximum payoff that would be available to a group of bargainers were they to act in accordance with an agreement strategy. The size of the total cooperative gain plays a crucial role in Gauthier's theory since, given their minimal cooperative utilities, it is only in relation to the cooperative gain that SFMs can advance their claims. It is then only in relation to these claims that they are able to make concessions in order to secure agreement.

What exactly determines the magnitude of the total cooperative gain in a bargaining situation? There is no set answer to this question. In the case of Tom and Dan the total cooperative gain is a gift they will receive contingent on their agreeing to a split. In other cases the total cooperative gain may be directly related to how hard each member of a group of cooperators is willing to work. For instance, in the absence of cooperation assume Larry and Mo can each prepare five cakes. If they work together they can manage fourteen. It is possible that Larry works better with Curly than he does with Mo. Although Curly may manage the same number of cakes as does Larry were each to work on his own, together they would produce seventeen. In these cases potential bargainers are cognizant of the size of the cooperative gain prior to cooperation. Furthermore, in such cases the decision to bargain in one way or another has no effect on the cooperative gain - the size of the gain is fixed at the outset. The problem bargainers face is to agree on a principle of distribution. Now it seems that there may be cases in which the very size of the cooperative gain may be influenced by knowledge of how the gain - whatever it is or will be - is to be distributed: Suppose Larry and Curly both stand to benefit from entering into an agreement. But Larry knows that Curly knows that he, Larry, will work much harder if he were to make a smaller concession than Curly. Larry just has this distinctive psychological or motivational trait. As a result of this peculiarity, assume the cooperative gain would be much larger were Larry to make a smaller concession than Curly than if they were to make equal relative concessions as required by MMRC in their particular case. To illustrate, assume for simplicity, that the initial prebargaining payoff of each is 0 units. Assume the cooperative gain (and hence the cooperative surplus in this initial case) is 500. Since each contributes equally to the production of this gain, MMRC awards to each 250 units. Suppose Larry and Curly agree not to split the gains equally; then Larry (but not Curly) works harder with a resulting cooperative gain of 1000. Suppose the two agree to divide this total so that Curly receives 300 units and Larry the rest. Assuming Curly's prebargaining payoff remains unchanged, Curly does better than he would have done had he not have agreed to unequal splits, and in consequence, had he settled for dividing the surplus of 500. Larry, in relation to the first option, does better as well. His gross payoff is 700; we may plausibly assume that his net gain is larger than 250 units. So cooperative gains do not need to be fixed by factors independent of the bargaining process itself. Knowledge of the very principle that bargainers are to use in distributing the cooperative gain may influence its magnitude. If this is all true, then in the relevant sorts of cases it may be SM-irrational for bargainers to minimax.

To conclude, I have tried to establish that an essential premise in Gauthier's argument that rational bargainers will act on MMRC is arbitrary. It is arbitrary in the sense that if this premise is construed as some kind of rational precept, there is no need to suppose that SFMs would prefer this rational precept to competitors. I have also tried to show, contrary to Gauthier, that there is little if any reason to assume that minimaxing is rational in every bargaining situation.

Are there more problems with Gauthier's ambitious contractarian project? Let's assume the bargaining problem solved: rational agreements are those sanctioned by MMRC. Assume Butch and Sundance know this. Will they escape the dilemma? Not unless they can overcome the "genuinely problematic element in a contractarian theory."

Notes

[1] See, for e.g., David Gauthier, "Justice as Social Choice," in Morality, Reason and Truth eds., David Copp and David Zimmerman (1984), Totowa, New Jersey: Rowman and Allanheld, 251-269, p. 260; David Gauthier, "Bargaining Our Way Into Morality: A Do-It-Yourself Primer," Philosophic Exchange 12 (1979), 15-27, p. 20; and David Gauthier, Morals By Agreement (1986), Oxford: Clarendon Press, p. 137.

[2] Gauthier's theory of rational bargaining is presented in Morals By Agreement, Chapter 5; "Justice as Social Choice," and "Bargaining Our Way Into Morality: A Do-It-Yourself Primer," pp. 19-20. An earlier version of the theory can be found in David Gauthier, "Rational Cooperation," Nous 8 (1974), 53-65.

[3] In Morals By Agreement, Chapter VII, Gauthier discusses constraints it would be rational to impose on prebargaining payoffs. In discussing bargaining I assume with Gauthier that prebargaining payoffs are not in question.

[4] "Justice as Social Choice," p. 260.

[5] The notion of an agreement outcome may be specified in this way: If IS is an interactive situation with n agents, then an agreement strategy is a set of n actions, one for each agent of IS. We can think of an agreement strategy as prescribing an action for each member of an interacting group of persons. An agreement strategy is effective if and only if each member of a group of interacting agents selects that strategy. A group of interacting agents acts in accord with an agreement strategy, AS, if and only if AS is effective and each agent does her AS-prescribed action. An agreement outcome is an outcome that would result were each member of a group of interacting persons to act in accord with an agreement strategy.

[6] Morals By Agreement, pp. 133-134.

[7] Morals By Agreement, pp. 154-155.

[8] See Morals By Agreement, for example, p. 136.

[9] Morals By Agreement, p. 155.

[10] Morals By Agreement, p. 137.

11. This is a simplifying assumption. In Morals By Agreement, pp. 154-155, Gauthier discusses how his theory can handle cases where there is "no single, transferable good, produced in fixed quantity and divisible at will among the cooperators."

12. Gauthier cautions:

> However, we must beware lest consideration of two person bargaining leads us to misunderstand the determination of claims. In a situation involving more than two persons, each person may not always claim all of the co-operative 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. For if someone were to press a claim to what would be brought about by the co-operative interaction of others, then those others would prefer to exclude him from agreement. (M by A, 139)

13. The example is a modification of one discussed by Gauthier in "Moral Artifice," The Canadian Journal of Philosophy 18 (1988), 385-418, pp. 390-391.

14. "Moral Artifice," p. 394.

15. "Justice as Social Choice," p. 263.

16. "Bargaining Our Way Into Morality: A Do-It-Yourself Primer," p. 20. Similar arguments appear in "Justice as Social Choice," pp. 263-264 and Morals By Agreement, pp. 141-145.

17. The rationale for (1) seems to be this: It is mutually advantageous and so rational to enter into an agreement. Bargainers can reach agreement if each agrees to choose the same agreement strategy. Agreement on the same agreement strategy can be realized only if each makes the concessions required of him by that strategy. If it is rational for any agent to make an agreement, each must believe there is an agreement strategy or a set of relative concessions that every rational person is willing to entertain. But every agreement strategy contains a relative concession at least as great as the minimax. So each person must suppose that there is an agreement strategy (or set of relative concessions) that every rational person is willing to entertain and that requires

some person to make a concession at least equal to the minimax.

See "Justice as Social Choice," p. 264 and Morals By Agreement, pp. 143-144.

8. (6) is implied by what Gauthier says in "Rational Cooperation," p. 56:

> Let us now formulate the basic condition of rational cooperation. First, a rational person will choose the greatest relative advantage compatible with that received by every other person. Second, he will reject a given relative advantage, if no person need receive such a small relative advantage. Third, he will expect any other rational person to reject a given relative advantage, if no person need receive such a small relative advantage. Hence, rational cooperation must secure an outcome affording the highest minimum relative advantage possible.....

I have extracted (6) from this passage replacing 'relative advantage' with 'relative concession' and making the other necessary changes. It must be emphasized that (6) is also implied by the discussion in "Justice as Social Choice," pp. 263-264 and by the discussion in Morals By Agreement, pp. 144-145.

19. Line (7) seems to be an assumption. See the passage cited in footnote 16 above. I question (7) in section 1.8.

20. "Rational Cooperation," p. 55.

21. Table 4.1 is a bit misleading. It assumes that utilities are linear with money values, and so interpersonally comparable. With no interpersonally comparable scale, what Abel and Mabel would do, I think, is this: To find Umax, each would locate the outcome in which the other assured his (her) partner that he (she) was getting his (her) base point. The partner would then see how much he (she) was getting there. That would be Umax for him (her). Since each would now have Umax and Umin, relative concessions could easily be calculated.

22. Morals By Agreement, p. 140.

23. I am assuming that there is an outcome, Oe, that affords each maximum equal shared gain. There may not be. VOICE may say to Tom and Dan, "I'll let you split $100.00

if you agree to have unequal shares." Perhaps this difficulty can be dealt with in some such way as this: In the absence of Oe, select the outcome nearest to Oe in the n-dimensional outcome space formed by listing the values of the possible outcomes to the interactors along the n dimensions, one dimension for each interactor's values.

CHAPTER 5

THE COMPLIANCE PROBLEM

1.1 The Issue

Gauthier tells us that

> The genuinely problematic element in a contractarian theory is not the introduction of the idea of morality, but the step from hypothetical agreement to actual moral constraint. Suppose that each person recognizes himself as one of the parties to agreement. The principles forming the object of agreement are those that he would have accepted *ex ante* in bargaining with his fellows, had he found himself among them in a context initially devoid of moral constraint. Why need he accept, *ex post* in his actual situation, these principles as constraining his choices? A theory of morals by agreement must answer this question. (M by A, 9)

He proceeds to "defend compliance with agreements based, explicitly or implicitly, on the principle of minimax relative concession." (M by A, 158) He says that

> If our defence fails, then we must conclude that rational bargaining is in vain and that co-operation, although on a rationally agreed basis, is not itself rationally required.....we must conclude that a rational morality is a chimera, so that there is no rational and impartial constraint on the pursuit of individual utility. (M by A, 158)

The classic PD nicely serves to illustrate the "compliance problem" to which Gauthier calls attention in these passages: Butch and Sundance contemplate with a sense of impending doom the matrix reproduced below.

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	1,1	10,0
Remains silent	0,10	9,9

Figure 5.1 Matrix 5.1

The numbers represent utilities with Sundance's preferences recorded first. As SFMs they know that they can do no better than attain a suboptimal outcome. The DA gives them more time. He even goes so far as to let them consult with each other. Perhaps by agreeing to do the cooperative thing, they will be able to escape their plight. But each soon realizes that such an agreement will not further their common end: Suppose the two have been convinced by Professor Gauthier's solution to the bargaining problem, in simplified terms, the problem of rationally "selecting" one of a number of non-equivalent mutually beneficial outcomes that could result from agreement.[1] Bargains reached in accord with the principle of minimax relative concession (MMRC), the solution recommends, are rational. Since this principle, according to Gauthier, impartially constrains the pursuit of direct self-interest and is "rationally justified," it also enjoys the status of being a moral principle.[2] So suppose Butch and Sundance, having been convinced by all this, agree during one of their consultation sessions to do the MMRC-prescribed cooperative thing - each will keep his lips sealed. However, although optimal, the ensuing outcome is not in equilibrium.[3] If one party does his part - if he complies with the agreement - the other party can do better by defecting. Since both are SFMs it appears that compliance with the agreement would be irrational even though making it were perfectly rational.

It is the failure of the outcome prescribed by rational bargaining as conceived by Gauthier, to exemply together, the properties of optimality and equilibrium, that contributes essentially to the compliance problem. The problem can be formulated this way: Suppose it is rational to enter into agreements which, if complied with, would result in outcomes that are fair and optimal but not in equilibrium. Why should you comply with these interest-constraining agreements?

A radio broadcast revives hope for the two felons. They hear that similarly situated prisoners are rejoicing all over. These others are celebrating Gauthier's answer to the compliance problem. That eminent philosopher has declared that it is rational to comply with the variety of agreement that is of concern to them.

But let's be more cautious than our revellers. Morals By Agreement suggests that this answer is open to a number of interpretations. Each merits scrutiny. Once we have carefully examined them, we will be in a better position to judge whether the felons really have cause for all their fanfare.

1.2 The PR-rationality of Compliance

In the introductory chapter of his book, Gauthier explains that

> A contractarian theory of morals, developed as part of the theory of rational choice, has evident strengths. It enables us to demonstrate the rationality of impartial constraints on the pursuit of individual interest to persons who may take no interest in others' interests. Morality is thus given a sure grounding in a weak and widely accepted conception of practical rationality [SM]. (M by A, 17)

The passage may be taken to indicate that Gauthier's response to the compliance problem is that compliance with the kind of agreement Butch and Sundance favor is SM-rational. But this, of course, could not be his considered view. In terms of our example, since confessing in the PD is the dominant strategy, Butch and Sundance could comply only on pain of being SM-irrational.

Gauthier's view is perhaps better explicated in the following set of passages.

> we defend compliance, not just with agreements, but with practices that would be agreed to or endorsed on the basis of the principle of minimax relative concession. (M by A, 158)

> We shall do this by demonstrating that, given certain palusible and desirable conditions, a rational utility maximizer, faced with the choice between accepting no constraints on his choices in interaction, and accepting the constraints on his choices required by minimax relative concession, chooses the latter. He makes a choice about how to make further choices; he chooses on utility-maximizing grounds, not to make further choices on those grounds. (M by A, 158)

> The received interpretation of practical rationality.....identifies rationality with utility-maximization 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-maximization 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. We shall consider whether particular choices are rational if and only if they express a rational disposition to choose. (M by A, 182-183)

Gauthier affirms that a SFM would adopt "constrained maximization, as his disposition for strategic behaviour." (M by A, 170)

These passages may leave the impression that Gauthier is defending the view that it is SM-rational merely to be disposed to comply with agreements made on the basis of MMRC. But I think his view is more interesting. For he tells us elsewhere that his theory of morals defends the "rationality of _actual_ _compliance_" with moral principles. (M by A, 182-183) To be disposed to comply with fair agreements is consistent with not complying with them on occasions when non-compliance is utility maximizing.

What, then, is the view advocated in this sequence of passages? It may be something like this (solution "PRS"): First, identify a rational disposition, D, for making choices. Next, establish the acceptability of a principle of practical reason analogous to PR.

PR: An act is rational if and only if it "expresses" D.

Finally, show that PR prescribes as rational acts of complying with agreements sanctioned by MMRC.

Notice that even if successful such an undertaking would at most sustain the PR-rationality of compliance. It would not, contrary to what Gauthier seems to have intended, ground morality "in a weak and widely accepted conception of practical rationality:" Assuming the extensional non-equivalence of SM and PR, presumably such a _bona_ _fide_ "grounding" would require establishing both the SM-rationality of making the kinds of agreement of interest to this discussion, and the SM-rationality of complying with such agreements. Let's, nevertheless, assess this interesting proposal directing attention primarily to its first component.

1.3 Constrained Maximization

Gauthier defends the position that the disposition to constrained maximization (CM) rather than to SM is the rational disposition for making choices. An understanding of his defense requires clarification of the two theories, SM and CM.[4] SM, as explained in Chapter I, is the theory that for any agent, s, and any act, a, a is SM-rational (for s) if and only if none of its alternatives has a higher expected utility (for s) than it has. CM is not as easily formulated. To facilate matters note that in parametric contexts where one's choices do not affect others' choices, SM and CM are extensionally equivalent. In strategic contexts where each interacting agent chooses her action partly on the basis of her expectations of others' choices, CM requires that

> Each person's choice must be a fair optimizing 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-maximizing response. (M by A, 157)

A fair optimizing response is

> one that, given the expected strategies of the others, may be expected to yield an outcome that is nearly fair and optimal - an outcome with utility payoffs close to those of the cooperative outcome, as determined by minimax relative concession.[5] (M by A, 157)

It appears that in strategic contexts where you expect your fellow interactors to cooperate in achieving an outcome that is fair and optimal, then provided such an outcome is possible, CM requires that you do the cooperative thing. In those strategic contexts where such an outcome is possible, but where you have no expectation of your fellow interactors cooperating to achieve it, as for instance could be the case were your fellow interactors SFMs, CM requires that you do the SM-rational thing. CM tries to ensure that those disposed to cooperate are not taken advantage of by potential exploiters. Finally, in strategic contexts where a fair and optimal outcome is not possible, CM again requires that you do what is SM-rational.[6]

1.4 The Choice Argument

Gauthier's alluring argument that the disposition to CM is rational is this:

> To demonstrate the rationality of suitably constrained maximization we.....consider what a rational individual would choose, given the alternatives of adopting straightforward maximization, and of adopting constrained maximization.....Taking others' dispositions as fixed, the individual reasons parametrically to his own best disposition.....[He ought to reason as follows:] Suppose I adopt straightforward maximization. Then I must expect the others to employ maximizing individual strategies in interacting with me; so do I, and expect a utility, u.
>
> Suppose I adopt constrained maximization.

> Then if the others are conditionally disposed to constrained maximization, I may expect them to base their actions on a co-operative joint strategy in interacting with me; so do I, and expect a utility u'. If they are not so disposed, I employ a maximizing strategy and expect u as before. If the probability that others are disposed to constrained maximization is p, then my overall expected utility is [pu' + (1-p)u].
>
> Since u'is greater than u, [pu' + (1-p)u] is greater than u for any value of p other than 0 (and for p=0, the two are equal). Therefore, to maximize my overall expectation of utility, I should adopt constrained maximization. (M by A, 172)

Although there is undoubtedly something very persuasive here, in the end I feel the argument is not utterly watertight.[7] To ascertain its shortcomings, let's recast it into a form that makes it easier to evaluate. To do so, I first list a number of assumptions that I believe it presupposes.

(1) The choice between CM and SM is to turn on the expected utility of adopting either conception. The selection of a particular conception is rational if and only if no alternative has a higher expected utility than it has for the agent in question. The expected utility of a conception of rationality is to be identified with the expected utility of the act of undertaking a permanent commitment to that conception. Suppose R and R* are conceptions of rationality and S an agent contemplating a choice between them. The expected utility of R for S is calculated as follows: Consider all the possible outcomes

of choosing R. For each, consider the utility of that outcome. For each, consider the probability of that outcome given that R is chosen. Multiply, for each, the probability and the utility. Add these products. The sum is the expected utility of R for S.

