RATIONAL INTENTION, RATIONAL ACTION

by

Joseph Mintoff

BMath(Hons), GDipPhil, MA(Hons), *W'gong*

A thesis submitted for the Degree of
Doctor of Philosphy
of the Australian National University
Division of Philosphy and Law
Research School of Social Sciences

30 April 1993

PART II

RATIONAL COOPERATION

Chapter Four

David Gauthier on Rational Cooperation

The most troubling form of moral deflationism, I claimed in the Introduction, is one granting that moral utterances have truth values, and even granting that some of them are true, but denying that they make any difference. Why should I be moral? – or so asks this most troubling of moral deflationists. There are (at least) two ways of answering this challenge. On the one hand, one may simply repudiate the demand that morality answer to each and every person's reason, and insist that this an inappropriate way of attempting to justify morality. On the other hand, one may accept the deflationist's demand, and try to argue for the possible rationality of morality within the bounds of each person's reason. David Gauthier, in his *Morals by Agreement*, attempts to take moral deflationism seriously, and to show why, under certain conditions, each person's reason does indeed counsel the way of morality. His argument has two basic parts. First is a contractarian analysis of morality: one morally ought to perform some action when it is what one would agree to do were one to employ a rational bargaining procedure, from a rational initial bargaining position, in a situation of perfect information. Second is a rationalistic justification for cooperation: when others are sufficiently cooperatively disposed and sufficiently knowledgeable about how each is disposed to behave, it is rational to do what one would in this manner rationally agree to do. In this chapter I will concentrate on the second part of Gauthier's project. We shall see that his argument is not without problems.

§1 The Self-Interest Theory, and rational cooperation

It seems clear that irrational persons may do things which fail to promote their interests, or what they value, and – worse still – things actually contrary to their interests; and equally as clear that rational persons may expect to do the best they could to promote their interests, or what they value. It seems, then, that a group of rational persons may expect to do better than a group of irrational persons in promoting their interests, or what they value. Yet this is not always the case. There are some situations in which rational persons can expect to do very poorly indeed, and certainly very much worse than their irrational cousins.

[1] You find yourself in Hobbes's state-of-nature.[1] There is a group of you, each armed to the hilt and suspicious of the others. None of you draws a distinction between what you morally may or may not do, and, in particular, each has the right, capability, and also no compunction against, using force upon others, if each perceives that this will benefit the pursuit of what they value most. There are clear advantages to be had by the possession of such a right, whether in the exploitation of others or in the defence against just such exploitation. You face a choice whether or not to relinquish your right to use force against the others. Should you do so or not?

This is not the world's most difficult decision problem. Pick any other person in the state-of-nature. Me, for instance. Either I will relinquish my right to use force against you, or I will not. On the one hand, if I will *not* relinquish my right, then the outcome of your also not doing so (namely, a nervous and somewhat tenuous balance of terror) would be much better for you than the outcome (namely, my ability to exploit you with threats of force) of your unilaterally relinquishing your right. On the other hand, if I *will* relinquish my right, then the outcome of your *not* doing so (namely, your ability to exploit me with threats of force) would again be much better for you than the outcome (namely, a calmer peace) of relinquishing your own right. Whatever I will do, then, it is obvious you rationally ought *not* to relinquish your right. And this is so even though the outcome of *joint* non-relinquishment of the right to use force (namely, a tenuous balance of terror) would be much worse for each of us than the outcome (namely, a somewhat calmer peace) of *joint* relinquishment of this right.

[1] I shall assume – solely for the sake of illustration – that Hobbes's state-of-nature is to be interpreted as a Prisoner's Dilemma. This is, by now, a common interpretation. See D. Gauthier, *The Logic of the Leviathan*, (Oxford: Clarendon Pr., 1969), pp, 14 ff., and pp. 76 ff.

We can justify all of these statements on the basis of the following two claims:

(S) If an agent is free to perform an action A, then the agent rationally ought to A if and only if the agent-relative expected-value of doing A exceeds that of doing any alternative to A.

(P1) Each of us faces an independent choice between the actions of relinquishing our right, or not; whatever the other person does, each values the outcome of not relinquishing to that of relinquishing; however, each of us values *joint* relinquishment of the right to *joint* non-relinquishment.

The first assumption, (S), is, of course, what I have been calling the standard formulation of the Self-Interest Theory. The second, (P1), is a description of the central features of the situation I have just described. Since it will be important later, it needs some explaining.

[2] Assumption (P1) defines what is in modern parlance is referred to as the *Prisoner's Dilemma*.[2] The three clauses of the assumption may be represented diagrammatically.[3]

[2] The name comes from the original example used to illustrate situations described by assumption (P1). For a detailed introduction to this example, and to the Prisoner's Dilemma more generally, see R. D. Luce & H. Raiffa, *Games and Decisions*, (New York: Wiley, 1957), pp. 94-102, and R. Campbell, 'Background for the Uninitiated,' in R. Campbell & L. Sowden, *Paradoxes of Rationality and Cooperation*, (Vancouver: Univ. Brit. Columbia Pr., 1985): 3-41.

[3] The three clauses are explicitely identified, amongst others, by G. Harman, 'Rationality in Agreement: A Commentary on Gauthier' "Morals by Agreement,"' *Soc Phil Pol* 5 (1988): 1-16. Since the sort of situation I will be examining is, in a number of ways, simplified, I will say a few words in justification.

(a) Only two agents are involved: you and I. Typically, though, these are not the central cases (P. Pettit, 'The Prisoner's Dilemma and Social Theory,' *Politics* 20 (1985): 1-11, and R. Hardin, 'Pragmatic Intuitions and Rational Choice,' in A. Diekmann & P. Mitter (ed.s), *Paradoxical Effects of Social behaviour: Essays in honour of Anatol Rapoport*, (Heidelberg: Physica Verlag, 1986): 27-36). In response, it needs to be noted that, even if they are not always central, two-person cases are widespread. Examples include: exchange (D. Mueller, *Public Choice*, (Cambridge: Cambridge Univ. Pr., 1975), Ch. 2); international relations (H. Wagner, 'The Theory of Games and the Problem of International Cooperation,' *Amer Pol Sci Rev* 77 (1983): 330-46, S. J. Brams, *Superpower Games – applying Game Theory to Superpower Conflict*, (New Haven: Yale Univ. Pr., 1985)); and the practises of promise-keeping, truth-telling and so on (D. Gauthier, 'Morality, Rational Choice and Semantic Representation: a Reply to my Critics,' *Soc Phil Pol* 5 (1988): 173-221).

(b) There is clearly only one outcome it would be rational for us to agree on: mutual cooperation. Typically, though, cooperating involves (at least) two separate problems: firstly, coming to some agreement concerning a joint course of

The first clause says that each of us faces an independent choice between the actions of relinquishing our own right or not. It will save words if, henceforth, we say that you *cooperate* (and denote this by 'C') if you relinquish this right to use force against me; and *defect* ('D'), if you do not. In short, the first clause says that *each agent faces an independent choice between cooperating and defecting*. Each pair of actions we may perform has an outcome, and each of us assigns values to these outcomes. Since both of us face a choice between two actions, there are four (= 2 x 2) possible outcomes. These outcomes, and the values each of us attaches to them, may be depicted as follows:

		You	
		C	D
Me	C	Civil Society c,c	You exploit Me s,t
	D	I exploit You t,s	War of all against all d,d

action (the 'bargaining' problem); and secondly, actually performing our assigned task, given the agreement we have made (the 'compliance' problem). (See D. Gauthier, *Morals by Agreement*, (Oxford: Clarendon Pr., 1986), p. 118, and K. Baier, 'The Conceptual Link Between Morality and Rationality,' *Nous* 16 (1977): 78-88.) In response, I choose in this thesis to concentrate only on the second problem – compliance – rather than the first – bargaining. Thus I concentrate on a case in which the resolution to the first, bargaining, problem is clear.

(c) The outcome it would be rational for us to agree on – mutual cooperation – is very unstable: *whatever the other person does*, it is better for each not to cooperate. Typically, though, the stability of an agreement is thrown into doubt by much less truculant factors (See M. Taylor, *Possibility of Cooperation*, (New York: Cambridge Univ. Pr.), Ch. 2, and A. K. Sen, 'Choice, Orderings and Morality,' in S. Korner (ed.), *Practical Reason*, (New Haven: Yale Univ. Pr., 1974): 54-66). In response it need only to be noted that if it can be shown cooperating is rational, even in conditions most hostile to the emergence of cooperation, then we may assume it can be rational under these less hostile conditions.

(d) Hobbes's state-of-nature is a particularly violent form of conflict, in which the concerns of each are simply with survival. Typically, though, the problems of cooperation are not as violent, and do not involve such a narrow conception of self-interest. In response, it is important to note that conflicts of the sort described in the Prisoner's Dilemma arise, strictly speaking, from the presence of what I have called agent-relative value, and not just from the presence of narrower forms of self-interest. As such, the type of situations addressed by the text is broader than just Hobbes's admittedly violent state-of-nature, which I use solely for the purposes of illustration. (D. Parfit, *Reasons and Persons*, (Oxford: Clarendon Pr,. 1984), pp. 95 ff., sec. 36, F. C. T. Moore, 'The Martyr's Dilemma,' *Analysis* 45 (1984): 29-33.)

The values I assign to possible outcomes are listed first, and are: my exploiting you (=t, Temptation payoff); mutual cooperation and the possibility of civil society (=c, Cooperation payoff); mutual defection and the war of all against all (=d, Defection payoff); and, your exploiting me (=s, Sucker payoff). Clearly, t > c > d > s, since I value most the outcome of my exploiting you (t), next mutual cooperation and the possibility of civil society (c), third the war of all against all (d), and, worst of all, your exploiting me (s).

This first clause means there can be no literally binding agreement between us.[4] Neither of us can perform any actions which will render us literally incapable of defecting. On the one hand, neither can force *the other* to cooperate, for the difference between the one of us and the other is not so considerable that the one can claim a benefit from the use of force over the other, that the other cannot also claim against the one. On the other, neither can force *themselves* to cooperate, for, unlike Odysseus neither has the rope and strong sailors to physically restrain themselves from not cooperating,[5] and cannot literally bind themselves to the cooperative action.

The second clause says that whatever the other person does, each values the outcome of not relinquishing to that of relinquishing their right. It will save words if, henceforth, we say that action A *dominates* action B for me if, whatever anyone else does, I value the outcome of doing A to that of doing B. Hence, the second clause says that *for each agent, defection dominates cooperation.* In terms of the diagram, if you were to cooperate, then defecting would get me an outcome (namely, my being able to exploit you with threats of force) I value most at t, which is greater than the value, c, of the outcome (namely, a calmer peace and the possibility of civil society) I would get were I to cooperate. On the other hand, if you were to defect, then defecting would get me an outcome (namely, a somewhat tenuous balance of terror) I value at

[4] That agreements in the Prisoner's Dilemma are not 'binding' is an assumption often made, but not often explained. See R. L. Cunningham, 'Ethics and Game Theory: the Prisoner's Dilemma,' *Papers on non-market Decision Making* 2 (1967): 11-26, who assumes that agreements are not 'binding'. D. Braybrooke, 'The Insoluable Problem of the Social Contract,' *Dialogue* 1 (1976): 3-37, and R. Campbell, 'Background for the Uninitiated,' in R. Campbell & L. Sowden (eds.), *Paradoxes of Rationality and Cooperation*, (Vancouver: Univ. Brit. Columbia Pr., 1985): 3-41, assume that coercion is not 'feasible'. J. Elster, *Ulysses and the Sirens*, (Cambridge: Cambridge Univ. Pr., 1979), ch. 2, is one who does discuss the different ways someone might be bound.

[5] See Homer, *The Odyssey*, 12.154-200

d, which is greater than the value, s, of the outcome (namely, your being able to exploit me) I would get were I to cooperate. Either way, the outcome is better for me if I defect, and the same applies, *mutatis mutandis,* to you.

This second clause means that there can be no binding agreement (in a different sense) between us. What is implied in particular is that there are no actions available to us which could set up incentives such that each player ends up valuing the outcome of their own cooperation to that of their own defection, whatever the other person does. In particular, it entails that coercive enforcement of cooperation, benevolence towards the other person, and mutual similarity of goals are either non-existent or limited to an extent consistent with still valuing the outcome of defection to that of cooperation, whatever others do.

The final clause of (P1) says that each of us, however, values *joint* relinquishment of the right to *joint* non-relinquishment. It will save words if, henceforth, we say that an outcome O_1 is *optimal* when there is no other outcome which everyone values at least as much as O_1, and someone values more than O_1. Hence, the final clause says that *the outcome of joint defection is not optimal.* In terms of the diagram, if both of us cooperate, then there would be an outcome – namely, the possibility civil society – which each of us values second-best, at c; if both of us defect, then there would be an outcome – namely, the war of all against all – which each of us values only third-best, at d. The possession of the right to use force in the state-of-nature leads, as Hobbes believed, and as I shall assume, to a war of all against all. There is the universal realisation that the right to use force might advantageously be employed in pre-emptive attacks upon others, and in this way is there universal uncertainty about whether or not one will be the object of just such an attack. Such a situation is a state of war, a state which 'consisteth not in battle only, or the act of fighting; but in a tract of time, wherein the will to contend to battle is sufficiently known'. For Hobbes, it is a state in which peace of mind is hardly possible, where the conveniences and securities of civil society surely lacking, and in which life is 'solitary, poor, nasty, brutish, and short'.[6]

[6] The relevant passages are taken from Hobbes, *Leviathan*, ed. C. B. MacPherson, (Harmondsworth: Penguin, 1985), chap. 13, pp. 185, 186.

In Hobbes's state-of-nature, a group of irrational agents – who cast off their right to use force against the others, and obtain as a result the possibility of civil society – would do much better than a group of rational agents – who retain their right, and with it the war of all against all. Many find this disturbing. Hobbes himself was certainly disturbed by this conclusion, and his solution to the problem involved our mutual relinquishment of the right to use force, and the setting up of a sovereign to enforce it. But is Hobbes, or anyone else for that matter, correct to suppose that the state of nature, as disturbing as it is, contains within it a justification for the rejection of the Self-Interest Theorist's conception of rationality?

§2 The Prisoner's Dilemma, and David Gauthier

Notoriously, Hobbes has problems explaining why persons initially disposed to act solely to produce outcomes best for themselves could, or would, come to relinquish their right to use force.[7] For everyone to *agree* to do so is, of course, not to solve the problem, since, whatever others do, any particular person will do better by reneging on such an agreement. However rational it might be for me to agree to cooperate, and cast off my right to use force against you, it remains irrational for me actually to do so. We are left with the residual problem of explaining why anyone would have reason to keep such an agreement. In this section I will introduce David Gauthier's attempt to show how it might be rational to keep just such agreements.

[1] Gauthier tries to show that, under certain conditions, it is rational to make such an agreement, and, if it is, then it is rational to keep it.[8] To be able do this, Gauthier changes the focus of discussion, and considers the question of the *type of agent* it is best to be. He distinguishes, in particular, between two types of agents: what he calls straightforward, and constrained, maximisers.

[7] See J. Hampton, *Hobbes and the Social Contract Tradition*, (New York: Cambridge Univ. Pr., 1986), ch.s 2 and 3 for a review of the secondary literature on Hobbes's discussion of this point.

