

MORALS BY AGREEMENT

DAVID GAUTHIER

CHAPTER VI.

COMPLIANCE: MAXIMIZATION CONSTRAINED

1.1 The just person is disposed to comply with the requirements of the principle of minimax relative concession in interacting with those of his fellows whom he believes to be similarly disposed. The just person is fit for society because he has internalized the idea of mutual benefit, so that in choosing his course of action he gives primary consideration to the prospect of realizing the co-operative outcome. If he is able to bring about, or may reasonably expect to bring about, an outcome that is both (nearly) fair and (nearly) optimal, then he chooses to do so; only if he may not reasonably expect this does he choose to maximize his own utility.

In order to relate our account of the co-operative person to the conditions on rational interaction stated in Chapter III, let us define a fair optimizing strategy (or choice, or response) as one that, given the expected strategies of the others, may be expected to yield an outcome that is nearly fair and optimal—an outcome with utility pay-offs close to those of the co-operative outcome, as determined by minimax relative concession. We speak of the response as nearly fair and optimal because in many situations a person will not expect others to do precisely what would be required by minimax relative concession, so that he may not be able to choose a strategy with an expected outcome that is completely fair or fully optimal. But we suppose that he will still be disposed to co-operative rather than to non-co-operative interaction.

A just person then accepts this reading of condition A: A': Each person's choice must be a fair optimizing response to the choice he expects the others to make, provided such a response is available to him; otherwise, his choice must be a utility-maximizing response. A just person is disposed to interact with others on the basis of condition A'.

A just person must however be aware that not all (otherwise) rational persons accept this reading of the original condition A. In forming expectations about the choices of others, he need not suppose that their choices will satisfy A'. Thus as conditions of strategic interaction, we cannot dispense with the original conditions A, B, and C; 'rational response' remains (at least until our theory has gained universal acceptance) open to several interpretations.

Our task in this chapter is to provide a utility-maximizing rationale for condition A'. We shall do this by demonstrating that, given certain plausible and desirable conditions, a rational utility maximizer, faced with the choice between accepting no constraints on his choices in interaction, and accepting the constraints on his choices required by minimax relative concession, chooses the latter. He makes a choice about how to make further choices; he chooses, on utility-maximizing grounds, not to make further choices on those grounds.

In defending condition A', we defend compliance with agreements based, explicitly or implicitly, on the principle of minimax relative concession. Indeed, we defend compliance, not just with agreements, but with practices that would be agreed to or endorsed on the basis of this principle. If our defence fails, then we must conclude that rational bargaining is in vain and that co-operation, although on a rationally agreed basis, is not itself rationally required, so that it does not enable us to overcome the failings of natural and market interaction. Indeed, if our defence fails, then we must conclude that a rational morality is a chimera, so that there is no rational and impartial constraint on the pursuit of individual utility.

In defending condition A', we uphold the external rationality of co-operation against the objections of the egoist. Whatever else he may do, the egoist always seeks to maximize his expected utility. Recognizing that co-operation offers the prospect of mutual benefit, he nevertheless denies that it is rational to behave co-operatively, where this would constrain maximization. This egoist makes his philosophical debut as the Foole in Thomas Hobbes's Leviathan, where we shall now observe him.

1.2 Hobbes begins his moral theory with a purely permissive conception of the right of nature, stating what one may do, not what one must be let do, or what must be done for one. The permission is rational, for as Hobbes says, 'Neither by the word right is anything else signified, than that liberty which every man hath to make use of his natural faculties according to right reason.'¹ And Hobbes claims that in the natural condition of humankind this liberty is unlimited,

so that 'every man has a Right to every thing; even to one anothers body.'² In so treating the right of nature, Hobbes expresses a straightforwardly maximizing view of rational action, subject to the material condition, central to his psychology, that each seeks above all his own preservation. For Hobbes each person has the initial right to do whatever he can to preserve himself, but there is no obligation on others, either to let him do or to do for him what is necessary to his preservation.

The condition in which this unlimited right is exercised by all persons is, Hobbes claims, one in which 'there can be no security to any man, (how strong or wise soever he be,) of living out the time, which Nature ordinarily alloweth men to live.'³ Persons who seek their own preservation find themselves locked in mortal combat. But if reason brings human beings to this condition of war, it can also lead them out of it. Hobbes says, 'Reason suggesteth convenient Articles of Peace, upon which men may be drawn to agreement. These Articles. . . are called the Lawes of Nature.'⁴ Laws of nature are precepts, 'found out by Reason, by which a man is forbidden to do, that, which is destructive of his life, or taketh away the means of preserving the same; and to omit, that, by which he thinketh it may be best preserved.'⁵

Since war is inimical to preservation, the fundamental or first law of nature is, 'That every man, ought to endeavour Peace, as farre as he has hope of obtaining it', to which Hobbes adds, 'and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre.'⁶ From this Hobbes immediately derives a second law, setting out, as the fundamental means to peace, 'That a man be willing, when others are so too, as farre-forth, as for Peace, and defence of himselfe he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other men, as he would allow other men against himselfe.'⁷ Since the unlimited right of nature gives rise to war, renouncing some part of this right is necessary for peace. The renunciation must of course be mutual; each person expects to benefit, not from his own act of renunciation, but from that of his fellows, and so no one has reason to renounce his rights unilaterally. What Hobbes envisages is a rational bargain in which each accepts certain constraints on his freedom of action so that all may avoid the costs of the natural condition of war. The defence of this second law is perfectly straightforward. Hobbes needs to say only that 'as long as every man holdeth this Right, of doing any thing he liketh; so long are all men in the condition of Warre.'⁸ And the mutuality required by the law is defended in an equally simple way: 'if other men will not lay down their Right, as well as he; then there is no Reason for anyone, to devest himselfe of his: For that were to expose himselfe to Prey, (which no man is bound to) rather than to dispose himselfe to Peace.'⁹ It is directly advantageous for each to agree with his fellows to a mutual renunciation or laying down of right, and so a mutual acceptance of constraint. Hobbes conceives such constraint as obligation, arising only through agreement, for there is 'no Obligation on any man, which ariseth not from some Act of his own; for all men equally, are by Nature Free.'¹⁰ Hobbes's theory, as our own, introduces morals by agreement.

Hobbes recognizes that it is one thing to make an agreement or covenant, quite another to keep it.. He does not suppose that the second law of nature, enjoining us to agree, also enjoins us to compliance. Thus he introduces a third law of nature, 'That men performe their Covenants made', which he considers to be the 'Originall of JUSTICE'.¹¹ A just person is one who keeps the agreements he has rationally made.

Hobbes's defence of this third law lacks the straightforwardness of his defence of the second. As he recognizes, without it 'Covenants are in vain, and but Empty words; and the Right of all men to all things remaining, wee are still in the condition of Warre.'¹² But this does not show that conformity to it yields any direct benefit. Each person maximizes his expected utility in making a covenant, since each gains from the mutual renunciation it involves. But each does not maximize his expected utility in keeping a covenant, in so far as it requires him to refrain from exercising some part of his previous liberty. And this opens the door to the objection of the Foole. We shall let him speak for himself. The Foole 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 to make, or not make; keep, or not keep Covenants, was not against Reason, when it conduced to ones benefit. He does not therein deny, that there be Covenants; and that they are sometimes . broken, sometimes kept; and that such breach of them may be called Injustice, and the observance of them Justice: but he questioneth, whether Injustice. . . may not sometimes stand with that Reason, which dictateth to every man his own good. . . .13

The Foole does not seriously challenge the second law of nature, for Hobbes assumes that each person will make only those covenants that he expects to be advantageous, and such behaviour the Foole does not question. What the Foole challenges is the third law, the law requiring compliance, or adherence to one's covenants, for let it be ever so advantageous to make an agreement, may it not then be even more advantageous to violate the agreement made? And if advantageous, then is it not rational? The Foole challenges the heart of the connection between reason and morals that both Hobbes and we seek to establish - the rationality of accepting a moral constraint on the direct pursuit of one's greatest utility.