(2) The utility of choosing a conception of rationality is to depend solely on the utility of the actions required by that conception.[8] This assumption restricts the outcomes that are to be taken into account in so choosing to actions. The fact, for example, that the choice of a particular conception, R, would maximize the satisfaction of an agent's preferences in virtue of his valuing acting in accord with the prescriptions of R, is to be considered irrelevant when choosing among conceptions. (The justification for this assumption will emerge shortly in section 1.5 below.)

(3) In every situation except Prisoner's Dilemma-like (PD-like) situations of a certain kind - "salient PD-like" situations - CM and SM prescribe identical actions. A salient PD-like situation is a situation (i) with agents who are directly aware of each others' rational disposition, or who at least have a good chance in identifying the rational dispositions of others; and (ii) in which the counterfactual independence condition is violated. This is the condition that no matter what choice

either player makes, he would still make that choice no matter what choice the other makes.[9]

The rationale for this assumption needs explaining. Gauthier claims that (a)

> To choose between SM and CM a person needs to consider only those situations in which they would yield different behaviour. (M by A, 171)

And (b),

> These situations must satisfy two conditions. First, they must afford the prospect of mutually beneficial and fair co-operation, since otherwise constraint would be pointless. And second, they must afford some prospect for individually beneficial defection, since otherwise no constraint would be needed to realize the mutual benefits. (M by A, 171)

Assumption (3), it may be supposed, takes (a) into account insisting that it is only with respect to salient PD-like situations that CM and SM "yield different behaviour." In addition, it may be thought that this assumption also accomodates (b) since both the conditions therein specified are met by salient PD-like situations, as the example soon to be presented might be taken to illustrate.

Clause (i) in the specification of a salient PD-like situation reflects Gauthier's admission that his argument takes choosers to be either transparent or translucent. They are transparent when "Each is directly aware of the disposition of his fellows, and so aware whether he is interacting with straightforward or constrained maximizers." (M by A, 174) Choosers are translucent when "their disposition to cooperate or not may be ascertained by others, not with certainty, but as more than mere guesswork."[10] (M by A, 174)

The need for clause (ii) may be rationalized in this way: Persons disposed to CM adopt a <u>conditional</u> disposition that makes their strategies dependent on each other. In a PD-like situation, for instance, CM-rational Butch will cooperate with CM-rational Sundance if and only if Butch is willing to cooperate in order to secure an outcome that is fair and optimal, and he <u>expects</u> Sundance to be so willing as well. As Gauthier reminds us, "the probability of the others acting co-operatively.....[is not] independent of one's own disposition." (M by A, 172) For suppose it were, and suppose the absence of all other possible influences like causal ones that would literally make agents' actions interdependent. Then the counterfactual independence condition would be satisfied. In that case, however, a dominance argument Gauthier and others have claimed, demonstrates that SFMs choosing between SM and CM would choose the former.[11] It is in consequence - if these philosophers are right - only in cases in which the counterfactual independence condition is not met that CM reputedly has a greater expected utility for an agent than SM. It may be urged that this, for example, would be the case if each player's actions were related to the other's actions in a way indicated by the probability matrix in Figure 5.2. Numbers in the matrix indicate conditional probabilities of a choice on a choice. So, for example, the conditional probability of Sundance's (Butch's) choosing SM if Butch (Sundance) chooses SM is 0.99. The matrix shows that either player will choose a disposition if and only if the other player chooses that very same disposition. Assuming that in the future they will be in the PD-like situation with payoffs specified by the matrix in Figure 5.1, each player's utility matrix, it may be proposed, would be matrix 5.3. (This matrix appears on page 150 below.) Expected utilities calculated on the basis of the information contained in these two matrices yield the desired results: The expected utility of SM for

	Butch chooses	
	SM	CM
Sundance chooses		
SM	0.99	0.01
CM	0.01	0.99

Figure 5.2 Matrix 5.2

each player would be = (1)(0.99) + (10)(0.01) = 1.09; whereas that of CM for each would be greater, (0)(0.01) + (9)(0.99) = 8.91.

	Butch chooses	
Sundance chooses	SM	CM
SM	1,1	10,0
CM	0,10	9,9

Figure 5.3 Matrix 5.3

Given assumptions (1), (2), and (3), in a world with no salient PD-like situations the expected utility of SM for an agent and that of CM for that agent are identical.

(4) (a) In every salient PD-like situation CM and SM prescribe different actions. (b) In a world with one (or more) salient PD-like situations, the expected utility of CM for an agent is greater than the expected utility of SM for that agent.

(5) Prior to choosing between CM and SM choosers know that they will be in a world with at least one salient PD-like situation.

On these assumptions it appears that the choice of CM is SM-rational. Gauthier's argument for this, now recast, is simple:

(1) Chooser's choice (Chooser is a SFM) of a conception of rationality is rational if and only if none of its alternatives has a higher expected utility than it has for Chooser.

(2) Among Chooser's choices, the expected utility of CM for Chooser is higher than the expected utility of SM for Chooser.

(3) Therefore, Chooser's choice of CM (and not SM) is rational.

1.5.1 The First Problem with the Choice Argument

Is this "choice argument" sound? I don't think so. Both premises are controversial.

First, assume premise (2) is true. Then despite its being the case that SM's expected utility for Chooser is not as great as that of CM's for Chooser, a SFM like Chooser should not select CM over SM. Premise (1) should be rejected. This is so since the premise fails to take into account the implications of the fact that rational choices are not made from a "rationality-neutral" perspective, but that they presuppose some such perspective: Consider what a choice of CM over SM would involve from the point of view of a SFM. It is obvious that a person's preferences in choosing a conception of rationality will be influenced by her knowledge of what that conception entails.[12] In her deliberations on whether to choose CM a significant and relevant piece of information is that sometimes CM will require that she restrict the pursuit of her own wants. Such restraint is required if one is to carry through with the kind of agreement that makes CM attractive. But to restrict the pursuit of her own wants is, for a SFM, to act contrary to reason. So available to such a maximizer will be the information that there will be occasions when CM requires her to act contrary to reason. But then it seems irrational for a SFM to adopt as her very conception of rationality, one that will sometimes require that she act contrary to the prescriptions of SM. Similarly, if one's initial conception of rationality were CM, it would be irrational to choose SM over CM. This would be so not because SM as of a particular time of choice, has a lower expected utility (as we are now supposing) than CM. Rather, the reason would be that when certain opportunities present themselves, as perhaps in some types of PD with trustworthy others, SM would prescribe exploitation. Such behavior is rationally unacceptable to a CM.

Friends of the choice argument may reply in this way: "It's true that SM requires that on each occasion of choice, you choose the alternative that is best-for-you as of that time. If you are a SFM and you want to choose rationally, you are commited to choosing in this way - you cannot choose otherwise. Maybe all this provides you with some reason to refrain from choosing CM - you know, for example, that if you now choose CM you won't be able to exploit gullible others in PDs. But if you are moved by this consideration, it can only be because you place a premium on acting consistently with SM over time. However, the choice of CM (by assumption) is utility maximizing. Forget about considerations of "temporal consistency." Think only about utility!"

The right response here, I think, should be this: "It might be true that the SFM places a premium on acting consistently with SM over time. After all, to be a SFM just is to do the best for yourself on each occasion of choice. But suppose, when confronted with the choice of deciding between SM and CM, you then choose CM. Then you must place a premium on maximizing utility as of <u>that</u> time: As a SFM you well know that your choice of CM as of then will sometimes prevent you from doing the utility maximizing thing in the future; in some PDs to come, you will have to forgo maximizing your advantage. So if you choose CM, you must place a premium on maximizing utility on one particular occasion of choice. You must think that it is more important to maximize utility on this occasion of choice, than it is to maximize utility on each (future) occasion of choice. Why is it any less rational, though, to place a premium on maximizing utility at each choice-point in time, than it is to place a premium on maximizing utility at one choice-point in time?"

The approach I have attributed to Gauthier in deciding among conceptions of rationality has a semblance of plausibility but is misleading. Its apparent strength derives from the assumption that the utility of CM for Chooser is greater than that of SM for Chooser. Chooser, as a SFM (and indeed as a constrained maximizer), need not (although she should as we will soon see) deny this. It isn't the case, as Gauthier reminds us, that the constrained maximizer "taking a larger view than her fellows" reasons "more effectively about how to maximize her utility," but she reasons in a <u>different way</u>.[13] (M by A, 170) From the point of view of a constrained maximizer, the SFM, in those PD situations in which he violates his rationally entered into agreements, acts irrationally. From the point of view of a SFM, the constrained maximizer in these same situations in which he complies with his rationally made agreements, acts irrationally. Premise (1) is insensitive to important differences to which Gauthier himself calls attention in the way in which these two kinds of maximizer reason.

1.5.2 A Possible Second Problem

This rejection of premise (1) may fail to convince the reader. Even so, Gauthier's argument still suffers defects. That is because whether it is sound partly depends on the cogency of assumptions (1) to (5). Both assumption (2) and assumption (4), I believe, are suspect. Let's consider each in turn. In discussing the second, I proceed as if the fourth is beyond reproach.

Assumption (2) may impose an arbitrary restriction on the outcomes that are to be considered in choosing among conceptions of rationality. Why restrict as it does the relevant outcomes to actions? An SM-rational agent, for instance, may value being a SFM just as she may value being charitable or kind. Choosing SM would maximize the satisfaction of this particular preference. To elaborate, assume that some SM-rational agent values acting in accord with the prescriptions of SM. If such an agent were to choose CM, she would know that in salient PD-like situations she would have to act SM-irrationally: assumption (4a) tells us that in such situations SM and CM prescribe different actions. As we have already remarked, from the point of view of a SFM, a constrained maximizer acts irrationally in some situations of this type. In light of this knowledge and given her values, the choice of

SM and not CM, it would seem, would maximize the satisfaction of her preferences.

Consider, secondly, a SFM whom we shall call "Reasonable." Reasonable values acting on the basis of rational precept RP.

RP: If S ought (SM-rationally) to do a at time t1, and S cannot do a at t1 without doing b at an earlier time t0, S ought (SM-rationally) to do b at t0.

Prior to choosing between SM and CM, suppose Reasonable knows that he will find himself in a salient PD-like situation. Suppose the SM-rational thing to do in such a situation is to confess whereas the CM-rational thing to do is to remain silent. In order both to act SM-rationally and to act consistently with RP in the dilemmatic situation to come, Reasonable must choose SM: If he were to commit himself to CM he could not then act consistently with these preferences. Yet again, choice of SM would maximize the satisfaction of Reasonable's preferences.

It is important not to loose sight of the fact that a requirement of SM is that one consider <u>all</u> possible outcomes of relevant alternatives in one's deliberations. So preferences like the ones Reasonable has cannot be discarded in deciding on the rational course of action in the absence of sound justification. What could this justification be, though, given that SM, at least in the fashion interpreted by Gauthier, does not constrain the content of agents' preferences?[14]

Gauthier would probably reply that I have muddied the waters here by adding extra payoffs in addition to those to be gained by action in the PD-like situation. He may readily concede that certain individuals like Reasonable with certain "nontraditional values" would not choose CM over SM. They would choose conversely. They would so choose because their idiosyncratic values would ensure that their choice situation is no longer any interesting variety of PD.

In light of this rejoinder, let's stipulate that there are no hidden payoffs: The numbers in matrix 5.1 exhaust everything that the choosers have at stake. But even with this proviso the choice argument, I think, does not work. It fails because assumption (4) is questionable. In fact, I think both its parts are false.

1.5.3 A Third Problem with the Choice Argument

Reconsider the case of Butch and Sundance. Assume that the PD-like matrix in Figure 5.1 correctly portrays the value of each possible outcome for each of these agents. Assume, in addition, that the two felons are transparent so that matrix 5.2 is an accurate depiction of how the actions of each are related to the actions of the other. Assume, further, that prior to being in the PD-like situation Butch and Sundance agree to do the cooperative thing. In this example, to do the cooperative thing is tantamount to keeping silent. Since each is transparent each will be aware of whether the other is disposed to cooperate. Transparency ensures that cheating is not possible. (M by A, 173-174)

If each is a constrained maximizer each will cooperate. Each will honor the pre-PD agreement and will receive 9 points. Now consider what will happen if each is straightforwardly rational. The felons know that they cannot cheat. Transparent Butch, for instance, cannot make the agreement with the intention of later failing to carry through. He cannot because being transparent, Sundance would know of the intention to be dishonest and would, from the start, refuse to make the agreement. Moreover, this must be emphasized: If all Butch knew about Sundance is that Sundance is disposed to CM, Butch would be foolish, merely on the basis of this information, to do the cooperative thing. As we saw earlier, to be disposed to CM is consistent with doing the utility maximizing thing on various occasions. If Butch is not to be deceived - if as of a particular time he is to have realistic expectations of achieving a fair and optimal outcome, he must know something over and above Sundance's rational disposition. He must know something about Sundance's intentions, and something about Sundance's intentions to act on those intentions. So the assumption of transparency is a strong one. We must assume that not only does transparency disclose the rational dispositions of agents, it also discloses agents' intentions to act in certain ways on specific ossasions of choice, and at least the likelihood that they will act on these intentions on those occasions. Now each knows that if he keeps silent so will the other and if he confesses so will the other. Such an "interdependency" of actions is an interesting consequence of transparency. So if each does the cooperative thing and remains silent each can expect 9 points. If each confesses each can expect 1 point. It would therefore be straightforwardly rational for each, in this PD-like situation, to remain silent.

Suppose, now, that Butch and Sundance are choosing between SM and CM. Suppose they are told that they will find themselves in a world with precisely one salient PD-like situation, a situation exactly like the one just described. The discussion in the last paragraph should make it clear that their utility matrix is matrix 5.4. (See page 160.) Expected utilities calculated by assuming probabilities recorded in matrix 5.2 tell us that each will fare no better as a constrained maximizer than as a SFM.

	Butch chooses	
	SM	CM
Sundance chooses		
SM	9,9	9,9
CM	9,9	9,9

Figure 5.4 Matrix 5.4

It won't help to assume that the PD situation in which Butch and Sundance anticipate finding themselves will be an authentic one in which the counterfactual independence condition is satisfied: In such a situation SM counsels confessing. But so does CM. In an authentic PD, we cannot assume any interdependency of action that could, for example, result from the supposition that agents are transparent. Introduce such interdependency and the counterfactual independence condition is violated. Then the actions of one partly "determine" the actions of the other. The resulting PD-like situation would consequently not be a real PD; it would in fact not be any different from the dilemmatic situation in which we first supposed the two felons to be. So assume, in this second case, that Butch and Sundance are not transparent - or at least assume that each has no idea about the rational disposition of the other. If agents' actions are not interdependent in any such way as they would be if agents were transparent, then the relevant utility matrix - I think - would be matrix 5.5. Each as a SFM would confess and would receive 1 point. Each as a constrained maximizer would also confess and would receive 1 point. Hence, no matter what your rational disposition, and no matter what the rational disposition of your partner, you would do best by confessing. The result once again is that SFMs do not do better than constrained maximizers in such authentic PDs.

	Butch chooses	
	SM	CM
Sundance chooses		
SM	1,1	1,1
CM	1,1	1,1

Figure 5.5 Matrix 5.5

To summarize results, in both inauthentic PD-like situations and in genuine PDs in which deception is not possible, transparent SFMs will do just as well as their constrained cousins. In the former sort of dilemmatic situation both types of maximizer will "cooperate." In the latter sort of situation they will refrain from cooperating. So it seems unlikely, as Gauthier believes, that "because they differ in their dispositions, straightforward and constrained maximizers differ also in their opportunities, [to cooperate] to the benefit of the latter." (M by A, 173)

1.5.4 A Fourth Problem with the Choice Argument

Finally, it's worth observing that assumption (5) forces a certain construal of the claim that CM is the rational disposition for making choices: Perhaps there are some dilemmatic situations in which a constrained maximizer would do better than a SFM. Assume there are. Even so, a SFM who expected to find himself in a world devoid of these kinds of situation would do no better by adopting CM than by adopting SM. If it is Gauthier's intention to convince any SFM - even a one like Reasonable - that she would do better under any conditions by becoming a constrained maximizer, the choice argument won't help him.

To skirt this difficulty one might retreat to the more cautious claim that it is SM-rational for agents with certain values in certain situations to choose CM over SM. I have no quarrel with this weaker view. But if the very justification for CM consists in showing that any SM-rational agent would do better by choosing CM over SM, then it is the stronger and not the weaker claim that needs to be sustained. Similarly, if SM is to be abandoned for the reason that SM-rational agents would not choose SM over a competitor,[15] it is once again the stronger claim that needs to be upheld: Why suppose a principle of rationality defective if choice of it is SM-rational for agents with certain values in certain situations, but not so rational for agents with a non-equivalent set of values in a different set of situations?[16]

The choice argument, it appears, is not free of worries.

1.5.5 A Problem with PR

Assume, contrary to what has just been concluded, that the disposition to CM is the rational disposition for making choices. Then the principle of practical reason, PR, appealed to by the second componet of PRS is to be interpreted as claiming that

An act is rational if and only if it "expresses" the disposition to constrained maximization.

Barring questions of just what PR so interpreted amounts to, Gauthier presents no argument in its defense.[17] He asserts that

> If one's dispositions to choose are rational, then surely one's choices are also rational. (M by A, 186)

But mere assertion constitutes no reason for accepting PR. Gauthier's defense of solution PRS is at best incomplete.