[8] There are three works in particular I will concentrate on in my discussion of Gauthier: 'Reason and Maximization,' *Can J Phil* 4 (1975): 411-433 (hereafter cited as 'RM'); 'Deterrence, Maximization, and Rationality', *Ethics* 94 (1984): 474-495 ('DMR'); and *Morals by Agreement*, (Oxford: Clarendon. Pr., 1986) ('MA'). DMR is reprinted in D. MacLean (ed.), *The Security Gamble: Deterrence Dilemmas in a Nuclear Age*, (Totowa, NJ: Rowman & Allanheld, 1984).

The first, straightforward, maximisers are just those agents who are disposed to do what the Self-Interest Theory tells them to do.[9] They are such that if the expected value to them of the outcome of doing A exceeds that of doing any alternative action, they do A. In particular, they keep a promise, or honour an agreement, if and only if it maximises expected-value for them to keep the promise, or honour the agreement.

The second, constrained, maximisers are disposed, in certain situations, to keep agreements for their own sakes. Says Gauthier:

> We shall therefore identify a constrained maximiser thus: (i) someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies, and approach what she would expect from the co-operative outcome determined by minimax relative concession; (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies.[10]

This is quite a general (and somewhat complex) definition, more so than I need for my discussion. It is amenable to useful simplification in the sorts of situations we are considering.

Take the first clause, which is concerned with the type of joint practices a constrained maximiser is conditionally prepared to adopt. These are ones (to take the second condition mentioned, first) where the utility she expects will "approach what she would expect from the co-operative outcome determined by minimax relative concession". Minimax relative concession is Gauthier's preferred rational bargaining procedure,[11] but we can replace any reference to it in the simple sorts of situations we are considering, since the advice it gives (and the advice any adequate bargaining procedure would give) is that we both cooperate, and cast off our right.[12] These practices are, further,

[9] See Gauthier, RM, pp. 428 ff, and MA, pp. 167 ff.

[10] Gauthier, MA, p. 167. Note that Gauthier uses the term 'utility' in the way I use the term 'expected-value'.

[11] This procedure is described in MA, ch. V, as well as 'Rational Cooperation,' *Nous* 8 (1974): 53-65, and 'Bargaining and Justice,' *Soc Phil Pol* 2 (1985): 29-47.

[12] This is not, strictly speaking, correct. Suppose t+s>2d, so that each of us values a half chance of being exploited and half chance of exploiting to the certainty of mutual defection. If, further, joint mixed strategies are introduced (Gauthier, MA,

ones such that "the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies." Since the outcome to be expected were everyone to employ individual strategies (that is, were everyone to follow the dictates of the Self-Interest Theory) is the war of all against all, it is plain that the joint strategy of mutual cooperation is (to say the least) no worse than this. In the sorts of situations we are considering, then, a constrained maximiser is conditionally disposed to cooperate.

Take now the second clause of the definition, which is concerned with the conditions under which a constrained maximiser will actually base her action on the rationally acceptable joint strategy. She acts in this way "should her expected utility be greater than what she would expect were everyone to employ individual strategies". What, exactly, does this mean? Gauthier makes this clear later:

> Her disposition to co-operate is conditional on her expectation that she will benefit in comparison with the utility she could expect were no one to co-operate. Thus she must estimate the likelihood that others involved in the prospective practice or interaction will *act co-operatively,* and calculate, not the utility she would expect were all to co-operate, but the utility she would expect if she co-operates, given her estimate of the degree to which others will *co-operate.* Only if this exceeds what she would expect from universal non-co-operation, does her conditional disposition to constraint actually manifest itself in a decision to base her actions on the co-operative joint strategy.[13]

It is clear Gauthier intends the cooperation of a constrained maximiser to be dependant upon her expectations regarding whether or not the others will cooperate. I make an estimate of the likelihood you will do your part of the joint strategy – casting off your right to use force – and then calculate the expected-value to me of my doing similarly. If this exceeds the expected-value to me of our war of one against other (*not* the expected-value to me of my doing otherwise), will I do similarly. Furthermore, when a constrained maximiser meets uncooperative persons, she 'does not play into their hands by basing

p. 120), then a reasonable bargaining procedure will *not* recommend the joint (pure) strategy [1(C,C)], but rather the joint (mixed) strategy [1/2.(C,D), 1/2.(D,C)]. I shall ignore this point, but see J. H. Sobel, 'Constrained Maximization,' *Can J Phil* 21 (1991), pp. 34 ff. for details.

[13] Gauthier, MA, p.169, emphasis added.

her actions on the joint strategy she would like everyone to accept, but rather, to avoid being exploited, she behaves as a straightforward maximizer.'[14] In the simple sort of situations we are considering, then, Gauthier's definition amounts to the following:

(CM) An agent is a constrained maximiser ('CM') if and only if: the agent cooperates if and only if the expected value to the agent of doing so exceeds that of mutual non-cooperation. The expected value of an action is calculated on an estimation of the probability that the other *will cooperate.*

An important implication of this definition, which we will need later, is that a constrained maximiser will cooperate if they believe the other person will cooperate, and will not if they believe the other will not.

[2] Gauthier then argues that, given certain conditions, one rationally ought to be a constrained, rather than a straightforward, maximiser. He imagines a situation, occurring before the issue of cooperation or not has arisen, in which an agent is to make a choice between becoming a straightforward, or a constrained, maximiser. He denotes by u the expected utility the agent could expect were each person to act on the basis of an individual strategy – in the case we are considering, this is the value, d, the agent assigns to the war of all against all. He denotes by u' the expected utility should all act on the cooperative joint strategy – in our case, the value, c, the agent assigns to civil society. Clearly, u' exceeds u. Gauthier then has the agent argue as follows. (For simplicity I will replace Gauthier's u by my *d*, and his u' by my *c).*

> 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, *[d]*.
>
> Suppose that 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 *[c]*. If they are not so disposed, I employ a maximizing strategy and expect *[d] as* before. If the probability that others are disposed to constrained maximization is *p*, then my overall expected utility is [pc + (1 –pd)].

[14] Gauthier, MA, p. 169

> Since *c* is greater than *d*, $[pc + (1 - p)d]$ is greater than *d* 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.[15]

Thus we see that Hobbes was correct to insist that rational persons in the state of nature could come to reason differently, and not act solely to produce outcomes the best for themselves. They would cease to reason as straightforward maximisers, and come instead to reason like constrained maximisers, cooperating when they have an expectation that others, too, will cooperate.

This particular argument, though, depends on a few quite strong assumptions, unstated but nevertheless very important. The first unstated assumption is that the agents, in choosing how they are to be disposed, face a choice only between constrained, and straightforward, maximisation, or that all other choices are inferior. What the argument shows though, if anything, is that under certain conditions it is better to be a constrained maximiser than a straightforward one. It follows from this that you *ought* to be a constrained maximiser only if these are the only choices or only if all other choices are clearly inferior. The second unstated assumption of this argument is that both the constrained and the straightforward maximiser will appear in their true colours.[16] On the supposition I have adopted straightforward maximisation, I must expect others not to cooperate only if I expect others to *know* I have adopted straightforward maximisation. And, on the supposition I have adopted constrained maximisation, I must expect other constrained maximisers to cooperate only if I expect them to *know* I have adopted constrained maximisation. Thus, it needs to be assumed that, after having adopted a disposition, the other will be able to tell what disposition I have adopted. It needs to be assumed, in other words, that dispositions are transparent.

The fact Gauthier's argument depends on such strong assumptions might (and has) provoked objections to the relevance of his argument.

[15] Gauthier, MA, p. 172

[16] Gauthier presents the argument in the text, and then himself points out that it depends on this second assumption, which he calls the assumption of *transparency* (MA, p. 173). In the following pages, he introduces a weaker form of the assumption, that of *translucency*, and proceeds to a second argument for the claim that, under certain conditions, it is rational to be a constrained maximizer. Since, in this thesis, I will not be concerned with this more complex second argument, or with the objection making it necessary, I will not introduce it.

After all, the dispositions amongst which one can choose presumably include others than just the two Gauthier considers,[17] and such dispositions will, in any case, be nowhere near as transparent as his argument requires.[18] Important as they are, I will not in this thesis pursue these points. In the first place, I believe (but will not in this thesis argue) that even when one does consider more realistic types of situations – in which there are more dispositions than just the two Gauthier introduces, and when these dispositions are not transparent – then it will still remain the case that some non-expected-value maximising disposition is rational (though, I should point out, it will not be that of constrained maximisation).[19] In the second place, and more importantly, even when we do restrict our attention to the sort of unrealistic case required for this argument to work, its conclusion is still sufficiently threatening to some. In particular, it is threatening to those who would defend the Self-Interest Theory, who would want, even in the unrealistic case we are considering, to deny it is rational to cooperate.

[17] On the restrictive nature of the disposition choice, see, for example, R. J. Arneson, 'Locke versus Hobbes in Gauthier's Ethics,' *Inquiry* 30 (1988), p. 313, A. Nelson, 'Economic Rationality and Morality,' *Phil Pub Affairs* 17 (1988), p. 156, D. Copp, 'Review of Morals by Agreement', *Phil Rev* 91 (1989), p. 413, J. Buchanan, 'The Gauthier Enterprise,' *Soc Phil Pol* 5 (1988), pp. 81 ff., P. Danielson 'The Visible Hand of Morality,' Can J Phil 18 (1988), p. 376, and H. Smith, 'Deriving Morality from Rationality,' in P. Vallentyne, *Contractarianism and Rational Choice: Essays on Gauthier's Morals by Agreement*, (New York: Cambridge Univ. Pr., 1991), pp. 238-9.

[18] On the assumption of transparency, and that of translucency, see, for example, A. Baier, 'Pilgrim's Progress,' *Can J Phil* 18 (1988), p. 328, A. Nelson, 'Economic Rationality and Morality,' *Phil Pub Affairs* 17 (1988), p. 160, R. J. Arneson, 'Locke versus Hobbes in Gauthier's Ethics,' *Inquiry* 30 (1987), p. 309, R. Hegselmann, 'Rational Egoism, Mutual Advantage, and Morality – a Review of D. Gauthier: Morals by Agreement,' *Erkenntnis* 31 (1989), sect. 2.1, D. Copp, 'Contractarianism and Moral Skepticism,' and G. Sayre-McCord, 'Deception and Reasons to be Moral,' in P. Vallentyne, *Contractarianism and Rational Choice: Essays on Gauthier's Morals by Agreement*, (New York: Cambridge Univ. Pr., 1991), pp. 220-1, and 191-5 respectively. In defense of the relevance of translucency assumptions, though, see J. Killcullen, 'Utilitarianism and Virtue,' *Ethics* 93 (1983): 451-66, J. Glover, *What Sort of People Should there be?* (Harmondsworth: Penguin, 1984), and R. H. Frank, *Passions within Reason*, (New York: Norton, 1988).

[19] I have in mind, here, the discussion of the so-called 'iterated' Prisoner's Dilemma. In an environment of many different types of agent who face the iterated Prisoner's Dilemma, it may very well be rational to adopt a disposition – so-called TIT-FOR-TAT – which reacts to previous cooperation of others with cooperation, and previous defection with defection, whether or not it is expected-value-maximising to do so. I shall not here argue that this is so, but see R. Axelrod, *The Evolution of Cooperation*, (New York: Basic Books, 1984) for details of the iterated Prisoner's Dilemma.

Such is Gauthier's attempt to argue it might be rational to cooperate, and keep agreements, even if one is free to do otherwise, and it has the best outcome to do otherwise. The strategy is an indirect one: first to argue for the rationality of cooperative dispositions, and then to move from there to the rationality of cooperative actions. There are thus two general counter-strategies available for those who would wish to deny his conclusion: first, one may deny he has shown it is rational to be *disposed* to cooperation; and second, one may claim that, even if he had shown this, he would not have shown it was rational *actually* to cooperate. I shall discuss these objections in turn.

§3 Rational Cooperative Dispositions?

There has been much discussion of Gauthier's argument and some have not been totally convinced.[20] Some objections centre of the first half of Gauthier's strategy: his argument for the rationality of being *disposed* to cooperating. We shall see it is an argument with a number of problems.

[1] The first problem concerns Gauthier's assumption that there is a fixed probability, p, to be assigned to the other agent's being a constrained maximiser.[21] In order to understand the objection, we need to introduce the distinction between *parametric* and *strategic* choice – a distinction Gauthier himself endorses. He says *parametric* choice occurs when 'the actor takes his behaviour to be the sole variable in a fixed environment. In parametric choice the actor regards himself as the sole centre of action.' Contrasted with this is *strategic* choice, in which 'the actor takes his behaviour to be but one variable amongst others, so that his choice must be responsive to his

[20] There are a number of books and journal issues devoted to a discussion of Gauthier's views, as well as numerous reviews of *Morals by Agreement*. These include: *Ethics* 97 (1988); *Can J Phil* 18 (1987); *Soc Phil Pol* 5 (1988); P. Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on Gauthier's Morals by Agreement*, (New York: Cambridge Univ. Pr., 1991).

[21] R. Arneson, 'Locke versus Hobbes in Gauthier's Ethics,' *Inquiry* 30 (1987), pp. 313, A. Nelson, 'Economic Rationality and Morality,' *Phil Pub Affairs* 17 (1988), p. 160 fn. 11, R. Hesgelmann 'Rational Egoism, Mutual Advantage, and Morality – a Review of D. Gauthier: Morals by Agreement,' *Erkenntnis* 31 (1989), sect. 2.1, and P. Danielson, 'Closing the Compliance Dilemma: how it is is Rational to be Moral in a Lamarkian World,' in Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on Gauthier's Morals by Agreement*, (New York: Cambridge Univ. Pr., 1991), p. 302.

expectations of others' choices, while their choices are similarly responsive to their expectations.'[22]

The problem is then as follows. The situation Gauthier is concerned to address is one in which there are a number of agents interacting, in the knowledge they are doing so. In the first instance, they face a choice about whether to cooperate or not. Gauthier understands *this* choice as one in which each agent takes his behaviour to be but one variable amongst others, so that his choice must be responsive to his expectations of others' choices. He understands the choice of cooperation, or not, to be a strategic one. If this is the only choice then, as we have seen, it seems that mutual non-cooperation and the war of all against all is the rational outcome. To handle this problem, Gauthier introduces dispositions, and he asks: which is the best disposition to choose? In the second instance, then, the agents in question face the question about whether or not to choose to be disposed to cooperating. Gauthier understands *this* issue, strangely, as one in which each agent regards himself as the *sole* centre of action. He understands the choice of a disposition of cooperation to be a parametric one. The problem, then, is simply this: it is inconsistent to suppose (as Gauthier does) that, when agents are choosing to cooperate, they should assume the choices of others are not fixed, but that, when they are choosing dispositions, they should assume the choices of dispositions of others are fixed.

Certain passages suggest that Gauthier may be able to respond to this problem. For the role that a *choice* of disposition plays in his argument is not, it turns out, a central one:

> the idea of a choice among dispositions to choose is a heuristic device to express the underlying requirement, that a rational disposition to choose be utility maximizing. In parametric contexts, the disposition to make straightforwardly maximizing choices is uncontroversially utility-maximizing. We may therefore employ the device of a parametric choice among dispositions to choose to show that in strategic contexts, the disposition to make constrained choices, rather than straightforwardly maximizing choices, is utility-maximizing.[23]

[22] Both quotes are from Gauthier, MA, p. 21. For a more detailed explanation of the distinction, he refers the reader to J. Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality*, (Cambridge: Cambridge Univ. Pr., 1979), pp. 18-9, 117-23.

[23] Gauthier, MA, p. 183

There is a distinction between the rationality of *choosing* to be a CM, and that of *being* a CM, and it is clear from this passage that Gauthier places most importance on the second. Gauthier does not explicitly address the objection we are now considering, but one may speculate that he could use this passage to claim that, since the idea of a choice of disposition is in any case only a heuristic one, he is not committed to any problems there may be with it.