1.3 In replying to the Foole, Hobbes claims that the question is, given sufficient security of performance by one party, 'whether it be against reason, that is, against the benefit of the other to performe, or not'.¹⁴ On the most natural interpretation, Hobbes is asking whether keeping one's covenant is a rational, that is utility-maximizing, response to covenant-keeping by one's fellows. If this is indeed Hobbes's view, then he is endeavouring to refute the Foole by appealing, in effect, to condition A for strategically rational choice, taking a rational response to be simply a utility-maximizing response. We may not be very hopeful about Hobbes's prospect of success.

Hobbes's first argument reminds the Foole that the rationality of choice depends on expectations, not actual results. It need not detain us. His second argument joins issue with the Foole at a deeper level.

He ... that breaketh his Covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any Society, that unite themselves for Peace and Defence, but by the error of them that receive him; nor when he is received, be retayned in it, without seeing the danger of their error; which errors a man cannot reasonably reckon upon as the means of his security.¹⁵

A person disposed to violate his covenants cannot be admitted as a party to co-operative arrangements by those who are both rational and aware of his disposition, and so such a person cannot rationally expect to reap the benefits available to co-operators. Even if his particular breaches of covenant would benefit him, yet the disposition that leads him to such breaches does not.

In effect Hobbes moves the question from whether it be against reason, understood as utility-maximization, to keep one's agreement (given sufficient security of others keeping their agreements), to whether it be against reason to be disposed to keep one's agreement. The disposition to decide whether or not to adhere to one's covenants or agreements by appealing to directly utility-maximizing considerations, is itself disadvantageous, if known, or sufficiently suspected, because it excludes one from participating, with those who suspect one's disposition, in those co-operative arrangements in which the benefits to be realized require each to forgo utility-maximization--or in Hobbes's terminology, require each to lay down some portion of his original, unlimited right of nature. The disposition to keep one's agreement, given sufficient security, without appealing to directly utility-maximizing considerations, makes one an eligible partner in beneficial co-operation, and so is itself beneficial. This will prove to be the key to our demonstration that a fully rational utility-maximizer disposes himself to compliance with his rationally undertaken covenants or agreements. But for Hobbes to take full advantage of this response to the Foole, he must revise his conception of rationality, breaking the direct connection between reason and benefit with which he began his reply. Hobbes needs to say that it is rational to perform one's covenant even when performance is not directly to one's benefit, provided that it is to one's benefit to be disposed to perform. But this he never says. And as long as the Foole is allowed to relate reason directly to benefit in performance, rather than to benefit in the disposition to perform, he can escape refutation.

Hobbes does suggest a revision in his conception of rationality in his discussion with Bishop Bramhall. Agreeing with Bramhall that 'moral goodness is the conformity of an action with right reason', he does not claim that what is morally good is conducive to one's benefit, but instead holds that

All the real good. . . is that which is not repugnant to the law. . . for the law is all the right reason we have, and... is the infallible rule of moral goodness. The reason whereof is this, that because neither mine nor the Bishop's reason is . . . fit to be a rule of our moral actions, we have therefore set up over ourselves a sovereign governor, and agreed that his laws shall. . . dictate to us what is really good.¹⁶

To the Foo1e's contention that injustice may 'sometimes stand with that Reason, which dictateth to every man his own good',¹⁷ Hobbes can reply that injustice may not stand with that reason that is constituted by the law of the sovereign. Just as it is unprofitable for each man to retain his entire natural right, so it is unprofitable for each man to retain his natural reason as guide to his actions. But Hobbes does not suppose that each man internalizes the right reason of the sovereign. His egoistic psychology allows the internalization of no standard other than that of direct concern with individual preservation and contentment. And so it is only in so far as the sovereign is able to enforce the law that compliance with it is rationally binding on the individual. But this is to propose a political, not a moral, solution to the problem posed by the Foo1e.

If the market acts as an invisible hand, directing the efforts of each person intending only his own benefit to a social optimum, the sovereign acts as a very visible foot, directing, by well-placed kicks, the efforts of each to the same social end. Each device performs the same task, ensuring the coincidence of an equilibrium in which each person maximizes his expected utility given the actions of his fellows, with an optimum in which each person gains the maximum utility compatible with the utilities of his fellows. Each device affects the conditions under which interaction occurs, leaving every individual free to maximize his utility given those conditions. Of course, the sovereign appears as a constraint on each person's freedom whereas the market does not, but this is the difference between visibility and invisibility; the sovereign visibly shapes the conditions that reconcile each person's interest with those of his fellows, whereas the market so shapes these conditions simply in virtue of its structure.

The sovereign makes morality, understood as a constraint on each person's endeavour to maximize his own utility, as unnecessary as does the market. Our moral enquiry has been motivated by the problems created for utility-maximizers by externalities. Adam Smith reminds us of the conditions in which externalities are absent, so that the market ensures that each person's free, maximizing behaviour results in an optimal outcome. Thomas Hobbes introduces the sovereign, who constrains each person's options so that maximizing behaviour results in a seemingly optimal outcome even when externalities are present. We may retain the idea of justice as expressing the requirement of impartiality for principles that regulate social interaction, but it no longer expresses a constraint on individual maximization. It would seem that between them, economics and politics resolve our problem with no need for morality.

But Hobbes's sovereign lacks the appeal of the market, and for good reason. The invisible hand is a costless solution to the problems of natural interaction, but the visible foot is a very costly solution. Those subject to the Hobbesian sovereign do not, in fact, attain an optimal outcome; each pays a portion of the costs needed to enforce adherence to agreements, and these costs render the outcome sub-optimal. Even if we suppose that power does not corrupt, so that the sovereign is the perfect instrument of his subjects, acting only in their interests, yet each would expect to do better if all would adhere voluntarily to their agreements, so that enforcement and its costs would be unnecessary. We pay a heavy price, if we are indeed creatures who rationally accept no internal constraint on the pursuit of our own utility, and who consequently are able to escape from the state of nature, in those circumstances in which externalities are unavoidably present, only by political, and not by moral, devices. Could we but voluntarily comply with our rationally undertaken agreements, we should save ourselves this price.

We do not suppose that voluntary compliance would eliminate the need for social institutions and practices, and their costs. But it would eliminate the need for some of those institutions whose concern is with enforcement. Authoritative decision-making cannot be eliminated, but our ideal would be a society in which the

coercive enforcement of such decisions would be unnecessary. More realistically, we suppose that such enforcement is needed to create and maintain those conditions under which individuals may rationally

expect the degree of compliance from their fellows needed to elicit their own voluntary compliance. Internal, moral constraints operate to ensure compliance under conditions of security established by external, political constraints. But before we can expect this view to be accepted we must show, what the Foole denies, that it is rational to dispose oneself to co-operate, and so to accept internal, moral constraints. Hobbes's argument that those not so disposed may not rationally be received into society, is the foundation on which we shall build.