1.6 The CM-rationality of Compliance

Although the choice argument does not succeed it is evident that CM as a theory of practical rationality enjoins compliance, under certain conditions, with fair optimal practices. So perhaps the solution to the compliance problem being sought is that it is CM-rational to comply with agreements that are rationally entered into.

In "Reason and Maximization," however, Gauthier tells us that CM is a moral principle.[18] Suppose, in addition, that CM is "rationally justified." Suppose it is interest-constraining, as it indeed seems to be; it requires compliance, under specific conditions, with agreements based on the principle of minimax relative concession. Suppose, finally, that it is impartial. Gauthier seems to rely on different accounts of impartiality. Sometimes he says that "a joint strategy.....is impartial because it is acceptable from every standpoint, by every person involved." (M by A, 151) At other times, he relies on a Rawlsian notion of impartiality - principles are impartial just in case they are selected from an "Archimedean Point" behind a "veil of ignorance."[19] If these suppositions are all true, then on the criteria proposed in Morals By Agreement, CM would again qualify as a moral principle. It would in fact be a moral principle that is also a principle for rational choice. But even if CM had this attractive feature, Gauthier would still encounter a difficulty: The rationality of compliance and so the rationality of being moral would on the present consideration be sustained on moral grounds. This would defeat Gauthier's project of generating morality "as a rational constraint from the <u>non-moral</u> premisses of rational choice." (M by A, 4, my emphasis.) Assume, then, that CM is not a moral principle so that this worry does not arise. This would not, unfortunately, permit Gauthier to circumvent a new problem: SM-rational agents will have been given no reason whatsoever to comply with agreements they rationally make. (Remember, the choice argument falls short of accomplishing what it is meant to.) Why should a SFM be in the least moved by being apprised that it is CM-rational but not SM so to comply with fair optimal practices?

A way out would be to argue that SM suffers a defect not shared by CM, that SM maybe, fails to satisfy a criterion any adequate principle of rationality must meet. "Reason and Maximization" intimates such a criterion. It's the criterion that a theory of rationality is adequate only if "self-supporting." Several versions of this criterion are amplified and evaluated in the next chapter.

For now, we conclude that on its own the choice argument fails to resolve the compliance problem. Perhaps the merrymaking of our felons is premature.

Notes

[1] Gauthier's theory of bargaining is presented in Chapter V, in David Gauthier, Morals By Agreement (1986), Oxford: Clarendon Press. We discussed this theory in the last chapter.

[2] Gauthier claims that "our concern is to validate the conception of morality as a set of rational, impartial constraints on the pursuit of individual interest." (M by A, 6)

[3] A Nash equilibrium outcome is the product of a set of actions, one for each interacting person, such that for each such person, there is no alternative action that this person would prefer, the actions of all the others being fixed.

[4] I think it is misleading to take CM to be a rational disposition. CM, like SM, is a theory of rationality. Just as one can be disposed to act in accord with SM, so one can be disposed to act in accord with CM.

[5] Gauthier explains that

> We speak of the response as nearly fair and optimal because in many situations a person will not expect others to do precisely what would be required by minimax relative concession, so that he may not be able to choose a strategy with an expected outcome that is completely fair or fully optimal. But we suppose that he will still be disposed to co-operative rather than to non-co-operative behaviour. (M by A, 157)

[6] If something more formal is desired in the characterization of CM, perhaps the following will suffice: Let's suppose that each act is either performed in a parametric context, in which case it is a "P-act," or in a strategic one, in which case it is an "S-act." Then CM can be formulated in this way:

CM: For any agent, R, and any act, a, a is CM-rational if and only if

(a) if a is a P-act, then a is SM-rational.
(b) if a is an S-act, then

(i) if there is an outcome, o, such that o is the product of a set, A, of actions, one for each interacting person, o is fair and optimal, and R expects each interacting person to be ready to cooperate in achieving o, then a is a member of A and R's doing a, provided the other interactors did the cooperative thing, would result in o; or
(ii) if there is an outcome, o1, such that o1=o in (i), and R does not expect each interacting person to cooperate in achieving oi, then a is SM-rational; or
(iii) if there is no outcome, o2, such that o2=o in (i), then a is SM-rational.

7. I here part company with what many have said on the choice argument: Richmond Campbell (Richmond Campbell, "Gauthier's Theory of Morals by Agreement," forthcoming in Philosophical Quarterly) believes that transparent SFMs, choosing between SM and CM, would choose CM. Edward McClennen expresses a similar belief in his paper "Constrained Maximization and Resolute Choice," forthcoming in Social Philosophy and Policy. While Peter Danielson takes the disposition to "reciprocal cooperation" and not the disposition to CM to be the rational disposition to adopt (Peter Danielson, "The Visible Hand of Morality," Canadian Journal of Philosophy 18 (1988), 357-384, pp. 375-378), he seems to agree that when choice is limited to CM and SM, the choice of CM is utility maximizing. Gregory Kavka says that it "may will be true" that CM is more rational than the disposition to maximize expected utility. (Gregory Kavka, "A Review of Morals By Agreement," Mind 96 (1987), 117-121, p. 120) L. W. Sumner thinks that Gauthier "has made a very strong putative case for thinking" that "a utility maximizer should not be a straightforward maximizer." (L. W. Sumner, "Justice Contracted," Dialogue 26 (1987), 523-548, p. 544)

Fred Feldman, in contrast, characterizing SM "in a slightly unorthodox way" believes that SFMs won't do worse than constrained maximizers in PD-like situations. See Fred Feldman, "On The Advantages Of Cooperativeness," forthcoming in Midwest Studies in Philosophy, especially section 7.

8. As Gauthier claims, "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" (my emphasis). See Morals By Agreement, pp. 182-183.

[9] I owe this formulation of the counterfactual independence condition to Professor Fred Feldman. See Feldman, "On The Advantages Of Cooperativeness," section 3.

[10] Gauthier is sensitive to the charge that to assume transparency is to rob the argument of much interest as it is not realistic to suppose that "real world" persons are transparent. For this reason, he explores the merits of the argument when the choosers in question are "translucent." I believe the argument has problems even when transparency is assumed. I will therefore restrict discussion to the ideal case in which this assumption holds.

For a critical discussion of the choice argument in "real world' contexts, see Richard J. Arneson, "Locke versus Hobbes in Gauthier's Ethics," Inquiry 30 (1987), 295-316, section V, pp. 304-315.

[11] See Morals By Agreement, p. 173 and Daniel M. Farrell, "Hobbes As Moralist," Philosophical Studies 48 (1985), 257-283, pp. 272-273. I remain unconvinced by these arguments. I think that in PDs in which the counterfactual independence condition is satisfied and in which deception is not possible, SFMs choosing between SM and CM would choose either. My reasons for so thinking are presented in section 1.5.3 in this chapter.

[12] Gauthier agrees with this. See, for example, David Gauthier, "Reason and Maximization," Canadian Journal of Philosophy 4 (1975), 411-432, p. 415.

[13] Morals By Agreement, pp. 169-170.

[14] Morals By Agreement, pp. 25, 34, 48.

[15] Gauthier suggests this in "Reason and Maximization," pp. 429-430; in David Gauthier, "The Irrationality of Choosing Egoism - A Reply to Eshelman," Canadian Journal of Philosophy 10 (1980), 179-187, pp. 184-185, and in Morals By Agreement, pp. 183-184.

[16] The interesting issue of assessing theories of practical rationality will be considered in much more detail in the next chapter.

[17] David Copp and Richmond Campbell draw similar observations in David Copp, "Contractarianism And Moral Scepticism," forthcoming and in Richmond Campbell, "Gauthier's Theory of Morals by Agreement," respectively. See, also, Holly Smith's discussion in her paper "Gauthier's Moral Contract," forthcoming.

18. In "Reason and Maximization," p. 432, Gauthier says that "The policy of agreed optimization [i.e. the policy of CM] may be identified with morality."

19. See David Gauthier, "Moral Artifice," Canadian Journal of Philosophy 18 (1988), 385-419, section 6, and Morals By Agreement, Chapter VIII, for Gauthier's account of Rawlsian or "Archimedean" impartiality.

David Copp conducts an interesting discussion of Archemedean impartiality in "Contractarianism And Moral Scepticism," section 5. One conclusion he there argues for is this:

> [U]nless everyone in society is roughly equal in power and productivity, schemes or arrangements that would pass the test of contractarian rationality [i.e. schemes or arrangements like the Lockian Proviso or the principle of minimax relative concession that would be rationally agreed to (let's suppose) by SFMs] would not pass the test of impartiality [they would not be selected by rational agents behind a Rawlsian veil of ignorance], at least not without qualification, and schemes that pass an unqualified impartiality test would not pass the rationality test. (pp. 28-29)

CHAPTER 6

ASSESSING THEORIES OF RATIONALITY

1.1 Introduction

In Morals By Agreement David Gauthier expresses an intriguing view about rationality when he claims that

> 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. (M by A, 183)

In a similar vein, in "Reason And Maximization" Gauthier affirms that

> 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. A person who is unable to submit his conception of rationality to critical assessment, indeed to the critical assessment which must arise from the conception itself, is rational in only a restricted and mechanical sense.[1]

The notion of subjecting our beliefs and acts to critical scrutiny is not too hard to grasp, but what about the notion of subjecting our standards of critical scrutiny themselves to critical scrutiny? The trouble with the latter notion seems to be that in assessing our standards we cannot appeal to any higher authority - for there is none higher than these very standards themselves. Nevertheless, as is evidenced by the passages just cited, Gauthier not only believes that it is possible to raise questions about the ultimate adequacy of these standards but that "full rationality" demands that one do so.

In this chapter I offer a number of interpretations of what I consider to be Gauthier's test for assessing theories - at least formally adequate ones - of rationality. I argue that none is cogent.

1.2 Two Sorts of Evaluative Test

It is important to distinguish between two different notions of an evaluative test for principles of practical reason. According to the "criterial notion" the test specifies a necessary condition any formally adequate theory of practical reason must satisfy. In contrast, the "comparative test" allows us to determine which of two competing theories, if any, is better. A theory's being better than a competitor is compatible with its failing to meet a necessary condition of adequacy for any such theory. I distinguish these two notions since both, I believe, are suggested in the relevant texts by Gauthier. What I call the "choice interpretation" of the evaluative test is itself succeptible to two interpretations. It can be construed as propounding either a criterial or a comparative test for assessing theories of rationality.

The "self-referential" interpretation, unlike the first, explicitly enunciates a criterial test.

To formulate the various interpretations, concede the intelligibility of the claim that a person's conception of rationality is something he can rationally choose to change. Assume that to choose among conceptions of rationality is simply to choose among acts of undertaking a permanent commitment to one of them. With this in mind we can now attempt to elucidate the variants of the choice interpretation.

1.3 A Criterial Test

Gauthier explains that a SFM, choosing among conceptions of rationality, will elect to abandon egoism in favor of an alternative, "constrained maximization" (CM).

> To demonstrate the rationality of suitably constrained maximization, we solve a problem of rational choice. We consider what a rational individual [i.e. an SM-rational individual] would choose, given the alternatives of adopting straightforward maximization, and of adopting constrained maximization, as his disposition for strategic behaviour. Although this choice is about interaction, to make it is not to engage in interaction. Taking others' dispositions as fixed, the individual reasons parametrically to his own best disposition. (M by A, 170-171)

> [W]e suppose it possible for persons, who may initially assume that it is rational to extend straightforward maximization from parametric to strategic contexts, to reflect on the implications of this extension, and to reject it in favour of constrained maximization. (M by A, 183-184)

In these passages Gauthier appears to be proposing that non-equivalent substantive theories of practical reason are to be evaluated on the basis of a rational choice among them. The favored alternative(s) is the one the choice of which is rational.

As a criterial test, CT, the choice interpretation says this:

CT: A principle of rationality is adequate only if choice of it (presumably by any agent) is rational.

CT faces an obvious objection. A rational choice of a theory of rationality is impossible without assuming some theory, itself an adequate one, as a basis for such a choice.

The force of this objection is better appreciated once it is elaborated into a dilemma. Let the theory under evaluation be T1. If choice of T1 is to be rational we must assume some theory of practical reason, T*, as a basis for our choice. Now either T* is identical to T1 or it is not. Suppose it is. Then CT is a self-referential criterion of adequacy. This criterion appropriately formulated, though, is flawed as I argue below. So CT itself is flawed. Suppose, alternatively, that T* is not identical to T1. Then T*'s ultimate adequacy may itself be questioned. On what basis do we establish _its_ credentials as an acceptable theory of practical reason? On the one hand, T*'s adequacy cannot be assumed since the very purpose of the test is to provide a criterion of adequacy for theories of rationality. On the other hand, a consistent application of the evaluative test now under consideration requisitions a rational choice of T* on the basis of some other non-identical theory, T**. But then of course the adequacy of T** itself may be queried. On pain of regress, it seems best to bow to this horn.

1.4 A Comparative Test

Although CT fails it's still open to a defender of the choice intepretation to hold that this interpretation enunciates a comparative test for theories of rationality. The attractions of such a test, it may be claimed, are evident: A comparative test avoids the unpromising or perhaps even impossible task of discovering and formulating criteria of adequacy for theories of practical reason. For what else other than our intuitions could we appeal to to discover such criteria if indeed there is anything to discover? But it is patent, it may be held, that our intuitions are unreliable when it comes to making such discoveries.[2] The most we can hope for in the face of competing theories like SM and CM is to try to ascertain which of them is better, and to attempt this by avoiding any appeal to our questionable intuitions.

To formulate the choice interpretation so that it captures the notion of a comparative test, let's limit our choice to two non-equivalent theories, T1 and T2. Then the comparative criterion, CC0, recommends that

CC0: T1 is better than T2 if and only if a choice of T1 is rational.

But rational on what basis? Again, there appear to be two possibilities. The choice of T1 may be rational on the basis of either one of T1 or T2, or T3, on the supposition that T3 is identical to neither T1 nor T2.

The second alternative, however, isn't really one since by assumption there are only two theories of choice, T1 and T2. We are assuming that there isn't a further theory that can be used as a basis for choosing between them. Even if there were, that theory would have to be sound: rational choices cannot, unless fortuitously, be forthcoming from defective theories of rationality. If sound, its being so could presumably be confirmed by a set of criteria of adequacy for theories of rational choice. But a powerful impetus for propounding a <u>comparative</u> test as opposed to a criterial one, derives from scepticism about there being any such set, and even if there were, of our being able to discover it.

So suppose T3 is identical to either T1 or T2. Then the test would have to be construed in this fashion:

CC1: T1 is better than T2 if and only if choice of T1 is rational given one of T1 or T2 as a basis for choice.

CC1 encounters a difficulty: Suppose the choice between T1 and T2 is to be made on the basis of T1. Then it seems that the supposition begs the question in favor of T1. It seems to presume that T1 is better than T2 for purposes of choosing between them. The arbitrariness of this sort of manoeuvre should be clear. One might attempt to resolve the problem in this manner: The context of choosing among conceptions of rationality, Gauthier tells us, is a parametric one. Assume the extensional equivalence of T1 and T2 in parametric contexts. Then these theories prescribe the very same actions in such situations of choice. Hence to use either as a basis for rationally choosing between them is not to presume one superior to the other and is therefore not to beg the question in favor of either.

This solution to the problem, however, incurs costs. The most obvious is that the criterion of adequacy for theories of rationality as summarized by the choice interpretation must now be construed as a restrictive and not a general one. The test limits itself to theories like SM and CM that are extensionally equivalent in parametric contexts and has no application whatsoever to theories that may not be so equivalent.[3] It may be possible to show, by independent argument, that all "maximizing" theories of rationality like SM and CM are in fact extensionally equivalent in parametric contexts so that there is no worry here. But then again it may not. In that case the objection needs to be defused.

Here's a suggestion: Why not first compare T1 and T2 relative to T1, then relative to T2, and then evaluate on the basis of CC2.

CC2: For any theories, Ti and Tk, Ti is better than Tk if and only if it is better as judged both by Ti and Tk.

Hence, if T1 and T2 are extensionally equivalent in parametric contexts, then one is chosen or they tie. If they are not so equivalent, then they tie.[4]

There is a problem, however, with this suggestion as well. The problem, by the way, would persist even if it could be shown that all "maximizing" theories are extensionally equivalent in the relevant context. In a nutshell, the problem is that one and the same theory of rationality may recommend different courses of action depending on the characteristics of the choice situation. The choice of theories of rationality, supposing a certain theory as a basis for choice, is not invariant across all parametric choice contexts. As we saw in the last chapter,[5] and indeed as Gauthier himself emphasizes, under certain conditions SM prescribes the choice of SM; under others, it may prescribe CM.[6] Confronted by this observation one must, I think, conclude one of two things. Either CC2 fails; or, eschewing this verdict, one must assume the burden of explaining why certain choice situations and not others are relevant to the testing of theories of rationality. We shall have more to say on this below.[7]

1.5 An Absolutist Self-support Test

Earlier works, in particular "Reason and Maximization," lend support to another criterial test for conceptions of rationality. Roughly, the idea is that theories of rationality are to be assessed self-referentially.

After explaining that a rational egoist will reject egoism in favor of CM, Gauthier tells us that

> straightforward maximization is not self-supporting; it is not rational for [a SFM] to choose to be a straightforward maximizer.[8]

On the plausible assumption that to abandon a theory of rationality is to reject it as untenable, the passage suggests that egoism is defective because it is not self-supporting. This, in turn, suggests that a necessary condition of a principle of rationality's being adequate is that the principle is self-supporting.

The notion of self-support is introduced by Gauthier in this way:

> It is rational to choose a conception of rationality if, given that conception of rationality, it is rational to choose it.[9]

The notion is ambigious. It can be construed in either an "absolutist" or a "relativist" fashion. According to the former,

ASS: Principle of rationality, R, is self-supporting =df. for any agent, under any conditions, the choice of R would be permitted by R.