If this is the way Gauthier would reply to the objection, then it is not clear he would have said enough to dispel the problem. For even though Gauthier's position does not, strictly speaking, commit him to a view about how *choosing dispositions* is to be understood, whether parametrically or strategically, it turns out that neither option would be congenial. He faces a dilemma: choice of disposition is to be understood as a parametric, or a strategic, choice. On the one hand, if it is to be understood as a parametric choice, then Gauthier encounters the objection with which we started: it seemed inconsistent to suppose that, when agents are choosing *to cooperate,* they should assume the choices of others are not fixed, but that when they are choosing *dispositions,* they should assume the choices (of dispositions) of others are fixed. These choices are either both strategic or both parametric. On the other hand, if the choice of disposition is to be understood as a strategic one, then Gauthier encounters a very similar objection: it seems inconsistent to suppose that, when agents are *choosing* dispositions, they should assume the choices of dispositions of others are not fixed, but when they are considering the rationality of *having* dispositions, they should assume the dispositions that others have are fixed. Choosing dispositions and having dispositions are either both parametric or both strategic.

[2] The second problem with Gauthier's argument for the rationality of constrained maximisation centres on the following crucial part of the above passage: "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."[24]

Even if we assume transparency, this is not quite right. Suppose both of us adopt constrained maximisation, and are both transparent.

[24] Gauthier, MA, p. 172.

Since I am transparent, you come to believe that I am a constrained maximiser. You come to believe I will cooperate if and only if I expect that you are sufficiently likely to cooperate. Similarly, since you are transparent, I come to believe that you are a constrained maximiser, and thus come to believe you will cooperate if and only if you expect that I am sufficiently likely to cooperate. How does it follow from this that each of us may expect that the other is sufficiently likely to cooperate? In fact, this does not follow.[25] From the mutually known facts that (a) I will cooperate if and only if I expect that you will, and that (b) you will cooperate if and only if you expect that I will, it does not follow that either of us will, in fact, come to expect the other to cooperate, and so does not follow that we both will, in fact, cooperate. It is as if each of us is waiting for the other to make the first move, but, simply because we are waiting, we may each fail to make any move at all. How might we solve this problem?

[2.1] The simplest solution is the following. If CM is defined in such a way that CMs are disposed to act on the basis of an estimation of the likelihood the other *will cooperate,* then it does not follow two CMs would cooperate. Since the agent's dispositions are assumed to be transparent, it would be better to define a CM in such a way that they are disposed to act on the basis of an estimation of the likelihood the other *is a constrained maximiser.* In particular, it would be simplest to define a constrained maximiser thus:

(CM2) A person is a constrained maximiser if and only if: they cooperate if and only if they believe the other person *is a constrained maximiser.*

If this is our definition, then the required inferences will indeed be valid. On the one hand, a CM will cooperate with a CM. When a CM faces another CM, she comes to believe the other is a CM (since they are transparent), and since she herself is a CM (and so cooperates with

[25] As some have pointed out. See, for example, R. Campbell, 'Critical Study: Gauthier's Morals by Agreement,' *Phil Quart* 38 (1988), sect. 3.1, P. Danielson, 'Closing the Compliance Dilemma,' and H. Smith, 'Deriving Morality from Rationality,' in P. Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on Gauthier's Morals by Agreement*, (New York: Cambridge Univ. Pr., 1991), pp. 306-15 and 239-41 respectively, and J. H. Sobel, 'Constrained Maximization,' *Can J Phil* 21 (1991), sect. IV.2.

all those she believes are CMs), then she will cooperate. On the other hand, a CM will not cooperate with an SM. When a CM faces an SM, she comes to believe the other is an SM (since they are transparent), and since she is a CM (and so cooperates only with those she believe are CMs), then she will not cooperate.

The problem with this solution, though, is that it renders the notion of constrained maximisation circular.[26] Of course, circularity is not necessarily a problem. It may, for example, be a conceptual truth that something is red if and only if it appears red to standard observers under standard conditions. This is not a problem – even if circular – since most sighted people have a grasp of the concept independent of this truth. Sometimes, though, circularity is a problem. In particular, (CM2) claims that someone is a constrained maximiser if, and only if, they cooperate with all and only those they believe are also constrained maximisers. This is a problem – because circular – since no-one has a grasp of this concept independent of this statement. We need to be provided with an independent handle on constrained maximisation, which this circular redefinition does not provide.

[2.2] Richmond Campbell argues that Gauthier does have a way out of the problem we have discovered, and proposes his own redefinition of constrained maximisation.[27] It goes like this:

(CM3) A person is a constrained maximiser if and only if: (i) they have property R, and (ii) they cooperate if and only if they believe that the other person *has property R*.

The task for this definition is to find a substitution instance for 'R' that will give the intuitively right answers.

First, one could take 'R' to be 'will cooperate.' If so, then a person is a constrained maximiser if and only if: (i) they will cooperate, and (ii) they cooperate if, and only if, they believe that the other person will cooperate. This form of the definition will not do. Its first clause is defective, since it is perfectly possible to be a CM without cooperating, when, for example, one is the sole CM in a population of transparent SMs. Sensibly, one does not cooperate, but this does not mean one

[26] H. Smith, in 'Deriving Morality from Rationality,' p. 242 fn. 18, rejects a very similar suggestion of Danielson's for this reason.

[27] R. Campbell, 'Gauthier's Morals by Agreement,' *Phil Quart* 38 (1988): 343-364.

would not be prepared to cooperate with those who are themselves more cooperative, and so this does not mean one is not a CM. Its second clause is also defective, since it suffers from the same problems as the original definition, (CM).

Second, one could take 'R' to be 'is a CM'. If so, then a person is a constrained maximiser if and only if: (i) they are a CM, and (ii) cooperate if, and only if, they believe that the other person is a CM. This form of the definition will also not do. It is, of course, equivalent to (CM2), and thus shares its circularity. What we need, then, is some property different from that of being a CM which nevertheless no SMs will have. If we can find such an 'R', then it will turn out that CMs will cooperate with one another, but not with SMs, just as the argument requires.

Third, one could adopt one of Campbell's own suggestions, and take 'R' to be the property of 'being ready to reciprocate cooperation when making the second move in sequential PDs'.[28] A *sequential* Prisoner's Dilemma ('PD') has the same structure as the Prisoner's Dilemma we have been considering in this chapter, except that one agent knows what the other did, typically because the other agent acted first. If this suggestion is adopted, then a person is a constrained maximiser if and only if: (i) they are ready to reciprocate cooperation when making the second move in sequential PD, and (ii) they cooperate if and only if they believe that the other person is ready to reciprocate cooperation when making the second move in sequential PD. This form of the definition will still not do. The term 'ready' is troublesome. If it is as non-committal as it sounds, then we will run into the same problem that generated the need for a redefinition of constrained maximisation in the first place. For, surely, our original constrained maximisers – those who would cooperate if they expected the other to cooperate – were *ready* to cooperate, but as ready as they were, they kept each other waiting and did not necessarily end up cooperating.

Fourth, and finally, let us slightly modify Campbell's definition, and take 'R' to be 'would reciprocate when making the second move in sequential PD'. If so, then a person is a constrained maximiser if and only if: (i) they would reciprocate when making the second move in sequential PD, and (ii) they cooperate if and only if they believe that the

[28] R. Campbell, 'Gauthier's Morals by Agreement,' p. 351.

other person would reciprocate when making the second move in sequential PD. On this suggestion, a CM would thus use the fact that the other person would reciprocate in a sequential Prisoner's Dilemma to ground his own cooperation.

This suggestion also has its difficulties.[29] For what a person *would* do in some counterfactual situation is sometimes not a good indicator of what they *will* do in a similar, but actual, situation. Suppose, for example, you know I am a trustworthy, but untrusting, person. Since I am trustworthy, I myself will not attempt to exploit anyone, and would reciprocate when making the second choice in a *sequential* Prisoner's Dilemma. I satisfy property R, and you know it. Since I am not trusting, though, I fear that you might try to exploit me, and will not cooperate in *simultaneous* Prisoner's Dilemmas which are the ones we are actually concentrating on in this chapter. I will not cooperate, and you know it. This shows that Campbell's definition is inadequate. Suppose you are a CM. It follows, on Campbell's definition, that you will cooperate with anyone you believe has property R. You know (and so believe) that I have property R. Therefore, CM that you are, you cooperate with me in the simultaneous Prisoner's Dilemma which is the focus of concern in this chapter. You cooperate, while knowing (and so believing) that I will not. The CM, as Gauthier meant to define her, would however *not* cooperate were she to believe, as you do, that I will not cooperate. Recall Gauthier says that when a CM has reason to suppose the other agents would not cooperate, then she "does not play into their hands by basing her actions on the joint strategy she would like everyone to accept, but rather, to avoid being exploited, she behaves as a straightforward maximizer".[30] This final substitution for 'R' is inadequate.

Gauthier, then, faces problems with his argument for the rationality of constrained maximisation. On the one hand, he supposes that the choice between the CM and SM dispositions is to be a parametric one; but it seems that, on pain of inconsistency, it should be understood as a strategic choice. On the other hand, the argument he provides is not valid, given his official definition of constrained maximisation; and, it seems, there need be no definition adequate to

[29] See H. Smith, 'Deriving Morality from Rationality,' p. 242 fn. 18

[30] Gauthier, MA, p. 169

the argument. Gauthier's indirect strategy seems to have fallen at the first hurdle.

§4 Rational Cooperative Actions?

Other objections centre of the second half of Gauthier's strategy: his move from the rationality of cooperative dispositions to the rationality of cooperative action. For even if we were to grant that Gauthier had shown each of us rationally ought to be disposed to cooperating, he still needs to conclude that it is rational actually to cooperate. We shall see it is a move open to a number of objections.

[1] It is a move Gauthier believes he is entitled to make. He states, in *Morals by Agreement*, that '[i]imperfect actors find it rational to dispose themselves to make less than rational choices. No lesson can be drawn from this about the dispositions and choices of the perfect actor. If her dispositions to choose are rational, then surely her choices are also rational.'[31] Gauthier, it seems, endorses a principle like the following:

(B2) If you rationally ought to adopt the disposition that if you believe that p then you perform A, and if you do believe that p, and if nothing relevant to the adoption of the disposition has changed save what must be changed with the coming about of p, then you are rationally permitted to perform action A.

This principle is prominent in practical reasoning, and it is surely plausible. If Gauthier's argument for the rationality of constrained maximisation is correct, then you rationally ought to adopt the disposition that if you believe the other person will cooperate then you cooperate as well. If, as seems likely, the other person becomes similarly disposed, then you will come to believe they will cooperate, and that (presumably) nothing relevant to the adoption of the disposition has changed. In this case you are rationally permitted to cooperate, even if you are free not to, and it has the best outcome for you not to.

[31] Gauthier, MA, p. 186. Similar comments occur in DMR, p. 487. The principle (B2), below, is based on a very similar principle Gauthier endorses in 'Afterthoughts,' in D. MacLean, (ed.), *The Security Gamble: Deterrence Dilemmas in a Nuclear Age*, (Totowa, NJ: Rowman & Allanheld, 1984), p. 159. The principle there refers to intentions rather than dispositions but it seems clear that Gauthier uses the terms 'intention', 'disposition', and 'policy' interchangeably.

Alas, like other bridging principles, there are more than enough people prepared to deny the validity of the principle, (B2), Gauthier needs. One nay-sayer is Gregory Kavka himself:

> [Gauthier says CM] is more rational that the disposition to maximize expected utility, because those who possess the former disposition will have more opportunities for mutually beneficial interaction. This may well be true, but it hardly follows, as Gauthier believes, that particular acts of constrained maximization are rational. It may be rational to dispose oneself to perform irrational acts, as Thomas Schelling has shown with examples like that of the small country which, for the purposes of deterrence, rationally disposes itself to resist – irrationally – any invasion by its much large neighbour. Is constrained maximization an instance of a (possibly) rational disposition to perform irrational actions? I believe that it is.[32]

This is a serious objection, for it is an important part of Gauthier's argument that he can move from the rationality of the disposition to cooperate to the rationality of the cooperative action itself.

Consistent with his discussion of the rationality of constrained maximisation, Gauthier is prepared to admit that being disposed, for example, to nuclear retaliation could be rational, when it maximises utility to be so disposed, and, if it is, then it is rational for the nation to carry out its failed threat, even if only a nuclear holocaust will result. These claims are, to say the least, far from plausible. Perhaps sensing in *Morals by Agreement* that the onus is now on him, Gauthier directs us to other arguments: "Deterrence, we have argued elsewhere, may be a rational policy, and non-maximising choices are then rational."[33] Since this is the crux of the issue before us, let's look at these other arguments.

[2] The reference in the above quote is to Gauthier's discussion of the rationality of deterrence, in his 1984 paper 'Deterrence, Maximisation,

[32] G. Kavka, 'Review of "Morals by Agreement",' *Mind* 96 (1987), p 120. Other nay-sayers include G. Harman, 'Rationality in Agreement: a Commentary on Gauthier's "Morals by Agreement",' *Soc Phil Pol* 5 (1988): 1-16; D. Parfit, *Reasons and Persons*, (Oxford: Oxford Univ. Pr.: 1986); D. Lewis, 'Devil's Bargains and the Real World', in D. Maclean, (ed.), *The Security Gamble: Deterrence Dilemmas in a nuclear age*, (Totowa: Rowman & Allanheld, 1984): 141-54, and S. Darwell, 'Rational Agent, Rational Action,' *Phil Topics* 14 (1986): 33-57.

[33] Gauthier, MA, p. 186.

and Rationality.' I will start with this paper, and, in particular, in section V, where Gauthier considers various objections to his claim that, if it is rational to adopt deterrent policies, then, if they fail, it is still rational to act on them.

Gauthier considers four objectors; it is the second who is of present concern. We are considering, remember, the person who admits 'the rationality of some deterrent policies, but nevertheless insists that these policies, although fully rational, involve the performance of irrational actions if certain conditions are satisfied.'[34] Gauthier's response to this person as follows:

> How then does his position differ from mine, in which I claim that deterrent policies may be rational, and if rational, involve performance of actions which, in themselves and apart from the context of deterrence, would be irrational, but which, in that context, result from rational intentions and so are rational? Surely he grants the substance of my argument but expresses his agreement in a misleading and even paradoxical way, insisting that actions necessary to a rational policy may themselves be irrational. To assess an action as irrational is, in my view, to claim that it should not be, or have been, performed.[35]

Gauthier suggests here that there is little difference between himself and his objector's position. And in this he is right: both agree, in the circumstances, that it is rational to adopt the deterrent policy, and both agree that, were such a threat not met with the required action, anyone who adopted this policy would carry out the threat.

As a response, however, this will not do. Gauthier asks how his objector's position differs from his own, and, as we have just seen, the two positions are very similiar. Where they differ, though, is that Gauthier thinks the resulting action would be rational, whereas his opponent thinks it would be irrational. David Lewis puts the point as follows:

> I am the second objector, the one who says that "it may be rational to adopt an intention even though it would be, and one knows that it would be, irrational to act on it"; I claim that it may be "rational to commit oneself to irrational behaviour" (and also that it may be good to commit oneself to evil behaviour).

[34] Gauthier, DMR, p. 487.

[35] Gauthier, DMR, p. 487.