2.1 The Foole, and those who share his conception of practical reason, must suppose that there are potentialities for co-operation to which each person would rationally agree, were he to expect the agreement to be carried out, but that remain unactualized, since each rationally expects that someone, perhaps himself, perhaps another, would not adhere to the agreement. In Chapter V we argued that co-operation is rational if each co-operator may expect a utility nearly equal to what he would be assigned by the principle of minimax relative concession. The Foole does not dispute the necessity of this condition, but denies its sufficiency. He insists that for it to be rational to comply with an agreement to co-operate, the utility an individual may expect from co-operation must also be no less than what he would expect were he to violate his agreement. And he then argues that for it to be rational to agree to co-operate, then, although one need not consider it rational to comply oneself, one must believe it rational for the others to comply. Given that everyone is rational, fully informed, and correct in his expectations, the Foole supposes that co-operation is actualized only if each person expects a utility from co-operation no less than his noncompliance utility. The benefits that could be realized through cooperative arrangements that do not afford each person at least his non-compliance utility remain forever beyond the reach of rational human beings—forever denied us because our very rationality would lead us to violate the agreements necessary to realize these benefits. Such agreements' will not be made.

The Foole rejects what would seem to be the ordinary view that, given neither unforeseen circumstances nor misrepresentation of terms, it is rational to comply with an agreement if it is rational to make it. He insists that holders of this view have failed to think out the full implications of the maximizing conception of practical rationality. In choosing one takes one's stand in the present, and

looks to the expected utility that will result from each possible action, What has happened may affect this utility; that one has agreed may affect the utility one expects from doing, or not doing, what would keep the agreement. But what has happened provides in itself no reason for choice, That one had reason for making an agreement can give one reason for keeping it only by affecting the utility of compliance. To think otherwise is to reject utility-maximization.

Let us begin our answer to the Foole by recalling the distinction introduced in V,1.3 between an individual strategy and a joint strategy,¹⁸ An individual strategy is a lottery over the possible actions of a single actor, A joint strategy is a lottery over possible outcomes, Co-operators have joint strategies available to them, We may think of participation in a co-operative activity, such as a hunt, in which each huntsman has his particular role co-ordinated with that of the others, as the implementation of a single joint strategy. We may also extend the notion to include participation in a practice, such as the making and keeping of promises, where each person's behaviour is predicated on the conformity of others to the practice.

An individual is not able to ensure that he acts on a joint strategy, since whether he does depends, not only on what he intends, but on what those with whom he interacts intend. But we may say that an individual bases his action on a joint strategy in so far as he intentionally chooses what the strategy requires of him. Normally, of course, one bases one's action on a joint strategy only if one expects those with whom one interacts to do so as well, so that one expects actually to act on that strategy, But we need not import such an expectation into the conception of basing one's action on a joint strategy.

A person co-operates with his fellows only if he bases his actions on a joint strategy; to agree to co-operate is to agree to employ a joint rather than an individual strategy, The Foole insists that it is rational to co-operate only if the utility one expects from acting on the co-operative joint strategy is at least equal to the utility one

would expect were one to act instead on one's best individual strategy. This defeats the end of co-operation, which is in effect to substitute a joint strategy for individual strategies in situations in which this substitution is to everyone's benefit.

A joint strategy is fully rational only if it yields an optimal outcome, or in other words, only if it affords each person who acts on it the maximum utility compatible in the situation with the utility afforded each other person who acts on the strategy. Thus we may say that a person acting on a rational joint strategy maximizes his utility, subject to the constraint set by the utilities it affords to every other person. An individual strategy is rational if and only if it maximizes one's utility given the strategies adopted by the other persons; a joint strategy is rational only if (but not if and only if) it maximizes one's utility given the utilities afforded to the other persons.

Let us say that a *straightforward* maximizer is a person who seeks to maximize his utility given the strategies of those with whom he interacts. A *constrained* maximizer, on the other hand, is a person who seeks in some situations to maximize her utility, given not the strategies but the utilities of those with whom she interacts. The first accepts the rationality of straightforward maximization. We, in defending condition A' for strategic rationality (stated in 1.1), accept the rationality of constrained maximization.

A constrained maximizer has a conditional disposition to base her actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility. But not all constraint could be rational; we must specify the characteristics of the conditional disposition. We shall therefore identify a constrained maximizer 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. Or in other words, a constrained maximizer is ready to co-operate in ways that, if followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial. In determining the latter she must take into account the possibility that some persons will fail, or refuse, to act co-operatively. Henceforth, unless we specifically state otherwise, we shall understand by a constrained maximizer one with this particular disposition.

There are three points in our characterization of constrained maximization that should be noted. The first is that a constrained maximizer is conditionally disposed to act, not only on the unique joint strategy that would be prescribed by a rational bargain, but on any joint strategy that affords her a utility approaching what she would expect from fully rational co-operation. The range of acceptable joint strategies is, and must be left, unspecified. The idea is that in real interaction it is reasonable to accept co-operative arrangements that fall short of the ideal of full rationality and fairness, provided they do not fall too far short. At some point, of course, one decides to ignore a joint strategy, even if acting on it would afford one an expected utility greater than one would expect were everyone to employ an individual strategy, because one hopes thereby to obtain agreement on, or acquiescence in, another joint strategy which in being fairer is also more favourable to oneself. At precisely what point one decides this we make no attempt to say. We simply defend a conception of constrained maximization that does not require that all acceptable joint strategies be ideal.

Constrained maximization thus links the idea of morals by agreement to actual moral practice. We suppose that some moral principles may be understood as representing joint strategies prescribed to each person as part of the ongoing co-operative arrangements that constitute society. These principles require each person to refrain from the direct pursuit of her maximum utility, in order to achieve mutually advantageous and reasonably fair outcomes. Actual moral principles are not in general those to which we should have agreed in a fully rational bargain, but it is reasonable to adhere to them in so far as they offer a reasonable approximation to ideal principles. We may defend actual moral principles by reference to ideal co-operative arrangements, and the closer the principles fit, the stronger the defence. We do not of course suppose that our actual moral principles derive historically from a bargain, but in so far as the constraints they impose are acceptable to a rational constrained maximizer, we may fit them into the framework of a morality rationalized by the idea of agreement.

The second point is that a constrained maximizer does not base her actions on a joint strategy whenever a nearly fair and optimal outcome would result were everyone to do likewise. 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 cooperate. 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-cooperation, does her conditional disposition to constraint actually manifest itself in a decision to base her actions on the co-operative joint strategy.

Thus, faced with persons whom she believes to be straightforward maximizers, a constrained maximizer 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, acting on the individual strategy that maximizes her utility given the strategies she expects the others to employ. A constrained maximizer makes reasonably certain that she is among like-disposed persons before she actually constrains her direct pursuit of maximum utility.

But note that a constrained maximizer may find herself required to act in such a way that she would have been better off had she not entered into co-operation. She may be engaged in a co-operative activity that, given the willingness of her fellows to do their part, she expects to be fair and beneficial, but that, should chance so befall, requires her to act so that she incurs some loss greater than had she never engaged herself in the endeavour. Here she would still be disposed to comply, acting in a way that results in real disadvantage to herself, because given her ex ante beliefs about the dispositions of her fellows and the prospects of benefit, participation in the activity affords her greater expected utility than non-participation.

And this brings us to the third point, that constrained maximization is not straightforward maximization in its most effective disguise. The constrained maximizer is not merely the person who, taking a larger view than her fellows, serves her overall interest by sacrificing the immediate benefits of ignoring joint strategies and violating co-operative arrangements in order to obtain the long-run benefits of being trusted by others.¹⁹ Such a person exhibits no real constraint. The constrained maximizer does not reason more effectively about how to maximize her utility, but reasons in a different way. We may see this most clearly by considering how each faces the decision whether to base her action on a joint strategy. The constrained maximizer considers (i) whether the outcome, should everyone do so, be nearly fair and optimal, and (ii) whether the outcome she realistically expects should she do so affords her greater utility than universal non-cooperation. If both of these conditions are satisfied she bases her action on the joint strategy. The straightforward maximizer considers simply whether the outcome he realistically expects should he base his action on the joint strategy affords him greater utility than the outcome he would expect were he to act on any alternative strategy--taking into account, of course, long-term as well as short-term effects. Only if this condition is satisfied does he base his action on the joint strategy.