According to the latter,

RSS: Principle of rationality, R, is self-supporting =df. for any agent, under certain conditions, the choice of R would be permitted by R.

It appears that it cannot be to the ASS conception of self-support that Gauthier subscribes: CM, unlike SM, is meant to be self-supporting. But on the absolutist conception, there are conditions under which this is not true. Here's one. Suppose a terrorist puts a gun to your head and threatens: "If you adopt CM, I'll kill you."

Under these conditions, CM prescribes that you ought not to adopt CM. So CM is not absolutely self-supporting.

1.6 Relativist Self-support Tests

Perhaps the self-referential principle endorsed by Gauthier is SR1:

SR1: A principle of rationality, R, is adequate only if R is relativistically self-supporting ("RSS").

SR1 is singularly unhelpful since it leaves unspecified the conditions under which adequate principles of rationality are to be self-supporting. To isolate these conditions, it will be helpful to remind ourselves that in situations of the relevant kind, Gauthier suggests that SM but not CM is not self-supporting.

First, notice that for agents with a particular psychological profile, SM may not fail in the required way.[10] This becomes evident if we reflect on rational agent Tom's disposition to maximize. Tom is a tenaciously ardent SFM. Ardent SFMs value being SFMs; they value being the sort of person that has as her conception of rationality SM. Just as one may value being generous, ardent SFMs value acting in accord with the prescriptions of rational egoism. Tenaciously ardent SFMs value being SFMs much more than they value being any other sort of maximizer. Presented with the choice between any other maximizing conception of rationality and SM, a tenaciously ardent SFM would choose to remain such a maximizer.[11]

In light of this counterexample, SR1 may be amended by a slight modification:

SR2: A principle of rationality, R, is adequate only if (i) given that R is one's principle of choice, and (ii) one does not value being an R-rational person, it is rational to choose R.

If one is a SFM, for example, clause (ii) of SR2 is to be construed as saying that one does not value being a SFM.

SR2 suffers at least two serious defects. To explain the first we need to distinguish between "teleological" and "non-teleological" theories of rationality. _Teleological theories_ place no constraints on the ends of rational agents or on what rational agents may value. SFMs, for example, espousing a teleological theory - SM - may value the happiness of others, cooperativeness in PDs, or being SFMs.[12] _Non-teleological theories_, in contrast, are ends-constraining branding as irrational certain ends. Some such theory, for instance, may rule that it is irrational to value being a tenaciously ardent SFM.

In adducing a principle of adequacy for a class of theories of rationality, it is desirable that the principle be applicable to every member of this class. Without such generality of applicability it would fail to be a principle of much interest. In particular, it should not turn out in the absence of any justification to the contrary, that such a principle preclude assessing theories of the relevant kind _merely_ in virtue of the values of those holding these theories. This is especially true if the theories in question are teleological - such theories do not in any way constrain the values of rational agents. It should be evident that SR2 fails in this respect. Since Tom is a tenaciously ardent SFM, SR2 cannot be used by him to evaluate SM. Suppose that unlike Tom Dan does not value being a SFM though he is such a maximizer. SR2 is then available to him as a principle of critical assessment. Given Tom's values SM is self-supporting. Given Dan's values it may not be self-supporting and hence Dan but not Tom may have to conclude that SM is defective. Suppose, now, that Tom undertakes "cognitive conversion therapy" and as a result comes to loathe being a SFM. In using SR2 to assess SM, he may now have to conclude like Dan that it is inadequate since it is not self-supporting. How can the mere fact of changing one's values call into question a _teleological_ theory of rationality? Of course, the kind of problem I am trying to articulate would not arise if the theories under scrutiny were non-teleological: according to these theories, it might well be irrational to value SM as Tom does.

It may be objected that this worry can easily be dispelled by stipulating that the utility of choosing a conception of rationality depend entirely on the utility produced by the actions required by the relevant conception.[13] So SR2 should be rejected in favor of SR2*.

SR2*: A principle of rationality, R, is adequate only if (i) given that R is one's principle of choice, and (ii) the utility of choosing a conception of rationality is to depend solely on the utility of the actions required by that conception, it is rational to choose R.

The trouble with SR2* is that clause (ii) lacks justification. Why accept the restriction it imposes especially if one is a tenaciously ardent R-rational person?

Alternatively, it might be rejoined that the problem would be avoided by wording SR2's right hand side counterfactually.

SR2CF: A principle of rationality, R, is adequate only if were R one's principle of choice for conceptions of rationality, and were one a person who did not value being R-rational, it would be rational to choose R.

SR2CF avoids the original worry only at the expense of having no relevance at all in assessing teleological theories given persons' actual values. Why be impressed by the results of an evaluative test if it requires us to indulge in the fiction that we are fundamentally different, in the relevant respect, from what we actually are?

I said above that SR2 has at least two defects. Its second defect (which also afflicts SR2*) is that it cannot be Gauthier's principle of self-support: On Gauthier's principle SM is not self-supporting. But SR2 allows for situations in which SM is self-supporting. Gauthier himself directs our attention to them.[14] These are PD-like situations that meet certain conditions. The most important of these conditions, for our purposes at least, is the counterfactual independence condition. The condition states that no matter what choice either player makes he would still make that choice no matter what choice the other makes.[15] SFMs who place no special value on being such maximizers, choosing between SM and CM in such PD-like situations, would choose the former. A dominance argument, Gauthier informs us, establishes this.[16] We need not here concern ourselves with CM nor with the details of the dominance argument.[17] For the point that needs recognition is simply that if Gauthier is right, SFMs in such PD situations who do not value SM, will choose SM over a competitor. In these situations SM is therefore self-supporting.

Under what conditions is SM non-self-supporting? An answer to this question will conceivably leave us in a better position to formulate Gauthier's principle. As I understand him, Gauthier argues that SM fails in the required way in PD-like situations with "causal" dependence and with agents who are directly aware of each others' rational disposition or who at least have a good chance of correctly identifying the rational disposition of others.[18] But merely specifying the condition in this way is not enough. One must also suppose, as we have seen, that the choosers in question must not value being SFMs. So perhaps Gauthier's principle amounts to this:

SR3: A principle of rationality, R, is adequate only if given that R is one's principle of choice, R must be such that in a PD-like situation with "causal dependence" and with parties (i) who do not value being R-rational, and (ii) are directly aware of each others' rational disposition or are at least in a good position to identify the rational disposition of others, R is self-supporting.

SR3 seems incredible. Why accept it as a condition of adequacy for theories of rationality? For starters, it suffers one of the same defects as does SR2: It implies that the acceptability of a theory of rationality - even a teleological one - crucially depends on the values of rational agents. But we saw that this is unacceptable.

For another thing, what's so special about PD-like situations of the specified sort? More specifically, why is it that situations of this kind have any significance when it comes to assessing theories of rationality? Evidently the reason cannot be that in such situations SM fails. That would just beg the question against this theory. Perhaps the thought is this: Persons in the real world often find themselves in PD-like situations of the type now under consideration. It may be urged that situations in which each benefits from mutual cooperation in relation to mutual non-cooperation, but benefits from non-cooperation whatever the other does, are situations of this sort. The relevance of "causal dependence" may be this: Assuming that each is directly aware of the rational disposition of his fellows, or at least has a good chance in identifying the rational disposition of his fellows, persons disposed not to abide by their agreements to cooperate when it is best for themselves to do so would be excluded from cooperative arrangements they would find advantageous. Those not so disposed may be expected to be included in such arrangements and so reap benefits which are not otherwise available.[19]

Even if all this is true, we have failed to be given any justification to believe that such PD-like situations are relevant to _testing the adequacy_ of principles of rationality. Such considerations _would_ be relevant were our condition of adequacy something like this:

SR4: A principle of rationality, R, is adequate only if R does not preclude R-rational agents who are directly aware of the rational disposition of others, and who do not value R, from cooperating when such agents find themselves in PD-like situations with "causal dependence."

Notice, though, that SR4 is not a self-referential criterion. It does not, for example, mention anything about self-support. It cannot therefore be Gauthier's principle. It is, moreover, as controversial as SR3.

A final comment on SR3 is worth making. Its idea that a theory of rationality's failing to be self-supporting in a specified situation is reason enough to reject that theory seems fundamentally misconceived. It is a mundane truism that a theory of rationality will probably prescribe as rational different choices in different situations. In certain situations, for example, it may be rational for a SFM to adhere to her rationally made agreements; in other situations - some PD-like ones perhaps - it may not be. From this truism nothing seems to follow about the adequacy or inadequacy of these sorts of theories. Why should things be any different when the objects of choice are conceptions of rationality? Under certain conditions a theory of rationality may recommend itself as the object of rational choice. Under other conditions it may not. None of this though, I think, should lead us to believe that there is anything out of the ordinary about the relevant theory, any more than a theory's telling us to do different things under different conditions should generate suspicion about that theory.

1.7 An "In Between" Self-support Test

So far I have argued that neither ASS nor RSS is tenable. ASS runs afoul of "terrorist counterexamples." RSS succumbs to, amongst other things, Tom-like cases. But it may now be objected that these two kinds of counterexample would probably impugn *any* theory of rationality. For this very reason such counterexamples, it could be urged, are not very interesting. They do not indicate anything particularly amiss with, for example, SM, or CM, or for that matter *any* theory of rationality. Furthermore, at least the second of them is highly unusual and perhaps even illicit: It is surely a reasonable demand, the objection may continue, that an evaluation of a theory of rationality should not be affected by our prior valuation of that theory. So although the rejection of ASS is justified, one need not go as far as RSS. Why not introduce a new "in between" notion of self-support, the AASS or "almost always self-supporting" notion?

AASS: A principle of rationality, R, is adequate only if for any agent under any _nice R_ condition, the choice of R is permitted by R.

A situation is nice R if and only if no agent in it values or disvalues any agent's being commited to R, and no agent tries to get any agent to be or not to be so commited. AASS, it may be claimed, escapes terrorist-like and Tom-like objections. It also satisfies the requirement that SM but not CM fail to be self-supporting.

The last claim, however, is not true: Consider SFM Gullible. Gullible is an "unconditional complier," a person who always complies with agreements rationally entered into. Assume Gullible values being such a person. Would she renounce SM in favor of CM? Probably not, since CM enjoins compliance with fair and rational agreements only with trustworthy others, and not with, for example, exploiters who will defect from such agreements when doing so is utility maximizing.

To cope with this problem, the specification of nice R could be broadened so that a situation is nice R if and only if no agent in it values or disvalues being commited to R, or to any non-trivial entailment of R, and no agent tries to get any agent to be or not to be so commited. But this won't do either. Reconsider friend Reasonable, another SFM. Reasonable values acting on rational precept RP.

RP: If S ought (rationally) to do a at time t1 and S cannot do a at t1 without doing b at an earlier time t0, S ought (rationally) to do b at t0.

Prior to choosing between SM and CM Reasonable knows that he will find himself in a PD-like situation with Gullible. He knows that the SM-rational thing to do in such a situation is to confess, and the CM-rational thing to do (let's suppose) is to remain silent. To act consistently with RP in this PD-like situation, Reasonable must choose SM: If he commited himself to CM, he could not then act SM-rationally nor in a manner consistent with RP in the expected dilemmatic situation. So the choice of SM and not CM would maximize the satisfaction of Reasonable's preference to act consistently with RP.

Of course, one may elect to deal with this hurdle in the same manner as one dealt with the former, by re-specifying nice R. But it should be evident, or at least highly suspect, that such a move would be ad hoc. In the absence of a non-question begging and non-ad hoc specification of nice R, AASS is implausible.

1.8 Does Self-support Matter?

In the last couple of sections I have proposed and evaluated various principles of self-support. Of these it is apparently only SR3 that reflects what Gauthier may have in mind. But even SR3, I argued, is controversial. In evaluating these principles I have not attacked the notion of self-support *per se*, the notion held in common by each of them. Rather, my criticisms have been directed to other specific features of each individual principle. I wish to conclude by extending criticism to the concept of self-support itself.

Why might it be thought that a principle of rationality's not being self-supporting is sufficient to impugn that principle? The question is pressing since the inference from a principle of rationality's not being self-supporting, to the conclusion that it is thereby inadequate, is obviously suspect if not clearly a *non-sequitur*. From the fact that a principle of rationality, R, is not self-supporting in certain circumstances, it does not seem to follow - at least not on first glance and not even on second - that R is not adequate. What follows, if anything, is that under those circumstances it is rational by R's own terms not to undertake a permanent commitment to R. But this fails to provide legitimate grounds for questioning R's adequacy. An extra premise is required to sustain the charge of inadequacy, one to the effect that

If it is R-rational not to undertake a permanent commitment to R, then R is not adequate.

This premise, however, is false or at least extremely contentious. If Gilligan, for example, were so psychologically constituted that were he to undertake a permanent commitment to R - supposing R to be his principle of choice - he would suffer paroxysms of mental anguish, it would be R-rational for him not to do so. Why should any of this, though, lead us to believe that R is defective?

A final reply to our question is this: It is desirable as "fully rational" agents to subject our theories of rationality themselves to rational assessment. However, it seems impossible to appeal to further standards to do so. But then, it might be urged, the <u>only</u> way in which we can assess our ultimate theories is self-referentially - is it rational by the theory's own terms for a person to choose it?

Grant it is not possible to assess utlimate principles of rationality by appealing to further standards. Even so, the response is flawed: Such impossibility does not entail the view advocated by the reply. Exactly <u>how</u> our basic principles of practical reason are to be evaluated, if at all, is a difficult and perplexing question. But without additional reasons than those provided by this reply, there seems to be nothing compelling about the recommendation that the sole means to assess ultimate principles is self-referentially.

1.9 The Unresolved Compliance Problem

I ended the last chapter on this note: Gauthier cannot solve the compliance problem by relying on the "choice argument." That argument fails. He might, nevertheless, insist (correctly) that (under specific conditions) it is CM-rational to comply with your rationally made agreements. If you have made a rational agreement with your trustworthy partner to do the cooperative thing in a PD-like situation, then CM requires that you comply with that agreement. If you are a SFM you should not be impressed by this fact, not unless you have been given reason to believe that SM but not CM is defective, or that CM is a better theory, in the light of an acceptable comparative test, than SM. The evaluative tests for conceptions of rationality that we have considered in this chapter won't help here. They do not show that SM is defective. Nor do they establish the superiority of CM to SM.

The compliance problem remains a problem.

[1] David Gauthier, "Reason And Maximization," Canadian Journal of Philosophy 4 (1975), 411-433, p. 431.

[2] Aware that his substantive theory of justice will conflict with our moral intuitions about concrete cases, Gauthier tells us to trust his theory over our intuitions. (David Gauthier, Morals By Agreement (1986), Oxford: Clarendon Press, p. 269) Presumably, he would adopt a similar position with respect to the role of our intuitions in assessing theories of practical rationality.

[3] The extensional equivalence of CM and SM in contexts of choice for theories of rationality may be explained in this way: CM, among other things, expresses a principle of interdependent action - action on the basis of agreement with others:

> [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 interdependently, with an expected outcome which affords each person other than himself at least as great a utility, and himself a greater utility." ["Reason And Maximization," p. 427]

Gauthier explains that "to act independently is to act interdependently with oneself alone." ("Reason And Maximization," p. 427) Independent action is therefore a special case of interdependent action. As a consequence, with respect to independent action, CM and SM are extensionally equivalent.

[4] Here's an explanation of these results. Suppose, first, that T1 and T2 are extensionally equivalent in parametric contexts. Then there are three possibilities: (i) Both recommend a choice of T1. Then T1 is chosen and so T1 is better. (ii) Both recommend a choice of T2. Then T2 is chosen, and so T2 is better. (iii) Both recommend either T1 or T2. Then neither is better than the other - they tie. Suppose, next, that T1 and T2 are not extensionally equivalent. Then T1 is not better than T2, and T2 is not better than T1. I'm calling this outcome "a tie."

5. See Chapter 5, section 1.5.2, this thesis.

6. In Morals By Agreement, pp. 181-182, Gauthier says this:

> If we fall into a society - or rather into a state of nature - of straightforward maximizers, then constrained maximization, which disposes us to justice, will indeed be of no use to us, and we must then consult only the direct dictates of our own utilities. In a world of Fooles, it would not pay to be a constrained maximizer, and to comply with one's agreements. In such circumstances it would not be rational to be moral.
>
> But if we find ourselves in the company of reasonably just persons, then we too have a reason to dispose ourselves to justice.

7. See sections 1.6 and 1.7 below.

8. "Reason And Maximization," p. 430.

9. "Reason And Maximization," p. 429.

10. If this is true, then in the interests of accuracy RSS should be interpreted thusly:

RSS: Principle of rationality, R, is self-supporting =df. for some agents, under certain conditions, the choice of R would be permitted by R.

11. Howard Sobel makes a very similar observation on page 685 in J. Howard Sobel, "Interaction Problems for Utility Maximizers," Canadian Journal of Philosophy 4 (1975), 677-688.

12. Gauthier holds that the content of a person's preferences is beyond rational assessment. (Morals By Agreement, pp. 25, 34, 48).

13. The concept of the expected utility of a conception of rationality may be elucidated in this way: The expected utility of a conception of rationality may be identified with the expected utility of the act of undertaking a permanent commitment to that conception. Suppose R and R* are conceptions of rationality. Suppose S is an agent contemplating a choice of conceptions. The expected

utility for S of R is calculated in this fashion: consider all the possible outcomes of choosing R. For each, consider the utility of that outcome. For each, consider the probability of that outcome given that R is chosen. Multiply, for each, the utility and the probability. Add these products. The sum is the expected utility of R for S.