> Gauthier claims that my position is no different from his own. Not so; I deny what he firmly asserts, that there may be actions which "in themselves and apart from the context of deterrence would be irrational, but which in that context result from rational intentions and so are rational."[36]

Gauthier believes the objector to be confused, and to set him right, he makes it clear that when he says that an action is rational, he means that it ought to be performed. But it seems Gauthier is confused, for what the objector says is exactly that you ought to have the disposition, but you ought not perform the action.

Gauthier has responded to Lewis's objection – he claims that Lewis adopts a position he had failed to consider. Gauthier understands Lewis to be saying that since the disposition and the retaliatory action expressing it are different things, opposed judgements about them are consistent. It's a view he finds 'schizophrenic',

> [b]ut suppose I accept it. ... Suppose that I am a rational actor, considering now what to do should I find myself faced with ADVANT [that is, faced with someone who has ignored my threat]. If I know, as Lewis supposes that I do, that it would be irrational for me to RETAL [that is, retaliate] given ADVANT, then is it *possible* for me to form the intention to RETAL? It seems clear to me that it is not possible. If Lewis were to say that it *would* be rational for me to form the intention to RETAL, if I *could*, then I could understand, although not accept, his position. But I find that I do not understand it.[37]

The claim here seems to be that if I believe that it would be irrational for me to form the intention to RETAL, then it is not possible for me to form that intention, and this is inconsistent with the assumption that I *can* form this intention.

This claim is ambiguous, but on either interpretation of no help to Gauthier.[38] On the one hand, it might mean that if I believe it would be irrational for me to form the intention to RETAL, then it is not

[36] D. Lewis, 'Devil's Bargains and the Real World,' p. 143.

[37] D. Gauthier, 'Afterthoughts,' in D. Maclean (ed.), *The security gamble: deterence dilemmas in a nuclear age,* (Totowa, NJ: Rowman & Allanheld, 1984), p. 160. In a similar vein, S. I. Benn, in 'Deterrence or Appeasement? Or, on Trying to be Rational about Nuclear War,' *J Applied Phil* 1 (1984): 5-20, suggests that if one thinks retaliation is grossly immoral then one cannot form the intention to retaliate.

[38] G. Kavka, 'A Paradox of Deterrence Revisited,' in his *Morals Problems of Nuclear Deterrence,* (New York: Cambridge Univ. Pr., 1987), pp. 44-5.

possible for me, if I am to remain rational, to form that intention. That is, it is not possible for me to remain rational and to form the intention. On this interpretation, what Gauthier says is true, but of no use to him. It is of no use to him since Kavka claims that adopting the deterrent intention is precisely to conditionally intend to perform an *irrational* action – when one forms the intention, Kavka will say, one thereby becomes irrational. On the other hand, Gauthier's claim might mean that if I believe it would be irrational for me to form the intention to RETAL, then it is not possible for me, whether I remain rational or not, to form that intention. On this interpretation, what Gauthier says is false, and still of no use to him. It is false because agents can promote the formation of the intention by exposing themselves to external influences which will render them irrational in the relevant respects. It is of no use to Gauthier, since what is to stop his objectors from claiming that, similarly, it is not possible for agents to adopt constrained maximisation once they realise it is a disposition to perform non-maximising, and so presumably irrational, actions? Gauthier's claim in response to Kavka is either false or not to the point.

[3] Gauthier has another answer to this so-called paradox of deterrence. In response to the third objector in 'Deterrence, Maximisation, and Rationality,' Gauthier claims that the rational agent is the one who takes the big picture in their aim to fulfil their values:

> The fully rational actor is not the one who assesses her actions from now but, rather, the one who subjects the largest, rather than the smallest, segments of her activity to primary rational scrutiny, proceeding from policies to performances, letting assessment of the latter be ruled by assessment of the former.[39]

The objector can almost, but not completely, agree with this claim. To see why, note that people such as Lewis and Kavka are claiming that since there are two things under discussion – the disposition to retaliate, and the action of retaliating – then there are two evaluations to be made. The objector can almost agree with this claim because they do not deny that an evaluation can be made of the larger segments of

[39] Gauthier, DMR, p. 488.

her activity, such as dispositions. On the rationality of the disposition to retaliate Gauthier and his opponents are not necessarily in disagreement. The objector, though, cannot completely agree with this claim, since Gauthier insists that the evaluation of the smaller segments of activity are to be determined by the evaluation of the larger, while the objectors insist that they require separate evaluation. If the rational dispositions are those which maximise expected-value, then why aren't the rational actions also the ones which maximise expected-value? Gauthier, it seems, has no answer. On the rationality of actually retaliating, then, Gauthier and his opponents are still in disagreement.

This is not the only objection one might have. Kavka admits that 'there may be something to' this wider segments view, and that there are clear advantages of agents acting according to rules, plans, and policies, than on a case-by-case basis. The advantages includes lower decision costs, and more efficient coordination and cooperation. Even so, Kavka believes that

> our normal view of rationality also implies being prepared to change previously formulated plans or intentions when there are significant stakes involved and relevant new information about outcome is available. This is precisely the situation that arises when deterrence fails in [a Special Deterrence Situation]. There is much harm to be done by retaliation, and the benefit that motivated formation of the intention to retaliate – prevention of the offence – is now unobtainable.[40]

Gauthier supposes, in bridging principle (B2), that if it is initially rational to adopt a disposition to A when p, and if p, then it is rational A, *unless* something relevant to the adoption of the disposition has changed *except what must have changed with the coming about of p.* Thus, for Gauthier, the fact the one's deterrence has failed is no reason to reconsider one's newly formed intention to destroy the world. Kavka offers a different picture, and suggests that if it is initially rational to adopt a disposition to A when p, and if p, then it is rational A, *unless* there are significant stakes involved and it is clear that the disposition cannot now do the job for which it was adopted. Thus, for Kavka the fact that one's deterrence has failed is more than enough

[40] Kavka, 'The Paradox of Deterrence Revisited,' 45-6.

reason to reconsider one's newly formed intention to destroy the world. The objector can, then, accommodate the intuition towards which Gauthier gestures, but can do so without having to admit – what seems totally implausible – that if it is rational to be disposed to nuclear retaliation, then, in the unlikely case that such deterrence fails, it is also rational to retaliate.

Gauthier, then, also faces problems with his move from the rationality of the cooperative disposition to that of cooperative actions. Even if we suppose, then, that Gauthier has managed to show that it is rational to be disposed to cooperating, he still will have failed to show that it is rational actually to cooperate. This is because there are situations – those involving deterrence – which clearly may very well involve rational intentions to perform irrational actions. Gauthier's indirect strategy seems also to have fallen at the second hurdle, and he joins the advocates of the other bridging principles which we met in Chapter Two.

Conclusion

Gauthier's argument for the rationality of cooperation in the state of nature is not without its problems. In the state of nature, the Self-Interest Theory unconditionally counsels agents to retain the right to use force against others, even though the mutual retention of this right leads to the war of all against all, and short and miserable lives for each. Though many, including now David Gauthier, think this fact can provide some sort of justification for rejecting the voice of such reason, it still seems they are mistaken. Even if we were to accept the contractarian analysis of morality – that one morally ought to perform some action when it is what one would agree to do in certain circumstances – Gauthier cannot show it is rational to be moral. It seems there can be no rational morals by agreement.

Chapter Five

Some Prisoner's Dilemma is a Rational Dilemma

The precise details of Gauthier's argument that it is rational to cooperate in some Prisoner's Dilemmas are, as we have seen, problematic. His argument for the rationality of cooperative dispositions, on the one hand, depended illegitimately on the assumption that disposition choice is parametric, and, on the other, employed a defective notion of constrained maximization. Furthermore, his use of a bridging principle connecting rational cooperative dispositions to rational cooperative action was also unjustified, as shown by the so-called paradox of deterrence. Rational morals by agreement seem not to be possible. But only seems not possible, I say. The idea behind the argument is, I believe, essentially sound. In this chapter I shall argue that given certain conditions (to be specified below), the actions resulting from rational agreements are also rational, even if one is free to do otherwise and it has the best outcome for one to do otherwise (though these actions may very well be irrational absent those conditions).

§1 Dealing with Two Objections against Gauthier

Gauthier's argument for the rationality of cooperative dispositions, on the one hand, depended illegitimately on the assumption that disposition choice is parametric, and, on the other, employed a defective notion of constrained maximization. The argument can, however, be reformulated to avoid both of these problems.

[1] Gauthier's argument for the rationality of constrained maximization suffered, as we saw, from the problem of assuming that each agent could, before decision-making commenced, assign a fixed probability to the other's disposition. In other words, that each agent takes their choice of disposition to be the sole variable in a fixed environment – that choice of disposition is a parametric one. This problem can be remedied.

A better argument for the rationality of cooperative dispositions would assume, rather, that each agent needs to determine, and not take as given, the likelihood of particular choices of dispositions on the part of the others. In other words, that each agent takes their choice of disposition to be but one variable amongst others, so that his choice must be responsive to his expectations of others' choices, and so on. A better argument would assume that our choice of disposition is a strategic one. Indeed, we need to assume that all choice – and not just that of dispositions – is strategic. We need to assume each agent takes all their behaviour to be but one variable amongst others.

The assumption that choice is strategic leads immediately to the question: how am I, for example, to derive an expectation concerning your behaviour, given that I do not have such an expectation to start with? The key to answering this question is that I will be able to come to some expectation about what you will do if I can put myself in your place, and can assume that you will act rationally. I will be able to come to some expectation about what you will do, that is, if, first, I can determine what you rationally ought to do, and, second, I have conclusive evidence to believe that you would do what (I see) you rationally ought to do. These assumptions need to be explained in more detail.[1]

It needs to be assumed, first, that each one of us can determine what the other rationally ought to do. Each needs to be able to put himself in the other's place and reason as they might reason. To be able to do this, each obviously needs to be acquainted with the options and values of the other. But that each is acquainted with the other's options and values is itself a part of the decision situation with which, presumably, each of us will also have to be acquainted. To be *completely* able to put ourselves into the other's shoes, then, it must further be assumed that *each has conclusive evidence to believe* that

[1] These two broad assumptions are what lies behind most formal characterisations of strategic decision situations. See, for example, R. D. Luce & H. Raiffa, *Games and Decisions*, (New York: Wiley, 1957), ch. 2, esp. the summary on pp. 53-5, and D. Gauthier, *Morals by Agreement*, (Oxford: Clarendon Pr., 1986), pp. 60-2. I will have no space in this thesis to discuss the legitimacy of this formulation of the notion of strategic decision. For a brief introduction to some of the problems, see J. Elster, *Ulysses and the Sirens: Studies in Rationality*, (New York: Cambridge Univ. Pr., 1979), pp. 117-23.

each has conclusive evidence to be acquainted with our decision situation. Continuing this line of reasoning, we see that we also need to assume that *we have conclusive evidence to believe that we have conclusive evidence to believe that* we are acquainted with our decision situation. And so on. What needs to be assumed, then, is that:

(P2a) (a) Each of us faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection, and (d1) *each of us has conclusive evidence to believe so, the other has conclusive evidence to believe that we have conclusive evidence to believe so, and so on.*

I simply add clause (d1) to assumption (P1) in the previous chapter to get this assumption, (P2a). Clause (a) of (P2a) merely reintroduces the assumptions underlying the basic decision situation we face – that of the state of nature. Clause (d1) introduces the fact, just discussed, that there is common knowledge of the circumstances we face. I say that there is *common knowledge* that p amongst a population S if and only if (a) everyone in S has conclusive evidence to believe that p, (b) everyone in S has conclusive evidence to believe that everyone in S has conclusive evidence to believe that p, (c) and so on...[2]

To be able to figure out what the other will do, it needs to be assumed, second, that each of us is rational. The first assumption, (P2a), entails that each of us has enough information to be able to figure out what it is rational for the other to do. But such figurings will be useless in predicting the other's actions unless each has some assurance the other is the sort of person who will come to the right conclusion about what to do, and will then do it. Such figurings will be useless unless it is assumed that

(P2b) (b) Each is doxastically and practically rational, and (d2) there is common knowledge between you and me this is so.

[2] Note I have assumed each has *conclusive evidence* to believe this information, and not that each *actually* believes it, or that each has *reason* to believe it. This definition differs from the one offered in D. Lewis, *Convention*, (Cambridge, Mass.: Harvard Univ. Pr., 1969), p. 56, in that where Lewis talks of *reason* to believe, I talk of *conclusive evidence* to believe.

Clause (b) of (P2b) introduces the assumption of *complete rationality,* This second assumption, (P2b), differs in two ways from the type of assumptions usually made.

First, when most authors assume the agents they are discussing are rational, they in fact explicitly assume that they are maximisers of (consequentalist) value.[3] I have chosen, however, to divide this usual assumption in two: (i) that each of us is practically rational and (ii) that an agent ought to do what maximises (consequentalist) value. I say that an agent is *practically rational* at some time if and only if they do at that time what they rationally ought to do at that time. What these actions are, of course, will depend on the circumstances, and the particular theory of rational action one most favours. According to the Self-Interest Theory – (ii) – assuming that each of us is practically rational amounts to making the usual assumption that each of us will perform that action with the greatest expected-value to ourselves.

Second, when most authors assume the agents they are discussing are rational, they in fact explicitly assume, in addition, that they are perfect reasoners, and that none of the arguments the authors in question provide in their texts would escape the attention of the agents themselves.[4] Again, I have chosen to split this assumption in two: (i') that each of us is doxastically rational, and that (ii') rational belief is determined by evidence:

(B) An agent rationally ought to believe that p if and only if they have conclusive evidence to believe that p.[5]

I say that an agent is *doxastically rational* at some time if and only if they believe at that time what they rationally ought to believe at that time. What these beliefs are, of course, will depend on the

[3] See, for example, R. D. Luce & H. Raiffa, *Games and Decisions,* (New York: Wiley, 1957), p. 50, and J. H. Sobel, 'The Need for Coercion,' in J. Pennock and J. Chapman, *Coercion,* (Chicago: Aldine-Atherton, 1972), p. 152.

[4] See, for example, D. Gauthier, *Morals by Agreement,* (Oxford: Clarendon Pr., 1986), p. 61, and J. H. Sobel, 'The Need for Coercion,' in J. Pennock and J. Chapman, *Coercion,* (Chicago: Aldine-Atherton, 1972), p. 152.

[5] This is what might be considered the standard theory of rational belief. As a theory, though, it needs further work: (i) what does 'conclusive evidence' mean? (ii) There are problems with sets of infinite beliefs. Given any particular prior belief set, there are infinite propositions one will have conclusive evidence to believe, and many of these it would serve no purpose at all to believe, and so would be propositions one was rationally permitted *not* to believe. Space prevents me from canvassing how one might deal with these problems.

circumstances, and the particular theory of rational belief one most favours. According to this theory of rational belief – (B) – assuming that each of us is doxastically rational amounts to assuming that each of us will believe what we have conclusive evidence to believe. The Self-Interest Theory is a theory of rational action; theory (B) one of rational belief.

What do I mean by saying that an agent *rationally ought* to believe that p? As with action, it is an 'ought' satisfying three requirements: (a) it is deliberative – it would be irrational to judge I ought, in this sense, believe that p and yet keep deliberating about whether or not to believe that p; (b) it is (strongly) belief-guiding – it would be irrational to judge I ought, in this sense, believe that p and yet not believe it. And, finally (c) it is absolute, rather than relative – to make a judgement I ought, in this sense, to believe that p is to make a judgement not relativised, for example, to evidence (though perhaps made on the basis of evidence).