Consider a purely isolated interaction, in which both parties know that how each chooses will have no bearing on how each fares in other interactions. Suppose that the situation has the familiar Prisoner's Dilemma structure; each benefits from mutual cooperation in relation to mutual non-cooperation, but each benefits from non-cooperation whatever the other does. In such a situation, a straightforward maximizer chooses not to co-operate. A constrained maximizer chooses to co-operate if, given her estimate of whether or not her partner will choose to co-operate, her own expected utility is greater than the utility she would expect from the non-cooperative outcome.

Constrained maximizers can thus obtain co-operative benefits that are unavailable to straightforward maximizers, however farsighted the latter may be. But straightforward maximizers can, on occasion, exploit unwary constrained maximizers. Each supposes her disposition to be rational. But who is right?

2.2 To demonstrate the rationality of suitably constrained maximization we solve a problem of rational choice. We consider what a rational individual would choose, given the alternatives of adopting straightforward maximization, and of adopting constrained maximization, as his disposition for strategic behaviour. Although this choice is about interaction, to make it is not to engage in interaction. Taking others' dispositions as fixed,

the individual reasons parametrically to his own best disposition. Thus he compares the expected utility of disposing himself to maximize utility given others' expected strategy choices, with the utility of disposing himself to co-operate with others in bringing about nearly fair and optimal outcomes.

To choose between these dispositions, a person needs to consider only those situations in which they would yield different behaviour. If both would be expressed in a maximizing individual strategy, or if both would lead one to base action on the joint strategy one expects from others, then their utility expectations are identical. But if the disposition to constraint would be expressed in basing action on a joint strategy, whereas the disposition to maximize straightforwardly would be expressed in defecting from the joint strategy, then their utility expectations differ. Only situations giving rise to such differences need be considered. These situations must satisfy two conditions. First, they must afford the prospect of mutually beneficial and fair co-operation, since otherwise constraint would be pointless. And second, they must afford some prospect for individually beneficial defection, since otherwise no constraint would be needed to realize the mutual benefits. We suppose, then, an individual, considering what disposition to adopt, for situations in which his expected utility is u should each person act on an individual strategy, u' should all act on a cooperative joint strategy, and u'' should he act on an individual strategy and the others base their actions on a co-operative joint strategy, and u is less than u' (so that he benefits from co-operation as required by the first condition) and u' in turn is less than u'' (so that he benefits from defection as required by the second condition). Consider these two arguments which this person might put to himself:

Argument (1): Suppose I adopt straightforward maximization. Then if I expect the others to base their actions on a joint strategy, I defect to my best individual strategy, and expect a utility, u'' . If I expect the others to act on individual strategies, then so do I, and expect a utility, u . If the probability that others will base their actions on a joint strategy is p , then my overall expected utility is $[pu'' + (1-p)u]$. Suppose I adopt constrained maximization. Then if I expect the others to base their actions on a joint strategy, so do I, and expect a utility u' . If I expect the others to act on individual strategies, then so do I, and expect a utility, u . Thus my overall expected utility is $[pu' + (1-p)u]$.

Since u'' is greater than u' , $[pu'' + (1-p)u]$ is greater than $[pu' + (1-p)u]$, for any value of p other than 0 (and for $p = 0$, the two are equal). Therefore, to maximize my overall expectation of utility, I should adopt straightforward maximization.

Argument (2): Suppose I adopt straightforward maximization. Then I must expect the others to employ maximizing individual strategies in interacting with me; so do I, and expect a utility, u . Suppose I adopt constrained maximization. Then if the others are conditionally disposed to constrained maximization, I may expect them to base their actions on a co-operative joint strategy in interacting with me; so do I, and expect a utility u' . If they are not so disposed, I employ a maximizing strategy and expect u as before. If the probability that others are disposed to constrained maximization is p , then my overall expected utility is $[pu' + (1-p)u]$.

Since u' is greater than u , $[pu' + (1-p)u]$ is greater than u for any value of p other than 0 (and for $p = 0$, the two are equal). Therefore, to maximize my overall expectation of utility, I should adopt constrained maximization. Since these arguments yield opposed conclusions, they cannot both be sound. The first has the form of a dominance argument. In any situation in which others act non-cooperatively, one may expect the same utility whether one is disposed to straightforward or to constrained maximization. In any situation in which others act co-operatively, one may expect a greater utility if one is disposed to straightforward maximization. Therefore one should adopt straightforward maximization. But this argument would be valid only if the probability of others acting co-operatively were, as the argument assumes, independent of one's own disposition. And this is not the case. Since persons disposed to co-operation only act co-operatively with those whom they suppose to be similarly disposed, a straightforward maximizer does not have the opportunities to benefit which present themselves to the constrained maximizer. Thus argument (1) fails.

Argument (2) takes into account what argument (1) ignores the difference between the way in which constrained maximizers interact with those similarly disposed, and the way in which they interact with straightforward maximizers. Only those disposed to keep their agreements are rationally acceptable as parties to agreements. Constrained maximizers are able to make beneficial agreements with their fellows that the straightforward cannot, not because the latter would be unwilling to agree, but because they would not be

admitted as parties to agreement given their disposition to violation. Straightforward maximizers are disposed to take advantage of their fellows should the opportunity arise; knowing this, their fellows would prevent such opportunity arising. With the same opportunities, straightforward maximizers would necessarily obtain greater benefits. A dominance argument establishes this. But because they differ in their dispositions, straightforward and constrained maximizers differ also in their opportunities, to the benefit of the latter.

But argument (2) unfortunately contains an undefended assumption. A person's expectations about how others will interact with him depend strictly on his own choice of disposition only if that choice is known by the others. What we have shown is that, if the straightforward maximizer and the constrained maximizer appear in their true colours, then the constrained maximizer must do better. But need each so appear? The Foole may agree, under the pressure of our argument and its parallel in the second argument we ascribed to Hobbes, that the question to be asked is not whether it is or is not rational to keep (particular) covenants, but whether it is or is not rational to be (generally) disposed to the keeping of covenants, and he may recognize that he cannot win by pleading the cause of straightforward maximization in a direct way. But may he not win by linking straightforward maximization to the appearance of constraint? Is not the Foole's ultimate argument that the truly prudent person, the fully rational utility-maximizer, must seek to appear trustworthy, an upholder of his agreements? For then he will not be excluded from the co-operative arrangements of his fellows, but will be welcomed as a partner, while he awaits opportunities to benefit at their expense-and, preferably, without their knowledge, so that he may retain the guise of constraint and trustworthiness.

There is a short way to defeat this manoeuvre. Since our argument is to be applied to ideally rational persons, we may simply add another idealizing assumption, and take our persons to be transparent.²⁰ Each is directly aware of the disposition of his fellows, and so aware whether he is interacting with straightforward or constrained maximizers. Deception is impossible; the Foole must appear as he is.

But to assume transparency may seem to rob our argument of much of its interest. We want to relate our idealizing assumptions to the real world. If constrained maximization defeats straightforward maximization only if all persons are transparent, then we shall have failed to show that under actual, or realistically possible, conditions, moral constraints are rational. We shall have refuted the Foole but at the price of robbing our refutation of all practical import.