Clause (ii) of SR2* restricts the outcomes of choosing a principle of rationality to actions.

14. Morals By Agreement, p. 173.

15. I owe this formulation of the counterfactual independence condition to Professor Fred Feldman.

16. Morals By Agreement, pp. 172-173. In the last chapter (section 1.5.3), I expressed reservations about this conclusion of Gauthier's. I there suggested that in PD situations that satisfy the counterfactual independence condition and in which deception is not possible, SFMs choosing between SM and CM will be indifferent between these two theories.

17. For an explanation of CM, see footnote 5 above, and section 1.3, Chapter 5, this thesis.

18. See Argument (2), in Morals By Agreement, p. 172, and the subsequent discussion on p. 173. Again, I disagree with Gauthier here: In PD-like situations with transparent SFMs, and in which deception is not possible, such maximizers will again choose either theory when confronted with a choice between SM and CM. See Chapter 5, section 1.5.3. I will proceed, though, as if Gauthier is right - that in such situations, SM fails to be self-supporting.

19. See Morals By Agreement, p. 173 and p. 183. I have different views about this matter that I present in section 1.5.3 in Chapter 5.

CHAPTER 7

THE RATIONAL CREDENTIALS OF THE CHOICE THEORY OF MORAL JUSTIFICATION

1.1 Introduction

The moral sceptic and his collegue - Hobbes' Fool, question the rational credentials of morality. In fact, they do more. They reject all moral requirements as "unjustified," because these requirements interfere with their pursuit of self-interest. Considerations of rational self-interest, they believe, provide the sole legitimate grounds for action. The contractarian strategy of Gauthier (and of others) to convince the amoralist otherwise, is to show that acknowledgment of at least some moral constraints is required by the very conception of reason endorsed by the amoralist.

The last couple of chapters have raised a question about the rational credentials of a theory of practical rationality itself. If we are to believe Gauthier, there are different standards of reason. There is the standard of SM accepted by the Fool and his sceptical friends and there is its rival, CM. In adopting this position, Gauthier does the Fool one better: While the Fool eschews morality, he acccepts reason as unproblematic. Gauthier's scepticism - at least his initial scepticism, extends even further - theories of practical reason themselves fall within its ambit.

To successfully deflate the Fool's amoralism, then, Gauthier must undertake at least a two-fold task. Not only must he show that morality can be "derived" from a theory of practical rationality, he must also establish the acceptability of the very theory of reason on which morality is to be "founded."

Morals By Agreement offers a unified contractarian scheme for the "justification" of both a morality and a theory of practical rationality. In this final chapter on Gauthier, I look more closely at this justificatory scheme. In so doing, I continue - in part - the discussion of the last chapter on assessing theories of practical rationality. But I now come at this issue from a slightly different angle.

1.2 Theories of Moral Justification

One of the central tasks, if not the central task, of Morals By Agreement is

> to seek to prove.....that principles of action that prescribe duties overriding advantage [i.e. moral principles] may be rationally justified. (M by A, 2)

The issue of what we mean when we say that a morality is "justified" is interesting because controversial. A few possible answers are these:

The epistemic answer says that a morality, M, is justified for a person, S, at time, t, if and only if S is epistemically justified in believing that M is true at t. In other words, a justified moral theory is one that we are epistemically justified in accepting.

Another answer is the moral one. A moral theory, M, is justified for S at t if and only if it is morally right for S to accept M at t. In this case, some higher order moral theory would be required to determine whether it would be morally right for S to accept M at t.

A third answer might be a sort of "best from the point of view of our intuitions" answer. The idea here is to try to see that you have landed on a set of beliefs that you will be able to "live with" come what may. Perhaps what is here being sought is a moral theory that seems to capture how we really feel about morality at the deepest level.

The "functional" answer assumes that a morality has a function and that it is justified for S at t if and only if it fulfils its function at t.

The last on the list of possibilities I canvass is the "abstract rationality" answer. According to this answer, moral theory M is justified for S at t if and only if it would be rational, in some sense of 'rational,' for S to accept M at t. Maybe S's life, for instance, would go better for S if S were to accept M than if S were to accept any other morality. Alternatively, perhaps S's preferences would be satisfied to a greater extent than they would be were S to choose some other theory. From the fact that it would be rational for you to accept M, it does not follow that you have any evidence that M is true. You may well have plenty of evidence that M would maximally satisfy your preferences if you were to choose it. But that is different from having evidence that M is true. It is in this respect that this "choice" conception of justification differs in an essential way from the epistemic one.

I believe the notion of justification appealed to in Morals By Agreement is the abstract rationality one. In fact, Professor Gauthier appears to use something like this notion to justify not only moralities but, paradoxical as it may seem, theories of rationality as well.[1] I argue in this chapter that this conception of justification generates problems for Gauthier and like-minded theorists who endorse a view of values as "subjective." It also engenders difficulties for theorists who espouse an "objectivist" axiology. More specifically, I attempt to show three things: (1) Choice theory in conjunction with an objectivist axiology runs afoul of the "morality is not itself justified" (MJ) objection. This is the objection, roughly, that from the fact that a person is justified in accepting a morality, possibly because choice of that morality for the person is rational, it does not follow that that morality is itself justified relative to that person. If values are objective, there may well be reasons for rejecting a morality as unjustified, even if choice of that morality for any person is rational. (2) When the object of justification is a theory of rationality, and if choice theory appropriately modified is held in conjunction with a subjectivist axiology and a criterion specifying a necessary condition of adequacy for theories of rationality, choice theory is also open to an MJ-like objection. (3) In cases in which the object of justification is a morality, and the choice of a morality is to be made from a set of theories, some of which are moral and some of which are not, choice theory in association with a subjectivist axiology is again vulnerable to an MJ sort of objection.

1.3 Gauthier's Abstract Choice Theory

This section elaborates the version of choice theory to which Gauthier subscribes. Although the passage does not explicitly mention anything about justification, it's fairly clear that Gauthier advocates choice theory when he says

> Rawls' idea, that principles of justice are the objects of a rational choice, is indeed one that we shall incorporate into our theory.....[W]e shall represent the choice as a bargain or agreement among persons.....[W]e claim to generate morality as a set of rational principles for choice. We are committed to showing why an individual, reasoning from non-moral premisses, would accept the constraints of morality on his choices. (M by A, 5)

The justification of a morality, Gauthier tells us, is to proceed in relation to a group of rational egoists whose goal is to maximize individual value. Somewhat surprisingly, Gauthier believes that goal must be achieved by validating "the conception of morality as a set of rational, impartial constraints on the pursuit of individual interest." (M by A, 8, my emphasis.) The rationale for constraint is contractarian:

> Morals by agreement offer a contractarian rationale for distinguishing what one may and may not do. (M by A, 9).....[T]he appeal to rational choice enables us to state, with new clarity and precision, why rational persons would agree *ex ante* to constraining principles, what general characteristics these principles must have as objects of rational agreement, and why rational persons would comply *ex post* with the agreed constraints." (M by A, 10)

The "weak and widely accepted" theory of rational choice on the basis of which individuals are to select principles is straightforward maximization (SM). Let's remind ourselves about SM. On this theory a person chooses rationally if and only if she maximizes her expected utility. (M by A, 182) The utility of an outcome, O, of some action, a, for some agent, S, is the value of O for S. The expected utility of an action, a, for some agent, S, can be conceptualized as a measure of the extent to which the various outcomes that could result were S to do a, would satisfy S's considered preferences for those outcomes. Considered preferences, Gauthier says, are those "that would pass the test of reflection and experience." (M by A, 31, 21-26) The expected utility of a for S is to be calculated by appeal to probabilities on S's evidence regarding the extent to which the various outcomes of a would generate satisfactions of S's considered preferences.

The abstract rationality conception of justification, CT, underlying Morals By Agreement can now be formulated more perspicuously in this fashion:

CT: A morality, M, is justified in relation to the members of a group if and only if it is rational for each member of that group both to accept M and to comply with M, given their considered preferences and their particular circumstances of choice.[2]

I said that Gauthier appears to use a version of choice theory to justify not only moralities but theories of rationality as well. He argues that "constrained maximization" (CM) is to be preferred to SM on the grounds that the former and not the latter would be chosen (under suitably specified conditions of choice) by SM-rational agents if they were to choose between them. Let PR be either a theory of morality or a theory of rationality. Then (abridging for convenience) Gauthier would probably accept principle CTR:

CTR: PR is justified in relation to the members of a group if and only if it is rational for each member of that group to accept PR.

The link between rational choice theory and the choice theory of justification should now be evident. There is, in addition, an interesting connection between rational choice theory - at least the kind of maximizing theory held by Gauthier - and value theory, and in consequence, an interesting connection between the choice conception of justification and axiologies. Let me first sketch this connection and then elucidate the theories of value commanding our interest.

I explained that Gauthier accepts a theory of rational choice that identifies rationality (at least in parametric contexts) with the maximization of a person's expected utility. Think of expected utility in the manner I recommended as a measure of a person's preferences for various possible outcomes realizable in action. Gauthier equates this measure with value. He holds that outcomes have value in virtue of being desired or preferred.

It appears to be Gauthier's view that any SM-rational agent, selecting among principles that are to govern his interaction with others, would select the principle of minimax relative concession (MMRC).[3] MMRC, he tells us, is a principle of distributive justice. Furthermore, he believes that any constrained maximizer but no SFM, would comply with MMRC's "interest-constraining" prescriptions when called upon to do so. So at least part of CT as conceived by Gauthier, the part saying that any rational agent would accept a moral principle, is to be understood as claiming that the selection of a particular morality would maximize the value of each member of that group. We can properly speak of each member's value since Gauthier believes that the expected utilities of individuals are not interpersonally comparable. (M by A, 134, 135) As far as compliance is concerned, matters are more complicated. Here I think CT must be construed as saying that it is CM-rational for each member of a group to comply with a moral principle that it is SM-rational for each to select, provided the others are ready to comply as well.

In contrast to Gauthier's "subjectivism," "objectivism" denies that value is "created" by preference. On this rival view, states of affairs can have value independently of being preferred or desired by anyone.

Rationality would then be concerned with the maximization of such "objective" value. The notion of maximizing objective value would, of course, have to be explained in an acceptable fashion. CT would then have to be interpreted as recommending that a morality is justified either _simpliciter_ or in relation to a group of persons if and only if if there is no alternative having a higher degree of objective value than it has.

1.4 Subjectivism and Objectivism

The distinction between subjectivism and objectivism needs refinement in order better to apprehend the contrast between the two. Gauthier explains that

> To conceive of value as dependent on affective relationships is to conceive of value as _subjective_. (M by A, 47)

> Value on this conception is a measure of [considered] preference. A measure depends for its existence on what it measures - no preference, no value.....[O]bjects or states of affairs may be ascribed value only in so far as, directly or indirectly, they may be considered as entering into relations of preference. Value is then not an inherent characteristic of things or states of affairs, not something existing.....in a manner quite independent of persons and their activities. Rather, value is created or determined through preference. (M by A, 46-47)

A consequence of this view is that

> a state of affairs is characterized not by a single value, but by a set of values, one for each person

> into whose preferences it does or may enter. Value does not afford a single uniform measure of preference but a measure relative to each valuer. (M by A, 25. See also p. 49 in this work.)

The passages show that the subjectivist theory in consideration is complex, having many tenets the most significant of which for our purposes are these:

S1: Value cannot exist independently of persons.

S2: States of affairs have value only in relation to the desires, feelings, hopes, fears, aversions etc. of persons. More succinctly, states of affairs have value only in relation to the affections of persons.

S2 specifies the particular feature of persons that according to this variety of subjectivism is essential to the existence of value.

S3: The value of a state of affairs for a person is identical to the value, if any, that that state of affairs has for that person. A state of affairs, S1, has value for a person, P, if and only if P has a considered preference for S1.

A variant of this kind of subjectivist theory holds that a state of affairs has value for a person if and only if it is desired by the person. On either variant, what is to be stressed is that a thing has its value relative to a person purely in virtue of the fact that the person values the thing. Preferences or desires, we might say, create value. States of affairs, if they exemplify what we could call "derived value" in relation to some person, do so in virtue of the fact that the person prefers them. On this view, values are to states of affairs as secondary properties like redness are to physical objects. There are other types of subjectivist theory. The kinds of theory, however, to which I confine attention are the ones just described. I shall hereafter take 'subjectivism' to denote theories of this sort.

Opposed to subjectivism is objectivism.

> To conceive of value as objective is to conceive of it as existing independently of the affections of sentient beings.....(M by A, 47)

> The objectivist may insist that there is a necessary linkage between sentient beings and value, holding that value provides the norm for our affections. The proper object of preference is, and is necessarily, the good. But this link, on the objectivist view, is not found within sentient affection, but ties that affection to something else which in itself affords the ground of the tie. This other terminal of the link (and we shall not impose on the objectivist a particular account of what it is) must be independent of sentient beings and their affections, even though linked to them. In our discussion, we shall designate this alleged other terminal, objective value. (M by A, 56)

Departing somewhat from the characterization of objectivism summarized in these passages, an objectivist need deny neither S1 nor S2. An objectivist hedonist, for example, may contend that if there were no sentient beings and their affections, nothing could have value. Other objectivists would reject both S1 and S2. Thus if one believed that the "form" of *The Good* endows states of affairs with value, and that there is no necessary connection between persons' affections and *The Good*, then one would have no reason to accept either S1 or S2.

The common strand of all objectivist theories is the denial of the "derived-value" thesis, S3. States of affairs, according to these theories, do not have value in virtue of being desired or preferred by anyone. This should be clear in relation to theories of the platonic variety, an example of which we just noted. But theories like objective hedonism also deny S3: A hedonist may hold that the state of affairs, S, *Tom takes pleasure in surfing at time t*, has intrinsic value. If S did in fact have value, if it were intrinsically good, it would be so independently of being desired or preferred by anyone, Tom included. That is of course not to deny that S's goodness may cause some person, perhaps Tom, to desire it. But some person's desiring S, or deriving pleasure from S, the objectivist insists, cannot "endow" S with value. We may now formulate objectivism in this way:

Objectivism: The value of a state of affairs, S, is an objective property of S. F is an *objective property* of S

=df. S has F, and S does not have F in virtue of any perons's desiring S, or wanting S, or deriving pleasure from S, or - generalizing somewhat barbarously - seeing to it that S has "derived value."

1.5 A Preliminary Observation

Consider the choice theory of justification, CT, held in association with a subjectivist account of value. CT (abbreviated for ease of exposition) states that a morality, M, is justified in relation to the members of a group, G, if and only if it is rational for each member of G to accept M. Embracing a subjective account of value, Gauthier imposes no substantive constraints on the states of affairs rational agents may prefer and so value. Indeed, he holds that the contents of persons' considered preferences cannot be rationally assessed. (M by A, 25, 26, 34) Given this freedom from constraint, it is unlikely, as I argued in Chapter 5, that SFMs will unanimously choose the same moral principles to constrain their interactions. The reason for this, roughly, being that their considered preferences for states of affairs, unless by an happy coincidence or as a mere matter of contingent fact, will not all be identical. In that case, no morality would be justified for the relevant group.

This preliminary observation aside, there is an objection to the necessary and sufficient conditions on moral justification enunciated by CT that I now wish to develop. I first show how this sort of objection - the MJ objection - afflicts CT held in combination with an objectivist theory of value. I then argue that when the object of justification is either a morality or a theory of rationality, choice theory in conjunction with a subjectivist axiology, together with certain additional considerations, succumbs to a similar problem.

1.6 Choice Theory and Objectivism

Consider a simplified version of choice theory, CTS, according to which a morality, M, is justified in relation to a person, S, if and only if the choice of M by S is rational. Does S's rational choice of M suffice to justify M itself? Is such rational choice even necessary to its justification? The answers are fairly evident if CTS is considered in conjunction with an objectivist account of value. Suppose "platonic" objectivism is true so that the form of The Good determines what is good. Suppose, further, that there is no necessary connection between person's preferences and The Good. In Gauthier's terminology, assume that value does not provide a "norm" for our preferences; the proper object of our preferences is not necessarily The Good. Then it is entirely possible that although S’s choice of M would be justified because rational, such a choice would not serve to justify M itself. There could well be independent reasons that undermined the tenability of M itself, reasons in one way or another having to do with objective value. An example might be useful. Since we are assuming objectivism is true, let’s also not implausibly assume that there is a true morality, act utilitarianism. Let act utilitarianism be the theory that an act is right if and only if none of its alternatives has a higher utility than it has. Let the utility af an act, a, be the result of subtracting all the units of intrinsic badness, if any, from all the units of intrinsic goodness, if any, that would result if a were performed. Ignoring certain complications, let relativism be the theory that an act is right if and only if it is permitted by the conventions of its society. Suppose choice of the latter and not the former theory would maximize S’s expected utility if S were to choose between the two. Then it should be clear that although S’s choice of relativism would be rational, this theory would not itself be (objectively) justified - relativism, by stipulation, is false. If values are objective, then it seems proper to require that a theory be justified only if it bears some important relation to objective value.