We will henceforth be dealing with two theories of rationality – one of rational action, (S), the other of rational belief, (B). I shall argue, however, that whenever we run into problems as a result of assuming both the Self-Interest Theory of rational action, and Theory (B) of rational belief, it is the first rather than the second which should be rejected. Assumptions (P2a) and (P2b), then, define the conditions of strategic choice in the state of nature.

Gauthier need not have assumed that disposition choice was parametric. As we shall see presently, the conclusion he desires will still follow from the above assumptions that such choice is strategic.

[2] Gauthier's notion of constrained maximization, we also have seen, is not unproblematic. In particular, he has difficulties identifying the grounds on which a constrained maximizer would actually cooperate. This problem can also be remedied.

What we need to do is to build the circularity we need into a promise, rather than into a disposition. First, we need to introduce the possibility that each of us can make such a promise to the other:

(P2c) (c) Each one of us, before we cooperate or defect, has the option of promising to cooperate or not; each choice is causally independent of that of the other; and each of us, after the other has agreed to or not, will have conclusive evidence to believe they have in fact done so or not. (d3) There is common knowledge between you and me that this is so.

Clause (c) of (P2c) introduces the possibility – consistent with all we have assumed to this point – that you and me can communicate, and, in particular, that *each of us can make promises to the other* before we do anything about relinquishing our right to use force.

Second, we need to determine the form a promise needs to take if it is to be effective. There are a number of possibilities. (1) I, for example, might simply say: 'I promise to lay down my right to use force'. But this is no good, since if I am disposed to keeping my promises, then this will only make me prey for you – I will lay down my right unconditionally, and you will then be able to exploit me. I need to make the promise conditional. (2) I might make such a promise, and say: 'I promise to lay down my right to use force, if you would lay down your right.'[6] But this is also no good, since even if both of us make this promise and are disposed to keep it, then we still will not have ensured that we will lay down our rights to use force – we might each be waiting for the other to go first. I need to make the conditional promise circular. (3) The proper form of the promise is in fact (P) 'I promise that: I will lay down my right to use force against the other if and only if the other says (P).'[7] If each of us were promise keepers, and each made this promise, then each of us would cooperate.

We have thus built the circularity we require into a promise, rather than into a disposition. And note that while statement (P) is self-referring, the circularity is not in this case vicious, since such self-reference is not unusual in language. The trick to surmount the second objection to Gauthier, therefore, is to replace a problematically circular notion of constrained maximization with an unproblematically circular promise of cooperation.

[6] This seems to be the form the promise takes in Hobbes. See his *English Works*, vol. ii, pp. 91-2, and D. Gauthier, *The Logic of the Leviathan*, (Oxford: Clarendon Pr., 1969), pp. 103-4. As we have also noted, it is the type of promise a constrained maximiser would make.

[7] This sort of promise is discussed by J. H. Sobel, 'The Need for Coercion,' in J. Pennock & J. Chapman (eds.), *Coercion*, (Chicago: Aldine-Atherton, 1972), pp. 171-176. In his discussion, Sobel does not consider (as I shall, later in the chapter) any type of person other than the ones I will call Agreement-Keepers, and does not argue (as I shall) that if it is rational to become an Agreement-Keeper, it is rational to actually keep one's agreements. My argument is an extension of Sobel's, itself partly inspired by Gauthier's discussion in 'Morality and Advantage,' *Phil Rev* 76 (1967): 460-75.

The first two objections I considered to Gauthier's position can, then, be adequately met. I will deal with the third – the paradox of deterrence – in the final chapter of the thesis. For the rest of this chapter, though, I want to continue with the reformulation of Gauthier's argument that, under certain conditions, it might be rational to cooperate in the Prisoner's Dilemma.

[3] Suppose, then that we find ourselves in the state-of-nature, are well-informed about our situation, are completely rational and can make promises. A situation such as this involves rational irrationality, for it is possible to argue that each of us rationally ought promise to cooperate, even though each of us rationally ought not actually to cooperate.

First, one can argue that we rationally ought to agree to cooperate.[8] (a) If I, for example, promise to cooperate then there are two possibilities: you will believe that I will carry out my promise, or you will not. If you believe that I will carry out my promise then there is a (small) chance that you (foolishly) will cooperate by laying down your arms. If you do cooperate, I will be able to exploit you, since even though I *promised* to cooperate I will not actually be so foolish as to do so. If you do not believe that I will carry out my promise, then you will presumably not cooperate, so we merely remain in the state of nature. Thus, on the one hand, if I promise to cooperate, then there is a (small) chance I will be able to exploit you, and a (large) chance it won't make any difference. (b) If, however, I do *not* promise to cooperate, there is presumably only one possibility: you will not cooperate. Thus, on the other hand, if I do *not* promise to cooperate, then it also won't make any difference. Considering (a) and (b) together, I might conclude I rationally ought to promise to cooperate.

Second, one can also argue that, even if we rationally ought to agree, it will still not be rational for us to keep the agreement.[9] (a) If you were to cooperate, then I would do better by not cooperating, since

[8] I should point out, though, that I do not necessarily endorse the argument – I introduce it simply for the sake of illustration.

[9] Thus, while I assume we can make agreements, I do not assume we can have them enforced. On this point, see W. G. Runciman & A. K. Sen, 'Games, Justice and the General Will,' *Mind* 74 (1965), p. 555, R. L. Cunningham, 'Ethics and Game Theory: the Prisoner's Dilemma,' *Papers on non-market decision making* 2 (1967), p. 12 and R. Campbell, 'Background for the Uninitiated,' in R. Campbell & L. Sowden, *Paradoxes of Rationality and Cooperation*, (Vancouver: Univ. British Columbia Pr., 1985), pp. 9-10.

I would then be able to exploit you. (b) If you were not to cooperate, then I would again still do better by not cooperating, by preventing the possibility that you will exploit me. Considering (a) and (b) together, I (once again) conclude I rationally ought not to cooperate.

The belief in this claim – that even if it it is rational for us to agree to cooperate, it is irrational for us actually to cooperate – depends on the Self-Interest Theory (S), Theory (B) of rational belief, as well as an extended version of the conditions, (P1) in the previous chapter, which defined the Prisoner's Dilemma:

(P2) (a) Each of us faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection, (b) *each of us is completely rational,* (c) *each of us, before we cooperate or defect, can make an agreement with the other,* and (d) *there is common knowledge between us that all this is so,*

I simply add, then, clauses (b), (c), and (d) to assumption (P1), in the previous chapter, to get this assumption, (P2). That is, I simply add, to the Prisoner's Dilemma, the stipulation that each of us is well-informed, completely rational, and can make each other promises. I shall call (P2) the *Promise Puzzle.*

In these situations, rational irrationality again rears its ugly head. It is rational irrationality, though, which any sensible person would not find conceptually troubling, since it is a commonplace it might be rational to make false promises, particularly if one has no assurance that others will do their part.

§2 Rational Agreement-Keeping Dispositions!

It may seem that promises, even if correctly formulated, are useless to induce cooperation in the state-of-nature. But this need not be so. In particular, if [1] each of us can choose how we are disposed to react to any agreement made between us, and if, after we have made our choice, it is possible for the other to tell how we have chosen to be disposed to react, then [2] it is rational to be the sort of person who keeps agreements.

[1] In this section, I want to introduce and motivate an extra assumption: that each of us can choose how we are disposed to react to any agreement made between us, and if, after we have made our choice, it is possible for the other to tell how we have chosen to be disposed to react.

As we have seen, the state of nature is, for Hobbes, a state of war of all against all, in which there can be no security for anyone (be they however strong or wise). As a consequence,

> it is a precept, or generall rule of Reason, *That every man, ought to endeavour Peace, as farre as he has hope of obtaining it; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre.* The first branch of which Rule, containeth the first, and Fundamentall Law of Nature; which is, *to seek Peace, and follow it.* The Second; the summe of the Right of Nature; which is, *By all means we can, to defend our selves.*[10]

Hobbes claims two things in the passage. On the one hand, he claims that a rational agent would cooperate, and renounce their right to use force, if they have the assurance that others would keep their side of the bargain, and renounce their right to use force. This Hobbes calls the Fundamental Law of Nature. On the other hand, he claims a rational agent would not cooperate, and renounce their right to use force, if they do not have the assurance others would keep their side of the bargain, and renounce their right to use force. This is what Hobbes calls the Right of Nature.

[1.1] If Hobbes is right, then any adequate theory of rationality will have to meet two requirements. On the one hand, an adequate theory will imply that rational agents would cooperate when they have the assurance that others would keep their side of the bargain. On the other hand, an adequate theory will imply that rational agents would not cooperate when they do not. Does the Self-Interest Theory meet these requirements?

First, the Self-Interest Theory implies rational agents would not cooperate when they have no assurance that others would keep their side of the bargain. Whatever the other does, it has the best outcome for one not to cooperate, and, according to the Self-Interest Theory, not

[10] T. Hobbes, *Leviathan*, ed. C. B. Macpherson, (London: Penguin, 1985), p. 190.

cooperating is what one rationally ought to do. This is so particularly if there is no assurance the other would cooperate. If one is a rational agent without such an assurance, then, the Self-Interest Theory correctly implies one would not cooperate.

Second, the Self-Interest Theory does not, however, imply rational agents would cooperate when they *do* have the assurance others would cooperate. Suppose, for example, we are two agents who can assure the other that we would cooperate. Suppose, in particular, that each of us, by taking a certain pill, can become *transparently trust-worthy*. We would cooperate with anyone similarly disposed, even if it did not have the best outcome for us to do so, and this would be obvious to anyone who met us. If you, for example, are like this, then you can give me an assurance that you would cooperate. To do so you would simply have to take the pill: you would become trustworthy; I would come to believe that you were (since you would be transparent); and I would have the assurance I need.

It is not clear, however, the Self-Interest Theory implies that we rational agents would cooperate when we have this type of assurance. To see why, consider two possibilities: that our mutual rationality will last until the time for cooperating, or that it will not.

(i) Suppose, on the one hand, that our mutual rationality lasts until the time comes for cooperating. Come the time for cooperating, it still remains that each of us would only obtain a loss from doing so, and thus, if the Self-Interest Theory is to be believed, each rationally ought not to cooperate. If our mutual rationality will endure to the time for cooperating, then, according to the Self-Interest Theory, each of us will *not* cooperate. In this first case, the Self-Interest Theory implies that initially rational agents would *not* cooperate, even if there is the indicated possibility of assuring the other that they will not be exploited.

(ii) Suppose, on the other hand, that our mutual rationality will *not* necessarily endure until the time for cooperation. In this case, the Self-Interest Theory would not have the implication described in case (i). But it would also fail to imply that two initially rational agents *would* cooperate. If we drop the assumption that we remain mutually rational up to and including the time for cooperation, there are no grounds for making any predictions about what we will do come this time, and thus no grounds for supposing that we, as two rational agents, would in fact cooperate. In this second case, the Self-Interest

Theory *fails* to imply that initially rational agents *would* cooperate, if there is the indicated possibility of assuring the other they will not be exploited.

On these grounds, the Self-Interest Theory should be rejected. On the one hand, it implies – reasonably enough – that rational agents would not cooperate when they have no assurance that others would keep their side of the bargain; but, on the other, it does *not* imply – surely unreasonably – that rational agents would (and may in fact imply that they would not) cooperate when – as above – they do have the requisite assurance. If Hobbes is right, then, an adequate theory implies that rational agents *would* cooperate when they have the assurance others would keep their side of the bargain, and so the Self-Interest Theory is inadequate. But even if Hobbes is not in general right, it should be clear in the above example that rational agents would indeed become transparently trustworthy, and thus avoid for themselves the ravages of the state-of nature. Whether or not Hobbes is in general right, the Self-Interest Theorist *will* want to argue two initially rational agents will come to cooperate in the situation I described above.

[1.2] The best response for the Self-Interest theorist at this point is to reject the assumption that two initially rational agents – you and I – will necessarily remain rational come the time for cooperation, and to assume instead we are rational up to, but not necessarily including, this time.[11] The Self-Interest Theorist need not be enamoured with rationality *per se,* and might suppose agents will hold onto rationality only up to the point it begins to get in the way of producing the best outcomes. The Self-Interest Theory advises one to perform those actions which produce the best outcomes for oneself, and if such actions involve making yourself disposed to perform later irrational actions (as it seems they will in the case I have introduced), then the initially rational agent is happy to treat their own rationality as a mere means.

To drop the assumption that our initial rationality will endure puts the Self-Interest Theorist on the other horn of the dilemma, since it then seems that their theory lacks any resources for showing that we

[11] On this response, see D. Parfit, *Reasons and Persons,* (Oxford: Clarendon Pr., 1984), pp. 45-49, where he discusses whether being rational could ever be a mere means.

will come to cooperate. This is not a problem, however. For, rather than assuming we will make our decisions about cooperating *when the time comes for action,* and assuming with this that we will be rational when this time comes, it would be better to assume instead that we make our decisions concerning cooperation *beforehand,* while we are still assumed to be rational. The Self-Interest Theorist may assume, as I shall follow them in assuming, that while we are rational beforehand we can commit ourselves to various *plans* of action, plans making later actions of cooperation contingent upon our believing that certain events have occurred.

What plans of action are available to us? There are two actions we will face later on: cooperating or defecting. There are two contingencies upon which each of us can base our action: our having agreed to cooperate, or our not having so agreed. There are thus four (=2x2) complete ways of reacting you, for example, might adopt:

(C | C) you make it now that you cooperate regardless of what (you believe that) we agreed to beforehand: if (you believe that) we agreed to cooperate then you cooperate, and if (you believe that) we did not agree then you cooperate,[12]

(C | D) you make it now that you cooperate if and only if we agreed to cooperate: if we agreed to cooperate then you cooperate, and if we didn't then you defect,

(D | C) you make it now that you cooperate if and only if we did not agree to cooperate: if we agreed to cooperate then you defect, and if we didn't then you cooperate,

(D | D) you make it now that you defect regardless of what we agreed to do beforehand: if we agreed to cooperate then you defect, and if we did not then you cooperate.

[12] These different plans are also called *strategies,* and are denoted by '(x | y)', which indicates that if (one believes) we have agreed to cooperate, then one will do x, and if (one believes) we have not agreed to cooperate, then one will do y. The notation is that of N. Howard, *The Paradoxes of Rationality: the Theory of Metagames and Political Behaviour,* (Cambridge, Mass.: MIT Pr., 1971). Note that since these strategies are concerned with internalised conditional commitments to act, they should, strictly speaking, be understood as conditionals relating two internal states of an agent. Thus, strictly speaking, if you adopt (C | D), then you are disposed to cooperate if *you believe* we have agreed, and not simply if *in fact* we have agreed. However, it gets tedious to include such a qualification, and I shall do so only when necessary.

Three of these strategies deserve particular mention. The first, (C | C), is the disposition someone we could call a 'Kantian' would have.[13] This person performs that action they would will to become a universal law. This, of course, is cooperation (whether or not it was agreed to), since universal cooperation is valued higher than universal defection. The second, (C | D), is the disposition of the person I will call the Agreement-Keeper. Such a person cooperates with those with whom they have agreed, but defects otherwise. Unlike the 'Kantian', they respond to a failure to agree with defection. The third, (D | D), is the disposition of the enduringly Self-Interested agent, who at all times performs that action which produces the best outcomes for themselves. Unlike the previous two agents, they respond to an agreement with defection.

Hobbes is right to suggest that without an assurance that the other is trustworthy, a rational agent would not cooperate, and may be right also to suggest that with such an assurance, a rational agent would cooperate. What he would be wrong to suggest, though, is that a person could have such an assurance from another only if there were some coercive power standing over both, enforcing the agreements made between them. If you, for example, can choose how to be disposed, and if you are transparent, you would be able to give me the requisite assurance without having to defer to some coercive power. To do so, you would simply have to become either a (C | C) or a (C | D): you would become the sort of person who would cooperate if we had agreed; I would come to believe you were (since you would be transparent); and I would thus have the assurance I need.