However, transparency proves to be a stronger assumption than our argument requires. We may appeal instead to a more realistic translucency, supposing that persons are neither transparent nor opaque, so that their disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork. Opaque beings would be condemned to seek political solutions for those problems of natural interaction that could not be met by the market. But we shall show that for beings as translucent as we may reasonably consider ourselves to be, moral solutions are rationally available.

2.3 If persons are translucent, then constrained maximizers (CMs) will sometimes fail to recognize each other, and will then interact non-co-operatively even if co-operation would have been mutually beneficial. CMs will sometimes fail to identify straightforward maximizers (SMs) and will then act co-operatively; if the SMs correctly identify the CMs they will be able to take advantage of them. Translucent CMs must expect to do less well in interaction than would transparent CMs; translucent SMs must expect to do better than would transparent SMs. Although it would be rational to choose to be a CM were one transparent, it need not be rational if one is only translucent. Let us examine the conditions under which the decision to dispose oneself to constrained maximization is rational for translucent persons, and ask if these are (or may be) the conditions in which we find ourselves.

As in the preceding subsection, we need consider only situations in which CMs and SMs may fare differently. These are situations that afford both the prospect of mutually beneficial co-operation (in relation to non-co-operation) and individually beneficial defection (in relation to co-operation). Let us simplify by supposing that the non-co-operative outcome results unless (i) those interacting are CMs who achieve mutual recognition, in which case the co-operative outcome results, or (ii) those interacting include CMs who fail to recognize SMs but are themselves recognized, in which case the outcome affords the SMs the benefits of individual defection and the CMs the costs of having advantage taken of mistakenly basing their actions on a co-operative strategy. We ignore the inadvertent taking . of advantage when CMs mistake their fellows for SMs.

There are then four possible pay-offs--non-co-operation, cooperation, defection, and exploitation (as we may call the outcome for the person whose supposed partner defects from the joint strategy on which he bases his action). For the typical situation, we assign defection the value 1, co-operation u'' (less than 1), non-cooperation u' (less than u''), and exploitation 0 (less than u'). We now introduce three probabilities. The first, p , is the probability that CMs will achieve mutual recognition and so successfully co-operate. The second, q , is the probability that CMs will fail to recognize SMs but will themselves be recognized, so that defection and exploitation will result. The third, r , is the probability that a randomly selected member of the population is a CM. (We assume that everyone is a CM or an SM, so the probability that a randomly selected person is an SM is $(1-r)$.) The values of p , q , and r must of course fall between 0 and 1.

Let us now calculate expected utilities for CMs and SMs in situations affording both the prospect of mutually beneficial cooperation and individually beneficial defection. A CM expects the utility u' unless (i) she succeeds in co-operating with other CMs or (ii) she is exploited by an SM. The probability of (i) is the combined probability that she interacts with a CM, r , and that they achieve mutual recognition, p , or rp . In this case she gains $(u'' - u')$ over her non-co-operative expectation u' . Thus the effect of (i) is to increase her utility expectation by a value $[rp(u'' - u')]$. The probability of (ii) is the combined probability that she interacts with an SM, $1-r$, and that she fails to recognize him but is recognized, q , or $(1-r)q$. In this case she receives 0, so she loses her non-co-operative expectation u' .

Thus the effect of (ii) is to reduce her utility expectation by a value $[(1-r)qu']$. Taking both (i) and (ii) into account, a CM expects the utility $\{u' + [rp(u'' - u')] - (1-r)qu'\}$.

An SM expects the utility u' unless he exploits a CM. The probability of this is the combined probability that he interacts with a CM, r , and that he recognizes her but is not recognized by her, q , or rq . In this case he gains $(1-u')$ over his non-cooperative expectation u' . Thus the effect is to increase his utility expectation by a value $[rq(1-u')]$. An SM thus expects the utility $\{u' + [rq(1-u')]\}$.

It is rational to dispose oneself to constrained maximization if and only if the utility expected by a CM is greater than the utility expected by an SM, which obtains if and only if p/q is greater than $\{(1-u')/(u'' - u') + [(1-r)u']/[r(u'' - u')]\}$.

The first term of this expression, $[(1-u')/(u'' - u')]$, relates the gain from defection to the gain through co-operation. The value of defection is of course greater than that of co-operation, so this term is greater than 1. The second term, $\{[(1-r)u']/[r(u'' - u')]\}$, depends for its value on r . If $r = 0$ (i.e. if there are no CMs in the population), then its value is infinite. As r increases, the value of the expression decreases, until if $r = 1$ (i.e. if there are only CMs in the population) its value is 0.

We may now draw two important conclusions. First, it is rational to dispose oneself to constrained maximization only if the ratio of p to q , i.e. the ratio between the probability that an interaction involving CMs will result in co-operation and the probability that an interaction involving CMs and SMs will involve exploitation and defection, is greater than the ratio between the gain from defection and the gain through co-operation. If everyone in the population is a CM, then we may replace 'only if' by 'if and only if' in this statement, but in general it is only a necessary condition of the rationality of the disposition to constrained maximization.

Second, as the proportion of CMs in the population increases (so that the value of r increases), the value of the ratio of p to q that is required for it to be rational to dispose oneself to constrained maximization decreases. The more constrained maximizers there are, the greater the risks a constrained maximizer may rationally accept of failed co-operation and exploitation. However, these risks, and particularly the latter, must remain relatively small.

We may illustrate these conclusions by introducing typical numerical values for co-operation and non-cooperation, and then considering different values for r . One may suppose that on the whole, there is no reason that the typical gain from defection over co-operation would be either greater or smaller than the typical gain from co-operation over non-co-operation, and in turn no reason that the latter gain would be greater or

smaller than the typical loss from non-cooperation to exploitation. And so, since defection has the value 1 and exploitation 0, let us assign co-operation the value $2/3$ and non-cooperation $1/3$.

The gain from defection, $(1 - u')$, thus is $2/3$; the gain through cooperation, $(u'' - u')$, is $1/3$. Since p/q must exceed $\{(1 - u') / (u'' - u') + [(1 - r)u'] / [r(u'' - u')]\}$ for constrained maximization to be rational, in our typical case the probability p that CMs successfully co-operate must be more than twice the probability q that CMs are exploited by SMs, however great the probability r that a randomly selected person is a CM. If three persons out of four are CMs, so that $r = 3/4$, then p/q must be greater than $7/3$; if one person out of two is a CM, then p/q must be greater than 3; if one person in four is a CM, then p/q must be greater than 5. In general, p/q must be greater than $2 + (1 - r)/r$, or $(r + 1)/r$.

Suppose a population evenly divided between constrained and straightforward maximizers. If the constrained maximizers are able to co-operate successfully in two-thirds of their encounters, and to avoid being exploited by straightforward maximizers in four-fifths of their encounters, then constrained maximizers may expect to do better than their fellows. Of course, the even distribution will not be stable; it will be rational for the straightforward maximizers to change their disposition. These persons are sufficiently translucent for them to find morality rational.

2.4 A constrained maximizer is conditionally disposed to cooperate 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. In the two preceding subsections, we have argued that one is rationally so disposed if persons are transparent, or if persons are sufficiently translucent and enough are like-minded. But our argument has not appealed explicitly to the particular requirement that co-operative practices and activities be nearly optimal and fair. We have insisted that the co-operative outcome afford one a utility greater than non-cooperation, but this is much weaker than the insistence that it approach the outcome required by minimax relative concession.

But note that the larger the gain from co-operation, $(u/l - u')$, the smaller the minimum value of p/q that makes the disposition to constrained maximization rational. We may take p/q to be a measure of translucency; the more translucent constrained maximizers are, the better they are at achieving co-operation among themselves (increasing p) and avoiding exploitation by straightforward maximizers (decreasing q). Thus as practices and activities fall short of optimality, the expected value of co-operation, u/l , decreases, and so the degree of translucency required to make cooperation rational increases. And as practices and activities fall short of fairness, the expected value of co-operation for those with less than fair shares decreases, and so the degree of translucency to make co-operation rational for them increases. Thus our argument does appeal implicitly to the requirement that co-operation yield nearly fair and optimal outcomes.