Suppose objective value did provide a norm for our preferences. In fact, assume these to be true: For any states of affairs, Si and Sj, and for any agent, p, p prefers Si to Sj if and only if Si is better than Sj. For any states of affairs, Si and Sj, Si is better than Sj if and only is Si has a higher degree of intrinsic goodness than Sj. Now, once again, suppose the choice of M were to maximize the satisfaction of S's, in fact anyone's, considered preferences and so were to qualify as rational. In this case wouldn't M itself be justified? I don't think so: There is a difference between the objective value of the choice of M, and the objective value of the actions M tells you to do or the objective value of M itself. It would be true that if the choice of M were maximally to satisfy anyone's considered preferences, its choice would be objectively better than the choice of any alternative. But that, I think, wouldn't support the further distinct desideratum that M itself were justified. I may be wrong here. Suppose, then, that in the case now under scrutiny M itself were justified. Even then there is a problem: M would be justified on what is now the supposition that it were beter than alternatives. That it would be better in this fashion, though, would be an objective fact, true independently of any choice of any person. Assume, now, that objective values do provide a norm for preferences, but assume also that most rational preferences happen to violate the norm. Then rational choice of a morality does not guarantee that the morality is objectively justified.[5]

So it seems that if values are objective, a rational choice of a morality does not show the morality itself to be justified. At best, it shows that the choice of the morality is justified. If I am right about this, then the abstract rationality account of justification is not neutral among conceptions of value.[6] It may be held in association with an objective theory of value only on pain of falling prey to the foregoing objection.

1.7 Choice Theory and Theories of Rationality

Gauthier uses the abstract rationality theory to defend not only moralities but theories of rationality as well. So let's evaluate the appropriate form of choice theory as it is used in this capacity. To do so, consider a simplified version of CTR, CTRS, which says that a theory of rationality, PR, is justified in relation to a person S if and only if the choice of PR by S is SM-rational. In addition to being a principle of rational choice, PR may also be a moral principle as it may satisfy criteria for being such a principle. So, for example, if PR constrains one's pursuit of self-interest, is "rationally justified," and is "impartial," Gauthier would probably acknowledge that it is a moral principle. (M by A, 3, 4) I introduce this complexity because Gauthier believes that the choice of CM and not SM (under specified conditions) is utility maximizing for any SM-rational agent. (M by A, 157-189) Furthermore, he believes that CM is a moral principle.[7]

There are two main variants of CTRS that demand attention. There is the variant that results if CTRS stands on its own, and the one generated if CTRS is held in tandem with a criterion of adequacy for theories of rationality. We examined something very similar to the first of these variants in sections 1.3 and 1.4 in Chapter 6. In these sections we discussed CT and various modifications of CT. CT says that a theory of rationality is adequate only if choice of it is rational. If we replace 'adequate' with 'justified,' then CT so amended is part of CTRS; CTRS implies that a necessary condition of a theory of rationality's being justified in relation to a person is that choice of the theory by that person is rational. But we saw that CT is defective. So the first sort of variant of CTRS must also be defective.

Now consider the second variant. Suppose CTRS is held in conjunction with (i) a subjectivist axiology, and (ii), a criterion of adequacy, C, for theories of rationality. The criterion may be something like this: A theory of rationality is adequate only if absolutely self-supporting.[8] The notion of absolute self-support is to be understood in this way: Principle of rationality, R, is absolutely self-supporting =df. for any agent under any conditions, the choice of R would be permitted by R. It is reasonable to require that a justified conception of rationality at least satisfy a criterion of adequacy for any such conception. How can a justified conception fail in this respect? In light of this constraint, CTRS may be modified to the view that

PR is justified in relation to S if and only if (i) the choice of PR by S is rational, and (ii), PR satisfies C.

Call this principle CTRS*. Formulated in this way, a problem with CTRS* may now be fairly obvious. There is no guarantee that a precept of practical rationality will satisfy both conditions (i) and (ii). A precept that satisfies condition (i), for instance, may fail to satisfy condition (ii). In such a case, although the choice of the relevant precept would be justified, the precept itself would not be justified. Let's consider an example. Let PR be CM and let C be the criterion of absolute self-support. Then CTRS* tells us that CM is justified in relation to some agent, S, just in case S's choice of CM maximizes S's utility, and CM is absolutely self-supporting. Gauthier, as I said before, argues that any SFM, and so S, would (under appropriate conditions) choose CM as his conception of rationality if presented with the choice of adopting either CM or SM. Assume he is right about this. But CM, as we saw in the last chapter (refer to section 1.5 of this chapter), is not absolutely self-supporting: If a terrorist puts a gun to your head and threatens: "If you adopt CM, I'll kill you," CM prescribes that you ought not to adopt CM. In such cases, the choice of CM would be justified because rational. CM, however, would not be justified, at least not by the lights of CTRS*, because CM is not absolutely self-supporting.

In addition, CTRS* has another problem. It has the same defect that undermines the first variant of CTRS, one to which we alluded above.

1.8 Choice Theory and Subjectivism

The criticism of choice theory just discussed might be thought to be of limited interest. It is, after all, primarily meant to show that Gauthier cannot consistently endorse the appropriate form of choice theory when the object of justification is a rationality, and some criterion of self support. It is also designed to caution us that theories of rationality cannot be justified in the way in which choice theory recommends. But the problems with choice theory do not end here. I want, now, to evaluate its credentials in the most general and interesting of cases.

Suppose one wishes to answer Hobbes' Fool who challenges the rationality of being moral. The Fool insists that there are certain situations in which one must act either irrationally or immorally. Some prisoner's dilemma situations may be situations of this sort. Suppose, further, that the challenge is to be met by showing that fully rational persons, SM-rational persons, would voluntarily agree to restrain their maximizing activity by choosing to be moral. In Gauthier's words, the challenge is to be met

> by demonstrating that.....a rational utility maximizer, faced with the choice between accepting no constraints on his choices in interaction, and accepting the constraints on his choices required by [a particular moral principle, MMRC], chooses the latter. (M by A, 158)

If this strategy to meet the Fool's challenge is to succeed, some of the principles among which rational agents are to choose, as the passage acknowledges, must be principles requiring no restraint on maximizing activity. That is, the alternatives among which rational agents are to choose must not be restricted to moral principles. In the sorts of cases when choice among alternatives is so restricted, choice theory loses much of its theoretical appeal. For much of this appeal, I think, derives from the hope that the challenge of the moral sceptic like Hobbes' Fool can be answered by showing that it would be rational for any individual, and so for the rational sceptic himself, to choose to be moral.

So consider the interesting case like the one discussed by Gauthier and arguably by Hobbes, where rational agents are to select principles of practical reason from an "impure set" that has both moral and non-moral principles as elements. CM and SM, for example, might be the members of one such set. Let's allow that some of the members in the impure set are principles that qualify as both moral principles and principles for rational choice so that we can accomodate the possibility that "to choose rationally, one must choose morally." (M by A, 4) Now let's evaluate choice theory, held in conjunction with a subjectivist axiology, in cases where the putative object of justification is a morality to be selected from an impure set.

It would be misleading to formulate the version of simplified choice theory here appropriate in this way:

CT*: A morality, M, is justified in relation to S if and only if S's choice of M is rational.

CT* is misleading because it fails to alert us to the fact that choice of a morality is to be made from an impure set. A better formulation would be this:

CTheory: P is a justified morality in relation to S if and only if (i) the choice of P by S from an impure set is rational, and (ii), P satisfies necessary and sufficient conditions for being a morality.

If we are concerned to answer Hobbes' Fool, choice of principles must be made from an impure set. Clause (i) records this requirement. Since principles are to be selected from an impure set, it is possible that a rational choice may result in a principle that is non-moral. In order to be a justified morality, though, the principle whose choice is rational must be a principle that is a moral one. Clause (ii) is sensitive to this demand.

Formulated in this appropriate fashion, however, it appears that CTheory is afflicted with one of the same sorts of difficulty that is damaging to CTRS*: There is no guarantee that a principle that satisfies the condition specified by clause (i) in CTheory also satisfies the condition specified by clause (ii). Again, examples will be helpful. Let the members of the impure set from which principles are to be selected be SM and CM. SM is not "interest-constraining." It is therefore, as Gauthier agrees, not a moral principle. If any morality, then, is justified CM must be justified, since it is the sole moral principle in the impure set now under consideration. However, in Chapter 5, we saw that under certain conditions the choice of SM and not CM is rational. A person, S, who values acting in accord with SM and who hates CM, for example, would do best if he were to choose SM. Here's another example: Suppose S knew that he would find himself in exactly one PD situation in which his partner, Gullible, would do the cooperative thing come what may. S would again do best if he were to settle for SM. By adopting SM but not CM, S could exploit Gullible. In these cases, although the choice of SM would be justified because rational, the object of choice would not itself be a justified morality. CTheory, it appears, yields incorrect results. To evade this worry, one might invoke Desperate:

Desperate: For any impure set, #, and any member, pi of #, and any agent s, if s's choice of pi is rational, then pi is a moral principle.

Desperate is false as the examples above demonstrate. Not only is it false, it cannot be appealed to with propriety without begging the question against Hobbes' Fool.

If CTheory is the correct version of abstract rationality theory underlying Morals By Agreement, then I think that this theory of moral justification is in need of serious amendment.

1.9 A Rumor

Gauthier has failed to provide Butch and Sundance with adequate justification to believe that they will be able to escape the PD. Even if they were to become constrained maximizers, they would do no better than they would have done had they remained straightforwardly rational. Even as constrained maximizers, the rational thing to do in the authentic PD in which they are is to confess.

A note from the underground surreptitiously reaches our felons. It urges them not to loose hope. Word has leaked out that resolute agents can overcome PDs. Is there any truth to this rumor? We'll find out in the next chapter.

CHAPTER 8

RESOLUTE CHOICE

1.1 Introduction

Edward McClennen has recently argued that the one-shot PD is "resolvable."[1] It is resolvable in the sense that fully rational agents are able to do whatever it takes to achieve an optimal outcome. His solution to the dilemma is meant to provide rational agents who face such a dilemma with a rationale for cooperation. It would also provide, if tenable, what we might take to be a vindication of the claim that it is rational to be moral.

McClennen tells us that (like David Gauthier) he has "become persuaded that there is a need for a reappraisal of the requirements of rational choice as typically presented." (PD and RC, 95) Unlike Gauthier who (according to McClennen) argues for cooperation by appealing to the notion of maximizing expected utility at the level of dispositions to choose, (Ibid.) he argues for it on the basis of "maximizing utility at the level of.....particular choice, but that this utility is contextually dependent on the nature of the choice situation." (Ibid.) His arguments are, consequently, of special interest to this thesis. But do they succeed? I believe not - for reasons that I will shortly present.

McClennen provides a diagnosis of what he conceives to be the central problem with SM, a problem he thinks that clearly surfaces when the theory is extended to "dynamic choice situations" in which agents are called upon to make a sequence of choices over time. It is this shortcoming of SM, he believes, that ultimately prevents SM-rational agents from escaping the dilemma. He then proposes his solution which appeals to the notion of rationality as "resolute choice."

In this chapter I begin with an explanation of the problem McClennen finds with SM - in a nutshell, that SM fails to satisfy a criterion of adequacy for conceptions of rationality. I then summarize the theory of resolute choice. Finally, I challenge McClennen's claim that rational agents who choose resolutely do better in PDs of a certain variety than do agents who choose in a straightforward fashion.

1.2 McClennen's Assessment of SM

The fault with SM, at least with the theory as it is normally construed, McClennen thinks, has to do with its requiring that

> on each occasion calling for decision,.....[persons or selves] maximize with respect to an antecedently and exogenously specified preference function given (again from the perspective of that same occasion for decision) the expected behavior of the other [persons or selves].

An example of decision making in dynamic contexts will help illuminate the difficulty. Consider the case of Ulysses and the Sirens.

> As Ulysses approaches the island of the Sirens, he has no desire to be detained by them; but if he acts on his present preferences (to get home as quickly and as inexpensively as possible), he faces a problem. He is informed that once he hears the Sirens, he will want to follow them. Since <u>here, now</u> he does not desire to have this happen, he precommits. He buys wax to stop up the ears of his sailors, good strong hemp with which to have himself bound to the mast, and (what is perhaps most costly of all) arranges for his first-mate to act as his agent.....
>
> It may be objected that.....[this tale] describes a case in which the self is thought to be temporarily overcome by some irrational (or non-rational) force. But.....[w]e have only to suppose that Ulysses realizes that he is in a situation in which he can predict that his preferences will undergo a specific change. (PD and RC, 98-99)

In tree diagram form, Ulysses' problem looks like this:

```
                                      ----(c) Ignore Sirens----O3
                                                           (Best)

   ----(a) Sail on--------O node 1 (n1)

                                      -----(d) Give in---------O2
                                                          (Worst)



O-------(b) Precommit------------------------------------------O1
node 0 (n0)                                         (Second best)
time = t0 at n0           time = t1 at n1
```

Figure 8.1 Ulysses' Tree Diagram

In such "dynamic" contexts of choice the SFM acts as what McClennen calls a "sophisticated chooser." Such an agent chooses at the outset - node 0 in our example - a plan consisting of a sequence of actions that maximizes the satisfaction of the preferences of his present self against the expected independent maximizing behavior of his future self. Assume Ulysses is a sophisticated chooser. What sequence of actions should he select at the first choice point in his decision tree? Ulysses then has three alternatives that we may represent as ac, ad, and b. If he were to choose the course of action that is best from the vantage point of n0, he would choose ac. But Ulysses is aware that this plan of action is not feasible. He is aware that if he were to reach node 1, he would be in the grips of the singing sisters and so would be unable to do c as the best-from-n0 plan requires. Expressed somewhat differently, Ulysses as a sophisticated chooser is cognizant that his future self at n1 will act in accord with the best-from-n1 plan available to it then. Knowing this, he now knows that ac is not feasible for his present self. But as of t0 each of ad and b is feasible. Since Ulysses is a maximizer, he does best by choosing b as of that time. McClennen summarizes the sophisticated approach to dynamic choice in this way:

> a [sophisticated] chooser regiments his _ex ante_ choice to his projected _ex post_ choices. That is, he makes a projection of how he will be disposed at n1 to choose from the options available at n1, a projection that is presumably independent of how he is disposed at n0 to choose from the plans available at n0, and he takes this projection to condition the feasibility of plans available at n0. (DC and R, 9)

It will be helpful to introduce some terminology: Assume that agent, s, is in a dynamic choice context in which s has to make a sequence of choices over time. Assume that s's problem to choose among the various courses of action can be represented by a tree diagram. Call the choice situation that this diagram describes "s's decision tree."

(1) A _plan_ is a set of directives that specifies what is to be done as of a time at each choice point in a decision tree.[2] _An agent implements a plan_ if and only if she acts in accord with each of its directives.

Let T be a decision tree and let A be a set of plans p1,.....,pi, in T. Let n1,.....,ni be the nodes in T and let t1,.....,ti be the times corresponding to n1,.....,ni. Assume in other words, that s will be at n1 at t1 and for any i>1, s will be, if s ever is, at ni+1 at ti+1, at ni+2 at ti+2 and so on. To facilitate matters, it will be helpful to think of a plan as an ordered n-tuple of actions, P, available at a time to s in T. The number of members in P = the number of nodes, as of a particular time, at which s could be in T. So, for example, if at t0 there are eleven nodes n0,.....,n10 at which s could be in T, then the plans available to s as of t0 will have eleven members, a0,.....,a10. Assume, in addition, that were s to implement such a plan at t0, s would do a0 at t0, a1 at t1, and so on.

(2) The *expected utility of plan p* for s in T is the sum of the s-expected utility of all the actions in p.[3]

(3) *Plan p is adequate* for s in T as of time t if and only if as of t, no alternative plan for s in T has a higher expected utility for s than p.

(4) *Plan p is best* for s in T as of t if and only if as of t, p's expected utility for s in T is higher than the expected utility for s of all other plans then available to s.

Assume, to simplify matters, that as of any time there are always plans that are best for s in T.

(5) *Plan p is feasible* for s in T as of t if and only if as of t, each of the actions that p specifies s do at each node in T is such that s can then do that action.

To ascertain whether a plan is feasible as of a time, we can (unrealistically) view s as determining whether at each choice point the plan prescribes an action that s is then able to perform. If s comes across some plan-prescribed action at some node that s believes s won't then be able to do, s discards the plan as not feasible.

Best plans need not be feasible. From the vantage point of n0, ac is Ulysses' best plan. But this plan is not feasible for him as it requires that he ignore the Sirens at t1, something he cannot then do.

These definitions permit a concise characterization of sophisticated choice. Let ni be a node in s's decision tree, T. Then as a sophisticated chooser, s selects and implements the best of the feasible plans available to s in T at ni as of ti.

We are now in a better position to understand McClennen's reservations about SM. The SFM acts as a sophisticated chooser in dynamic contexts. Sophisticated choice presupposes a separability principle, "SEP." This principle governs a rational agent's subsequent preferences in a decision tree. It requires that

> what determines preference at any point in a sequence of decisions to be made is what would govern preference at that point, were the agent to confront that choice de novo, as a new decision problem, i.e., not against the background of any previous decisions. (CM and RC, 20)

SEP has implications for what plans are feasible for agents in sequential choice situations. It is SEP, for instance, that constrains the set of plans available to Ulysses at his point of initial choice.

As I understand McClennen, SEP is unacceptable. The constraint it imposes on rational dynamic choice is therefore also unacceptable. SEP is not acceptable because it violates a criterion of adequacy for conditions that constrain rational choice. In most general terms the criterion is this:

> rational choice.....[is] choice of efficient means to given ends.....I propose to adopt a strong form of this criterion, according to which a given condition does not qualify as a constraint on rational choice <u>unless</u> it can be shown that violating the condition involves the agent in choice of means insufficient to his ends. That is, I take the existence of a pragmatic argument for a given axiom as not merely sufficient for its qualifying as a rationality condition but as <u>necessary</u>. (DC and R, 12)

Reformulated, the criterion says that a condition, C, on rational choice is adequate if and only if it involves the agent in a choice of means sufficient to his ends. The criterion can be amplified, I think, in this way: Let R be a theory of rationality and let C be a condition presupposed by R. C, that is, is a necessary condition of R.