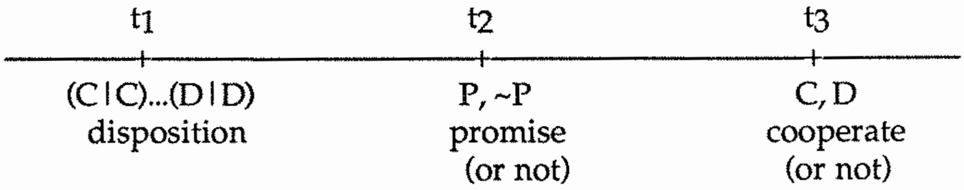
Any adequate theory of rationality will imply that two agents will cooperate if they find themselves in a situation where each can offer the other some assurance that they will not be exploited. A situation in which our dispositions are transparent is just such a situation, and the Self-Interest Theory *can* imply that we two initally rational agents would cooperate only if, it seems, we assume that:

[13] I do not know if Kant is really committed to endorsing (C | C). A. K. Sen, in 'Choice, Orderings and Morality,' in S. Korner, *Practical Reason*, (New Haven: Yale Univ. Pr., 1974), p. 57, seems to think he is.

(P3) (x) *Each of us, before we promise to cooperate or not, has the option of adopting one of the strategies (C | C), (C | D), (D | C) or (D | D); the choice each makes is causally independent of that of the other; and each of us, after the other has adopted one of their possible strategies, will have conclusive evidence to believe they have in fact done so;* (a) each faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection; (b) each of us is completely rational; (c) each of us, before we cooperate or defect, can make an agreement with the other; and (d') there is common knowledge between us that all this is so.[14]

I simply add clause (x) to assumption (P2), introduced above, to get this assumption, (P3). That is, I simply add, to the Promise Puzzle, the stipulation that each can choose how they would react to an agreement, though, after the fact, such a choice will be transparent to the other. I will call (P3) the *Third Counterexample* (the first two, of course, occuring in the discussion of the Toxin Puzzle in Chapter Two).

[2] I now want to argue that if each of us can choose how we are disposed to react to any agreement made between us, and if, after we have made our choice, it is possible for the other to tell how we have chosen to be disposed to react – that is, if assumption (P3) obtains – then it is rational to become an Agreement-Keeper. In showing this, three times are important. They are indicated on the following line:



Initally, at time t_1, each is free to adopt one of the dispositions (C | C) to (D | D). At a time later than this, t_2, each is free to make a promise or

[14] This is, of course, a very restictive assumption. The contigencies upon which one might base the action of cooperating are presumably numerous, and the plans or dispositions one has adopted will almost never be transparent. In response, I reiterate my earlier (unsubstantiated) claim that even when one does consider more realistic types of situation, then it will still remain the case that some non-expected-value maximizing cooperative disposition is rational. Substantiating this claim is work for another time.

not. And at a time still later, t_3, each is free to keep the promise, and cooperate, or not, and defect. The situation we both initially face at time t_1 (while we are still well-informed and completely rational) can be depicted as follows:

			You							
			C|C		C|D		D|C		D|D	
			P	~P	P	~P	P	~P	P	~P
Me	C|C	P	c,c	c,c	c,c	s,t	s,t	c,c	s,t	s,t
		~P	c,c	c,c	s,t	s,t	c,c	c,c	s,t	s,t
	C|D	P	c,c	t,s	**c,c**	**d,d**	s,t	t,s	s,t	d,d
		~P	t,s	t,s	**d,d**	**d,d**	t,s	t,s	d,d	d,d
	D|C	P	t,s	c,c	t,s	s,t	d,d	c,c	d,d	s,t
		~P	c,c	c,c	s,t	s,t	c,c	c,c	s,t	s,t
	D|D	P	t,s	t,s	t,s	d,d	d,d	t,s	d,d	d,d
		~P	t,s	t,s	d,d	d,d	t,s	t,s	d,d	d,d

Key: x|y = making it at t_1 that: one does x at t_3 if we agree at t_2 and y if not; P = promising at t_2 to cooperate at t_3; t = the value of exploiting the other at t_3; c = the value of mutual cooperation at t_3; d = the value of mutual defection at t_3; s = the value of being exploited by the other at t_3. (NB: $t > c > d > s$.)

Each of us is free to adopt one of the strategies (C|C) to (D|D), and the outcomes of all combinations of these are as shown. The *immediate* outcome of a pair of strategies is, itself, a situation in which both of us are free later with respect to another action, this time about whether or not we are going to make the promise, (P), to cooperate.

Take, for example, the outcome if I adopt (C|D), and you also adopt (C|D). Whether or not cooperation will immediately result is not yet given. What is given is that we will face a later situation of the form:

			You	
			C|D	
			P	~P
Me	C|D	P	c,c	d,d
		~P	d,d	d,d

It is straightforward to see this is so. Suppose that each has made themselves an Agreement-Keeper – you and I have each performed action (C|D). Then I (ditto: you) am such that if I promise to cooperate, and believe that you have promised, then I will cooperate. Later, each of us is free to promise or not: P or ~P, and the outcomes of all combinations of these are shown. On the one hand, if each of us promises (I do P, and you do P), then I will believe that we have done so, and I will cooperate. Since, similarly, you will also cooperate, the result is an outcome (mutual renunciation of our right to use force) which each values at 'c'. And this is what is in the top left-hand cell. On the other hand, if one of us does not promise (I do ~P or you do ~P or both), then each will know, and neither will cooperate; the result being an outcome (mutual retention of the right to use force) which each values at 'd'. And this is what is in the remaining cells. Performing this sort of reasoning with all other possible combinations of strategies gives the complex decision matrix indicated above.

The situation we both initially face (while we are still well-informed and completely rational) may be simplified, to obtain:

<table>
<tr><th></th><th></th><th colspan="4">You</th></tr>
<tr><th></th><th></th><th>C|C</th><th>C|D</th><th>D|C</th><th>D|D</th></tr>
<tr><td rowspan="4"><b>Me</b></td><td>C|C</td><td>c,c</td><td>s,t</td><td>c,c</td><td>s,t</td></tr>
<tr><td>C|D</td><td>t,s</td><td>c,c</td><td>t,s</td><td>d,d</td></tr>
<tr><td>D|C</td><td>c,c</td><td>s,t</td><td>c,c</td><td>s,t</td></tr>
<tr><td>D|D</td><td>t,s</td><td>d,d</td><td>t,s</td><td>d,d</td></tr>
</table>

For even though the *immediate* outcome of a pair of strategies is, as we have seen, a situation in which each of us is free to promise or not, we can determine now what it would be rational then to do, and so determine (if we assume that we will at the time of promising still be rational) what we will then do, and so determine the outcome of our adoption now of various strategies.

Take again, for example, the resulting situation if I adopt (C|D), and you also adopt (C|D). The Self-Interest Theory says to each of us, in this case, to promise, for each knows that the other will adopt either P or ~P, and know that if they choose P in this case then they will never do any worse, and may very well do better, than if they were to adopt any other strategy. Of either choice, in this case, P is the only one they have nothing to lose by performing. Each is initially uncertain about what the other will do. Therefore, each rationally ought to do P, and (if we assume that we will then be rational), it follows that each of us will do P, and the result will be mutual renunciation of our right to use force, and an outcome we value at c,c. Performing this process with the remaining fifteen possible combinations of strategies gives the simplified matrix indicated.

This inference employs the following dominance, or 'sure-thing', principle: if action A weakly dominates action B over partition P, and one is uncertain concerning P, then one rationally ought not to B. This principle employs two key notions: (a) A *weakly dominates* B over P when, for all p in P, the expected value of A&p is no less than that of B&p, and, for at least one p in P, it is greater; (b) one is *uncertain* concerning P when, for all p in P, one assigns no probability to p. I am initially uncertain about what you will do, since ours is a strategic situation, in which we do not take as given the likelihood of particular behaviour on the part of the other.

Note, we have to be careful how we apply dominance reasoning with regard to the complex matrix with which we started.[15] I may not, for example, eliminate from consideration the line (C|D)|P – where I become an Agreement-Keeper and then promise to cooperate – because it is dominated, as it is, by the line (D|D)|P – where I remain a Straightforward Maximiser, but then agree to cooperate. The reason is simple. I may use dominance reasoning amongst my actions A_1, A_2, ..., A_n, relative to your actions B_1, B_2, ..., B_n, only if your choice of B_i is causally independent of my choice of A_i. However, your choice, for example, of (C|D)|P is not causally independant of my choice, for example, of (D|D)|P. If I were to choose (D|D)|P you would not be so foolish as to choose P, and, *a fortiori,* would not be choosing (C|D)|P. Hence, I may not use dominance reasoning amongst my actions (C|C)|P, ..., (D|D)|~P relative to your actions (C|C)|P, ..., (D|D)|~P. I may, however, use dominance reasoning (as I will) amongst my (C|C), .., (D|D) relative to your (C|C), ..., (D|D), since these are causally independent. (This is assumption P3(x).) And I may use dominance

[15] See M. Bar-Hillel & A. Margalit, 'Newcomb's Paradox Revisited,' *Brit J Phil Sci* 23 (1972): 295-304 for a discussion of the limitations on the application of the dominance principle.

reasoning (as I have) amongst my P, ~P relative to your P, ~P since these are also causally independent. (This is assumption P3(a).)

On the basis of this simplified matrix, we can see that each of us rationally ought to become Agreement-Keepers. Each knows that the other will adopt one of the strategies (C | C) to (D | D), and know that if they choose (C | D) then they will never do any worse, and may very well do better, than if they were to adopt any other of their own strategies. Strategy (C | D) is the only one they have nothing to lose by adopting, and so each ought to adopt (C | D). Each is initially uncertain about what the other will do. Each of us ought to make ourselves Agreement-Keepers, and, if we do, then we each ought to make the promise, thus agreeing to cooperate, and thus eventually cooperating by casting off our right to use force against the other, and escaping from the war of the one against the other.

If we are well-informed and completely rational agents in the state-of-nature who can make agreements with each other, and who can choose how to be disposed with regard to keeping such agreements, though such a choice is, after the fact, transparent to the other person, then we rationally ought to become agreement-keepers, would become agreement-keepers and would thus cooperate. But would it be *rational* for us to cooperate? Or would it just be another instance of rational irrationality?

§3 Rational Agreement-Keeping Actions!

Even if Gauthier's argument, introduced in the previous chapter, for the rationality of constrained maximization had established it might be rational to be *disposed* to cooperating, some objected this would not imply it was rational *actually* to cooperate. In the remainder of the chapter I will argue, firstly, that – contrary to the Self-Interest Theory – we are indeed rationally permitted to cooperate in the Third Counter-example, and, secondly and more generally, that given certain conditions (to be specified below), the actions resulting from rational agreements are also rational, even if one is free to do otherwise and it has the best outcome for one to do otherwise (though these actions may very well be irrational absent those conditions).

[1] The first half of the task for this section consists in demonstrating that in the Third Counterexample, we are rationally permitted to cooperate. I shall need to reply to the Foole, who

> hath sayd in his heart, there is no such thing as Justice; and sometimes also with his tongue; seriously alleaging, that every mans 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 make, and not make; keep, or not keep Covanents, was not against Reason, when it conduced to ones benefit. He does not therein deny, that there be such 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;[16]

Since, whatever the other does with their right to use force, each of us does worse for ourselves to cast off our own, the Foole will say, whether in his heart or in print,[17] that it is not rational for us to do so, that we rationally ought not to do so.

My answer to the Foole takes the form of the following argument. [1.1] In the Third Counterexample, if the Self-Interest Theory (S) and Theory (B) of rational belief are both true, then so are all the following:

(1) Each rationally ought (at t_1) to become an Agreement-Keeper;
(2) Each rationally ought (at t_2) to believe we have agreed to cooperate; and
(3) Each rationally ought (at t_3) to not cooperate.

Yet, since one rationally ought to S_1, rationally ought to S_2, and rationally ought to S_3 only if it is logically possible that one S_1, S_2, and

[16] T. Hobbes, *Leviathan*, ed. C. B. Macpherson, (London: Penguin, 1985), p. 203.

[17] Doubts concerning the inference from the rationality of cooperative dispositions to the rationality of cooperative acts are expressed by M. Perkins & D. C. Hubin, 'Self-Subverting Principles of Choice,' *Can J Phil* 16 (1986), §III, G. Kavka, 'Review of 'Morals by Agreement',' *Mind* 96 (1987), pp. 120-1, R. Arneson, 'Locke versus Hobbes in Gauthier's Ethics,' *Inquiry* 30 (1987), §V(b), G. Harman, 'Rationality in Agreement: a Commentary on Gauthier's 'Moral's by Agreement',' *Soc Phil Pol* 5 (1988), pp. 5-6, R. Campbell, 'Critical Study: Gauthier's Theory of Morals by Agreement,' *Phil Quart* 38 (1988), §3.2, D. Copp, 'Contractarisnism and Moral Skepticism', 204-7 and H. Smith, 'Deriving Morality from Rationality', 244-9, both in P. Vallentyne, *Contractarisnism and Rational Choice*, (New York: Cambridge Univ. Pr., 1991), pp. 204-7 and 244-9 respectively.

S_3 – that is, since (OP_3) is true – and since it is not logically possible to become an Agreement-Keeper, believe we have agreed to cooperate, but not cooperate, then one of these claims must be false. [1.2] We have no good reason against thinking – (2) – that each rationally ought (at t_2) to believe we have agreed. [1.3] And we have no good reason against thinking – (1) – each rationally ought (at t_1) to become an Agreement-Keeper. Therefore, the final claim – (3) – must be false, and it must indeed be rationally permitted for each to cooperate. The Foole is wrong.

[1.1] The first premise of the argument is that, in the Third Counterexample, if the Self-Interest Theory (S) and Theory (B) of rational belief are both true, then so are (1), (2) and (3).

(1) I spent all of section §2[2] arguing that each rationally ought (at t_1) to become an Agreement-Keeper. I shall not repeat the argument here.

(2) Each, then, rationally ought to become Agreement-Keepers. Since each is completely rational at the time of choosing dispositions, it follows each becomes an Agreement-Keeper. Each is transparent, and so each will have conclusive evidence to believe that the other is an Agreement-Keeper, and so (according to (B)) each rationally ought to believe so. Since each is completely rational at the time of promising, it follows each believes the other is an Agreement Keeper. In this case it has the best outcome to promise and so (according to (S)) each rationally ought to promise. Since each is completely rational at the time of promising, it follows each promises to cooperate. Each can tell whether the other has made a promise or not, and so each will have conclusive evidence to believe we have made an agreement, and so (according to (B)) each rationally ought to believe we have agreed.

(3) However, it still has the best outcome not to cooperate. If theories (S) and (B) are true, then each rationally ought not to cooperate.

Yet, since one rationally ought to S_1, rationally ought to S_2, and rationally ought to S_3 only if it is logically possible that one S_1, S_2, and S_3 – that is, since (OP_3) is true – and since it is not logically possible to become an Agreement-Keeper, believe we have agreed, and yet not cooperate, then one of the claims (1)-(3) must be false. But which one?

[1.2] The second premise of my argument is that we have no good reason against thinking – (2) – that each rationally ought (at t_2) to believe we have agreed. We have no reason for rejecting the second statement.