But there is a further argument in support of our insistence that the conditional disposition to co-operate be restricted to practices and activities yielding nearly optimal and fair outcomes. And this argument turns, as does our general argument for constraint, on how one's dispositions affect the characteristics of the situations in which one may reasonably expect to find oneself. Let us call a person who is disposed to co-operate in ways that, followed by all, yield nearly optimal and fair outcomes, *narrowly compliant*. And let us call a person who is disposed to co-operate in ways that, followed by all, merely yield her some benefit in relation to universal non-cooperation, *broadly compliant*. We need not deny that a broadly compliant person would expect to benefit in some situations in which a narrowly compliant person could not. But in many other situations a broadly compliant person must expect to lose by her disposition. For in so far as she is known to be broadly compliant, others will have every reason to maximize their utilities at her expense, by offering 'co-operation' on terms that offer her but little more than she could expect from non-co-operation. Since a broadly compliant person is disposed to seize whatever benefit a joint strategy may afford her, she finds herself with opportunities for but little benefit.

Since the narrowly compliant person is always prepared to accept co-operative arrangements based on the principle of minimax relative concession, she is prepared to be co-operative whenever cooperation can be mutually beneficial on terms equally rational and fair to all. In refusing other terms she does not diminish her prospects for co-operation with other rational persons, and she ensures that those not disposed to fair co-operation do not enjoy the benefits of any co-operation, thus making their unfairness costly to themselves, and so irrational.

In the next chapter we shall extend the conception of narrow compliance, so that it includes taking into account not only satisfaction of minimax relative concession, but also satisfaction of a standard of fairness for the initial bargaining position. We shall then find that for some circumstances, narrow compliance sets too high a standard. If the institutions of society fail to be both rational and impartial, then the narrowly compliant person may be unable to effect any significant reform of them, while depriving herself of what benefits an imperfect society nevertheless affords. Then--we must admit--rationality and impartiality can fail to coincide in individual choice.

But we suppose that among fully rational persons, institutions, practices, and agreements that do not satisfy the requirements of minimax relative concession must prove unstable. There would, of course, be some persons with an interest in maintaining the unfairness inherent in such structures. But among the members of a society each of whom is, and knows her fellows to be, rational and adequately informed, those who find themselves with less than they could expect from fair and optimal co-operation can, by disposing themselves to narrow compliance, effect the reform of their society so that it satisfies the requirements of justice. Reflection on how partiality sustains itself shows that, however important coercive measures may be, their effectiveness depends finally on an uncoerced support for norms that directly or indirectly sustain this partiality, a support which would be insufficiently forthcoming from clearheaded constrained maximizers of individual utility.

2.5 To conclude this long section, let us supplement our argument for the rationality of disposing ourselves to constrained maximization with three reflections on its implications--for conventional morality, for the treatment of straightforward maximizers, and for the cultivation of translucency.

First, we should not suppose that the argument upholds all of conventional morality, or all of those institutions and practices that purport to realize fair and optimal outcomes. If society is, in Rawls's words, 'a cooperative venture for mutual advantage', then it is rational to pay one's share of social costs--one's taxes. But it need not be rational to pay one's taxes, at least unless one is effectively coerced into payment, if one sees one's tax dollars used (as one may believe) to increase the chances of nuclear warfare and to encourage both corporate and individual parasitism. If tax evasion seems to many a rational practice, this does not show that it is irrational to comply with fair and optimal arrangements, but only, perhaps, that it is irrational to acquiesce willingly in being exploited.

Second, we should not suppose it is rational to dispose oneself to constrained maximization, if one does not also dispose oneself to exclude straightforward maximizers from the benefits realizable by co-operation. Hobbes notes that those who think they may with reason violate their covenants, may not be received into society except by the error of their fellows. If their fellows fall into that error, then they will soon find that it pays no one to keep covenants. Failing to exclude straightforward maximizers from the benefits of co-operative arrangements does not, and cannot, enable them to share in the long-run benefits of co-operation; instead, it ensures that the arrangements will prove ineffective, so that there are no benefits to share. And then there is nothing to be gained by constrained maximization; one might as well join the straightforward maximizers in their descent to the natural condition of humankind.

A third consideration relates more closely to the conceptions introduced in 2.3. Consider once again the probabilities p and q , the probability that CMs will achieve mutual recognition and cooperate, and the probability that CMs will fail to recognize SMs but will be recognized by them and so be exploited. It is obvious that CMs benefit from increasing p and decreasing q . And this is reflected in our calculation of expected utility for CMs; the value of $\{u' + [rp(u'' - u')] - (1 - r)qu'\}$ increases as p increases and as q decreases.

What determines the values of p and q ? p depends on the ability of CMs to detect the sincerity of other CMs and to reveal their own sincerity to them. q depends on the ability of CMs to detect the insincerity of SMs and conceal their own sincerity from them, and the ability of SMs to detect the sincerity of CMs and conceal their own insincerity from them. Since any increase in the ability to reveal one's sincerity to other CMs is apt to be offset by a decrease in the ability to conceal one's sincerity from SMs, a CM is likely to rely primarily on her ability to detect the dispositions of others, rather than on her ability to reveal or conceal her own.

The ability to detect the dispositions of others must be well developed in a rational CM. Failure to develop this ability, or neglect of its exercise, will preclude one from benefiting from constrained maximization. And it can then appear that constraint is irrational. But what is actually irrational is the failure to cultivate or exercise the ability to detect others' sincerity or insincerity.

Both CMs and SMs must expect to benefit from increasing their ability to detect the dispositions of others. But if both endeavour to maximize their abilities (or the expected utility, net of costs, of so doing), then CMs may expect to improve their position in relation to SMs. For the benefits gained by SMs, by being better able to detect their potential victims, must be on the whole offset by the losses they suffer as the CMs become better able to detect them as potential exploiters. On the other hand, although the CMs may not enjoy any net gain in their interactions with SMs, the benefits they gain by being better able to detect other CMs as potential co-operators are not offset by corresponding losses, but rather increased as other CMs become better able to detect them in return.

Thus as persons rationally improve their ability to detect the dispositions of those with whom they interact, the value of p may be expected to increase, while the value of q remains relatively constant. But then p/q increases, and the greater it is, the less favourable need be other circumstances for it to be rational to dispose oneself to constrained maximization. Those who believe rationality and morality to be at loggerheads may have failed to recognize the importance of cultivating their ability to distinguish sincere co-operators from insincere ones.

David Hume points out that if 'it should be a virtuous man's fate to fall into the society of ruffians', then 'his particular regard to justice being no longer of use to his own safety or that of others, he must consult the dictates of self-preservation alone'.²¹ If we fall into a society—or rather into a state of nature—of straightforward maximizers, then constrained maximization, which disposes us to justice, will indeed be of no use to us, and we must then consult only the direct dictates of our own utilities. In a world of Fools, it would not pay to be a constrained maximizer, and to comply with one's agreements. In such circumstances it would not be rational to be moral.

But if we find ourselves in the company of reasonably just persons, then we too have reason to dispose ourselves to justice. A community in which most individuals are disposed to comply with fair and optimal agreements and practices, and so to base their actions on joint co-operative strategies, will be self-sustaining. And such a world offers benefits to all which the Fools can never enjoy.