Crit: C is adequate if and if only if there is no alternative to R, R*, such that (i) R* does not presuppose C, and (ii), R*-rational agents do better in certain situations - "hard case situations" - than R-rational agents, and (iii) R*-rational agents do at least as well as R-rational agents in all situations other than hard case ones.

Understood in this way, Crit really specifies a standard of adequacy for theories of rationality. The general intuition underlying Crit may still be unclear. The following passage, I believe, sheds some light on this matter.

> such [a sophisticated] agent, unlike the resolute agent, must forego certain opportunities or expend valuable resources. This is to say that parametric reasoning works against such continuing interests of the agent--that agents who are capable of adopting and carrying through on plans will do better, over time, than do those who always reason parametrically. (CM and RC, 16-17)

Taking our cue from this passage, we can now finally formulate McClennen's worry about SM: SEP violates Crit. Since SM is commited to SEP, SM is inadequate. There are hard case situations in which those who can only choose in an SM-rational manner will fail to do as well as those who are capable of being resolute.

To assess this central claim of McClennen's, we need to say something about hard case situations, and we need to compare and contrast resolute choice with sophisticated choice.

1.3 Hard Case Situations

PD-like situations of a certain sort and relevantly similar sequential choice situations are the two varieties of hard case situations. Reconsider Ulysses' case. Think of his present and future selves as discrete agents each with "its own agenda of preferences." If we think of one-person sequential cases in this way, then it is fairly simple to see that such cases display important similarities to PDs. As a basis for comparison, consider PD matrix 8.2. Numbers represent expected utilities with

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	1,1 (a)	10,0 (b)
Remains silent	0,10 (c)	9,9 (d)

Figure 8.2 Matrix 8.2

Sundance's utilities ranked first. Think of (a), (b), (c), and (d) as "agreement outcomes" that would result were Butch and Sundance - both SM-rational agents - to agree to act in a certain fashion and then act in the agreed-on manner when called upon to do so. Assume Gauthier is right in supposing that the principle of minimax relative concession (MMRC) governs the ex ante agreement that underlies a fair and rational cooperative venture.[4] Then Butch and Sundance, acting on MMRC will both elect to remain silent; they will do the cooperative thing. On the assumption that MMRC "expresses the principle of utility maximization in the context of bargaining,"[5] choosing in accord with the principle should satisfy Butch's and Sundance's strongest ex ante preferences - their preferences to cooperate. Now suppose the time comes when each is called upon to comply with his ex ante agreement. Each has as his strongest preference at this later time, the preference to do whatever will maximize his own utility regardless of the behavior of the other. That is, the preference of each at this time is determined by SM. Matrix 8.2 illustrates that no matter what the other does, each in light of the preferences each now has does best by not complying. So an argument from dominance yields the result that each should now confess.

In both this case and the Ulysses' case agents or selves (whatever the case may be), as SFMs, maximize with respect to the preferences they have at discrete times. In addition, in both cases there is a need to coordinate the preferences between past and future selves if the gains of an optimal outcome are to be enjoyed: If Butch and Sundance each has as his strongest ex ante preference and as his strongest ex post preference the preference to cooperate, each would be able to comply with the mutually beneficial interest-constraining agreement. Similarly, if it were possible for the preferences of Ulysses' present self and those of his future self (when the latter existed) to be coordinated, Ulysses' would be able to disregard the singing Sirens and sail right by the island.

McClennen's claim is that resolute choosers do better than SFMs in cases similar to the ones just considered that meet these conditions: (1) Each agent (or self) knows or has reasonable grounds for supposing that every agent (or self) is rational. (2) Agents (or selves) are "transparent" to each other. That is, an agent is directly aware not only of the rational disposition of others, but is also aware, or at least has good assurance, of how the others will act, given their estimation of how the agent will act. Deception is therefore impossible. Cases that satisfy condition (2) are those in which the counterfactual independence condition is violated. Recall, the condition states that no matter what choice agents (or selves) make, an agent (or self) still makes that choice no matter what choices the others make. This condition would be violated, for example, if Butch's actions were related to those of Sundance's in a way depicted by matrix 8.3. (See page 237 below. The numbers indicate conditional probabilities of a choice on a choice.) In cases such as these, Butch will cooperate with Sundance if and only if Butch wants to do the cooperative thing and he expects Sundance to want to do the same as well.

	Butch	
	Confesses	Remains silent
Sundance		
Confesses	0.99	0.01
Remains silent	0.01	0.99

Figure 8.3 Matrix 8.3

1.4 *Resolute Choice*

McClennen tells us that resolute choice involves

> 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*, i.e., sequences of choices to be made over time and subject to various contingencies, 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. If such an agent is successful in this regard, then it can be said that what he chooses *ex*

> <u>post</u> to do is consistent with what he resolved (or planned) to do. Such an agent can be described as a resolute chooser. (CM and RC, 16)

> While the sophisticated agent disciplines <u>ex</u> <u>ante</u> choice to his projection of <u>ex</u> <u>post</u> independently based choice, the resolute agent disciplines <u>ex</u> <u>post</u> choice to <u>ex</u> <u>ante</u> choice of a plan. (DC and R, 10)

> Sophisticated choosers do the very best they can, given the constraints imposed by the need to adjust present choice to future given behavior. This is what Ulysses does and what most of us do. But.....such an approach to choice involves a retreat to second-best.....If Ulysses manages to save himself from the cursed isle by means of wax, hemp, and agency arrangements, still he could have done even better if he had simply <u>resolved</u> to sail right by the island and pay the singing sisters no mind. But so, also, if rational players who know each other to be such do well enough by devising various precommitment schemes or other devices to provide one another with incentives to cooperate, they could do better still by simply resolving to so cooperate. (PD and RC, 101)

From these passages it is possible to extract an account of <u>resolute</u> <u>choice</u>: Let ni be a node in s's decision tree, T. Then as a resolute chooser, s selects and implements the best plan at s's disoposal in T at ni as of ti.

Compare resolute choice with sophisticated choice. As a sophisticated chooser, s selects and implements the best of the <u>feasible</u> plans available to s in T at ni. In terms of our examples, sophisticated Ulysses at n0 implements plan b. In contrast, resolute Ulysses saves himself the cost of precommitment, resolves to ignore the Siren's song, and then ignores their song. He implements the best-from-n0 plan, ac.[6] Similarly, each of resolute Butch and resolute Sundance selects the plan which calls for an agent to make and keep the interest-constraining agreement, and then proceeds to implement that plan. SFMs Butch and Sundance, however, each end up confessing. In this way resolute agents do better than SFMs similarly situated, at least so seems to claim McClennen.

1.5 Straightforward Maximizers versus Resolute Agents

This last claim of McClennen's can be challenged. Consider a hard case PD-like situation with SM-rational agents Butch and Sundance. Each knows that the other is SM-rational and each is transparent to the other. Transparency ensures that the counterfactual independence condition is violated. We may consequently take matrix 8.3 as accurately portraying the way in which Butch's actions are related to Sundance's actions: Each will confess if and only if the other does. Each will remain silent if and only if the other does. Under conditions of transparency, as we remarked earlier, it is not possible to dissemble. It is not possible, for example, for Butch to make an agreement with Sundance to remain silent with the intention of not later doing his part. Being transparent, Sundance would be aware of this intention of Butch's and would refuse to make the agreement in the first instance. Assume that matrix 8.1 is the relevant utility matrix, but this time assume that numbers represent not expected utilities but just plain utilities with Butch's values listed first. The expected utility of confessing for each party is (1)(0.99)+(0.01)(10) = 1.09. This is less than the expected utility of cooperating for each which works out to (0)(0.01)=(9)(0.99) = 8.91. On the assumption that it is SM-rational for each to make an agreement to do the cooperative thing, SM, under the present assignment of probabilities and utilities, requires that each keep the agreement. SFMs will therefore do the "cooperative" thing in this hard case situation. But then they would do no worse than they would have done had they chosen resolutely. For presumably the best plan available to each at the outset calls for each to make and keep an agreement to do the "cooperative" thing.

Consider the case of Ulysses. If Ulysses' selves are transparent to each other, then Ulysses' present self (let's imaginatively suppose) is aware of the rational disposition of his future self and (let's suppose again) is aware of what his future self will do. Ulysses will therefore sail on and ignore the Sirens if and only if he now has the desire to do this and he now expects his future self to have a similar desire as well. But then choosing resolutely or choosing in a sophisticated manner will once again yield identical results.

Consider, now, a really hard case situation in which though Butch and Sundance know each other to be SM-rational, they are not transparent to each other. So assume that in this authentic PD situation the counterfactual independence condition is satisfied. The SM-rational thing to do here is to violate the agreement to do the cooperative thing. As I understand McClennen, resolute choosers in this situation would also fail to adhere to the interest-constraining agreement. They would renege since the best plan at the disposal of these agents sanctions violation: Suppose Butch is debating what plan to implement. Assume that of the interesting possibilities, he could implement either the plan that calls for cooperation, or the one that prescribes reneging. Which of these is best for him? Suppose Sundance does the cooperative thing. Then Butch does best by implementing the latter plan. Suppose Sundance reneges. Then again Butch does best by implementing the second plan. No matter what Sundance does, Butch does best by implementing the second plan. This plan, it seems, is therefore best for Butch. Since positions are symmetric, the plan is best for Sundance as well.

It appears, to collect results, that transparent SM-rational agents do no worse than transparent resolute agents in PD-like situations in which the counterfactual independence condition is violated. The same is true of "opaque" SM-rational agents in authentic PDs.

It is evident that something has gone wrong here. These are not the results anticipated by McClennen. Suppose, in light of this observation, McClennen is construed as recommending that even in these very hard PD situations, resolute choosers would adopt and implement a plan that requires making and keeping an agreement to remain silent. The notion of what plan is best, assuming such agents still select and implement best plans as of a time and a place, would require reinterpretation. It might then be claimed that such resolute choosers would do better than would SFMs in a similar position. This claim, however, would once again be mistaken: Resolute choosers of this sort would not, contrary to what this proposal assumes, find themselves in an authentic PD. The "preference problem" ensures that this is so.

1.6 The Preference Problem

The preference problem arises because of certain special features of authentic PDs. If agents are to be in this sort of PD, at least two things must be true. First, the counterfactual independence condition must be satisfied. Second, assuming there is a "cooperative" outcome, agents must rank their preferences for outcomes from most preferred to least in this fashion: (1) Unilateral noncooperation; (2) mutual cooperation; (3) mutual noncooperation; (4) unilateral performance of the act the mutual performance of which would result in the cooperative outcome. Rational agents who face an impending PD may prefer _ex ante_ to cooperate. They may, for example, find it advantageous to make an interest-constraining agreement. But if their _ex post_ preferences to keep the agreement are stronger than their _ex post_ preferences to violate unilaterally, then it seems that they would not be in a genuine PD. The problem of the PD, after all, has its roots in the fact that agents have preferences that they rank in the manner just delineated. We may now formulate a necessary condition for being in an authentic PD:

NCPD: If agents are to be in an authentic PD, these agents must have have PD preferences that are "standardly ranked."

McClennen's discussion of rational cooperation generates a problem about preference because McClennen sometimes gives the impression that resolute choosers could be in an authentic PD even though they fail at the relevant time to have preferences that are standardly ranked.

Responding to a worry of the economist Sen, for example, McClennen says

> One can mark here what appear to be two distinct, although closely connected issues. One concerns whether certain ways of arguing that rational agents should cooperate in a Prisoner's Dilemma situation might be dismissed on the grounds that the reasons cited imply that those who are so disposed do not face a genuine Prisoner's Dilemma. The second is whether some particular account of how persons might come to cooperate in such a situation could be faulted on the grounds that it "resolves" the problem in an purely *ad hoc* manner. It would seem, for example, that those who are concerned about the welfare of others, and who are led thereby to act in a cooperative manner can be said, in so behaving, to reveal their concern for others. However, if one supposes them to cooperate for that reason, then they do not really face a Prisoner's Dilemma. Moreover, such a "resolution" of the problem has the air of being *ad hoc*.
>
> The model I have proposed is not subject to either of these objections. Following Sen, I start with the assumption that the agents' preferences for *outcomes* conform to the pattern of a classic Prisoner's Dilemma situation. I simply move from there to challenge the distinct assumption that preferences for outcomes, abstractly considered, must be taken as controlling for preferences over actions and, hence, for choice. (CM and RC, 18-19)

He further informs us that

> Again, it may be argued that what counts against my model is that the very concept of preference itself implies that the agent will be disposed, at each choice point in time, to choose an alternative that is maximally preferred and that the plan the agent adopts must have the property that it calls upon the agent to make choices at each choice point that are consistent with the preferences he has at that point in time. Call this the principle of Dynamic Consistency. This principle is clearly one to which my proposed model is faithful. (CM and RC, 19-20)

Finally, in a footnote to a sentence in a passage in which McClennen gives a reinterpretation, in terms of choosing resolutely, of Gauthier's claim that an agent can behave as a constrained maximizer, McClennen explains that

> Since on the account offered here a rational agent does not ever act contrary to the preferences he has for actions at the time of choice, and, correspondingly, does not choose other than a utility maximizing action, it is perhaps misleading to describe him as a constrained maximizer. He maximizes in an unconstrained sense his preferences for available actions, although not his preferences over (separably considered) outcomes. (CM and RC)

How do we interpret "maximally preferred" and "utility maximizing action" in the last two passages if we are to suppose, consistent with what the first passage seems to suggest, that the relevant resolute agents would in fact be in an authentic PD? There are a number of interesting possibilities intimated by McClennen himself. These appeal to preferences that have as their objects different outcomes.

Assume that a number of agents expect to find themselves (together) in a PD at some future time and are now at the first nodes in their decision trees. When faced with the problem of coordinating with each other, they may have preferences for the implementation of plans whose consequences are judged to be best. Such agents treat their preferences "among feasible sets of actions as sensitive to more than just their preferences for the corresponding outcomes of those actions." (CM and RC, 15) On the other hand, they may have preferences for outcomes "abstractly considered" - outcomes of actions that could result at a particular point in time if those actions were performed at that time. Such preferences are shaped by a concern for outcomes that still remain realizable in the future and are not, unlike the former preferences, sensitive to any decisions that have been made in the past. Clearly we have here two different sorts of preference, the first sort having as object the implementation of plans, and the second, having as object "outcomes abstractly considered." Call the first sort "p preferences" and the second sort "o preferences."

In the passages just cited, McClennen suggests that resolute agents like resolute Butch and resolute Sundance, facing a coordinating problem, choose "an alternative that is maximally preferred." What sort of preference is being referred to here, a p preference or an o preference? Alternatively, is it p preferences or o preferences that such resolute agents take "as controlling for preferences over actions and, hence, for choice"?

Suppose it is o preferences. Suppose, that is, that Butch and Sundance act in conformity with their o preferences. Then if they are in an authentic PD they will end up not cooperating. As a result they fare no better than they would if they were sophisticated choosers. Suppose, alternatively, that the relevant preferences are p preferences. Suppose, that is, that they discipline their preferences "to conform to a plan whose consequences [they] judge to be superior." (CM and RC, 20) On the possibility that interests us here, the best plan in some sense of 'best' calls for resolute agents to cooperate. So assume that Butch and Sundance, acting to satisfy maximally their p preferences, do the cooperative thing. Then NCPD will not be satisfied. It won't be satisfied because Butch and Sundance have as "controlling" or as maximally preferred _ex post_ preferences to cooperate and not _ex post_ preferences to do the noncooperative thing. They have "non-standardly ranked" PD preferences. Since both are rational, I am assuming that their _ex post_ preferences to cooperate are stronger than their _ex post_ preferences to defect. I am not supposing the converse is true and that they simply fail to act on their strongest preferences at the time of decision. If they were to fail in this way, they would not be rational. Since NCPD specifies a condition that is necessary if agents are to be in an authentic PD, these resolute agents won't be in such a PD.

There is a third possibility. Why not suppose that agents have both p preferences and o preferences? Having o preferences ensures that they are in an authentic PD. But when it comes time for action, these agents act on their p preferences.

I believe McClennen would not endorse this position, as the last two passages cited strongly intimate. Further evidence that he would reject this third option is provided by the following passage:

> the suggestion that a rational agent will, at a certain point in a decision tree, choose other than what she prefers at that point plays hob with the whole notion of revealed preference: if the agent chooses A rather than B, then there is a perfectly appropriate sense in which what the agent really prefers (the preference revealed by choice) is A rather than B. (PD and RC, 102-103)

If I am right about the preference problem, then a heavy explanatory burden is placed on a theory of rationality, R, according to which R-rational agents can escape an authentic PD. Presumably, if such agents are able to escape this type of dilemma, and not something that resembles but is not an authentic PD, they must have non-standardly ranked PD preferences. If they have non-standardly ranked PD preferences, an explanation is owed of how these agents would be in an authentic PD in the first place. Such an explanation is not impossible to come by. In fact there is a way to interpret resolute choice that provides just what is needed:

Maximizing conceptions of rationality such as SM are often held in conjunction with the thesis that the contents of individuals' preferences are beyond rational assessment. This is, for example, Professor Gauthier's conviction.[7] Since this "non-constraining" thesis and a theory of the maximizing variety are logically independent, the one may be held without the other. Suppose one renounces the non-constraining thesis. Then one might try to defend a notion of "rational preferences" or "context sensitive preferences."[8] It might, for instance, be urged that not all preferences are rational and that what preferences it is rational to have partly depends on the situation one is in. Whatever this notion of rational preference amounts to, if it is to contribute to a resolution of an authentic PD, it must minimally entail both that an _ex ante_ preference for cooperation is rational as is an _ex post_ preference to act cooperatively. Now one could identify rationality with the maximization of one's rational preferences. This alternative to SM - "ASM" - might be some such theory according to which an act is rational if and only if none of its alternatives maximizes the satisfaction of its agent's rational preferences to a greater extent than it does. ASM could then be used to demonstrate the rationality of cooperation in one-shot PDs. The connection between ASM and resolute choice might not unreasonably be claimed to be this: Not all preferences, according to this way of conceptualizing matters, including - presumably - "standard" _ex post_ PD preferences to act noncooperatively, are rational. But an *ex ante* PD preference for cooperation and an *ex post* PD preference to act cooperatively are rational. Agents who resolve to act cooperatively and then act on that resolve when the time comes can be understood to be acting on their rational preferences. The solution conceives agents as having standard *ex ante* and standard *ex post* PD preferences. ASM-rational agents act, however, in conformity with the subset of these preferences that are rational, and not in conformity with the preferences that are strongest. The strongest preferences an agent has at a time need not be those that are rational at that time. The solution would therefore be one to an authentic PD.