The Foole, in order to save the Self-Interest Theory, (S), of rational action, may propose to reject theory (B), of rational belief. A response such as this, it might be suggested, is indeed the one the defender of the Self-Interest Theory would be inclined to make, since this commonly held view about the nature of rational belief is no part of the Self-Interest theory and the Foole can offer his own, alternative, account of rational belief. Roughly, this is that a person rationally ought to believe that p if and only if the expected-value to them of the outcome of believing that p exceeds that of not believing.[18] Such a theory of rational belief is particularly *a propos* to the situations we are considering, for again, no matter what you do, it has the best outcome for me not to believe we have agreed, since then I will not cooperate.

There are two reasons for dismissing this response. First, this Self-Interest theory of rational belief is not without independent problems. It is difficult to know whether or not real persons could actually follow the dictates of such a theory, and it may well be constituitive of belief that in a large proportion of cases conclusive evidence for believing results in the relevant belief. Furthermore, it may well be impossible to believe what maximises expected-value without having a significant proportion of one's beliefs sensitive to evidence rather than expected-value. Second, and more importantly, rejecting theory (B) is more than we need to do in order to free ourselves from our dilemma. I have already argued in the first part of the thesis that the Self-Interest Theory is false, and thus should be rejected. This argument in no way implicated Theory (B) of rational belief, and so gave us no reason for rejecting this theory. So it is in this case. The conservative strategy – to reject in this case what needs to be rejected anyway – is to reject the Self-Interest Theory, (S). I shall consider no further the possibility that Theory (B) is false, and the possibility that (2), above, is false.

This leaves only statements (1) and (3), one of which must be rejected. But which one?

[18] For some discussion of this type of theory, see S. Nathenson, 'Nonevidential Reasons for Belief: a Jamsian View,' *Phil Phenom Res* 42 (1981-2): 44-54, M. Fisher, 'Truth as a Problem for Utilitarianism,' *Mind* 89 (1980): 249-255.

[1.3] The third premise of my argument is that we have no good reason against thinking – (1) – that each rationally ought (at t_1) to become an Agreement-Keeper. We have no reason to reject the first statement, and reason to reject the third.

I will show this by providing an argument directly for the claim each is rationally permitted to cooperate. In the Third Counterexample, there are eight ways you, for example, could be:

(a)	A	B	C
(b)	A	B	~C
(c)	A	~B	C
(d)	A	~B	~C
(e)	~A	B	C
(f)	~A	B	~C
(g)	~A	~B	C
(h)	~A	~B	~C

where: A=becoming an Agreement-Keeper; B=Believing we have agreed to cooperate; and C=Cooperating. Maximally rational persons believe all they rationally ought to believe: if you are maximally rational, and rationally ought to believe that p, then you would believe that p. We have already established that you rationally ought to believe we have agreed – hence if you were maximally rational, then the only ways you might be are: (a), (b), (e), and (f). However, whether or not you are maximally rational, it is not logically possible for you to be an Agreement-Keeper, believe we have agreed, and yet not cooperate – therefore (b) is not really a way you might be. Hence, if you were maximally rational, there are only three ways you might be:

(a)	A	B	C
(e)	~A	B	C
(f)	~A	B	~C

Furthermore, maximally rational agents are those who can expect to most promote what they value. Of these three ways you might be: best is (a) – since you are an Agreement-Keeper, I will cooperate, and civil society will be possible; next is (f) – since you are not an Agreement-Keeper, I will not cooperate, and it would be foolish for you to do so; worst is (e). Hence, if you were maximally rational, then you would become an Agreement-Keeper, believe we have agreed, and therefore cooperate. But anything a rational person would do is something ordinary mortals, like you, are rationally permitted to do: if you would A were you maximally rational, then you are rationally permitted to A. You would cooperate were you maximally rational, and so you are rationally permitted to do so.

In this way, then, we see that in the Third Counterexample you are rationally required to become an Agreement-Keeper, rationally required to believe we have agreed, and thus rationally permitted to cooperate, even though you are free not to cooperate (P3(a)) and it has the best outcome for you not to cooperate (P3(a)). This claim may be generalised.

[2] The second half of the task for this section is to demonstrate that, GIVEN that you rationally ought to adopt the enduring disposition to do what we have agreed to do, THEN if you rationally ought to believe we have agreed to do something, you are rationally permitted to perform your part of the bargain, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise (though doing so may very well be irrational absent this condition). The demonstration relies again on the deontic principles I introduced above.

The argument for this claim is as follows. The deontic principle (OP_3) says that one rationally to S_1, rationally ought to S_2, and rationally ought to S_3 only if it is logically possible that one S_1, S_2, and S_3. Letting S_1 be 'adopt the enduring disposition that if you believe we have agreed to A then you A', S_2 be 'believe we have agreed to A' and S_3 be 'not A', it follows from (OP_3) that

(1) You rationally ought to adopt the enduring disposition that if you believe we have agreed to A then you A, you rationally ought to believe we have agreed to A, and you rationally ought not to A only if it is logically possible for you to adopt the disposition, have the belief, and yet not A.

Yet, as we have seen, this is not possible. On the basis of some propositional logic (in particular, the principle that $D \,\&\, B \,\&\, A \rightarrow P$ and $-P$ entail $D \rightarrow (B \rightarrow -A)$), it follows from (1) that that GIVEN that you rationally ought to adopt the enduring disposition to do what you believe we have agreed to do, THEN if you rationally ought to believe we have agreed to perform some action, then you are rationally permitted to perform that action. The above argument depends on no particular assumptions about the action in question – A. Hence, our conclusion obtains EVEN IF one is free not to A, and it has the best outcome not to A (though, I should point out, it may not be rationally permitted to A if the above conditions are not given).

Conclusion

The Foole, then, is wrong, and under some (but not all) conditions, rational agreements make for rational actions. The Self-Interest Theory is false. In the state of nature, it unconditionally counsels agents to retain the right to use force against others, even though the mutual retention of this right leads to the war of all against all, and short and miserable lives for each.

I started this thesis with the worry morality might be irrational. But if we can defend the contractarian analysis of morality – that one morally ought to perform some action when it is what one would agree to do in certain circumstances – and if (as I believe) Gauthier's type of argument applies to more realistic types of situations – involving more than four dispositions from which to choose, and dispositions not transparent – then we may perhaps be able to show it is rational to be moral.

It seems, after all, that there can be rational morals by agreement, and that progress might perhaps be made on what Sidgwick called 'the profoundest problem of ethics.'

Chapter Six

The Prisoner's Dilemma – Reply to Objections

I have argued that the Foole is wrong, and that given the rationality of a certain disposition, rational agreements do indeed make for rational actions. My argument for these claims, and particularly the first, depends (amongst other things) on a central assumption:

(P3) *Third Counterexample:* (x) Each of us, before we promise to cooperate or not, has the option of adopting one of the strategies (C | C), (C | D), (D | C) or (D | D); the choice each makes is causally independent of that of the other; and each of us, after the other has adopted one of their possible strategies, will have conclusive evidence to believe they have in fact done so; (a) each faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection, (b) each of us is completely rational, (c) each of us, before we cooperate or defect, can make an agreement with the other, and (d) there is common knowledge between us that all this is so.

The Foole, however, might claim that when I add clause (x) to the Promise Puzzle, (P2), to get (P3), I change the situation in a crucial way. As we will see, this is to claim, in effect, that assumption (P3) is inconsistent.[1] There are two broad types of objections to these assumptions, and in this chapter I shall argue that neither of them succeed.

§1 Is it really still possible for you not to cooperate?

I constructed the Third Counterexample by adding to the Promise Puzzle, (P2), the stipulation that each of us, before we promise, is free to choose how we are disposed, and that this choice is, afterwards,

[1] Some who would make this charge – though not defenders of the Self-Interest Theory – include R. Hardin & B. Barry (ed.s), *Rational Man and Irrational Society?* (Beverly Hills: Sage, 1982), p. 382, and D. Braybrooke, 'The Insoluable Problem of the Social Contract,' *Dialogue* 1 (1976): 3-37.

transparent to the other person. In effect, I added the stipulation that it maximises expected-value for each to adopt the enduring disposition of always doing what we agree to do, to get (P3). I claimed that, in these modified situations, you are rationally permitted to keep your agreement, and cooperate, even though you are free not to do so, and it maximises expected-value for you not to do so. The Foole, however, may object that this stipulative addition means each is no longer free not to cooperate. As I shall argue in this section, though, this objection is mistaken, and in the Third Counterexample each may indeed still be free not to cooperate.

[1] There are some[2] who suggest that constrained maximisers – who will cooperate if they believe others are likely to cooperate – are the sort of persons who *cannot but* cooperate if they believe others are likely to cooperate:

> If CM is a permanent mechanism inducing co-operation in spite of defection being [value]-maximizing, and therefore, presumably being the more preferable action at the time of actual choice of action, it would seem that there is a crucial sense in which, at that time, one is *not* acting voluntarily – that is, from an immediate preference so to act – but merely at the behest of the mechanism. ... On the mechanism interpretation, Gauthier seems wrong to think of co-operation as voluntary, free, and rational action.[3]

> We need to know more about the disposition to cooperate. We do know that it is not a disposition to consciously *decide* on economic grounds to comply with agreements. If the CM were in the position to deliberate, then utility maximization would sometimes dictate noncompliance, and it would not be economically rational to adopt the disposition. Internalizing the disposition to cooperate entails binding oneself to comply strongly enough to overcome the inclination to straightforwardly maximize. In the absence of rope, mast, and sturdy sailors at one's command, the bonds must originate in one's own psyche.[4]

[2] Apart from MacIntosh and Nelson, below, others making this suggestion include R. Campbell, 'Moral Justification and Freedom,' *J Phil* 85 (1988): 192-213.

[3] D. MacIntosh, 'Two Gauthiers?' *Dialogue* 28 (1989), pp. 47, 48. See also his 'Libertarian Agency and Rational Morality: Action Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem,' *S J Phil* 26 (1988): 499-525.

[4] A. Nelson, 'Economic Rationality and Morality,' *Phil Pub Affairs* 17 (1988), p. 157.

MacIntosh claims a person disposed reliably to cooperate, even if it is not value-maximising to do, has a 'mechanical' disposition; Nelson compares the actions of such a person to those of Odysseus 'bound' to the mast. We see here the suggestion that, come the time for co-operating, a constrained maximiser – and so too, presumably, an Agreement-Keeper – *cannot but* cooperate.

[2] What might justify claims such as these? I will deal with the forms such a justification would take were it to address Agreement-Keeping, rather than constrained maximisation.

A first suggestion is that you, for example, are reliably disposed to keep agreements only if you *could not but* cooperate were you to believe we had agreed. On this view of the matter, all reliable dispositions are dispositions resulting in unfree acts. Yet this is surely not the case (and, I suspect, a view neither MacIntosh nor Nelson had in mind). My butcher is so disposed that were he to believe I asked him for a pound of ground beef, then he would give me a pound of ground beef. I have asked him for as much and, seeing me, he believes I have asked him for a pound of ground beef. He gives me a pound of ground beef, but this is not to say that he was compelled to do so. Not at all. He was free not to give me a pound of ground beef, but, thankfully, he is disposed not to exercise this freedom. It makes perfect sense to suppose that one has a certain capacity one is disposed not to exercise, and so makes sense to suppose that you have a capacity not to cooperate which you are disposed not to exercise.

A better suggestion (and the one I think MacIntosh and Nelson actually have in mind) is that you are reliably disposed to keep agreements, *even if it is not expected-value maximising to do so,* only if you could not but cooperate were you to believe we had agreed. On this view of the matter, all reliable dispositions *to act in a non-expected-value maximising manner* are dispositions resulting in unfree acts. Yet this is still not quite correct. From the fact that I am reading the day's newspapers in the library, rather than doing what I believe it is best for me to do – looking up some references – it does not follow that my action of reading the papers is an unfree one; it only follows (if at all) that it is irrational. Thus, from the fact that I do something non-expected-value maximising it does not follow that my action is unfree. Similarly from the fact I am disposed to doing something not expected-value maximising – like cooperating – it does not follow that I am disposed to performing unfree actions. It makes perfect sense to suppose one has a certain capacity which one is (if at all) *irrationally* disposed not to exercise, and so makes sense to suppose that I have a capacity not to cooperate which I am (if at all) irrationally disposed not to exercise.

The Foole might claim that the stipulative addition I made to the Promise Puzzle, (P2), to the Third Counterexample means that neither of us is free to renege on our agreement, but neither of the two possible interpretations of this claim is convincing. More convincing would it be to claim that you are reliably disposed to keep agreements only if you are reliably disposed to value cooperation over non-cooperation whenever you believe we have agreed. On this view of the matter, all reliable dispositions cannot but be dispositions resulting in (free) expected-value maximising actions. But to make this suggestion is just to object that in the situation described by (P3) – where you are actually disposed to cooperating, since you ought to be disposed to cooperating, and are rational – it must no longer have the best outcome for you not to cooperate. And this is the second objection with which I want to deal.

§2 Does it really still have the best outcome for you not to cooperate?

I constructed the Third Counterexample by adding to the Promise Puzzle, (P2), the stipulation that each of us, before we promise, is free to choose how we are disposed, and that this choice is, afterwards, transparent to the other person, to get (P3). I claimed that, in these modified situations, you are rationally permitted to keep your agreement, and cooperate, even though you are free not to do so, and it maximises expected-value for you not to do so. The Foole, however, may now object this stipulative addition means it *no longer has the best outcome for you* not to cooperate. As I shall argue in this section, however, this second objection is also mistaken, and that in the Third Counterexample it may indeed still have the best outcome not to cooperate.

[1] The problem, again, is that it seems the action of your cooperating is unintelligible, unless we suppose that, somehow, it now has the best outcome for you to cooperate. There are some philosophers[5] who suggest that a constrained maximiser – who will cooperate if they believe others are likely to cooperate – must be the sort of person now valuing the outcome of cooperation to that of non-cooperation:

> Gauthier assumes that two people totally *unconcerned* with each other's interests can have equally strong commitments to keep an agreement, even though either of them could increase his utility payoff by breaking the agreement. I find this assumption simply *incoherent* – unless these two people do develop some *concern* from each other's interest.[6]

> The problem is set in the following passage: "[a] constrained maximiser is conditionally disposed to co-operate in ways that, followed by all, would yield nearly optimal and fair outcomes, and *does* co-operate in such ways when she may actually expect to benefit" (177). The problem is that Gauthier does not say exactly *why* she *does* co-operate (at least not in that context). Certainly she would benefit *more* by defecting. ... It seems to me then that the disposition merely *determines* co-operation, for it simply cannot rationalise it.[7]

Harsanyi's claim implies a coherent notion of a constrained maximiser will be one which, for example, says that the constrained maximiser develops some concern for the other agent. MacIntosh insists that if one would benefit more from defection, then there is nothing to rationalise cooperation were it to occur.

[2] What might justify claims such as these? I will deal with the forms such a justification would take were it to address Agreement-Keeping, rather than constrained maximisation. At base, the issue is whether or

[5] Apart from Harsanyi and MacIntosh, below, others making this suggestion include J. S. Kraus & J. L. Coleman, 'Morality and the Theory of Rational Choice,' *Ethics* 97 (1987), p. 722, fn. 13, J. Mendola, 'Gauthier's 'Morals by Agreement' and Two Kinds of Rationality,' *Ethics* 97 (1987): 765-74, D. Copp, 'Contractarianism and Moral Skepticism,' in P. Vallentyne (ed.), *Contractarianism and Rational Choice: Essays on David Gauthier's 'Morals by Agreement'*, (New York, Cambridge Univ. Pr., 1991), pp. 206-7, D. Braybrooke, 'The Insoluable Problem of the Social Contract,' in R. Campbell & L. Sowden, *Paradoxes of Rationality and Cooperation*, (Vancouver: Univ. Brit. Columbia Pr., 1985), p. 289, and E. F. McClennen, 'Constrained Maximization and Resolute Choice,' *Soc Phil Pol* 5 (1988), §5.