Hume finds himself opposed by 'a sensible knave' who claimed that 'honesty is the best policy, may be a good general rule, but is liable to many exceptions; and he . . . conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions.'²² Hume confesses candidly that 'if a man think that this reasoning much requires an answer, it would be a little difficult to find any which will to him appear satisfactory and convincing'.²³ A little difficult, but not, if we are right, impossible. For the answer is found in treating honesty, not as a policy, but as a disposition. Only the person truly disposed to honesty and justice may expect fully to realize their benefits, for only such a person may rationally be admitted to those mutually beneficial arrangements—whether . . . actual agreements or implicitly agreed practices—that rest on honesty and justice, on voluntary compliance. But such a person is not able, given her disposition, to take advantage of the 'exceptions'; she rightly judges such conduct irrational. The Fool and the sensible knave, seeing the benefits to be gained from the exceptions, from the advantageous breaches in honesty and compliance, but not seeing beyond these benefits, do not acquire the disposition. Among knaves they are indeed held for sensible, but among us, if we be not corrupted by their smooth words, they are only fools.

3.1 In defending constrained maximization we have implicitly reinterpreted the utility-maximizing conception of practical rationality. The received interpretation, commonly accepted by economists and elaborated in Bayesian decision theory and the Von Neumann-Morgenstern theory of games, identifies rationality with utility-maximization at the level of particular choices. A choice is rational if and only if it maximizes the actor's expected utility. We identify rationality with utility-maximization at the level of dispositions to choose. A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition. We shall consider whether particular choices are rational if and only if they express a rational disposition to choose.

It might seem that a maximizing disposition to choose would express itself in maximizing choices. But we have shown that this is not so. The essential point in our argument is that one's disposition to choose affects the situations in which one may expect to find oneself. A straightforward maximizer, who is disposed to make maximizing choices, must expect to be excluded from co-operative arrangements which he would find advantageous. A constrained maximizer may expect to be included in such arrangements. She benefits from her disposition, not in the choices she makes, but in her opportunities to choose.

We have defended the rationality of constrained maximization as a disposition to choose by showing that it would be rationally chosen. Now this argument is not circular; constrained maximization is a disposition for strategic choice that would be parametrically chosen. But 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. We must however emphasize that it is not the choice itself, but the maximizing character of the disposition in virtue of which it is choiceworthy, that is the key to our argument.

But there is a further significance in our appeal to a choice among dispositions to choose. For we suppose that the capacity to make such choices is itself an essential part of human rationality. We could imagine beings so wired that only straightforward maximization would be a psychologically possible mode of choice in strategic contexts. Hobbes may have thought that human beings were so wired, that we were straightforwardly-maximizing machines. But if he thought this, then he was surely mistaken. At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection. Thus we suppose it possible for persons, who may initially assume that it is rational to extend straightforward maximization from parametric to strategic contexts, to reflect on the implications of this extension, and to reject it in favour of constrained maximization. Such persons would be making the very choice, of a disposition to choose, that we have been discussing in this chapter.

And in making that choice, they would be expressing their nature not only as rational beings, but also as moral beings. If the disposition to make straightforwardly maximizing choices were wired in to us, we could not constrain our actions in the way required for morality. Moral philosophers have rightly been unwilling to accept the received interpretation of the relation between practical rationality and utility-maximization because they have recognized that it left no place for a rational constraint on directly utility-maximizing behaviour, and so no place for morality as ordinarily understood. But they have then turned to a neo-Kantian account of rationality which has led them to dismiss the idea that those considerations that constitute a person's reasons for acting must bear some particular relationship to the person.²⁴ They have failed to relate our nature as moral beings to our everyday concern with the fulfilment of our individual preferences. But we have shown how morality issues from that concern. When we correctly understand how utility-maximization is identified with practical rationality, we see that morality is an essential part of maximization.

3.2 An objector might grant that it may be rational to dispose oneself to constrained maximization, but deny that the choices one is then disposed to make are rational.²⁵ The objector claims that we have merely exhibited another instance of the rationality of not behaving rationally. And before we can accuse the objector of paradox, he brings further instances before us.

Consider, he says, the costs of decision-making. Maximizing may be the most reliable procedure, but it need not be the most cost-effective. In many circumstances, the rational person will not maximize but satisfice—set a threshold level of fulfilment and choose the first course of action of those coming to mind that one expects to meet this level. Indeed, our objector may suggest, human beings, like other higher animals, are natural satisficers. What distinguishes us is that we are not hard-wired, so that we can choose differently, but the costs are such that it is not generally advantageous to exercise our option, even though we know that most of our choices are not maximizing.

Consider also, he says, the tendency to wishful thinking. If we set ourselves to calculate the best or maximizing course of action, we are likely to confuse true expectations with hopes. Knowing this, we protect ourselves by choosing on the basis of fixed principles, and we adhere to these principles even when it appears to us that we could do better to ignore them, for we know that in such matters appearances often deceive. Indeed, our objector may suggest, much of morality may be understood, not as constraints on maximization to ensure fair mutual benefit, but as constraints on wish-fulfilling behaviour to ensure closer approximation to maximization.

Consider again, he says, the benefits of threat behaviour. I may induce you to perform an action advantageous to me if I can convince you that, should you not do so, I shall then perform an action very costly to you, even though it would not be my utility maximizing choice. Hijackers seize aircraft, and threaten the destruction of everyone aboard, themselves included, if they are not transported to Havana. Nations threaten nuclear retaliation should their enemies attack them. Although carrying out a threat would be costly, if it works the cost need not be borne, and the benefit, not otherwise obtainable, is forthcoming.

But, our objector continues, a threat can be effective only if credible. It may be that to maximize one's credibility, and one's prospect of advantage, one must dispose oneself to carry out one's threats if one's demands are not met. And so it may be rational to dispose oneself to threat enforcement. But then, by parity of reasoning with our claims about constrained maximization, we must suppose it to be rational actually to carry out one's threats. Surely we should suppose instead that, although it is clearly irrational to carry out a failed threat, yet it may be rational to dispose oneself to just this sort of irrationality. And so similarly we should suppose that although it is clearly irrational to constrain one's maximizing behaviour, yet it may be rational to dispose oneself to this irrationality.

We are unmoved. We agree that an actor who is subject to certain weaknesses or imperfections may find it rational to dispose himself to make choices that are not themselves rational. Such dispositions may be the most effective way of compensating for the weakness or imperfection. They constitute a second-best rationality, as it were. But although it may be rational for us to satisfy, it would not be rational for us to perform the action so chosen if, cost free, the maximizing action were to be revealed to us. And although it may be rational for us to adhere to principles as a guard against wish-fulfilment, it would not be rational for us to do so if, beyond all doubt, the maximizing action were to be revealed to us.

Contrast these with constrained maximization. The rationale for disposing oneself to constraint does not appeal to any weakness or imperfection in the reasoning of the actor; indeed, the rationale is most evident for perfect reasoners who cannot be deceived. The disposition to constrained maximization overcomes externalities; it is directed to the core problem arising from the structure of interaction. And the entire point of disposing oneself to constraint is to adhere to it in the face of one's knowledge that one is not choosing the maximizing action.

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.