It is premature, of course, to pass judgment on whether such a solution would be acceptable. What is significant is that a solution along these lines stresses that if rational agents - 'rational' in some sense of the term - are to escape an authentic PD, it must not be the case that the strongest preferences these agents have at a time of choice are those that are "controlling for choice." This is something, however, that McClennen does not seem to accept. He seems commited to the view that at each choice point in time, rational agents choose alternatives that are maximally preferred.

In conclusion, contrary to what McClennen appears to believe, SM does not violate Crit. In hard cases in which each agent is transparent to all agents and knows that every agent is rational, SFMs do no worse than resolute choosers. In situations of this kind, it is straightforwardly rational to do the "cooperative" thing. In really hard authentic PD situations, SFMs again do no worse, nor do they do any better, than resolute choosers who have non-standard *ex post* PD preferences. The reason this time being that such resolute choosers would not in fact be in authentic PDs.

Notes

[1] See Edward F. McClennen, "Prisoner's Dilemma and Resolute Choice," in _Paradoxes of Rationality and Cooperation_ eds., R. Campbell and L. Sowden (1985), Vancouver: The University of British Columbia Press, 94-104; Edward F. McClennen, "Dynamic Choice and Rationality," forthcoming; and Edward F. McClennen, "Constrained Maximization and Resolute Choice," forthcoming in _Social Philosophy and Policy_. Page references to these works appear in parentheses in the text, beginning from the first to the third, as follows: (PD and RC, page number); (DC and R, page number); (CM and RC, page number).

[2] In pages 3 and 4 in "Dynamic Choice and Rationality," McClennen tells us that

> A plan specifies what is to be done at each subsequent choice point in the decision tree that might be reached, given previous choices specified in the plan, and various contingent events.

[3] There may be a problem with this definition, as Professor Feldman indicates: Suppose my sequence of action is like this: a1-->a2-->a3-->outcome o. Suppose these are all true: If I were to do a1, I would get outcome o. If I were to do a2, I would get o. If I were to do a3, I would get o. Suppose the value of o for me is 10 points. Then the expected utility of the plan prescribing the above sequence of actions is 30 points. This seems wrong.

To evade this problem, let's sitipulate - arbitrarily I realize - that no such sequences of action occur in s's decision tree.

[4] David Gauthier, _Morals By Agreement_ (1986), Oxford: Clarendon Press, p. 14.

[5] See _Morals By Agreement_, p. 145 and p. 151.

[6] A potential worry here is this: Some best plans, unlike feasible plans, seem to require that agents act in ways that they cannot. Plan ac, for example, requires that Ulysses ignore the Sirens. How can rationality require that one do what one is incapable of? McClennen responds in this way:

> [R]esolute choice enjoys two advantages not enjoyed by sophisticated choice. First, while precommitment is typically costly, resolute choice has the attractive property that it can achieve whatever precommitment can achieve without the costs in question. Second, the specific use to which resolute approach has here been put turns on the idea of disciplining future choice to holistically oriented evaluation of the whole sequence of choices to be made. In contrast the sophisticated approach involves tailoring our choice of a whole plan to projected independent determined choices at subsequent stages. The former, it seems to me, is the more promising view: it invites us to think of ourselves as more than merely (passive) predictors of our own future choices, as capable of disciplining future choice, when to do so can be shown to be in our interest. In such cases, a retreat to a sophisticated approach seems very much like an admission of weakness of will. While the inability to carry through in certain situations may be a fact about human nature, it is unclear why a theory of how to behave in the face of such an inability (namely, to adopt a sophisticated approach) should be taken as other than a theory of how best to proceed under conditions of imperfect rationality. ["Dynamic Choice and Rationality," p. 17]

7. *Morals By Agreement*, pp. 25, 26, 34.

8. Richard Brandt, for example, argues for something like this. See Richard Brandt, *A Theory of the Good and the Right* (1979), Oxford: Clarendon Press, especially Chapter 6. Derek Parfit has an interesting discussion on whether desires can be "intrinsically rational" or "rationally required" in section 46 in Derek parfit, *Reasons And Persons* (1986), New York: Oxford University Press.

CHAPTER 9

CONCLUSION

1.1 Doom

I am the source of sad news: Butch and Sundance read Chapter 8. Each ends up spending ten years in jail.

1.2 The Contractarian's Dilemma

I want to conclude with some general comments on why I think the contractarian approach to "justifying" morality is unlikely to succeed. When I speak about the "contractarian approach," I have in mind primarily contractarian theories of the type advanced by Gauthier. So let's review the essential features of that approach.

On Gauthier's scheme a morality is to be justified in relation to the actual (considered) preferences of rational agents. The task is to show that no matter what the preferences of these persons, they have reason to be moral: Each, concerned with furthering her own good, would do best for herself by conforming to the morality that is the outcome of a rational bargain among them. Rational bargaining specifies the terms of rational agreement - it specifies the content of a particular morality. Gauthier must demonstrate that compliance with this morality, the morality that is the "product" of rational bargaining, is advantageous to all parties concerned. Toward this end, the strategy is to show that these persons would comply with the interest-constraining requirements of the agreed-upon morality as, roughly, it is in their long-term interest to do so.

This type of contractarian approach is particularly engaging because, if successful, it promises to accomplish two very noteworthy goals: (i) It generates morality "as a rational constraint from the non-moral premises of rational choice." (M by A, 4) (ii) It shows that the apparent conflict between considerations of morality and those of rational self-interest is merely apparent. There is no real clash here because on this approach moral principles are a subset of rational principles of choice: "To choose rationally, one must choose morally." (M by A, 4)

If we keep these two goals in mind, I think it is easy to see why this type of contractarianism will probably founder. Gauthier begins his project by assuming that the agents in relation to whom a morality is to be justified are SM-rational. This is essential if "goal (i)" is to be accomplished: SM as a "weak and widely accepted" precept of rational choice makes no pretense to being a moral one. Furthermore, it is presupposed at the outset, or it must be so, that SM is not a defective standard of rationality. If it were, there would be little reason to be interested in the outcome of a bargain among _SM_-rational agents.

Moreover, since CM and SM are extensionally equivalent in "non-strategic" contexts of choice, any defect in SM would also infect CM. It's evident, though, that Gauthier believes CM to be beyond reproach.

The major hurdle, however, that such SM-rational persons face, as Gauthier tells us, is not rational bargaining, but rational compliance with the injunctions of the morality that is the outcome of rational bargaining. The problem of compliance arises because these injunctions, being moral ones, are interest-constraining. The PD serves as a nice device to show that it is not SM-rational to comply with the prescriptions of the agreed-upon morality: MMRC requires that you do the cooperative thing in a (real) PD, SM proscribes such action.

Gauthier endeavors to solve the compliance problem by "reinterpret[ing] the utility-maximizing conception of practical rationality" (M by A, 182): It's certainly CM-rational to comply with the requirements of MMRC (under specified conditions) provided your fellow interactors are trustworthy constrained maximizers. But this way of resolving the compliance problem introduces a complexity. There are now two rival standards of reason, SM and its competitor. I suggested earlier that if you are a SFM, you will not be moved by this solution to the problem of compliance, not unless you have been given reason to believe that CM is superior to SM. Gauthier tries to supply such a reason by arguing that as a SFM it would be rational for you to change your very conception of rationality - you would do best for yourself if you were to become a constrained maximizer. Perhaps we can take the choice argument as indicative of some sort of evaluative test for assessing theories of rationality. I expressed reservations about the choice argument. But the worry with Gauthier's brand of contractarianism that I am now interested in formulating arises, I think, even if we suppose the choice argument sound. In fact, I think Gauthier countenances a dilemma:

Either there is a sound argument, or an acceptable evaluative test, that establishes the superiority of CM to SM, or there is not. Suppose there is. Then CM is better than SM - *it's* the legitimate standard of reason. But CM is a moral principle, it's interest-constraining. I think it's precisely because CM has this feature that compliance with the interest-constraining requirements of MMRC is CM-rational. On the supposition that CM is better than SM, or that CM is really *the* standard of rationality, Gauthier can sustain the claim that moral principles are a subset of rational principles of choice. But the cost, of course, of arguing in this way should be obvious: The contractarian argument would loose much of its interest since it would fail to accomplish "goal (i);" it would fail to derive morality from the non-moral premises of a theory of rational choice: SM but not CM is such a non-moral theory.

Assume there were some alternative to SM - a principle of rationality, R, that had these features: R requires compliance with the prescriptions of MMRC, R is not a moral principle, and R is superior to SM by the lights of some cogent evaluative test.[1] Wouldn't Gauthier then be able to evade the first horn? The answer, unhappily, is "No." If R were not a moral principle, then "goal (ii)" that Gauthier's contractarian theory strives to achieve would have to be abandoned. It would not be true that to choose rationally one must choose morally.

Suppose, then, that there is no evaluative test, or if you want, suppose the choice argument fails. Then once again Gauthier's contractarianism collapses: It may be CM-rational to comply with the interest-constraining prescriptions of MMRC, but as a SFM, you will have been provided with no reason whatsoever to adhere to these interest-thwarting requirements.

1.3 Conclusion

"Should I be moral?" Glaucon, or Hobbes' Fool, or the moral sceptic, might ask. If we take these people to be asking "Is it rational to be moral?" or "Is it prudentially obligatory to do what's morally obligatory?" the answer is clear. It is not. The PD, I think, is inescapable.

Notes

[1] I'm not sure a theory of this sort is possible. But assume, for the sake of argument, that such a theory exists.

BIBLIOGRAPHY

Arneson, Richard J. "Locke versus Hobbes in Gauthier's Ethics," Inquiry 30 (1987), 295-316.

Baier, Annette C. "Piligram's Progress," Canadian Journal of Philosophy 18 (1988), 315-330.

Baier, Kurt. The Moral Point of View (1958), Ithaca, New York: Cornell University Press.

Baier, Kurt. "Rationality and Morality," Erkenntnis 11 (1977), 197-223.

Barry, Brian B. and Hardin, Russell., eds. Rational Man and Irrational Society (1982), Beverly Hills: Sage Publications.

Brandt, Richard B. A Theory of the Good and the Right (1979), Oxford: Clarendon Press.

Braybrooke, David. "Social Contract Theory's Fanciest Flight," Ethics 97 (1987), 759-764.

Braybrooke, David. "The Insoluble Problem of the Social Contract," Dialogue 15 (1976), 3-37. This paper also occurs in Paradoxes of Rationality and Cooperation, eds. Campbell, R. and Sowden, L. (1985), Vancouver: The University of British Columbia Press, 277-306.

Campbell, R. "Gauthier's Theory of Morals by Agreement," forthcoming in Philosophical Quarterly.

Campbell, R. "Moral Justification and Freedom," The Journal of Philosophy 85 (1988), 192-213.

Campbell, R. and Sowden L., eds. Paradoxes of Rationality and Cooperation (1985), Vancouver: The University of British Columbia Press.

Copp, D. "Contractarianism and Moral Scepticism," forthcoming.

Copp, D. and Zimmerman, D., eds. Morality, Reason and Truth (1984), Totowa, New Jersey: Rowman and Allanheld.

Danielson, Peter. "The Visible Hand of Morality," Canadian Journal of Philosophy 18 (1988), 357-384.

Danielson, Peter. "The Moral and Ethical Significance of TIT FOR TAT," *Dialogue* 25 (1988), 449-470.

Darwall, Stephen L. *Impartial Reason* (1983), Ithaca, New York: Cornell University Press.

Farrell, D. M. "Taming *Leviathan*: Reflections on Some Recent Work on Hobbes," *Ethics* 98 (1988), 793-805.

Farrell, D. M. "Hobbes as Moralist," *Philosophical Studies* 48 (1985), 257-283.

Farrell, D. M. "Reason and Right in Hobbes' *Leviathan*," *History of Philosophy Quarterly* 1 (1984), 297-314.

Feldman, Fred. "On The Advantages of Cooperativeness," forthcoming in *Midwest Studies in Philosophy*.

Gauthier, David. "Moral Artifice," *Canadian Journal of Philosophy* 18 (1988), 385-419.

Gauthier, David. *Morals By Agreement* (1986), Oxford: Clarendon Press.

Gauthier, David. "Bargaining Our Way Into Morality: A Do-It-Yourself Primer," *Philosophical Exchange* 3 (1982a), 15-27.

Gauthier, David. "No Need for Morality: The Case of the Competitive Market," *Phiosophic Exchange* 3 (1982b), 41-54.

Gauthier, David. "The Irrationality of Choosing Egoism - A Reply to Eshelman," *Canadian Journal of Philosophy* 10 (1980), 179-187.

Gauthier, David. "David Hume: Contractarian," *The Philsophical Review* 88 (1979), 3-38.

Gauthier, David. "Thomas Hobbes: Moral Theorist," *The Journal of Philosophy* 76 (1979), 547-561.

Gauthier, David. "Economic Rationality and Moral Constraints," in *Midwest Studies in Philosophy 3, Studies in Ethical Theory*, eds. French, Peter A., Vehling, Theodore, E., Wettstein, Howard. (1978), Morris: The University of Minnesota Press.

Gauthier, David. "The Social Contract as Ideology," *Philosophy and Public Affairs* 6 (1977), 130-164.

Gauthier, David. "Coordination," Dialogue 14 (1975), 195-221.

Gauthier, David. "Reason and Maximization," Canadian Journal of Philosophy 4 (1975), 411-432.

Gauthier, David. "The Impossibility of Rational Egoism," The Journal of Philosophy 71 (1974a), 439-456.

Gauthier, David. "Rational Cooperation," Nous 8 (1974b), 53-65.

Gauthier, David. The Logic of Leviathan (1969), Oxford: Oxford University Press.

Gauthier, David. "Morality and Advantage," The Philosophical Review 76 (1967), 460-475.

Hampton, Jean. "Can We Agree on Morals?," Canadian Journal of Philosophy 18 (1988), 331-355.

Hampton, Jean. Hobbes and the Social Contract Tradition (1986), Cambridge: Cambridge University Press.

Hampton, Jean. "Hobbes State of War," Topoi 4 (1985), 47-60.

Hardin, Russell. "Bargaining For Justice," forthcoming.

Held, Virginia. "Rationality and Reasonable Cooperation," Social Research 44 (1977), 708-744.

Kavka, G. "A Review of Morals By Agreement," Mind 96 (1987), 117-121.

Kavka, G. Hobbesian Moral and Political Theory (1986), Princeton, New Jersey: Princeton University Press.

Kavka, G. "Right Reason and Natural Law in Hobbes's Ethics," The Monist 66 (1983), 120-123.

McClennen, E. F. "Prisoner's Dilemma and Resolute Choice," in Paradoxes of Rationality and Cooperation, eds. Campbell, R. and Sowden, L. (1985), Vancouver: The University of British Columbia Press, 94-104.

McClennen, E. F. "Constrained Maximization and Resolute Choice," forthcoming in Social Philosophy and Policy.

McClennen, E. F. "Dynamic Choice and Rationality," forthcoming.

Mendola, Joseph. "Gauthier's Morals by Agreement and Two Kinds of Rationality," Ethics 97 (1987), 765-774.

Molesworth, W., ed. The English Works of Thomas Hobbes (1939), London: John Bohn.

Morris, Christopher W. "The Relation Between Self-Interest And Justice In Contractarian Ethics," forthcoming in Social Philosophy and Policy.

Nelson, Alan. "Economic Rationality and Morality," Philosophy & Public Affairs 17 (1988), 149-166.

Parfit, Derek. Reasons And Persons (1986), New York: Oxford University Press.

Paul, Ellen Frankel., Miller, Fred D. Jr., Paul, Jeffrey., eds. Ethics & Economics (1985), Oxford: Basil Blackwell Publisher Limited.

Rescher, Nicholas. "Rationality And Moral Obligation," Synthese 72 (1987), 29-43.

Resnik, Michael, D. CHOICES An Introduction To Decision Theory (1987), Minneapolis: University of Minnesota Press.

Ripstein, Arthur. "Gauthier's Liberal Individual," forthcoming.

Ripstein, Arthur. "Foundationalism in Political Theory," Philosophy & Public Affairs 16 (1987), 115-137.

Smith, Holly. "Gauthier's Moral Contract," forthcoming.

Sobel, J. H. "The Need for Coercion," in Pennock, J. R. and Chapman, J. W., eds. Coercion: Nomos XIV (1972), Chicago and New York: Aldine & Atherton, 148-177.

Sobel, J. H. "Interaction Problems for Utility Maximizers," Canadian Journal of Philosophy 4 (1975), 677-688.

Sobel, J. H. "Utility Maximizers in Iterated Prisoner's Dilemmas," Dialogue 15 (1976), 38-53.

Sumner, L. W. "Justice Contracted," Dialogue 26 (1987), 523-548.

Wallace, James D. Virtues And Vices (1978), Ithaca, New York: Cornell University Press.