[6] J. Harsanyi, 'Review of 'Morals by Agreement',' *Econ Phil* 3 (1987), p. 343.

[7] D. MacIntosh, 'Two Gauthiers?', *Dialogue* 28 (1989), p. 46. Again, see his 'Libertarian Agency and Rational Morality: Action Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem,' *S J Phil* 26 (1988): 499-525.

not the following claim, a similar one to which we met in Chapter Three, is true:

(G3) In the Promise Puzzle, you cooperate intentionally only if the expected-value of doing so is no less than that of not cooperating.

If statement (G3) were true, then my opponent would have established that my counterexample is, indeed, incoherent. It is this statement which lies at the heart of the dispute.[8]

Statement (G3), though, is not true. Suppose, for example, you are aware of all the relevant outcomes: if you were to cooperate, then the outcome would be that you either forgo the opportunity to exploit the other or relinquish the ability to defend yourself; if not, not. Suppose further you value the outcome of not cooperating to that of cooperating. Nevertheless, seeing the benefits to be obtained by agreement, you agree to cooperate, and so do I. You come to believe, correctly, that we have made an agreement. Suppose finally, *you believe (rightly or wrongly, but surely as many believe) that the fact we have an agreement to cooperate is a sufficient reason, in and of itself, to cooperate*. Come the time for cooperating, you do so. In this case, (G3) is false. On the one hand, the antecedent is true: you cooperate, and do so intentionally, since (roughly) cooperation is caused, presumably in the right way, by there being a consideration (namely, that we agreed to do so) you took to be the case, and took to be (rightly or wrongly) a sufficient reason for you to cooperate. From your perspective, you had all the reason you needed to cooperate. On the other hand, the consequent is false: the expected-value of cooperating is *less* than that of not cooperating, since you are completely aware of the fact that the only outcome of cooperating will be the loss of the ability to defend yourself or to exploit the other, and you would rather not suffer this outcome. From your own perspective, however, this reason against cooperating is not of greater strength than the reason you take yourself to have for cooperating – namely, that we agreed. Intentional action, then, is not necessarily action maximising (agent-

[8] For other repsonses to this dispute, see E. F. McClennen, 'Constrained Maximization and Resolute Choice,' *Soc Phil Pol* 5 (1988), pp. 108 ff., and D. Gauthier, 'Morality, Rational Choice and Semantic Representation: a Reply to my Critics,' Soc Phil Pol 5 (1988), §V.

relative) expected-value, but is, if anything at all, action which maximises (agent-relative) expected-strength-of-reasons.

Again, my opponent may not be impressed. I claimed you cooperate intentionally, even though this is not expected-value maximising, if the fact we agreed gives you (at least by your own lights) an extra, non-consequentalist, reason to cooperate. My opponent may claim, though, that while this intention does really give you a reason, it is in fact a *consequentalist* reason.

[3] The suggestion is that if you cooperate, your action manifests the value you place on keeping promises.[9] My opponent's response is as follows: if you are aware of the details of the situation, then you will be aware that the outcome of cooperating is that you fulfil your promise, and the outcome of not cooperating that you do not. However, if you are aware that these are the relevant outcomes, and you cooperate, then you must (implicitly or explicitly) value keeping your promises no less than the sacrifice that may be involved in doing so. Furthermore, in a Promise Puzzle you *are* aware of the details of the situation. Thus, in a Promise Puzzle, if you cooperate then (a) you must be aware that the outcome of cooperating is the fulfilment of your promise and the outcome of not doing so is the non-fulfilment of promise, and (b) you must (implicitly or explicitly) value fulfilling your promises no less than the sacrifice involved in doing so. However, if (a) and (b) are true, then, by definition, the expected-value of cooperating is in fact no less than that of not cooperating. It follows that, in the Promise Puzzle case, if you cooperate, then the expected-value of doing so must be no less than that of not doing so. Statement (G3) is true.

This objection to my claim that agreements provide extra reasons for action is very similar to the objection to my claim that intentions provide extra reasons for action, but it seems the responses will have to be different. The (not so short) response in the case of intentions was that the relevant objection did not do justice to the causal role intention plays in guiding action. Intention enters into the guidance of

[9] My suspicion is that both Harsanyi and MacIntosh, above, would want to insist on this response. See also S. Darwell, 'Rational Agent, Rational Act,' *Phil Topics* 14 (1986): 33-57, and E. F. McClennen, 'The Prisoner's Dilemma and Resolute Choice,' in R. Campbell and L. Sowden, *Paradoxes of Rationality and Cooperation*, (Vancouver: Univ. British Columbia Pr., 1985): 94-104

action not because its presence becomes just one more fact to be considered in deliberation, but rather because it typically pre-empts any further consideration of actions which are inconsistent with the fulfilment of that intention, even though such actions may be ones the agent is perfectly free to do. But it seems this response is not available in the case of agreements, since the belief that we have made an agreement can, and it seems will always, guide action in virtue of the fact that its presence becomes just one more fact to be considered in deliberation, rather than by pre-empting the consideration of options inconsistent with keeping the agreement. It may be that intention guides action via pre-emption, but unlikely that belief about agreements does.

Can the belief that we have agreed play an intention-like role in the guidance of action? I think the answer to the question is 'yes', and to see this, consider the following account – that of Michael Bratman's – of the functional role of intention.[10] The account requires an internal state to satisfy three functional conditions if it is to count as an intention to A. First, the state in question must have a characteristic stability, in that, unexpected conditions to one side, it will continue to exist even when not causally interacting with other internal states. Second, if the agent is in the state, then they will view the issue of whether or not to A as a closed one, and will be disposed not to engage in further thought about whether to A. Finally, the state plays a causal role in deliberation about how to go about doing A. For example, a person in this state, and believing the only way they can A is if they to B, will typically arrive at an intention to B.

It is clear that your belief that we have agreed *can* play each of these functional roles. First, it has the required stability, since we have in fact agreed, and you may be expected not to forget this. Second, in some persons, but not necessarily all, the belief they have agreed to do something will render closed the question of whether or not to do it. To be sure, there will be some (untrustworthy) people for whom this will not be true, but these people do not tell against my suggestion that your belief *can* play this second causal role. Finally, in some persons,

[10] See M. Bratman, *Intentions, Plans, and Practical Reasoning,* (Cambridge, Mass.: Harvard Univ. Pr., 1987). Though I concentrate on Bratman's account, I believe any adequate theory of intention will yield the conclusions of this paragraph. See also J. Raz, 'Reasons for Actions, Decisions and Norms,' *Mind* 84 (1975), pp. 486, 492 ff., and J. Kilcullen, 'Utilitarianism and Virtue', *Ethics* 93 (1983), pp. 458 ff.

though again not necessarily all, a belief that they have agreed to do something will lead to deliberations about how to keep the agreement. For example, your belief that we have agreed to relinquish the use of our arms, and your belief (suppose) that the only reliable way to do this is to turn your swords into ploughshares, leads to an intention to do precisely this. If Bratman is right, then it is a conceptual truth that an intention to A satisfies these three requirements. It may not be a conceptual truth, but rather just an empirical truth about you, that the belief we have agreed also satisfies these three requirements. Such a belief state, then, is playing an intention-like role in guiding your cooperation, and, if it is, then we may suppose, just as with intention, that it guides your cooperation by pre-empting the consideration of options inconsistent with keeping the agreement, rather than by just becoming one more fact to be considered in deliberation about whether to keep the agreement or not.

The Foole might claim that the stipulative addition I made to the Promise Puzzle, (P2), to get the Third Counterexample means that it no longer maximises expected-value for each not to cooperate, but, as we have now seen, such a claim is false. In agreeing to cooperate, each commits themselves to cooperating, but there is no reason to suppose that a commitment to such an action occurs only when each is incapable of doing otherwise, or only when it has the best outcome to perform the action. (Indeed, if only these could provide us with the motive to perform some action, it would be tempting to say we *lacked* commitment to that action.) For all the Foole has said, in the Third Counterexample it may indeed still have the best outcome not to cooperate.

§3 What reason is there for you to cooperate?

The conclusion of the previous chapter stands: you are rationally permitted to cooperate, even though you are free not to do so, and even though it has the best outcome for you not to do so. This, though, raises a question: if you really are rationally permitted to cooperate, then what reason could you possibly have to do so? I shall argue in this section that in the Third Counterexample, and given some conditions, the fact that we have agreed to cooperate is reason enough for you to cooperate.

[1] I claim that if you become an Agreement-Keeper in the situation described by (P3), then the fact that we have agreed to cooperate is (sufficient) reason for you to cooperate. It is important to note two things about this claim. First, I am making a claim about the reasons you have *only in those situations described by assumption (P3),* and am not claiming that, in all cases, the fact that one has made an agreement to do something is a sufficient reason to do that thing. Second, even in this restricted class of cases, I claim only that you have reason to cooperate *if the specified conditions obtain,* and am not claiming you have reason to cooperate, even if we have not agreed, or you have not become an Agreement-Keeper.

Since the Foole is wrong, and the Self-Interest Theory false, there must be a non-consequentalist reason for cooperating; my further speculation, then, concerning this non-consequentalist reason is that it is the fact we have agreed to cooperate. The idea that agreements provide one with extra, non-consequentalist, reasons for action is hardly new, but I want to add my own examples in support of it.

It is important to be clear about the form of my argument. It does not depend on an inference from *the rationality, or the fact, of your believing that* our agreement is a reason to cooperate to that of its *actually being* a reason. This would be to invoke just another bridging principle, which Parfit claims[11] (and I agree) is false. Rather, I am arguing that (a) we may use your belief that it is a reason to defeat an objection to my position, allowing me (b) to reassert my claim that you are rationally permitted to cooperate, and thus leading to my speculation (c) that the fact we have agreed is in indeed a reason to cooperate. There are, though, those who would claim there is no reason to cooperate. In the remainder of this chapter, I will consider what one of them has to say.

[2] Gregory Kavka, in a review of *Morals by Agreement,*[12] comes at one point to consider the reasons a constrained maximiser might have for doing the non-maximising thing, and cooperating. Kavka canvasses a number of possibilities:

[11] D. Parfit, *Reasons and Persons,* (Oxford: Clarendon Pr., 1984), pp. 21 ff.

[12] G. Kavka, 'Review of 'Morals by Agreement',' *Mind* 96 (1987): 117-121. All page references in this subsection will be to this review.

> B correctly calculates that she would maximize long-run expected utility by not carrying out her part. What valid reason could she have to comply rather than maximize? (1) *Compliance is morally right.* This reason doesn't help Gauthier because it gets things backwards from his perspective. ... (2) *Compliance will increase B's future opportunities for cooperation.* By hypothesis this factor is outweighed in B's calculation of expected utility. (3) *If B had not been disposed to comply, she would not have reached a beneficial agreement with A.* True, perhaps, but the agreement has been made and A has already complied. The past cannot be changed; at present non-compliance maximizers expected-utility. ... (4) *B acts contrary to her rationally adopted disposition if she does not comply.* True, but the question is whether this disposition is one to perform rational or irrational acts. Asserting that the rationality of dispositions establishes the rationality of acts in accordance with those dispositions begs the question at issue. ... (5) *If B always acted in this way (non-compliance), she would have few opportunities to cooperate.* This is simply a combination of reasons (2) and (3) and is correspondingly outweighed, confused, or both. (6) *If everyone acted like this, B (or everyone) would lose the benefits of cooperation.* This has no relevance to B who is concerned only to promote her own utility, ... Being unable to think of other plausible reasons B might have to comply, I am inclined to doubt the rationality of acts of constrained maximization.[13]

Kavka is right, it seems to me, to claim that none of (1) to (6) provides sufficient reason for a constrained maximiser to cooperate. Kavka would be right, also, were he to claim that no statements similar to (1) to (6) provide sufficient reason for an Agreement-Keeper to cooperate. Kavka would be wrong, though, to infer from this that there is no reason to cooperate. Having been quoted in length, we can see that he has been thorough in his search for reasons, but, unfortunately, not thorough enough.

Kavka has neglected to consider the possibility that the reason for cooperating is simply that one has agreed to do so. But it may seem that he has a more general argument up his sleeve. He says

> in the absence of sufficient forward-looking reasons for complying, he would not find it rational to comply. Unless, that it, the fact that he has in the past committed himself to constrained maximization itself constitutes a reason for compliance. But, as argued in the previous paragraph [when discussing reason

[13] Kavka, p. 120

> (4), above], it is hard to see how this prior commitment could be such a reason, independent of moral considerations.[14]

In this passage, Kavka suggests that the only way one could argue it is rational to cooperate is to assume – what I agree it would be question begging to assume – that actions expressing a rational disposition are themselves rational. His point may be well taken against Gauthier, but not against the argument I provided in the previous chapter. I argued that it might be rational to cooperate by assuming instead – and more reasonably – that any adequate theory of rationality will not imply that an agent rationally ought to perform each of a number of actions they cannot perform together. We faced a choice, then, whether to reject the claim that you rationally ought to adopt the cooperative disposition, or reject the claim that you rationally ought not to cooperate, and I argued we should reject the latter. This means that you are rationally permitted to cooperate, and this in turn means that there *must be* a reason for you to cooperate, which the consequentalist aspect of the Self-Interest Theory omits. I have suggested (entirely plausibly in my view) that this extra reason is the fact that we have agreed to cooperate.

In response to Kavka, then, I do not assume that this is a reason in order to argue that it is rational for you to cooperate; rather, I argue that it is rational for you to cooperate, conclude that there must be some reason or other for you to cooperate, and speculate that the reason is that we have agreed to cooperate. My argument runs in exactly the opposite order to that which Kavka finds objectionable.

In the Third Counterexample, and given some conditions, the fact we have agreed to cooperate is reason enough for one to cooperate. My argument for this claim has been that, in these situations, the deontic principle (OP_n) entails there is some non-consequentalist reason for you to drink the toxin, and I simply make the further speculation that this reason is simply the fact we have agreed. This is just as one would expect. In common parlance, the fact that one has made an agreement to do something is often taken to be a reason, in and of itself, for one to do that thing, a reason independent of any consequentalist considerations. We see that, in some circumstances at least, agreements are indeed reasons for action.

[14] Kavka, p. 121

Conclusion

The Foole might claim the Third Counterexample is, in effect, incoherent, but he would be mistaken. We end the second part of the thesis with the conclusions of the previous chapter standing. First, that the Foole's claim that it is rational to keep an agreement only if it maximises expected-value to do so is mistaken. Second, that GIVEN that you rationally ought to adopt the enduring disposition to do what we have agreed to do, THEN if you rationally ought to believe we have agreed to do something, you are rationally permitted to perform your part of the bargain, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise (though doing so may very well be irrational absent this condition). It was Nietzsche who said that '[m]an himself must first of all have become *calculable, regular, necessary,* even in his own image of himself, if he is to be able to stand security for *his own future,* which is what one who promises does!'[15] If I am right, then reason, at least, need not stand in the way of such security.

[15] F. Nietzsche, *On the Genealogy of Morals*, trans. by Walter Kaufman and R. J. Hollingdale (New York: Random House, 1967), second essay, sec. 1, p. 58.