But what of the threat enforcer? Here we disagree with our objector; it may be rational for a perfect actor to dispose herself to threat enforcement, and if it is, then it is rational for her to carry out a failed threat. Equally, it may be rational for a perfect actor to dispose herself to threat resistance, and if it is, then it is rational for her to resist despite the cost to herself. Deterrence, we have argued elsewhere, may be a rational policy, and non-maximizing deterrent choices are then rational. 26

In a community of rational persons, however, threat behaviour will be proscribed. Unlike co-operation, threat behaviour does not promote mutual advantage. A successful threat simply redistributes benefits in favour of the threatener; successful threat resistance maintains the status quo. Unsuccessful threat behaviour, resulting in costly acts of enforcement or resistance, is necessarily non-optimal; its very *raison d'être* is to make everyone worse off. Any person who is not exceptionally placed must then have the *ex ante* expectation that threat behaviour will be overall disadvantageous. Its proscription must be part of a fair and optimal agreement

among rational persons; one of the constraints imposed by minimax relative concession is abstinence from the making of threats. Our argument thus shows threat behaviour to be both irrational and immoral.

Constrained maximizers will not dispose themselves to enforce or to resist threats among themselves. But there are circumstances, beyond the moral pale, in which a constrained maximizer might find it rational to dispose herself to threat enforcement. If she found herself fallen among straightforward maximizers, and especially if they were too stupid to become threat resisters, disposing herself to threat enforcement might be the best thing she could do. And for her, carrying out failed threats would be rational, though not utility-maximizing.

Our objector has not made good his case. The dispositions of a fully rational actor issue in rational choices. Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices.

3.3 To conclude this chapter, let us note an interesting parallel to our theory of constrained maximization—Robert Trivers' evolutionary theory of reciprocal altruism. We have claimed that a population of constrained maximizers would be rationally stable; no one would have reason to dispose herself to straightforward maximization. Similarly, if we think of constrained and straightforward maximization as parallel to genetic tendencies to reciprocal altruism and egoism, a population of reciprocal altruists would be genetically stable; a mutant egoist would be at an evolutionary disadvantage. Since she would not reciprocate, she would find herself excluded from co-operative relationships.

Trivers argues that natural selection will favour the development of the capacity to detect merely simulated altruism. This of course corresponds to our claim that constrained maximizers, to be successful, must be able to detect straightforward maximizers whose offers to co-operation are insincere. Exploitative interactions between CMs and SMs must be avoided.

Trivers also argues that natural selection will favour the development of guilt, as a device motivating those who fail to reciprocate to change their ways in future.²⁸ In our argument, we have not appealed to any affective disposition; we do not want to weaken the position we must defeat, straightforward maximization, by supposing that persons are emotionally indisposed to follow it. But we may expect that in the process of socialization, efforts will be made to develop and cultivate each person's feelings so that, should she behave as an SM, she will experience guilt. We may expect our affective capacities to be shaped by social practices in support of cooperative interaction.

If a population of reciprocal altruists is genetically stable, surely a population of egoists is also stable. As we have seen, the argument for the rationality of constrained maximization turns on the proportion of CMs in the population. A small proportion of CMs might well suffer more from exploitation by undetected SMs than by cooperation among themselves unless their capacities for detecting the dispositions of others were extraordinarily effective. Similarly, a mutant reciprocal altruist would be at a disadvantage among egoists; her attempts at co-operation would be rebuffed and she would lose by her efforts in making them.

Does it then follow that we should expect both groups of reciprocal altruists and groups of egoists to exist stably in the world? Not necessarily. The benefits of co-operation ensure that, in any given set of circumstances, each member of a group of reciprocal altruists should do better than a corresponding member of a group of egoists. Each reciprocal altruist should have a reproductive advantage. Groups of reciprocal altruists should therefore increase relative to groups of egoists in environments in which the two come into contact. The altruists must prevail—not in direct combat between the two (although the co-operation possible among reciprocal altruists may bring victory there), but in the indirect combat for evolutionary survival in a world of limited resources.

In his discussion of Trivers's argument, Jon Elster notes two points of great importance which we may relate to our own account of constrained maximization. The first is, 'The altruism is the more efficient because it is not derived from calculated self-interest.'²⁹ This is exactly our point at the end of 1.1—constrained maximization is not straightforward maximization in its most effective guise. The constrained maximizer genuinely ignores the

call of utility-maximization in following the co-operative practices required by minimax relative concession. There is no simulation; if there were, the benefits of co-operation would not be fully realized.

The second is that Trivers's account 'does not purport to explain specific instances of altruistic behaviour, such as, say, the tendency to save a drowning person. Rescue attempts are explained by a general tendency to perform acts of altruism, and this tendency is then made the object of the evolutionary explanation.'³⁰ In precisely the same way, we do not purport to give a utility-maximizing justification for specific choices of adherence to a joint strategy. Rather we explain those choices by a general disposition to choose fair, optimizing actions whenever possible, and this tendency is then given a utility-maximizing justification.

We do not, of course, have the competence to discuss whether or not human beings are genetically disposed to utility-maximizing behaviour. But if human beings are so disposed, then we may conclude that the disposition to constrained maximization increases genetic fitness.

NOTES

1. Hobbes, *De Cive*, ch. I, para. 7; in *Man and Citizen*, p. 115.

2. Hobbes, *Leviathan*, ch. 14, p. 64.

3. Ibid.

4 Ibid., ch. 13, p. 63.

5. Ibid., ch. 14, p. 64.

6. Ibid.

7. Ibid., ch. 14, pp. 64-5.

8 Ibid., ch. 14, p. 65.

9 Ibid.

10. Ibid., ch. 21, p. 111.

11. Ibid., ch. 15. P. 71.

12. Ibid.

13. Ibid., ch. 15, p. 72.

14. Ibid., ch. 15, p. 73.

15. Ibid.

16. Hobbes, *The Questions Concerning Liberty, Necessity, and Chance*, 1656, no. xiv; in Sir William Molesworth (ed.), *The English Works of Thomas Hobbes*, 11 vols. (London, 1839-45), vol. 5, pp. 193-4.

17. Hobbes, *Leviathan*, ch. 15, p. 72.

18 Our answer to the Foole builds on, but supersedes, my discussion in 'Reason and Maximization', *Canadian Journal of Philosophy* 4 (1975), pp. 424--33.

19. Thus constrained maximization is not parallel to such strategies as 'tit-for-tat' that have been advocated for so-called iterated Prisoner's Dilemmas, Constrained maximizers may co-operate even if neither expects her choice to affect future situations. Thus our treatment of co-operation does not make the appeal to reciprocity necessary to Robert Axelrod's account; see 'The Emergence of Cooperation among Egoists', *American Political Science Review* 75 (1981), pp. 306-18.
20. That the discussion in 'Reason and Maximization' assumes transparency was pointed out to me by Derek Parfit. See his discussion of 'the self-interest theory' in *Reasons and Persons* (Oxford, 1984), esp. pp. 18-19. See also the discussion of 'Reason and Maximization' in S. L. Darwall, *Impartial Reason* (Ithaca, NY, 1983), esp. pp. 197-8.
21. Hume, *Enquiry*, iii. i, p. 187.
- 22 Ibid., ix. ii, pp. 282-3
- 23 Ibid., ix. ii, p, 283.
24. See, for example, T. Nagel, *The Possibility of Altruism* (Oxford, 1970), pp. 90-124.
25. The objector might be Derek Parfit; see *Reasons and Persons*, pp. 19-23. His book appeared too recently to permit discussion of his arguments here.
26. See 'Deterrence, Maximization, and Rationality', *Ethics* 94 (1984), pp. 474-95; also in D. MacLean (ed.), *The Security Gamble: Deterrence Dilemmas in the Nuclear Age* (Totowa, NJ, 1984), pp. 101-22.
27. See R. L. Trivers, 'The Evolution of Reciprocal Altruism', *Quarterly Review of Biology* 46 (1971), pp. 35-57.
28. Ibid., p. 50.
29. J. Elster, *Ulysses and the Sirens: Studies in rationality and irrationality* (Cambridge, 1979), p. 145.
- 30 Ibid., pp. 145-6.